ATTRIBUTE ORIENTED INDUCTION HIGH LEVEL EMERGING PATTERN (AOI-HEP)

H L H SPITS WARNARS

PhD 2013

ATTRIBUTE ORIENTED INDUCTION HIGH LEVEL EMERGING PATTERN (AOI-HEP)

HARCO LESLIE HENDRIC SPITS WARNARS

A thesis submitted in partial fulfilment of the requirement of

the Manchester Metropolitan University for the degree of Doctor of Philosophy

School of Computing, Mathematics and Digital Technology

the Manchester Metropolitan University

2013

Acknowledgements

First of all I would like to acknowledge my PhD director of study and as first PhD supervisor, Dr. Maybin K. Muyeba for his time, patient, guidance, help and advice during my PhD study. I appreciate for the wisdom of his supervision, to find the research idea in second year of my PhD study, made me independent in my PhD research and gave me challenge in my research progression. He guided me in writing international conference papers,which were published in Lecture Notes in Computer Science (LNCS). I am also thankful to my reviewer Dr. Keeley A. Crockett for her constructive opinion on my PhD study. Thanks to the chair of my Viva Prof. Nicholas Bowring, internal examiner Dr. Liangxiu Han and second examiner Dr. M. Saraee for their input and improvement upon this thesis.

I thank the Research, Enterprise and Development (RED) office at Manchester Metropolitan University for supporting my PhD academic research skills with training and workshops. Also thanks to staff in the Research Degree Administration office, Faculty of Science and Engineering, Manchester Metropolitan University. They gave quick response when I encountered problems during the four years of my PhD study. Also, my acknowledgements go out to Indonesian government for the scholarship. Thanks to the chairman of Budi Luhur foundation Mr. Kasih Hanggoro, MBA.

I thank all my colleagues, Dr. Alma Adventa for her proof reading and help in writing, Dr. Gindo Tampubolon, Dr. Yanuar Nugroho, Dr. Delvac Oceandy and Dr. Danny Pudjianto for all PhD matters. Also thanks to Indonesian Students Association Greater Manchester for the friendships and cheerfulness. Thanks to Indonesian Fellowship Manchester for the spirit, care and prayer.

Finally, heartful thanks go to my family, my mum and dad, my parents in law, my wife Jen Nie for her delicious cuisines, my daughters Michell and Olivia, my sons Leonel, Laurens and Leandro for their patience, love and prayer.

Abstract

Attribute-Oriented Induction of High-level Emerging Pattern(AOI-HEP) is a combination of Attribute Oriented Induction (AOI) and Emerging Patterns (EP). AOI is a summarisation algorithm that compact a given dataset into small conceptual descriptions, where each attribute has a defined concept hierarchy. This presents patterns are easily readable and understandable.Emerging patterns are patterns discovered between two datasets and between two time periods such that patterns found in the first dataset have either grown (or reduced) in size, totally disappeared or new ones have emerged. AOI-HEP is not influenced by border-based algorithm like in EP mining algorithms. It is desirable therefore that we obtain summarised emerging patterns between two datasets. We propose High-level Emerging Pattern (HEP) algorithm. The main purpose of combining AOI and EP is to use the typical strength of AOI and EP to extract important high-level emerging patterns from data.

The AOI characteristic rule algorithm was run twice with two input datasets,to create two rulesets which are then processed with the HEP algorithm. Firstly, the HEP algorithm starts with cartesian product between two rulesets which eliminates rules in rulesets by computing similarity metric (a categorization of attribute comparisons). Secondly, the output rules between two rulesets from the metric similarity are discriminated by computing a growth rate value to find ratio of supports between rules from two rulesets. The categorization of attribute comparisons is based on similarity hierarchy level. The categorisation of attributes was found to be with three options in how they subsume each other. These were Total Subsumption HEP (TSHEP), Subsumption Overlapping HEP (SOHEP) and Total Overlapping HEP (TOHEP) patterns. Meanwhile, from certain similarity hierarchy level and values, we can mine frequent and similar patterns that create discriminant rules.

We used four large real datasets from UCI machine learning repository and discovered valuable HEP patterns including strong discriminant rules, frequent and similar patterns. Moreover, the experiments showed that most datasets have SOHEP but not TSHEP and TOHEP and the most rarely found were TOHEP. Since AOI-

HEP can strongly discriminate high-level data, assuredly AOI-HEP can be implemented to discriminate datasets such as finding bad and good customers for banking loan systems or credit card applicants etc. Moreover, AOI-HEP can be implemented to mine similar patterns, for instance, mining similar customer loan patterns etc.

Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1. Motivations

Recent developments in data minng show that single algorithms are no longer feasible to be used in isolation when looking for patterns that span different datasets, users and application. This research is motivated by a hybrid approach involving Attribute Oriented Induction (AOI) [30-48,55,58,61-63] and Emerging Patterns (EP) [65-72,75-84] that extracts high-level and emerging patterns, respectively from data. AOI was proposed in 1989 by Han and his colleagues [31], integrates a machine learning paradigm especially learning from examples technique with database operations, extracts generalized rules from an interesting set of data and discovers high-level data regularities. AOI uses concept hierarchy as background knowledge or taxonomies of every attribute domain. The attribute concepts are ordered level by level from specific (low-level) into general or higher level concepts. AOI then performs generalization of attribute values by ascending to the next higher level concepts along the paths of the concept hierarchy [40-42]. Meanwhile, Emerging Patterns (EP) was proposed in 1999 [80]. EP captures emerging trends when applied to time stamped datasets or capture useful contrasts between data classes when applied to datasets with classes. EP captures significant changes and differences between datasets defined as itemsets whose supports increase significantly from one dataset to another [67, 77]. The increasing of supports for itemsets from one dataset to another is called growth rate [67, 77].

The strength of AOI is in the use of concept hierarchies for generalization in order to generalize from low level data into high-level data. Moreover, AOI has been tested successfully against large relational datasets [44] and is able to learn different kinds of rules such as characteristic, discrimination, classification, data evolution regularities [39], association and cluster description rules[40]. Meanwhile, the strength of EP is in discriminating between datasets using growth rate functions (the ratio of the supports in one dataset D_1 to another dataset D_2 [75,79]. Combining AOI and EP proposes a technique that can strongly discriminate high-level data and learn different kinds of rules, thus creating a novel technique called Attribute-Oriented Induction High-level Emerging Pattern (AOI-HEP). This technique uses growth rate of patterns between datasets to discriminate high-level data resulting from concept hierarchy generalizations. Furthermore, this presents AOI-HEP as a new data mining framework relevant to the management level decision making process..

Previously, EP was proposed with border-based algorithm [79] and has had influences on other extended EP mining algorithms such as Classification by Aggregating EPs (CAEP) [84], information-based approach for Classification by Aggregating EPs (iCAEP)[67] and Classification by Aggregating Essential EPs (CAEEP)[100]. Others include Decision making by EPs (DeEP) [65], Bayesian Classification by EPs (BCEP)[78], Constrained Emerging Patterns (CEP) [77], Jumping Emerging Pattern classifier (JEP-Classifier)[83], JEP space[82], Knowledge Trends Data Analysis (KTDA)[89], Prediction by Likelihoods (PCL) [102] and Gtree[90]. Border-based algorithm avoid the long process naive algorithms do to get the counts of all itemsets in a large collection of candidates by manipulating only borders of some two collections and derive all EPs whose support satisfies a minimum support threshold [79]. Border-based algorithms define borders<L,R> where L is the sets of the minimal itemsets (superset or the most general EPs on the left) and R is the sets of the maximal itemsets (subset or the most specific EPs on the right) [71,72]. Border-based algorithm defines border <L,R> where each element L is a subset of some elements in Rand each element of R is a superset of some elements in L [79].

Border-based algorithms have limitations as follows:

1. Differential procedure algorithm as part of border-based algorithm has to be called in multiple numbers of times when discovers all EPs [79].

2. Border-based algorithm has to be called twice when discovers all Jumping EP (JEPs) in both datasets (from target to contrasting datasets and vice versa) [70,78]. Meanwhile, Essential JEP (EJEP) and EJEP Classifier (EJEPC) use tree structure called Pattern-tree (P-tree) algorithm which efficiently mine EJEPs and EJEPCs in both datasets (from target to contrasting datasets and vice versa) without calling the algorithm twice [70,78].

The new AOI-HEP algorithm is not influenced by border-based approaches but is similar to comparisons with the decision tree technique CART-based method, which again, is not influenced by border-based algorithm. The CART-based approaches discover relevant EPs for classification using a CART tree [91].

AOI-HEP is similar to DeEP algorithm in terms of reduction the number of instances and attributes. DeEP was influenced by a border-based algorithm in EP to access low level data and has advantages on accuracy, speed and dimensional scalability over CAEP and JEP-C [83]. DeEP reduces number of instances and attributes from the training data with instance-based approach [65,72,81] whilst AOI-HEP uses AOI characteristic rule algorithm [30-31]. DeEP reduces the number of attributes with intersection operation using neighbourhood-based intersection method [65,72, 81] while AOI-HEP reduces number of attributes by attribute generalization and removal of redundant tuples as a second step [30-31]. Moreover, DeEP reduces number of instances by selecting the maximal itemsets from intersection operation [65,72,81] while AOI-HEP reduces number of instances with AOI generalization until distinct instances are less or equal to an instance threshold[30-31].

Moreover, this research is also motivated in the following ways and findings:

1. Total Subsumption HEP (TSHEP – those rules that are completely subsumed).
2. Subsumption Overlapping HEP (SOHEP – those rules that overlap and are subsumed).
3. Total Overlapping (TOHEP – those rules that are completely overlapping).
4. Frequent patterns.
5. Similar patterns.

In this thesis, AOI-HEP has been successfully implemented using four large real datasets from UCI machine learning repository., and discovered TSHEP, SOHEP, TOHEP, frequent and similar patterns. The experiments showed that most datasets have SOHEP but not TSHEP and TOHEP, and the most rarely found were TOHEP. Frequent patterns that show synonymy with large pattern are interesting to be mined since with frequent patterns we can have strong discrimination, whilst similar patterns are interesting to be mined which can show equality of patterns that represent similar behaviours. The experiments showed that TSHEP tend to frequent patterns and TOHEP tend to similar patterns. Meanwhile, SOHEP occur between frequent and similar patterns (which are based on frequent similarity value i.e. the frequent similarity subsumption for frequent patterns and frequent similarity overlapping for similar patterns).

From frequent and similar patterns, we can create discrimination rules which show discrimination for each dataset influenced by learning high-level concepts in one of attributes of dataset. From frequent and similar patterns, we can get strong discriminant rules if the patterns have large growth rates where there are large supports in target dataset and small supports in contrasting dataset [69,71,79].

1.2. Contributions

The main contributions of this thesis are:

1) Presenting a new framework that combines Attribute-Oriented Induction (AOI) and Emerging Pattern (EP).
2) Discriminating high-level data from two different rulesets which are from two different datasets.
3) Mining different types of High-level Emerging Patterns (HEP) i.e. Total Subsumption HEP (TSHEP), Subsumption Overlaping HEP (SOHEP) or Total Overlapping HEP (TOHEP).
4) Mining frequent and similar patterns.

.

5) Creating discriminant rules including strong discriminant rules from frequent and similar patterns.
6) Finding the interesting dataset to be mined for frequent and/or similar patterns.

1.3. Organization of thesis

The rest of the thesis is organized as follows: Chapter 2 illustrates research literatures in data mining and in particular two data mining techniques, Attribute Oriented Induction (AOI) and Emerging Patterns (EP); In chapter 3, Attribute Oriented Induction High-level Emerging Pattern (AOI-HEP) mining framework is described and defines TSHEP, SOHEP and TOHEP including theory to mine TSHEP, SOHEP, TOHEP, frequent and similar patterns; Chapter 4 defines AOI-HEP experiments using four datasets from UCI machine learning repository [56] to mine TSHEP, SOHEP, TOHEP, frequent and similar patterns which can create discriminant rules. Finally, conclusion and possible future research are described in chapter 5.

Chapter 2: Literature Review

2.1. Introduction

This chapter presents an overview of the research literature in general of data mining and in particular two data mining techniques, Attribute Oriented Induction (AOI) and Emerging Patterns (EP). These data mining techniques are combined to propose a new algorithm called AOI-HEP. In the next section, we discuss data mining theory and discovery of patterns from data, the interestingness of discovery patterns, Knowledge Discovery in Databases (KDD) methodology and data mining algorithm (methods or techniques). Moreover, section 2.3 specifies the AOI data mining technique, kinds of knowledge rules that be can learned, concept hierarchy as AOI background knowledge and AOI characteristic and discrimination rules algorithms including eight generalization strategy steps. Furthermore, section 2.4 defines EP data mining technique which can captures the differences between classes, the support for class, growth rate, Jumping Emerging Pattern (JEP) and EPs algorithms include EP-based classifier for classification more than two classes. Meanwhile, section 2.5 illustrates the powerful AOI-HEP as combination AOI and EP data mining techniques, the differences between AOI-HEP and the famous border-based algorithm, similarity AOI-HEP with DeEP algorithm and previous researches which combining EP with other technique. Finally, the summary for this chapter is presented in section 2.6.

2.2. Data Mining

Businesses need information which can be used as data for the decision making process. Data when used effectively can be very valuable in the competitive business world. On the other hand when data is not used effectively, it could be less competitive for a business. Data mining is useful for processing data and then extracting patterns that are valuable for decision making. For instance, customer patterns are valuable information for banking systems to secure bank loan and precious knowledge for retail systems to increase profit and customer loyalty. As shown in figure 2.1, the data mining algorithm is the process

of discovering patterns and models from data for the decision making. Data mining as a particular step in Knowledge Discovery in Databases (KDD) is a non trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns of the data [4]. Data mining uses specific algorithms such as discrimination, classification, association, clustering and etc. to produce different patterns and models. The discovered patterns should have the following criteria [1,4] :

1) Valid on new data with some degree of certainty.
2) Novelty where at least to the system and preferably to the user.
3) Usefulness that lead to some benefits to the users or for the tasks.
4) Understandable which can be estimated through simplicity, and if not immediately then after some post processing.

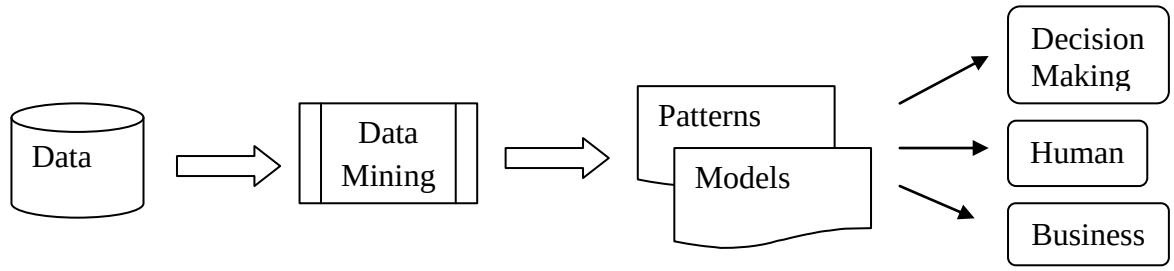


Figure 2.1. Transformation of data to become patterns or models with data mining

Discovery patterns or knowledge will yield many patterns and as the huge number of pattern are difficult to understand then elimination process can be applied with the use of threshold value, in order to find the most interesting patterns. The interestingness of discovered patterns is measured by combining the four patterns criteria such as validity, novelty, usefulness and simplicity for understandable estimation [16]. The two methods commonly used to find interesting pattern are [28] :

1) Subjective methods, which are user driven and domain dependent and for instance, the user should specify the rule to be considered interesting.
2) Objective methods, which are data driven and domain independent and for instance, the interestingness of rule depends on the quality of the rule and its similarity to other rules.

The subjective and objective methods are similar with two types of KDD goals [1,4] and they are :

1) Verification, where the system is limited to verify the user's hypothesis.
2) Discovery, where the system autonomously finds new patterns. The discovery goal is then subdivided into i.e.:

2.a) Prediction, where the system finds patterns for predicting the future behaviour of some entities.

2.b) Description, where the system finds patterns for presentation to users in a human understandable form. Predictivemodels can be descriptive model and vice versa.

As a scientific discipline, data mining intersects with other disciplines [2,4] i.e. databases, statistics, machine learning, Artificial Intelligence, expert system and pattern recognition, neural network, data visualization, information retrieval, image and signal processing, and spatial data analysis. Data mining has been applied for some real world problems in many industries such as spatial data mining, musical data mining, text data mining, visual data mining, privacy preserving data mining and etc [21]. In 2011 KDnuggets poll surveyed industries that have applied data mining in their operations and the top ten industries were Customer Relationship Management (CRM), banking, health care, education, fraud detection, science, social networks, credit scoring , direct marketing/fundraising and insurances[14].

Spatial data mining is a process of mining the knowledge from spatial data such as image and movie in order to find patterns. It has wide applications in Geographic Information Systems (GIS), remote sensing, image and video database, medical imaging, robot navigation, and etc[29]. Musical Data Mining use data mining techniques, including co occurrence analysis in order to discover similarities between songs and classify songs into correct genre's or artists [24,25]. Meanwhile, text data mining where use of large online text collections to discover new facts and trends about the words itself[10-12]. Visual data mining apply data mining by using information visualization technology to improve data analysis [5]. Finally, in privacy preserving data mining is extended or preserved user privacy in order to letting the users to provide a modified value for sensitive attributes, where the modified value may be generated using custom code, a browser plug-in or extension to products in order to mask sensitive information [6-9].

Data mining is capable of handling the huge data overload problem as data continues to grow. KDnuggets poll showed the largest database/datasets that have been used with 21.4% voters used over 1 Terabyte database/dataset, 4% voters used over 1 Petabyte database/dataset and 19.5% voters used 1.1 to 10 Gigabyte database/dataset [19]. Data mining extracts knowledge from different kinds of databases e.g. relational databases, transaction databases, object oriented databases, deductive databases, spatial databases, temporal databases, multimedia databases, heterogeneous databases, active databases, legacy databases, and the internet information-base[3]. KDnuggets poll showed the most popular

data types to be mined in 2011, and the top ten were table data, time series, itemset/transactions, text (free-form), anonymized data, location/geo/mobile data, other, social network data, email and web content [17].

2.2.1. Knowledge Discovery in Databases

The efforts in the industries mainly concern on the definition of methodologies that can guide the implementation of Data Mining applications. KDD as one of the most popular methodologies [22,23] has focus on the overall process of knowledge discovery from data, including store and access data, the efficient algorithms which deal with huge data, the interpretation and visualization of the knowledge discovery results. The term of KDD was coined in 1989 at the first KDD workshop [13]. As shown in figure 2.2., data mining as an essential step in KDD process consisting of an interactive and iterative of the following nine steps [1,4,13,27]:

1) Developing an understanding of application domain and identifying the goals of the KDD process from the customer's viewpoint.
2) Creating a target dataset, selecting dataset and focusing on a subset of variables or data samples on which discovery is to be performed.
3) Data cleaning and pre processing, which include removing noise, collecting the necessary information, handling missing data fields and accounting for time sequence information as well as DBMS issues such as type, schema and mapping of missing and unknown values.
4) Data reduction and projection by finding the useful features to represent the data.
5) Matching the goals of KDD process in step one to particular data mining method through summarization, classification, regression, clustering and so on.
6) Exploratory analysis and model and hypothesis selection by choosing the data mining algorithm and selecting the method to be used for pattern searching.
7) Data mining by searching the patterns of interest in a particular representational form.
8) Interpreting the mined patterns which can also involve visualization of the extracted patterns and models or visualization of the data given by the extracted models.
9) Acting the discovered knowledge by using the knowledge directly, incorporating the knowledge into another system for further action, or simply documenting it and reporting

it to the interested parties. This process also includes checking and resolving the potential conflicts with the previously believed (or extracted) knowledge.

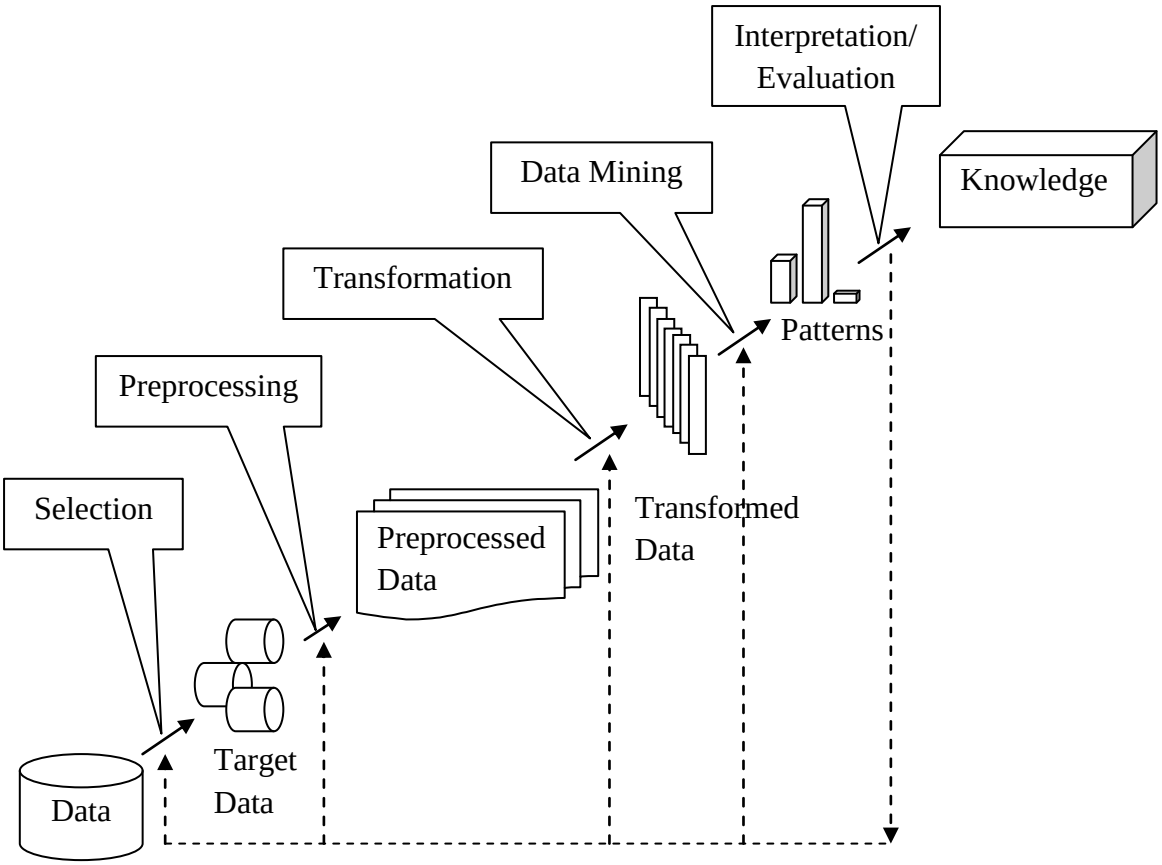


Figure 2.2. Phases of the KDD methodology

2.2.2. Data Mining Methods

There are many data mining methods/techniques/algorithms and a survey by KDnuggets showed the data mining algorithms which were used in 2011, and the top ten algorithms were decision tree/rules, regression, clustering, statistic (descriptive), visualization, time series/sequence analysis, support vector (SVM), association rules, ensemble methods and text mining [15]. Other survey showed the top ten algorithms were C4.5, K-Means, Support Vector Machine (SVM), Apriori, Expectation Maximization (EM), PageRank, AdaBoost, k-Nearest Neighbors (kNN), Naive Bayes and Classification and Regression Tress(CART) [20]. The following algorithms were in the top ten lists of both surveys i.e. decision tree/rules with C4.5, regression with CART, clustering with K-Means, support vector (SVM) with Support Vector Machine (SVM), association rules with Apriori, ensemble methods with Ada Boost.

Data mining algorithms tasks can be classified into [18,26] :

1) Supervised learning, with a known output variable in dataset and input labelled data which include classification, fuzzy classification, regression, decision tree, Support Vector Machine (SVM), artificial Neural Network, Naive Bayes and K-nearest Neighbor .
2) Unsupervised learning, without known output variable in dataset and input unlabeled data which include clustering,Expectation Maximization (EM), association rule and Self-Organizing Map (SOM).

A data mining algorithm should consists of three primary components [1,4] i.e. :

1) Model representation, where the language is used to describe discoverable patterns.
2) Model evaluation, where quantitative statements meet the goals of KDD process.
3) Search method, which consists of two components :
   3.a) Parameter search, where the algorithm must search for the parameters, which optimize the model evaluation criteria, based on the observed data and fixed model representation.
   3.b) Model search, where a loop occurs over the parameter search method.

2.3. Attribute Oriented Induction

Attribute Oriented Induction (AOI) method was first proposed in 1989 integrates a machine learning paradigm especially learning-from-examples techniques with database operations, extracts generalized rules from an interesting set of data and discovers high level data regularities [31]. AOI provides an efficient and effective mechanism for discovering various kinds of knowledge rules from datasets or databases. The AOI method has been implemented in a data mining system prototype called DBMINER [36,37,43,45,55] which previously called DBLearn and been tested successfully against large relational database. DBLearn [32,38,62,63] is a prototype data mining system which was developed in Simon Fraser University. DBMINER was developed by integrating database, OLAP and data mining technologies [34,55].

AOI approach is developed for learning different kinds of knowledge rules such as characteristic rules, discrimination rules, classification rules, data evolution regularities [39], association rules and cluster description rules[40].

1) Characteristic rule is an assertion which characterizes the concepts which satisfied by all of the data stored in database. This rule provides generalized concepts about a property

that can help people to recognize the common features of the data in a class. For example the symptom of the specific disease [47].

2) Discriminant rule is an assertion, which discriminates the concepts of one (target) class from another (contrasting). This rule give a discriminant criterion which can be used to predict the class membership of of new data, for example to distinguish one disease from the other [47].
3) Classification rule is a set of rules, which classifies the set of relevant data according to one or more specific attributes. For example, classifying diseases into classes and provide the symptoms of each [30].
4) Association rule is association relationships among the set of relevant data. For example, discovering a set of symptoms frequently occurring together [35,50].
5) Data evolution regularities rule is general evolution behaviour of a set of the relevant data (valid only in time-related/temporal data). For example, describing the major factors that influence the fluctuations of stock values through time [33,41]. Data evolution regularities can then be classified into characteristic rule and discrimination rule [41].
6) Cluster description rule is used to cluster data according to data semantics [50], for example clustering the university student based on different attribute(s).

2.3.1. Concept hierarchies

One advantage of AOI is that it has concept hierarchy as the background knowledge which can be provided by the knowledge engineers or domain experts [40,41,42]. Concept hierarchy stored a relation in the database provides essential background knowledge for data generalization and multiple level data mining. Concept hierarchy represents taxonomy of concept of the attribute domain values. Concept hierarchy can be specified based on the relationship among database attributes or by set groupings and be stored in the form of relations in the same database [45]. Concept hierarchy can be adjusted dynamically based on the distribution of the set of data relevant to the data mining tasks. The hierarchies for numerical attributes can be constructed automatically based on data distribution analysis [45]. Concept hierarchy for numeric will be treated differently for the sake of efficiency [58,59,60,61,64]. For example if there are a range of value between 0 and 1.99, then there willbe 199 values start from 0.00 until 1.99, but for efficiency there will be only 1 record created with 3 fields rather than with 200 records with 2 fields.

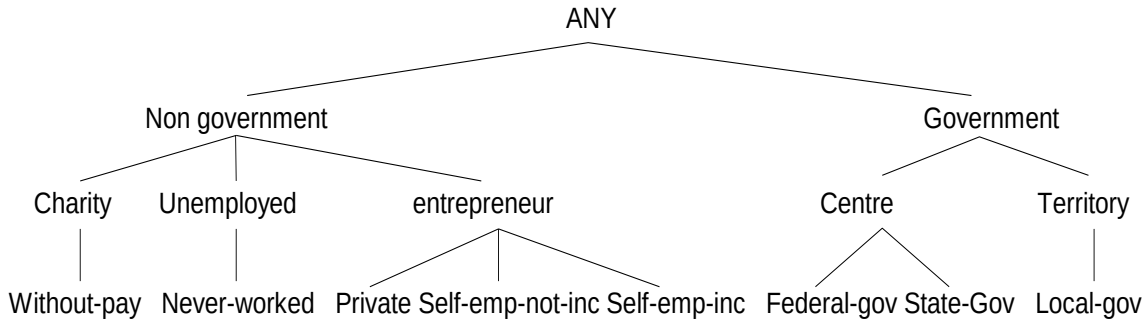


Figure 2.3. A concept hierarchy tree for attribute workclass in adult dataset[56]

Without-pay	⊂ Charity
Never-worked	⊂ Unemployed
{Private,self-emp-not-inc,self-emp-inc}	⊂ entrepreneur
{federal-gov,state-gov}	⊂ Centre
Local-gov	⊂ Territory
{Charity,Unemployed,entrepreneur}	⊂ Non government
{Centre, Territory}	⊂ Government
{Non government, Government}	⊂ ANY(workclass)

Figure 2.4. A concept hierarchy for concept hierarchy tree attribute workclass in adult dataset[56]

In concept hierarchy, concepts are ordered by levels from specific or low level concepts into general or higher level. Generalization is achieved by ascending to the next higher level concepts along the paths of the concept hierarchy. The most general concept is the null description as the most specific concepts correspond to the specific values of the attributes in the database, which described as ANY. Concept hierarchy can be balanced or unbalanced, where unbalanced hierarchy then must be converted to a balanced hierarchy. Figure 2.3 shows the concept hierarchy tree for attribute workclass in adult dataset[56], which has three levels. The first level as the low level has 8 concepts and they are without-pay, never-worked, private, self-emp-not-inc, self-emp-inc, federal-gov,state-gov and local-gov concepts. The second level has 5 concepts and they are charity, unemployed, entrepreneur, centre and territory concepts. The third level as the high level has two concepts and they are non government and government concepts. For example, the concept of non government at the high level has 3 sub concepts in the second level: charity, unemployed and entrepreneur concepts. The concept entrepreneur at the second level has three sub concepts in the low level: private, self-emp-not-inc and self-emp-inc concepts. The concept hierarchy tree in figure 2.3 can be represented in figure 2.4 where symbol ⊂ indicates generalization, for

example, Without-pay $\subset$ Charity indicates that Charity concept is a generalization of Without-pay concept.

2.3.2. AOI characteristic and discriminant rules

AOI can be implemented with an architecture design shown in figure 2.5 where characteristic rule and discriminant rule can be learned directly from the transactional database (OLTP) or Data warehouse (OLAP) [44,46] with the help of the concept hierarchy as the knowledge generalization. Concept hierarchy can be created from OLTP database as a direct resource.

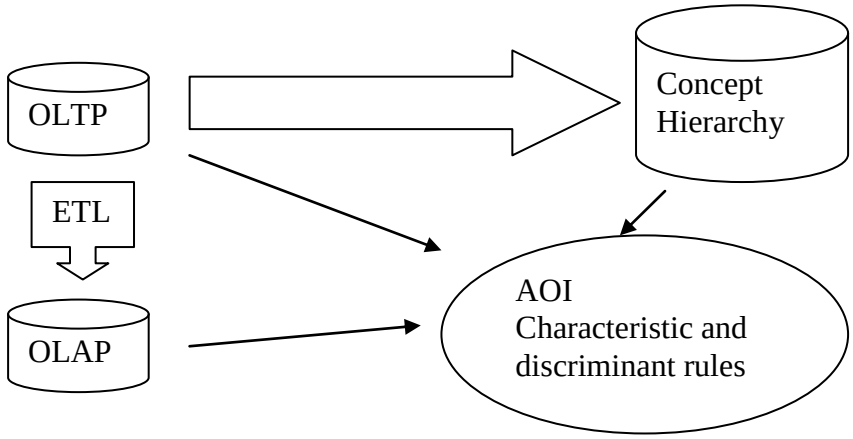


Figure 2.5. AOI Characteristic and discriminant rules architecture

From a database, we can identify two types of learning:

1) Positive learning as the target class where the data are tuples in the database, which are consistent with the learning concepts. Positive learning/target class will be built when learn characteristic rule
2) Negative learning as the contrasting class in which the data do not belong to the target class. Negative learning/contrasting class will be built when learn discrimination or classification rule.

Characteristic rule has been used by AOI in order to recognize, learning and mining as a specific character for each of attribute as their specific mining characterization. Characteristic rule process the generalization with help of concept hierarchy as the standard saving background knowledge to find target class as a positive learning. Mining rule cannot be limited with just only one rule, as the more rules can be created the more mining can be

done. This has been proven as an intelligent system, which can help human to make a system that has ability to think like a human [3]. Rules often can be discovered by generalization in several possible directions [47].

Relational database as resources for data mining with AOI can be read with data manipulation language select sql statement [51,52,53,54]. Using a query for building rules gives an efficient mechanism for understanding the mined rules [49,50]. In the current AOI, a query is processed with SQL-like data mining query language DMQL at the beginning of the process [57].It collects the relevant sets of data by processing a transformed relational query, generalizes the data by AOI and then presents the outputs in different forms [45].

AOI generalizes and reduces the prime relation further until the final relation can satisfy the user expectation based on the set threshold. One or two thresholds can be applied, where one threshold is used to control both of number of distinct attributes and tuples in the generalization process, whilst two thresholds are used to control the number of distinct attributes and tuples in the generalization process. Threshold as a control for the maximum number of tuples of the target class in the final generalized relation can be replaced with group by operator in sql select statement which will limit the final result of generalization. Setting different threshold will generate different generalized tuples as the needed of global picture of induction repeatedly as time-consuming and tedious work [48]. All interesting generalized tuples as multiple rules can be generated as the global picture of induction by using group by operator or distinct function in the sql select statement.

AOI can perform datawarehouse techniques by doing generalization process repetitively in order to generate rules at different concepts levels in a concept hierarchy, enabling the user to find the most suitable discovery levels and rules. This technique performs rollup (progressive generalization [44]) or drill down (progressive specialization [44]) and operation [40,45] have been recognized as datawarehouse techniques. Finding the most suitable discovery levels and rules would add multidimensional views to a database using generalization process repetitively at different concepts level.

There are eight strategy steps that must be done [41] in the process of generalization. Here, step one until seven are for characteristic rule and step one to eight 8 are for discriminant rule.

1) Generalization on the smallest decomposable components, generalization should be performed on the smallest decomposable components of a data relation.

2) Attribute removal, if there is a large set of distinct values for an attribute but there is no higher level concept provided for the attribute, the attribute should be removed during generalization.
3) Concept tree Ascension, if there exists a higher level concept in the concept hierarchy for an attribute value of a tuple, the substitution of the value by its higher level concept would generalize the tuples.
4) Vote propagation, the value of the vote is the value of accumulated tuples where the vote will be accumulated when merging identical tuples in the generalization.
5) Threshold control on each attribute, if the number of distinct values in a resulting relation is larger than the specified threshold value, further generalization on this attribute should be performed.
6) Threshold control on generalized relations, if the number of tuples is larger than the specified threshold value, further generalization will be done based on the selected attributes and the merging of the identical tuples should be performed.
7) Rule transformation, change final generalization to quantitative rule and qualitative rule from a tuple (conjunctive) or multiple tuples (disjunctive).
8) Handling overlapping tuples, if there are overlapping tuples in both target and contrasting classes, these tuples should be marked and eliminated from the final generalized relation.

AOI characteristic rule algorithm [41] is given as follow:

AOI characteristic rule algorithm

Input: dataset, concept hierarchies, learning task, attribute threshold, rule threshold
Output: characteristic rule of the learning task

```
1   For each of attribute Ai (1 ≤ i ≤ n, where n= # of attributes) in the generalized relation GR
2   { While #_of_distinct_values_in_attribute_Ai > threshold
3   { If no higher level concept in concept hierarchy for attribute_Ai
4   Then  remove attribute Ai
5           Else   substitute the value of Ai by its corresponding minimal generalized concept
6   Merge identical tuples
7   }
8   }
9   While #_of_tuples in GR > threshold
10  { Selective generalize attributes
11  Merge identical tuples
12  }
```

Figure 2.6. AOI characteristic rule algorithm

This AOI characteristic rule algorithm is the implementation of step one to seven of the generalization strategy steps. The algorithm shows two sub processes i.e. control number of distinct attributes and control number of tuples.

1) Control number of distinct attributes is a vertical process which checks every per attribute vertically. This is done by checking all attributes in the learning results of a dataset until the number of distinct attributes less equal than the threshold. This first sub process is just applied to the attributes which the number of distinct attributes greater than threshold. Each of attribute which the number of distinct attribute greater than threshold will be checked if it has a higher level concept in the concept hierarchy. If it has no higher level concept then the attribute will not be used. On the other hand if it has higher level concept then the attribute value will be substituted with the value of the higher level concept. Merging identical tuples will be done in order to summarize generalization and accumulate the value of the vote of the identical tuples by eliminating the redundant tuples. Eventually, after this first sub process all the attributes in generalization will have number of distinct attributes less equal than the threshold. This first sub process is implementation of step one to five of the generalization strategy steps.
2) Control number of tuples is a horizontal process, which checks per rule horizontally. This is carried out for those attributes, which passed the first sub process where each attribute will have the number of distinct attributes less equal than the threshold. This second sub process is only done while the number of rules is greater than threshold. Selective generalization of the attributes and merging of the identical tuples will reduce the number of rules. Selecting candidate attribute for further generalization can be done by preferences with finding the ratio on the number of tuples or the number of distinct attribute values. Selecting candidate attribute for further generalization can be examined by user based on the non interesting one, either non interesting attribute or non interesting rule. As with first sub process merging the identical tuples will be done in order to summarize generalization and accumulate the vote value of identical tuples by eliminating the redundant tuples. Eventually, after this second sub process the number of rules is less equal than the threshold. This second sub process is the implementation of step three, four and six of the generalization strategy steps.

AOI discriminant rule algorithm is the implementation of step one until eight of generalization strategy steps. Since AOI discriminant rule and AOI characteristic rule algorithms have the same generalization strategy steps between steps one and seven, then

literally they have the same process and the difference is just only in step eight. They also have the same sub processes i.e. control number of distinct attributes as the first sub process and control number of tuples as the second sub process. The step handling overlapping tuples as the eight generalization strategy step is process in the beginning before the first sub process and both in first and second processes before merge identical tuples.

AOI discriminant rule algorithm [39] is shown below:

AOI discriminant rule algorithm
Input: dataset, concept hierarchies, learning task, attribute threshold, rule threshold Output: discriminant rule of the learning task
For each of attribute Ai ($1 \leq i \leq n$, where n= # of attributes) in the generalized relation GR { Mark the overlapping tuples While #_of_distinct_values_in_attribute_Ai > threshold { If no higher level concept in concept hierarchy for attribute_Ai Then remove attribute Ai Else substitute the value of Ai by its corresponding minimal generalized concept Mark the overlapping tuples Merge identical tuples } } While #_of_tuples in GR > threshold { Selective generalize attributes Mark the overlapping tuples Merge identical tuples }

Figure 2.7. AOI discriminant rule algorithm

2.4. Emerging patterns

Emerging Patterns (EPs) as discovery knowledge from database capture emerging trends when applied in time stamped databases or capture useful contrasts between data classes when applied to datasets with classes[80]. Moreover, EPs capture significant changes and differences between datasets are defined as itemsets whose supports (frequencies) increase significantly from one dataset to another. The changing of supports for itemsets from one dataset to another (the ratios of the two supports) is called growth rates. Furthermore, EPs use user-defined threshold in order to reduce large candidate patterns, then can be said EPs are itemsets whose growth rates are larger than a given threshold. Finally, EPs are similar to discriminant rules or evolution rules in Attribute Oriented Induction (AOI) [40] but

different since EPs do not limited by exclusiveness constraint and because the extra information of growth rate [79].

Those EPs with very large growth rates are notable differentiating characteristic between 2 datasets and have been useful for building powerful classifiers [69,79]. Thus, Those EPs with very large growth rates are frequent in one class but rare in another class. Meanwhile EPs with low to medium support such as 1% until 20% can give very useful new insights and guidance to experts, in even "well understood" applications [79]. Hence, the low supports EPs such as 0.1 until 5% may be new knowledge to the dataset and discover small support EPs is interesting [79]. The interestingness of discovery small support EPs due to reason too many EPs candidates and make naive algorithms too costly to examine all itemsets in dataset. For example if there are 350 itemsets in dataset then naive algorithm would need to process 2^{350} (Cartesian product) itemsets in order to find their supports in datasets D1 and D2 and then determine their growth rates.

2.4.1. Growth rate and Jumping Emerging Patterns

Let $I = \{i_1,i_2,...,i_N\}$ be a set of items. A dataset is a set D of transactions. An itemset X is a subset of I. The support of an itemset X in a dataset D, denoted as $supp_D(X)$ in equation 2.1.

$$supp_D(X)= \frac{count_D(X)}{|D|} \tag{2.1}$$

where : $supp_D(X)$ = support in dataset D containing itemset X

$count_D(X)$ = the number of transactions in dataset D containing itemset X where $count_D(X) = t \in D$ and $X \subseteq t$, where t is instance in D

$|D|$ = total number of instances in dataset D

D = Dataset

X = Itemset or pattern

Given a positive number σ, we say an itemset X is σ-large in dataset D if $supp_D(X) \geq \sigma$, and X is σ-small in dataset D otherwise. Assume there are given an ordered pair of datasets D1 and D2 then growth rate of an itemset X from datasets D1 to D2 denoted in equation 2.2 as $GrowthRate_{D1 \to D2}(X) = \frac{supp_{D2}(X)}{supp_{D1}(X)}$ if $supp_{D1}(X) \neq 0$, = 0 if $supp_{D1}(X)=supp_{D2}(X)=0$, and = ∞ if $supp_{D1}(X)=0 \neq supp_{D2}(X)$. For EPs are associated with two datasets, dataset D1 will be called

background dataset or can be called negative class of the EPs and dataset D2 will be called target dataset or can be called positive class. Given ρ> 1 as growth rate threshold, an itemset X is said to be an ρ-emerging pattern (ρ-EPs or simply EPs) from D1 to D2 (sometimes states as an EPs in/of D2) if GrowthRate(X)≥ρ.

$$\text{GrowthRate(X)} = \begin{cases} 0 & if\ \text{supp}_{\text{D1}}(\text{X}) = 0 \text{ and } \text{supp}_{\text{D2}}(\text{X}) = 0 \\ \infty & if\ \text{supp}_{\text{D1}}(\text{X}) = 0 \text{ and } \text{supp}_{\text{D2}}(\text{X}) \neq 0 \\ \dfrac{\text{supp}_{\text{D2}}(\text{X})}{\text{supp}_{\text{D1}}(\text{X})} = \dfrac{\frac{count_{D2}(X)}{|D2|}}{\frac{count_{D1}(X)}{|D1|}} & otherwise \end{cases} \quad (2.2)$$

where : ∞ = infinity, when $\frac{n}{0}$, Jumping EPs (JEPs)

suppD1(X) = support in dataset D1 containing itemset X (equation 2.1)

suppD2(X) = support in dataset D2 containing itemset X (equation 2.1)

Jumping EPs (JEPs) are special EPs and also special discriminant rule whose supports increase abruptly from zero support in one dataset to non-zero support in another. JEPs is EPs with infinite (∞) growth rate value whose support is zero in dataset D1 ($\text{supp}_{\text{D1}}(\text{X})=0$) and support is non-zero in dataset D2 ($\text{supp}_{\text{D2}}(\text{X})\neq0$). For discovering JEPs, HORIZON-MINER algorithm is used to find the large border (horizontal border) of all itemsets with non-zero support and MBD-LLBORDER is used to find JEPs using the two large borders derived by HORIZON-MINER as inputs [79]. Tree-based algorithms for computing JEPs that are 2 until10 times faster than previous methods, which combination two novel features [75] :

1) Tree-based data structure for storing the raw data which is similar to Frequent Pattern (FP-tree) [73,74].
2) Development of a mining algorithm operating directly on the data contained in the trees.

2.4.2. EPs algorithms

There are a lot of EPs algorithms and the previous algorithms border-based MBD-$\text{LL}_{\text{BORDER}}$ algorithm and ConsEPMiner. The EP mining with border-based MBD-$\text{LL}_{\text{BORDER}}$ algorithm avoid the long process naive algorithms to get the counts of all itemsets in a large collection of candidates, by manipulating only borders of some two collections and derive all EPs whose support satisfies a minimum support threshold in dataset D2 [79]. The border-based MBD-$\text{LL}_{\text{BORDER}}$ algorithm discovers all EPs by calling differential procedure B_{ORDER}-D_{IFF} algorithm in multiple numbers of times. ConsEPMiner is a constraint based EPs

Miner that utilize two types of constraints (support and growth rate treshold) which efficiently mining EPs and use another constraint called growth rate improvement to eliminate the uninteresting EPs [68]. Beside three external constraints (support, growth rate and growth rate improvement), there are another three inherent constraint which are not user given, namely same subset support, top growth rate and same origin.

The EPs algorithms has been extended to classification called EP-based classifier where the process of finding a set of models can describe and distinguish between two or more data classes or concepts. For handling classification where distinguish more than 2 classes then each instance in dataset D is associated with p class labels: C_1,C_2, ...,C_p and partition dataset D into p sets: D_1,D_2, ...,D_p with D_i containing all instances of class C_i.

There are many EP-based classifier and they are:

1) Classification by Aggregating EPs (CAEP).

   Its first application for EPs classification, employs ConsEPMiner algorithm and has three steps [84]:

   1.a) For each class C, all the EPs meeting some support and growth rate thresholds, from the opponent set of all none-C instances to the set of all C instances.

   1.b) Aggregating the power of the discovered EPs for classifying an instance s. Aggregating differentiating score for each class C by summing the differentiating power of all EPs of class C that occur in instance s.

   1.c) Normalizing score for class C by dividing it by some base score of the training instances of class C.

   The accuracy and performance CAEP can be improved with Score Behaviour Knowledge Space (SBKS) which to record the behaviour of training data on scores to make final classification decision. SBKS is an m-dimensional space where each dimension corresponds to the score of the class [66].

2) Information-based approach for classification byaggregating EPs (iCAEP).

   A variant of CAEP and compare to CAEP, iCAEP has better predictive accuracy and shorter time for training and classification [67].

3) The Decision making by EPs (DeEP).

   DeEP is instance-based classifier which makes decisions through EPs [65,72,81]. As a lazy EP-based classifier, instance-based approach creates remarkable reduction on both volume (the number of instances) and dimension (the number of attributes) of the training

data. DeEPs have advantages on accuracy, speed and dimensional scalability over CAEP [84] and JEP-Classifier [83].

DeEPs need three main steps to determine the class of a test instance:

3.a) Discovering border representation of EPs.

The step aims to learn discriminating knowledge from training data, reducing the data and discovering all JEPs. Assume we have classification set $D_p=\{P_1,...,P_m\}$ of positive instances and set $D_n=\{N_1,...,N_n\}$ of negative instances.

3.b) Selecting the more discriminating EPs.

Since the number of JEPs is usually large then the most general JEPs among all JEPs will be reduced. By the most general JEPs is mean that the proper subsets are not JEPs anymore.

3.c) Determining collective scores based on the selected EPs for classification.

Determine the collective score of T instance for any specific class C by aggregating the supports of the selected EPs in class C using compact summation method.

4) Bayesian Classification by EPs (BCEP).

As a hybrid of the EP-based classifier and Naive Baiyes (NB) classifier[71,78] is superior than CAEP. There are 2 kinds of interesting EPs when mining with BCEP and they are :

4.a) Essential EPs (eEP), are EPs with very large growth rate (typically more than 1000), enough(large) supports in the target class (usually threshold 1%) and that are contained in the left bound of the border representing EPs collection. Large growth rate show sharp discriminating power, large supports show enough coverage on the training dataset which EPs are more resistant to noise

4.b) Essential JEPs (EJEP) [70], are subset of JEPs which removing JEPs that contain noise and redundant information.

BCEP utilize tree-based algorithm [75] to efficiently mine the complete eEP and EJEP for each class.

5) Constrained Emerging Patterns (CEP).

CEP the same with border based MBD-LLBORDER algorithm [83] to find itemsets which support $\geq\alpha$ threshold in target (D2) dataset and support $\leq \beta$ threshold in background (D1) dataset [76,77]. CEP mining can be accomplished by an extension of JEP mining in two steps. Step 1 is to represent border based algorithm where one border represent target (D2) dataset with support $\geq\alpha$ threshold and the other border represent background (D1) dataset with support $\geq \beta$ threshold. Method for mining JEPs can be applied, once the borders are computed to gain the desired patterns in the next step [77].

Thereafter, step two is to mine the CEP by operating on the relevant borders. When β=0, CEP become JEP and when β>0 will have greater robustness CEP. Pair-wise classification strategy is used to mine CEP with more than two datasets, where each of dataset will be treated as target class and will be compared with unioning other datasets. For example CEP for dataset D1 are found by comparing D1 against the background dataset D2∪ D3∪... Dn. The CEP for dataset D3 are found by comparing D3 with respect to the background dataset D1∪D2∪D4∪....∪Dn etc.

For handling JEP, there are many EP-based classifier and they are :

1) JEP-Classifier (JEP-C).

JEP-Classifier is JEPs classification which partially influenced by CAEP and uses exclusively JEPs. JEP-Classifier uses datasets with more than two classes in an ordered way with pair-wise feature concept [83]. JEP-Classifier utilize border based MBD-LL_{BORDER} algorithm to discover border of all JEPs in order to identify the most expressive JEPs. Border based MBD-LL_{BORDER} algorithm is used to find JEPs in large databases and using semi naive $_{JEP}$PRODUCER algorithm to find JEPs in small databases. The most expressive JEPS is the most frequency JEPs with large support that build accurate classification. The most expressive JEPs is the left bounds of the border.

2) JEP spaces.

JEP space with respect to target (D2) dataset and background (D1) dataset is defined as the set of all JEPs from background(D1) to target (D2) datasets ($\frac{\text{supp}_{D_2}(X)}{\text{supp}_{D_1}(X)}$). JEP space is collection where element only occurs in target dataset but not in background dataset [72]. JEP space satisfies the property of convexity and can be represented by two bounds, left bound and right bound, consisting respectively of the most general JEPs and the most specific JEPs [72,82]. There are 3 border operations for algorithm maintaining JEP spaces and they are :

2.a) Border difference (-).

Border difference is similar with MBD-LL_{BORDER} algorithm [79] and using B_{ORDER}-D_{IFF} algorithm [79] with a slight different in output. The same like inputs for $_{JEP}$PRODUCER algorithm [83], JEP space is represented with two horizontal borders (horizontal spaces or convex space) from datasets D1 of positive instances denoted $<\{\emptyset\},R_1>$ and D2 of negative instances denoted $<\{\emptyset\},R_2>$. In other words, JEP space is represented with border <L,R> which have 2 bounds, they are Left bound/the most general JEPs/positive instances/$<\{\emptyset\},R_1>$ and Right bound/the most

specific JEPs/negative instances/<{Ø},R_2>. Horizontal border or horizontal space is non-zero support itemsets in the dataset. JEP space to D1 and D2 is present the set difference [{Ø},R_1] - [{Ø},R_2], where is subtracting all non-zero support itemsets in dataset D2 from all non-zero support itemsets in dataset D1.

2.b) Border union (∪).

Border union is union of old JEP space and some JEP space created by new data. Suppose old JEP spaces D_1 and D_2 are positive and negative instances respectively. Assume a set i1 ($_iR_1$) of new positive instances are inserted then JEP space (D_1+i1) and D_2 or new JEP space is the union of the previous JEP space and a JEP space associated with i1. Insertion of new Left bound/the most general JEPs/positive instances/<{Ø},R_1> has set: ([{Ø},R_1] ∪ [{Ø},$_iR_1$]) - [{Ø},R_2] = ([{Ø},R_1] - [{Ø},R_2]) ∪ ([{Ø},$_iR_1$]) - [{Ø},R_2])

2.c) Border intersection (∩).

Border intersection is intersection of old JEP space and some JEP space created by new data. Suppose old JEP spaces D_1 and D_2 are positive and negative instances respectively. Assume a set i2 ($_iR_2$) of new negative instances are inserted then JEP space D_1 and (D_2 +i2) or new JEP space is the intersection of the previous JEP space and a JEP space associated with i2. Insertion of new Right bound/the most specific JEPs/negative instances/<{Ø},R_2> has set: [{Ø}, R_1] – ([{Ø},R_2] ∪ [{Ø},$_iR_2$])=([{Ø},R_1] - [{Ø},R_2]) ∩ ([{Ø}, R_1]) - [{Ø},$_iR_2$])

3) Essential JEP (EJEP) and EJEP-Classifier (EJEP-C).

EJEP is discrimination between two classes and EJEP-Classifier (EJEPC-C) is classification for more than two classes by aggregating EJEPs with adopting pair-wise features concept. EJEP-C uses two parameters: the minimum support threshold and the percentage of top ranking items used for mining EJEPs. EJEP uses tree structure called Pattern-tree (P-tree) algorithm to mine EJEPs and the method advantage is a single-scan algorithm which efficiently mine EJEPs of both data classes (from D1 to D2 and from D2 to D1) at the same time [70,78]. Whilst border-based and ConsEPMiner algorithms will call the algorithm twice using target classes D2 and D1 separately.

2.5. Critical Analysis of Literatures and New Approach

AOI and EP have been recognized as powerful mining technique in order to extract important knowledge from data. For level data processing, AOI usually are represented at high abstraction level in the concept hierarchy but EP concerns to distinguish properties at low conceptual level. AOI is recognized as a powerful mining technique since has been tested successfully against large relational database [44] and can learn different kinds of rules [39]. Process of generalization steps in AOI which produce high level data based on concept hierarchy as background knowledge, become the typical strength of AOI. Moreover, EP is recognized as a powerful mining technique to discriminate datasets [75,79]. Growth rate as ratio of the supports in one dataset to another dataset is justification for powerful discrimination, become the typical strength of EP. The main purpose to combine AOI and EP become AOI-HEP is to use typical strength of AOI and EP.

AOI-HEP unite the powerful AOI and EP by applying growth rate as a standard function in EP and using concept hierarchy as background knowledge in AOI. AOI-HEP apply growth rate for ratio of the supports at the same or different high level itemsets instead of the same low level itemsets as used in EP. Mining high level data with EP was ever proposed by:

1) Using brute-force approach with optimisations to mining generalised EPs called GTree algorithm[90]. GTree algorithm was influenced with border-based algorithm and a pattern is considered an EP if superset EP in Left border and has leaf-level as subset in Right border[90].
2) Using local-recording algorithm to hide sensitive information that is EPs in dataset for Privacy Preserving Data Publishing (PPDP) purposes and recoding is a process of grouping existing values to some new generalized values from concept hierarchies [104]. The algorithms is measuring the reduction of growth rate EPs with recording all attributes in frequent itemsets and hide all EPs with a minimal distortion in frequent itemsets [104].

Therefore, AOI-HEP as a new technique has powerful distinguishing features between datasets at high abstraction level. AOI-HEP can mines frequent and similar patterns. AOI-HEP proved to learn knowledge rules such as discriminant rule from frequent and similar patterns. AOI-HEP is influenced with EP which is recognized closely relate to frequent pattern[89] and in EP, patterns will be recognized as EP if patterns have a high support (frequent) in one class and low support (infrequent) in other one[71,89].

EP was proposed with border-based algorithm [79], which influences most other EP mining algorithms such as CAEP[84], CAEEP [100], DeEP[65], BCEP[78], CEP[77], JEP-C[83], JEP space[82] as EP-based classifier[103] algorithm, KTDA[89], PCL [102] and GTree [90] algorihtms. Border-based algorithm influences many other algorithms by implementing the border-based algorithm in their algorithms. Border-based algorithm define border <L,R> where L is the sets of the minimal itemsets (superset or the most general EPs) and R is the sets of the maximal itemsets (subset or the most specific EPs) [71,72]. Border-based algorithm defines border <L,R> where each element L is a subset of some elements in Rand each element of R is a superset of some elements in L [79].

AOI-HEP is not influenced by border-based algorithm and the differences between AOI-HEP and border-based algorithm are :

1) Border-based algorithm which influence EP-based classifier such as CAEP,DeEP, BCEP, JEP-C can do classification task by learning more than two datasets. Whilst AOI-HEP not yet implemented, but future research needs to be undertaken to extend AOI-HEP ability by learning more than two datasets in order to mine not only classification task, but other knowledge rules such as association rule, data evolution regularities and cluster description rules. Moreover, other future research in order to find interested HEP in AOI-HEP, the discovery is not just only from D1 to D2 datasets, but will be extended from D2 to D1 datasets as refer in the third step of discovery the interesting EPs in DeEP [65,72,81].
2) AOI-HEP discovers high level pattern instead of low level pattern in border-based algorithm.
3) AOI-HEP will have less High level EP (HEP) because of discovery in high level as cartesian product between rulesets. Whilst border-based algorithm will have the huge number of EPs because of discovery in low level.
4) AOI-HEP uses similarity hierarchy level and value between attributes in rules and growth rate threshold to reduce the number of HEP. However, border-based algorithm uses border and growth rate threshold to reduce the huge number of EPs, which consists the minimal and maximal EPs [80].
5) AOI-HEP discovers with the same or different itemset, but border-based algorithm with the same itemset.
6) There is no infinite (∞) growth rate or Jumping High level EP (JHEP) in AOI-HEP since all rules (high level pattern) in the ruleset have number of instances. Whilst in border-

based algorithm there is Jumping EP (JEP) where EP which is having infinite (∞) growth rate with support is 0 in one dataset and more than 0 in the other dataset [79].

7) For finding frequent pattern, AOI-HEP uses similarity hierarchy level (if different level and has subsumption LV=0.4 or 0.5) between attributes where are totally subsumed in Total Subsumption HEP (TSHEP) or frequent subsumption in Subsumption Overlapping HEP (SOHEP). In contrast, KTDA algorithm uses four parameters to find frequent patterns i.e. minimal EP growth rate, minimal EP support in target class, minimal EP support increase per iteration and reduce discovered EPs[89]. Whilst the border-based MBD-LL_{BORDER} algorithm [79,84] uses support which satisfies a minimum support threshold as frequent pattern or large border in target class.
8) Support for pattern is counted if the pattern is frequent or similar pattern in AOI-HEP, but in border-based algorithm, support for pattern is counted first then check if satisfies a minimum support threshold as frequent pattern or large border in target class[71].
9) AOI-HEP algorithm is not use support threshold, but border-based algorithm uses support threshold to mine frequent or large support in target class[71].
10) Superset and subset terms are used in AOI-HEP to discriminate frequent pattern between superset and subset rules, whilst in border-based algorithm Left border as superset(the most general) and Right border as subset (the most specific).
11) Strong/sharp discriminating power for both of algorithms is expressed by large support in target class that make large growth rate [71]. However, different with border-based algorithm, there is one requirement for strong/sharp discriminating power in AOI-HEP where subsumption similarity attributes (LV value=0.4 or 0.5) are the same in TSHEP or the same frequent in TSHEP or SOHEP.

AOI-HEP has similarity with the Decision making by EPs (DeEP) algorithm in term of reduction the number of instances and attributes, but DeEP as influenced by border-based algorithm which is discovering low level data. DeEP uses instance-based approach to reduce on both the volume (the number of instances) and the dimension (the number of attributes) of the training data[65,72,81] . In contrast, AOI-HEP uses AOI characteristic rule to reduce the number of instances and attributes. Meanwhile, DeEp reduces the number of attributes with intersection operation using neighbourhood-based intersection method which determine continuous attribute value are relevant to a given testing instance [65,72,81]. Whilst AOI-HEP reduce the number of attributes with attribute removal as second step of AOI generalization process and the number of attributes will depend on the number of concept hierarchies in AOI. Moreover, DeEP reduces the number of instances with selecting the

maximal itemsets from intersection operation [65,72,81]. Meanwhile, AOI-HEP reduces the number of instances with AOI generalization and the number of instances will depend on rule threshold in AOI.

AOI-HEP that combines between AOI and EP techniques has similarity with previous researches in terms of combining EP with other techniques such as:

1) Decision tree technique CART-based method is used to discover relevant EPs for classification that consist six steps where CART trees replace border-based algorithm function [91]. A user-friendly tool KTDA system is CART-based method implementation with some extensions and improvements [89].
2) EPs are used to construct weighted Support Vector Machines (weightedSVMs) by calculating numeric scores for each instances based on EPs then use scores to assign weights for training [92].
3) Weighting the training instances with EP for the class memberships for fuzzy SVM classifier[93].
4) Generalize decision tree and weighted classes assigned to the training data instances and discovering weights for the training instances for decision tree[94]
5) Proposed EP-weighting scheme as visual word weighting scheme by finding EPs of visual keywords in training dataset and adaptive weighting assignment is performed for each visual word according to EPs[95].
6) Proposed Contrast Pattern tree (CP-tree) was inspired by FP-Tree which mining frequent patterns without candidate generation, for mining Strong Jumping EPs (SJEPs), Noise-tolerant EPs (NEPs) and Generalized Noise-tolerant EPs (GNEPs) for classification task [101].
7) EP and Decision Tree are used in rare-class classification (EPDT) where EP is used to improve the quality of rare-case classification. [96,99].
8) EP is used in rare-class classification (EPRC) with three stages such as generating new undiscovered EPs for the rare class, pruning low support EPs and increasing the supports of the rare-class EPs [97,99].
9) Expanding the training data space (ETDS) using EPs and genetic methods (GM) with four methods such as generation by superimposing EPs, generation by Crossover, generation by Mutation and generation by Mutation and EPs [98].

2.6. Conclusion

In this chapter two, we have an understanding about data mining as a part of KDD. Two data mining techniques AOI and EP are combined to produce a novel AOI-HEP (Attribute Oriented Induction High level Emerging Pattern). AOI is a data mining technique which mines high level patterns whilst EP algorithm is concerned with emerging low level patterns. AOI-HEP is a combination the strength of AOI and EP respectively and there were previous researches that combine EP with other techniques. The strength of AOI is able to learn different kind of rules and using process generalization steps to produce high level data. Meanwhile, the strength of EP is recognized as powerful discrimination using growth rate as supports ratio between datasets. AOI-HEP do not implement border-based algorithm and there are some differences with it. Moreover, AOI-HEP has similarity with the Decision making by EPs (DeEP) algorithm in term of reducing on both the number of instances and attributes of the training data. In the next chapter, the proposed novel idea AOI-HEP as combination AOI and EP will be presented.

Chapter 3: AOI-HEP Mining Framework

3.1. Introduction

The aim of this chapter is to describe Attribute Oriented Induction High level Emerging Pattern (AOI-HEP) framework. AOI-HEP combines two data mining techniques i.e. Attribute Oriented Induction (AOI) and Emerging Patterns (EP). Section 3.2 presents the AOI-HEP framework and describes the AOI-HEP algorithm. Moreover, section 3.3 describes representation rules and rulesets, TSHEP, SOHEP and TOHEP definitions. Section 3.4 presents HEP algorithm as implementation part of AOI-HEP framework. Furthermore, section 3.5 defines metric similarity function C{ R_i^1 , R_j^2 } for attribute comparison by summing attribute comparisons as some type of number similarity between the two rules from different rulesets. This then determines HEP patterns either as TSHEP, SOHEP or TOHEP. Besides this, HEP can be categorized as frequent or similar patterns. Meanwhile, section 3.6presents the growth rate function GR{ R_i^1, R_j^2 } for HEP by using the same or different high level itemset instead of the same low level itemset as used in EP. Finally, the conclusion of this chapter is given in section 3.7.

3.2. AOI-HEP framework

As mention in section 2.5 at chapter 2, AOI-HEP unites the powerful of AOI and EP mining techniques by applying AOI characteristic rule algorithm and improvement EP growth rate respectively. AOI-HEP framework in figure 3.1 shows unification of AOI and EP mining techniques where AOI characteristic rule algorithm is influenced by AOI mining technique, whilst High level Emerging Pattern (HEP) algorithm is influenced by EP mining technique.

AOI characteristic rule algorithm is run twice with input two datasets D1 and D2 which are horizontal partitions from one dataset where each dataset is divided into two sub datasets based on learning high level concept in one of chosen dataset attribute. Concept the horizontal partitions will be explained in section 4.2 at chapter 4 where the horizontal

partition is implemented to each dataset of four datasets experiment from the UCI machine learning repository. Moreover, AOI characteristic rule algorithm has concept hierarchy and attribute and rules thresholds. Concept hierarchy is background knowledge for data generalization which generalise from low level data into high level data. Whilst attribute and rules thresholds are thresholds to eliminate distinct attributes and tuples until they are less or equal than attribute and rules thresholds respectively [30]. Attribute and rules thresholds can be seen at screen display for AOI-HEP application in figure 4.1 at chapter 4. Furthermore, in AOI-HEP framework, AOI characteristic rule algorithm has output rulesets $\{ R_i^1 \}$and $\{ R_j^2 \}$ from datasets D1 and D2 respectively. Number of rules in rulesets $\{ R_i^1 \}$ and $\{ R_j^2 \}$ will be decided by rules threshold for AOI characteristic rule algorithm as will be explained in section 3.3. Experiment for Rulesets $\{ R_i^1 \}$ and $\{ R_j^2 \}$ as final ruleset result from AOI characteristic rule algorithm, can be seen in section 4.3 at chapter 4. Finally, Rulesets $\{ R_i^1 \}$ and $\{ R_j^2 \}$ are input for High level Emerging Pattern (HEP) algorithm in AOI-HEP framework.

Meanwhile, High level Emerging Pattern (HEP) algorithm in AOI-HEP framework as shown in figure 3.1 has two functions i.e. similarity function $C\{ R_i^1, R_j^2 \}$ and growth rate function $GR\{ R_i^1, R_j^2 \}$. The $C\{ R_i^1, R_j^2 \}$ function is a metric similarity function which applies cartesian product between rulesets $\{ R_i^1 \}$and $\{ R_j^2 \}$, and eliminates the cartesian product by determining type of HEP i.e. either TSHEP, SOHEP or TOHEP as explained in section 3.5.Determining HEP types is applied by summing categorization of attribute comparison value and hierarchy level based on subsumption and overlap thresholds as explained in section 3.5 as well. Moreover, The $C\{ R_i^1, R_j^2 \}$ function has five controls which grouped into two groups where the first group is frequent and similar controls and the next group is TSHEP,SOHEP and TOHEP controls. These five controls can be seen at screen display for AOI-HEP application in figure 4.1 at chapter 4. The frequent and similar controls are radio buttons or option buttons where condition is true then the mining process either as frequent or similar pattern as explained in section 3.5.2 and 3.5.3 respectively. The experiments for mining frequent and similar patterns will be explained in section 4.4 and 4.6 at chapter 4 respectively. Meanwhile, TSHEP,SOHEP and TOHEP controls are checklists to filter out type of HEP based on where condition is true. Mining TSHEP, SOHEP and TOHEP will be

explained in section 3.5.1 and the experiments for mining TSHEP, SOHEP and TOHEP will be explained in section 4.3 at chapter 4.

Moreover, the GR{ R_i^1, R_j^2 } function is ratio of the supports between rulesets { R_i^1 }and { R_j^2 } has GrowthRate threshold which eliminates the type of HEP which are outputs from C{ R_i^1, R_j^2 } function with growth rate less equal than GrowthRate threshold as explained in section 3.6. The growth rate threshold can be seen at screen display for AOI-HEP application in figure 4.1 at chapter 4. Meanwhile, the experiments for implement the GR{ R_i^1, R_j^2 } function will be explained in section 4.3 at chapter 4.

Finally, AOI-HEP framework in figure 3.1 has four outputs i.e. GrowthRate, HEP pattern, SLV and HEP pattern% where GrowthRate is output from GR{ R_i^1, R_j^2 } function and the others are outputs from C{ R_i^1, R_j^2 } function. The experiments for AOI-HEP framework outputs can be seen between tables 4.9 to 4.15 in section 4.3 at chapter 4. GrowthRate output is ratio between two rulesets, whilst HEP pattern output is a summing categorization of attribute comparison values and hierarchy levels (SLV value in equation 3.1) whose options are between total subsumption HEP (TSHEP - those rules that are completely subsumed), subsumption overlapping HEP (SOHEP - those rules that overlap and are subsumed) and total overlapping HEP (TOHEP - those rules that are completely overlapping) [85]. Meanwhile SLV output determines the similarity of patterns on the same hierarchy level based on the attribute similarity hierarchy levels in equation 3.1. Furthermore, HEP pattern% output is the percentage of composition attributes comparison or LV value in equation 3.1.

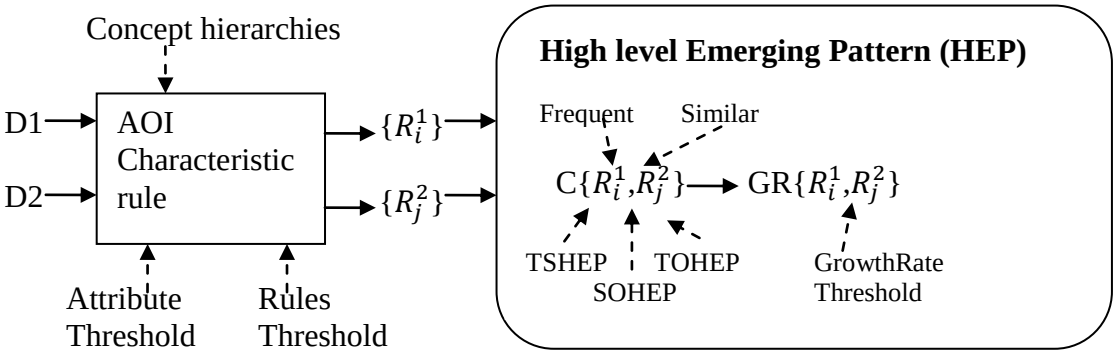


Figure 3.1. AOI-HEP Framework

3.3. HEP definitions

For High level Emerging Patterns (HEP), let D1 and D2 be horizontal partitions of some dataset $D^x = \{A_1,...,A_p\}$ with p attributes $1 \le i \le p$ and $1 \le x \le 2$ .Rulesets $\{R_i^1\}$ and $\{R_j^2\}$ from datasets D1 and D2 are represented as $R^x = \{r_1^x, r_2^x, ..., r_n^x\}$ in figure 3.2. In figure 3.2 each ruleset Rx consists of n rules where n ≤ R.Thr, a rules threshold. Each rule in a ruleset Rx is represented by attributes $r_n^x = \{A_1^x, A_2^x, A_{...}^x, A_m^x, |r_n^x|\}$, where $|r_n^x|$ is number of tuples forming the rule and m is the number of attributes in a ruleset as in equation 3.1. Figure 3.2 shows the representation of rulesets $R^x = \{r_1^x, r_2^x, ..., r_n^x\}$ vertically where $r_n^x \in R^x$ and each rule $r_n^x = \{A_1^x, A_2^x, A_{...}^x, A_m^x, |r_n^x|\}$, horizontally where $A_m^x \in r_n^x$. Based on the previous paragraph and figure 3.2, as an example we have used rule r_1^1 in ruleset 1 and rule r_1^2 in ruleset 2. $A_m^1 \in r_1^1$ where all attributes A_m^1 are member of rule r_1^1 in ruleset 1 and $A_m^2 \in r_1^2$ where all attributes A_m^2 are member of rule r_1^2 in ruleset 2. For example, if there are four attributes (m=4 in equation 3.1) then rule $r_1^1 = \{A_1^1, A_2^1, A_3^1, A_4^1, |r_1^1|\}$ and rule $r_1^2 = \{A_1^2, A_2^2, A_3^2, A_4^2, |r_1^2|\}$.

$$
\begin{aligned}
D^x \rightarrow R^x \rightarrow r_1^x &= \{A_1^x, A_2^x, A_{...}^x, A_m^x, |r_1^x|\} \\
r_2^x &= \{A_1^x, A_2^x, A_{...}^x, A_m^x, |r_2^x|\} \\
r_{...}^x &= \{A_1^x, A_2^x, A_{...}^x, A_m^x, |r_{...}^x|\} \\
r_n^x &= \{A_1^x, A_2^x, A_{...}^x, A_m^x, |r_n^x|\}
\end{aligned}
$$

Figure 3.2. Representation rules and rulesets

3.3.1. TSHEP definition

For Total Subsumption HEP (TSHEP) we say rule r_1^1 is totally subsumed by rule r_1^2 if $A_{[1..m]}^1 \subseteq A_{[1..m]}^2$ then $r_1^1 \subseteq r_1^2$. This means rule r_1^1 is TSHEP by rule r_1^2 ($r_1^1 \subseteq r_1^2$, rule r_1^1 is a subset of rule r_1^2) if each attribute in A_m^1 is subsumed by each attribute in A_m^2 ($A_{[1..m]}^1 \subseteq A_{[1..m]}^2$). Based on example four attributes for rules r_1^1 and r_1^2 , if each attribute in A_m^1 is subsumed by each attribute in $A_m^2 \{A_1^1 \subseteq A_1^2, A_2^1 \subseteq A_2^2, A_3^1 \subseteq A_3^2, A_4^1 \subseteq A_4^2\}$ then $r_1^1 \subseteq r_1^2$.

3.3.2. TOHEP definition

For Total Overlapping HEP (TOHEP) we say ruler_1^1 totally overlaps with ruler_1^2 if $A^1_{[1..m]} \cap A^2_{[1..m]}$ then $r_1^1 \cap r_1^2$. This means ruler_1^1 is TOHEP with ruler_1^2 ($r_1^1 \cap r_1^2$, ruler_1^1 is overlap with ruler_1^2) if each attribute in A^1_m is overlap with each attribute inA^2_m ($A^1_{[1..m]} \cap A^2_{[1..m]}$). Based on example four attributes for rules r_1^1andr_1^2, if each attribute in $A^1_m$is overlap with each attribute inA^2_m $\{A^1_1 \cap A^2_1, A^1_2 \cap A^2_2, A^1_3 \cap A^2_3, A^1_4 \cap A^2_4\}$ then $r_1^1 \cap r_1^2$.

3.3.3. SOHEP definition

For Subsumption Overlapping HEP (SOHEP) we say rule r_1^1 is subsumed by and overlaps with ruler_1^2: if $A^1_{[1..m1]} \subseteq A^2_{[1..m1]}$ and $A^1_{[m1+1..m]} \cap A^2_{[m1+1..m]}$ then $r_1^1 \subset r_1^2$ and $r_1^1 \cap r_1^2$. This means rule$r_1^1$is SOHEP with ruler_1^2($r_1^1 \subset r_1^2$,ruler_1^1 is a proper-subset of rule r_1^2) and ($r_1^1 \cap r_1^2$, ruler_1^1 overlaps with rule r_1^2), if some attributes from 1 to m1 in $A^1_m$are subsumed by some attributes from 1 to m1 inA^2_m($A^1_{[1..m1]} \subseteq A^2_{[1..m1]}$) and if some attributes from m1+1 to m in $A^1_m$are overlap with some attributes from m1+1to m inA^2_m ($A^1_{[m1+1..m]} \cap A^2_{[m1+1..m]}$), where m1 is the number subsumption attribute and m is the number of attributes in a ruleset as in equation3.1. Based on example four attributes for rules r_1^1andr_1^2, if the first two attributes in $A^1_m$are subsumed by the first two attributes in $A^2_m$and certainly the last two attributes in A^1_m are overlap with the last two attributes in $A^2_m\{A^1_1 \subseteq A^2_1, A^1_2 \subseteq A^2_2, A^1_3 \cap A^2_3, A^1_4 \cap A^2_4\}$then $r_1^1 \subset r_1^2$and $r_1^1 \cap r_1^2$.

3.4. HEP algorithm

Figure 3.3 shows the HEP algorithm as part of AOI-HEP framework in figure 3.1. The HEP algorithm has inputs such as rulesets R_i^1 and R_j^2 , TSHEP, SOHEP, TOHEP, GR_threshold, num_attr, |D2| , |D1|, Frequent and Similar. The HEP algorithm inputs are in accordance with inputs for HEP in AOI-HEP framework figure 3.1 where for HEP in figure 3.1 there are rulesets $R_i^1$and R_j^2 inputs, TSHEP, SOHEP, TOHEP, Frequent and Similar for $C\{R_i^1, R_j^2\}$function, GR_threshold for $GR\{R_i^1, R_j^2\}$function. The GR_threshold threshold has

default value 0 and maximum value 100. TSHEP, SOHEP and TOHEP will be explained in section 3.5.1. Moreover, num_attr input is the number attributes in rulesets R_i^1 and R_j^2 as m in equation 3.1. The outputs from HEP algorithm are in accordance with the HEP outputs shown in figure 3.1 and they are GrowthRate, HEP pattern, SLV and HEP pattern%. The outputs are printed in line 17 in HEP algorithm.

HEP algorithm

Input :$\{R_i^1\}$, $\{R_j^2\}$, TSHEP, SOHEP, TOHEP, GR_threshold, num_attr,|D2|,|D1|, Frequent, Similar

Output :growth rate, HEP pattern, SLV, HEP pattern%

1. While(noAllANY(R_i^1))
2. While (noAllANY(R_j^2))
3. SLV=0, over=0, subs=0, F=0,S=0
4. for x=1 to num_attr
5. If R_i^1[x]== R_j^2 [x] and R_i^1[x]==”ANY” SLV=SLV+2.1, over=over+1,S++
6. If R_i^1[x]== R_j^2[x] and R_i^1[x]!=”ANY” SLV=SLV+2, over=over+1
7. If R_i^1[x]!= R_j^2[x] and R_j^2[x] subsump by R_i^1[x] SLV=SLV+0.4, subs=subs+1
8. If R_i^1[x]!= R_j^2[x] and R_i^1[x] subsump by R_j^2[x] SLV=SLV+0.5, subs=subs+1,F++
9. subs_=subs/num_attr*100
10. over_=over/num_attr*100
11. If TSHEP and/or SOHEP and/or TOHEP
12. If subs>0 and over==0 and TSHEP HEP pattern=”TSHEP”, HEP pattern%=subs_
13. If subs>0 and over>0 and SOHEP HEP pattern=”SOHEP”, HEP pattern%=subs_+over_
14. If subs==0 and over>0 and TOHEP HEP pattern=”TOHEP”, HEP pattern%=over_
15. growth rate=(R_j^2[x+1]/|D2|) / (R_i^1[x+1]/|D1|)
16. If growth rate > GR_threshold and/or (Frequent and F==x or F==x-1) and/or(Similar and S<x-1)
17. Print growth rate, HEP pattern,SLV,HEP pattern%

Figure 3.3. HEP algorithm

In the HEP algorithm, line number 1 and 2 are used to exclude rule with ANY values in all attributes in rulesets R_i^1 and R_j^2 respectively. Rules with ANY values are less meaningful and do not offer meaningful interpretation. $C\{R_i^1, R_j^2\}$ and $GR\{R_i^1, R_j^2\}$functions in figure 3.3 are shown between line number 3 and 14, and between line number 15 and 16 in HEP algorithm respectively.

3.5. Metric similarity

This section presents the metric similarity function $C\{R_i^1, R_j^2\}$between rulesets $\{R_i^1\}$ and $\{R_j^2\}$. As mention in section 3.2, the $C\{R_i^1, R_j^2\}$function is a metric similarity function which apply cartesian product between rulesets R_i^1 and R_j^2, and eliminate the cartesian product by determining type of HEP. The determining type of HEP is applied by summing categorization of attribute comparison value and hierarchy level based on subsumption and overlap thresholds. To derive similarity hierarchy level value (SLV) in the HEP algorithm, firstly, we determine categories of attribute values between the rulesets as shown in figure 3.4. The categorization is based on similarity hierarchy level and the values shown in equation 3.1 as LV. Secondly, by summing the attribute categorizations or LV values, we get SLV (equation 3.1) as the similarity between the two rules. The two steps described above are shown between line numbers 4 and 8 in the HEP algorithm of figure 3.3.

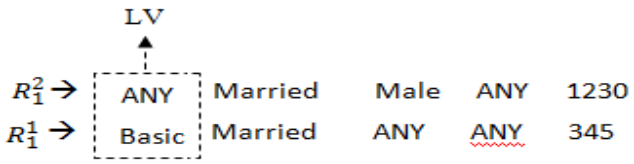


Figure 3.4. Comparing rule 1of ruleset 2 $\{R_i^2\}$ and rule 1 of ruleset 1 $\{R_i^1\}$

$$SLV=\sum_{i=1}^{m} LV_i \quad (3.1)$$

where:

SLV=similarity value based on the similarity of attributes hierarchy level and values

m= number of attributes in a ruleset, where m > 1

(number of attributes in concept hierarchies - 1)

i=attribute position

LV_i = categorization of attributes comparison based on similarity hierarchy level and values, and the options are

1. If hierarchy level is different and the attribute in rule of ruleset R2 is subsumed by the attribute in rule of ruleset R1, LV=0.4.
2. If hierarchy level is different and the attribute in rule of ruleset R1 is subsumed by the attribute in rule of ruleset R2, LV=0.5.
3. If hierarchy level and values are the same and the attributes values are not ANY, LV=2.

4. If hierarchy level and values are the same and the attributes values are ANY, LV=2.1.

The four categorization of attribute comparisons or LV in equation 3.1 is based on two main categorizations i.e. subsumption (LV=0.4 or LV=0.5) and overlapping (LV=2 or LV=2.1). Thus, the attributes will be categorized as subsumption when attributes comparison has different hierarchy level and value (LV=0.4 or LV=0.5). On the other hand, the attributes will be categorized overlapping when comparison between attributes has the same hierarchy levels and values(LV=2 or LV=2.1). For each LV option values 0.4,0.5,2 and 2.1 are user defined number, where option numbers 0.4 and 0.5 as values for subsumption categorization (minimum categorization) and option numbers 2 and 2.1 as values for overlapping categorization (maximum categorization). LV=0.4 is minimum value for subsumption categorization and if ruleset R2 is subsumed by ruleset R1. On the other hand LV=0.5 is maximum value for subsumption categorization and if ruleset R1 is subsumed by ruleset R2. LV=2 is minimum value for overlapping categorization and if the attributes values are not ANY, on the other hand LV=2.1 is maximum value for overlapping categorization and if the attributes values are ANY. Finally, LV=0.4 and LV=2.1 are taken as the minimum and maximum values of LV values respectively.

After the similarity between the two rules (SLV) has been derived, then we can determine type of HEP between TSHEP, SOHEP or TOHEP and mining frequent and similar patterns. Finally, from frequent and similar patterns we can create discriminant rules which show the discrimination between two rules in rulesets. The five mining patterns type i.e.: TSHEP, SOHEP, TOHEP, frequent and similar patterns have minimum and maximum SLV values which can be derived with equations 3.2 and 3.3. The minimum and maximum SLV values for TSHEP and TOHEP can be derived with equation 3.2 since they are not combination between subsumption and overlapping, where TSHEP for subsumption and TOHEP for overlapping. Meanwhile, the minimum and maximum SLV values for SOHEP, frequent and similar patterns can be derived with equation 3.3 since they are combination between subsumption and overlapping. In equations 3.2 and 3.3, m is the number of attributes in ruleset similar as m in equation 3.1, c and c1 are LV value in equation 3.1 which has options between 0.4, 0.5, 2.0 and 2.1. The equation 3.2 indicates the frequency c for m times where c as LV value has similar frequency (subsumption or overlapping) m times. The equation 3.3 indicates the frequency c for m-1 times plus c1 where c as LV value has similar frequency (subsumption or overlapping) m-1 times plus c1 as combination c. The

implementation for equations 3.2 and 3.3 will be explored in the next sub section to find minimum and maximum SLV values for the five mining patterns type.

$$m*c \tag{3.2}$$

$$(m-1)*c + c1 \tag{3.3}$$

where: m = m in equation 3.1

c = LV options (0.4, 0.5, 2.0 and 2.1) in equation 3.1

c1 = combination c, LV options (0.4, 0.5, 2.0 and 2.1) in equation 3.1

3.5.1. Mining TSHEP, SOHEP and TOHEP

Determining type of HEP between TSHEP, SOHEP or TOHEP is shown between line 12 and 14 in figure 3.3 which is categorized with variables over and subs. Variable over represents the overlapping (LV=2 or LV=2.1) and variable subs represents the subsumption (LV=0.4 or LV=0.5) which are possibly having increment as shown between line number 5 and 6, and number 7 and 8 in figure 3.3 respectively. The mining between TSHEP, SOHEP or TOHEP can be filtered when checklists TSHEP, SOHEP and TOHEP are in condition true as shown in line number 11figure 3.3. TSHEP and TOHEP are composition subsumption (LV=0.4 or LV=0.5) and overlapping (LV=2.0 or LV=2.1) respectively, whilst SOHEP as composition between subsumption (LV=0.4 or LV=0.5) and overlapping (LV=2.0 or LV=2.1) have minimum and maximum SLV values as shown in figure 3.5.

The two arrow lines in figure 3.5 show the influence of two main categorizations subsumption and overlapping. The overlapping arrow line shows the influence overlapping from LV=2.1 (maximum value for overlapping categorization) until LV=0.5 (maximum value for subsumption categorization). Whilst subsumption arrow line shows the influence subsumption from LV=0.4 (minimum value for subsumption categorization) until LV=2 (minimum value for overlapping categorization). SLV is categorized as TSHEP, SOHEP or TOHEP. SLV is categorized as TSHEP when have all subsumption LV values (LV=0.4 or LV=0.5) or those rules that are completely subsumed. While SLV is categorized as SOHEP when have combination subsumption and overlapping LV values (LV=0.4 or LV=0.5 and LV=2 or LV=2.1) or those rules that are overlap and subsumed. Moreover, SLV is categorized as TOHEP when have all overlapping LV values (LV=2 or LV=2.1) or those

rules that are completely overlapping. For similarity between rulesets $\{R_i^1\}$ and $\{R_j^2\}$, since SLV has LV with minimum and maximum values 0.4 and 2.1 then SLV in equation 3.1 has minimum and maximum values of m*c with equation 3.2 where c=0.4 and c=2.1 then m*0.4 and m*2.1 respectively. Thus, SLV in equation 3.1 has different minimum and maximum values for TSHEP, SOHEP and TOHEP as shown in figure 3.5.

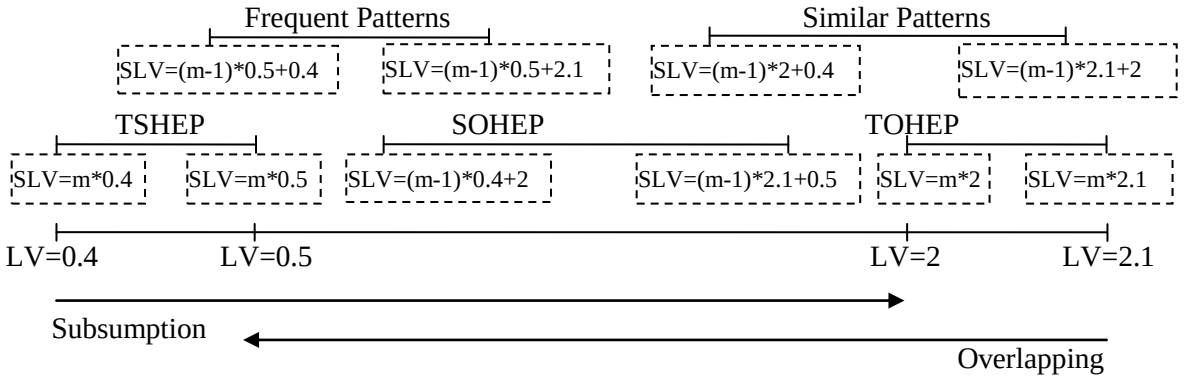


Figure 3.5. Composition subsumption and overlapping for mining patterns

Since TSHEP has SLV value with all subsumption LV values between minimum and maximum values 0.4 and 0.5 then TSHEP has minimum and maximum SLV values of m*c with equation 3.2 where c=0.4 and c=0.5 then m*0.4 and m*0.5 respectively. Meanwhile, since TOHEP has SLV value with all overlapping LV values between minimum and maximum values 2.0 and 2.1 then TOHEP has minimum and maximum SLV values of m*c with equation 3.2 where c=2.0 and c=2.1 then m*2.0 and m*2.1 respectively. Since SOHEP has SLV value which combination between subsumption (LV=0.4 or LV=0.5) and overlapping (LV=2.0 or LV=2.1) categorizations, then SOHEP has minimum and maximum SLV values of (m-1)*c+c1 with equation 3.3 where c=0.4,c1=2 and c=2.1,c1=0.5 then (m-1)*0.4+2 and (m-1)*2.1+0.5 respectively. SOHEP minimum SLV value is SLV=(m-1)*0.4+2 shows the frequency of minimum subsumption categorization (LV=0.4) in m-1 times((m-1)*0.4) plus 2.0 as minimum overlapping LV value categorization. Whilst SOHEP maximum SLV value is SLV=(m-1)*2.1+0.5 shows the frequency of maximum overlapping categorization (LV=2.1) in m-1 times((m-1)*2.1) plus 0.5 as maximum subsumption LV value categorization. Thus, Minimum SLV value for SOHEP shows frequent minimum subsumption (LV=0.4) and minimum overlapping (LV=2.0), whilst maximum SLV value for SOHEP shows frequent maximum overlapping (LV=2.1) and maximum subsumption (LV=0.5).

3.5.2. Mining Frequent pattern

Frequent pattern is a combination of feature patterns that appear in dataset with frequency not less than a user-specified threshold[86] and the frequent pattern synonym with large pattern was first proposed for market basket analysis in the form of association rules[105]. With frequent pattern we can have strong/sharp discrimination power where have large growth rate and support in target (D2) dataset and other support in contrasting (D1) dataset is small [69,71,79]. In AOI-HEP, the frequent pattern is shown by the subsumption LV=0.4 or LV=0.5 and as mention previously when LV=0.4 then ruleset R2 is subsumed by ruleset R1 (R2 $\subset$ R1) where R2 is subset rule and R1 is superset rule. On the other hand when LV=0.5 then ruleset R1 is subsumed by ruleset R2 (R1 $\subset$ R2) where R1 is subset rule and R2 is superset rule. R2 is in target (D2) dataset and R1 is in contrasting (D1) dataset ($\frac{D2}{D1} = \frac{target}{contrasting} = \frac{R2}{R1}$) , and it is as accordance with HEP growth rate in equation 3.4. Superset rule is a frequent pattern since subset rule is part of the superset rule and for instance when SLV has the same LV values (SLV=0.5+0.5+0.5+0.5=2) then certainly the number of instances in superset rule is larger than in its subset rule. Thus, that instance condition SLV=0.5+0.5+0.5+0.5=2 shows that superset rule (frequent pattern) has high support (large pattern) and subset rule (infrequent pattern) has low support. in EP, patterns will be recognized as EP if have high support (frequent pattern) in one class and low support (infrequent pattern) in other one [71,89]

From frequent patterns, we can create a discrimination rule and are interested in mining the frequent pattern with strong/sharp discrimination power. In EP, the strength of discrimination power is expressed by its large growth rate and support in target (D2) dataset [69,71,79]. This is called an essential Emerging Patterns (eEP) [71]. In AOI-HEP, the strength of discrimination power is expressed by its large growth rate and support in target (D2) dataset as well. Certainly, to make large growth rate can be happened when have large support in target (D2) dataset and low support in contrasting (D1) dataset. Indeed, in EP, patterns will be recognized as EP if have high support in one class and low support in other one [71,89]. Moreover, support in contrasting (D1) dataset must be less than support in target (D2) dataset where by the end will create large growth rate.

In AOI-HEP, the strength of discriminant power is expressed by subsumption LV=0.5 (explained in previous paragraph) where R2 in target (D2) dataset is superset and R1 in contrasting (D1) dataset is subset. The strength of discrimination power with subsumption LV=0.5 shows that have large support in target (D2) dataset and low support in contrasting (D1) dataset, where by the end will create large growth rate. TSHEP may have SLV value with subsumption LV=0.5, particularly for SLV value of TSHEP with all subsumption LV=0.5 (SLV value with similarity subsumption LV=0.5, for instance SLV=0.5+0.5+0.5+0.5=2). Thus, for discriminant rule from frequent pattern which SLV value of TSHEP with similarity subsumption LV=0.5 will have frequent pattern with strong discrimination power. However, not all TSHEP have SLV value with all subsumption LV=0.5, but there is TSHEP have SLV value with nearly all subsumption LV=0.5 and recognized as TSHEP with frequent subsumption LV=0.5. Moreover, SOHEP as combination subsumption and overlapping is interested to be explored since there are SOHEP with frequent subsumption LV=0.5.

Two parts of objects are similar if they are similar in all features (full matching similarity) or if the percentage of similar features is greater than the 80%[87] or if they are similar in at least 90% of the features[88]. Since there are TSHEP with all subsumption LV=0.5 where have full similarity subsumption LV=0.5, then there are frequent pattern with strong discrimination power for TSHEP or SOHEP with frequent similarity subsumption LV=0.5 at percentage value of (m-1)/m*100 where m as in equation 3.1. Since the strength of discriminant power is expressed by subsumption LV=0.5 and frequent pattern which can be mined from TSHEP or SOHEP has minimum and maximum SLV values of (m-1)*c+c1 with equation 3.3 where c=0.5,c1=0.4 and c=0.5,c1=2.1 then (m-1)*0.5+0.4 and (m-1)*0.5+2.1 respectively as shown in figure 3.5. Minimum and maximum SLV value for frequent pattern are SLV=(m-1)*0.5+0.4 and SLV=(m-1)*0.5+2.1 show the frequent similarity subsumption (LV=0.5) in m-1 times at percentage value of (m-1)/m*100 ((m-1)*0.5) plus 0.4 as minimum subsumption and 2.1 as maximum overlapping LV value categorization respectively. Thus, minimum and maximum SLV value for frequent pattern show frequent similarity subsumption (LV=0.5) (in TSHEP and SOHEP) at percentage value of (m-1)/m*100 which express discrimination power plus minimum subsumption LV=0.4 (show influences TSHEP) and maximum overlapping LV=2.1 (show influences SOHEP) respectively.

The HEP algorithm in figure 3.3 shows the process of mining frequent pattern with strong discrimination power, which is executed by giving condition true to input frequent variable. Moreover, variable counter F, will be incremented when have subsumption LV=0.5 as shown in line number 8. In line number 16, if input Frequent variable is true and variable F=x or F=x-1 then the output will be categorized as frequent pattern with strong discrimination power, where x is m in equation 3.1. F=x represents to TSHEP with full similarity subsumption LV=0.5, while F=x-1 represents to TSHEP or SOHEP with frequent similarity subsumption LV=0.5. Finally, with AOI-HEP we can mine two conditions frequent pattern with strong discrimination power and they are:

1. TSHEP with full similarity subsumption LV=0.5 or TSHEP with frequent similarity subsumption LV=0.5 at percentage value of (m-1)/m*100.

2. SOHEP with frequent similarity subsumption LV=0.5 at percentage value of (m-1)/m*100.

3.5.3. Mining Similar patterns

Similar patterns are interesting to mine because similarity pattern between datasets show the equality pattern which can represent similar behavior patterns. There are many examples of the important similar patterns in data mining process. In business, it is important to discover companies with similar patterns such as similar growth patterns, similar product selling patterns and etc. In education, it is important to discover students with similar patterns such as similar student behavior patterns, similar student progress patterns and etc. In banking system, it is important to discover customer with similar patterns such as similar customer behavior patterns, similar customer loan patterns and etc. Searching similar patterns are important and can be used for segmentation or prediction. For example in banking system, banking segmentation and banking prediction with similar banking transaction could help to show banking transaction prediction, with similar customer behavior patterns could help to uncover fraud, and loan prediction [106]. The similarity patterns can be measured with similarity two or more attributes or by calculating distance with euclidean distance or manhattan distance [3].

In AOI-HEP similar patterns are shown by the overlapping LV=2.0 or LV=2.1 and as shown in equation 3.1 LV=2.0 when hierarchy level and values are the same and the attributes values are not ANY. Whilst LV=2.1 when hierarchy level and values are the same and the attributes values are ANY. Since TOHEP is rules that are completely overlapping where SLV with LV=2.0 or LV=2.1 and show the similarity between rules then TOHEP is a similar pattern. TOHEP has full similarity overlapping LV=2.0, TOHEP with combination overlapping LV=2.0 and LV=2.1, but not for TOHEP with full similarity overlapping LV=2.1 since LV=2.1 is ANY and means nothing. Moreover, AOI-HEP framework is not interested in TOHEP with full similarity overlapping LV=2.1 since line number 1 and 2 in figure 3.3 show the exclusion for rule with ANY values in all attributes in rulesets. The same like frequent pattern, we can create discrimination rule from similar pattern and SOHEP as combination subsumption and overlapping is interested to be explored since there are SOHEP with frequent overlapping LV=2.0 and LV=2.1.

As mentioned in sub section 3.5.2, two parts of objects are similar if their features are similar (full matching similarity) or if the percentage of similar features is greater than the 80%[87]. Further, if they are similar in at least 90% of the features[88]. SOHEP has frequent similarity overlapping LV=2.0 or frequent combination overlapping LV=2.0 and LV=2.1 at percentage value of (m-1)/m*100 where m as in equation 3.1. However, for SOHEP with frequent similarity overlapping LV=2.1 at percentage value of (m-1)/m*100 is not interesting, since LV=2.1 is ANY and means nothing. Since overlapping LV=2.0 and LV=2.1 show the similar patterns and can be mined from SOHEP and TOHEP, and similar patterns have minimum and maximum SLV values of (m-1)*c+c1 with equation 3.3 where c=2.0,c1=0.4 and c=2.1,c1=2.0 then (m-1)*2.0+0.4 and (m-1)*2.1+2.0 respectively as shown in figure 3.5.Minimum and maximum SLV value for similar patterns are SLV=(m-1)*2.0+0.4 and SLV=(m-1)*2.1+2.0 show the frequent similarity minimum and maximum overlapping (LV=2.0 and LV=2.1) in m-1 times at percentage value of (m-1)/m*100 ((m-1)*2.0 and (m-1)*2.1)plus 0.4 and 2.0 as minimum subsumption and overlapping LV value categorization respectively. Thus, minimum and maximum SLV value for similar patterns show frequent similarity minimum and maximum overlapping (LV=2.0 and LV=2.1) (in SOHEP and TOHEP) at percentage value of (m-1)/m*100 plus minimum subsumption LV=0.4 (show influences SOHEP) and minimum overlapping LV=2.0 (show influences TOHEP) respectively.

HEP algorithm in figure 3.3 shows the process mining similar pattern is executed by giving condition true to input similar variable. Moreover, variable counter S will be incremented when have overlapping LV=2.1 as shown in line number 5. In line number 16, if input Similar variable is true and variable S<x-1 then the output will be categorized as similar pattern, where x is m in equation 3.1. S<x-1 represents to SOHEP with frequent similarity overlapping LV=2.1 < x-1 where SOHEP with frequent similarity overlapping LV=2.1 at percentage value of (m-1)/m*100 is not interesting (for instance SOHEP with SLV=2.1+2.1+2.1+0.5). Finally, with AOI-HEP we can mine two conditions similar pattern and they are:

1. TOHEP with full similarity overlapping LV=2.0 or TOHEP with combination overlapping LV=2.0 and LV=2.1, but not for TOHEP with full similarity overlapping LV=2.1.

2. SOHEP with frequent similarity overlapping LV=2.0 or SOHEP with frequent combination overlapping LV=2.0 and LV=2.1 at percentage value of (m-1)/m*100, but not for SOHEP with frequent similarity overlapping LV=2.1 at percentage value of (m-1)/m*100 .

3.6. HEP Growth Rate

Besides eliminating patterns with similarity function $C\{R_i^1, R_j^2\}$, the large number of HEP (Cartesian product between rulesets) is eliminated by the growth rate function $GR\{R_i^1, R_j^2\}$ with given a GrowthRate threshold. Growthrate is a standard function used in Emerging Patterns (EP) [79], and the difference in our approach is discovering high level emerging pattern with the same or different itemset instead of low level pattern with the same itemset. As mentioned in section 3.3, rulesets are AOI outputs and each of rule in ruleset has $|r_n^x|$ as the number of tuples forming the rule (figure 3.2). Because of rule in ruleset has $|r_n^x|$ as the number of tuples, then there is no Jumping High level Emerging Patterns (JHEP), where JHEP is related as a term of JEP. JEP is EP with support is 0 in one dataset and more than 0 in the other dataset or EP as special type of EP which is having infinite growth rate (∞).

Growth rate GR{ R_i^1, R_j^2 } is shown in figure 3.1 and in line number 15 in the HEP algorithm in figure 3.3 is used to discriminate between datasets D2 and D1. This growth rate can be calculated using equation (3.4). We can define that a HEP is a ruleset whose support changes from one ruleset in dataset D1 to another ruleset in dataset D2. In other words, HEP is a ruleset whose strength of high level rule Y of ruleset R1 in dataset D1 changes to high level rule X of ruleset R2 in dataset D2. Conventionally, this is defined as follows:

$$GR(X,Y) = \frac{SupportD2(X)}{SupportD1(Y)} = \frac{CountR2(X)/\ |D2|}{CountR1(Y)/\ |D1|} \quad (3.4)$$

where:

X = High level rule of ruleset R2 in dataset D2.

Y = High level rule of ruleset R1 in dataset D1.

D2 = Dataset D2.

D1 = Dataset D1.

|D2| = Total number of instances in dataset D2.

|D1| = Total number of instances in dataset D1.

Count R2(X) = Number of high level rule X of ruleset R2 in dataset D2.

Count R1(Y) = Number of high level rule Y of ruleset R1 in dataset D1.

Support D2(X) = Composition number of high level rule X of ruleset R2 in D2.

Support D1(Y) = Composition number of high level rule Y of ruleset R1 in D1.

3.7. Conclusion

In this chapter, Attribute Oriented Induction High level Emerging Pattern (AOI-HEP) framework is presented. AOI-HEP framework shows the detail about this proposed mining technique where AOI characteristic rule algorithm is combined with High level Emerging Pattern (HEP) algorithm. HEP pattern as an output from HEP algorithm is categorized as TSHEP, SOHEP or TOHEP based on two main categorizations subsumption and overlapping. HEP pattern will be categorized as TSHEP where have attribute subsumption, and will be categorized as TOHEP where have attribute overlapping. Meanwhile, HEP pattern will be categorized as SOHEP where have attribute subsumption and overlapping.

Moreover, HEP pattern can be categorized between frequent and similar pattern. HEP pattern can be categorized as frequent pattern when HEP pattern is TSHEP or SOHEP with frequent attribute subsumption at percentage value of (m-1)/m*100. Meanwhile, HEP pattern can be categorized as similar pattern when HEP pattern is TOHEP or SOHEP with frequent attribute overlapping at percentage (m-1)/m*100. Finally, HEP growth rate is used to discriminate frequent and similar patterns with the same or different itemset. In chapter 4, AOI-HEP framework is implemented which used four datasets from UCI machine learning repository.

Chapter 4: AOI-HEP Experiments

4.1. Introduction

This chapter presents an experimental evaluation with four datasets from the UCI machine learning repository. We mine TSHEP, SOHEP and TOHEP patterns include frequent and similar patterns. From frequent and similar patterns we can find discriminant rules. In section 4.2, we show preliminaries and definitions with the five chosen attributes for each dataset. Each of the chosen attributes will have a concept hierarchy for AOI generalization purposes. Next, section 4.3 explains execution time and the SLV value results per dataset between TSHEP, SOHEP and TOHEP. Section 4.4 shows the process of mining frequent patterns with strong discriminating power from the AOI rulesets which can be used to build discriminant rules. Moreover, section 4.5 shows strong discrimination rules which were created from frequent patterns with strong discriminating power. Meanwhile, section 4.6 explains mining similar patterns from the AOI rulesets which can be used to build discriminant rules. Furthermore, section 4.7 shows discriminant rules which were created from similar patterns. Section 4.8 shows observations with confidence equations to justify frequent or similar pattern for each dataset. Finally, section 4.9 gives a summary for this chapter.

4.2. Preliminaries on datasets

Experiments used four datasets from the UCI machine learning repository: adult, breast cancer, census, and IPUMS datasets with the number of instances being 48842, 569, 2458285 and 256932 respectively [56]. The programs were run with attribute and rule thresholds of 6. Threshold 6 was chosen based on the preliminary experiments done on a dataset [56] such that to get meaningful numbers of rules, a higher threshold is preferable after trial experiments. The experiments showed that TSHEP as rare patterns and are numerous if using attribute thresholds between 4 and 6, and rules thresholds between 5 and 10. Since it was rare to find TSHEP, we decided to use a bigger attribute threshold of 6 for all other experiments. Similarly, 6 was chosen for the rules threshold, since 6 is median between 2 and 9. Moreover,

we obtained numerous TSHEP rules for thresholds between 5 and 10 as expected when thresholds are bigger.

Each dataset has concept hierarchies built from five chosen attributes with a minimum concept level of three. The attributes in concept hierarchies for adult dataset include workclass, education, marital-status, occupation, and native-country attributes as shown between appendices 1 and 5 respectively. The attributes in concept hierarchies for the breast cancer dataset contains attributes i.e. clump thickness, cell size, cell shape, bare nuclei and normal nucleoli attributes as shown between appendices 6 and 10 respectively. Meanwhile, class, marital status, means, relat1 and yearsch attributes, were given to concept hierarchies for the Census dataset as shown between appendices 11 and 15 respectively. Finally, the attributes in concept hierarchies for the IPUMS dataset consists of relateg, marst, educrec, migrat5g and tranwork attributes as shown in appendices 16 and 20 respectively.

Each dataset was divided into two sub datasets based on learning the high level concept in one of their attributes. Learning the high level concept in one of their five chosen attributes for concept hierarchies, makes the parameter m in equation 3.1 at chapter 3 have value 4, where value 4 comes from five chosen attributes for concept hierarchies minus 1 and 1 is the attribute for the learning concept. In the adult dataset, we learn by discriminating between the “government” (4289 instances) and “non government” (14 instances) concepts of the “workclass” attribute (appendix 1) in datasets D2 and D1 respectively. In the breast cancer dataset, we learn by discriminating between “aboutaverclump” (533 instances) and “aboveaverclump” (289 instances) concepts of the “clump thickness” attribute (appendix 6) in datasets D2 and D1 respectively. Meanwhile Census dataset learns “green” (1980 instances) and “no green” (809 instances) concepts of the “means” attribute (appendix 13) for datasets D2 and D1 respectively. Finally, the IPUMS dataset learns “unmarried” (140124 instances) and “married” (77453 instances) concepts of the “marst” attribute (appendix 17) as datasets D2 and D1 respectively.

4.3. Experiments

Experiments were carried out by a java application as shown in figure 4.1. The experiments were tested on Intel(R) Atom(TM) CPU N550 (1.50 GHz) with 1.00 GB RAM. The AOI-HEP application has an input dataset and corresponding concept hierarchies in the form of flat files respectively. The AOI-HEP application was run 4 times as the number of experimental datasets and with the attribute and rule thresholds 6 as mentioned in section 4.2.

By running AOI-HEP application with input adult, breast cancer, census and IPUMS datasets, we have rulesets R2 and R1. Tables 4.1 and 4.2 are result from adult dataset with running time approximately 3 seconds, whilst tables 4.3 and 4.4 are result from breast cancer dataset with running time approximately 3 seconds too. Meanwhile, tables 4.5 and 4.6 are result from census dataset with running time approximately 4 seconds, whilst tables 4.7 and 4.8 are result from IPUMS dataset with running time approximately 13 seconds. Incredibly, the extraordinary running time of 13 seconds with the input IPUMS dataset happened because IPUMS has huge instances learning dataset's unmarried and married concepts with 140124 and 77453 instances respectively. Tables 4.1 and 4.2 are result from learning government and non government concepts of "workclass" attribute of adult dataset, whilst tables 4.3 and 4.4 are result from learning AboutAverClump and AboveAverClump concepts of "clump thickness" attribute of breast cancer dataset. Meanwhile tables 4.5 and 4.6 are result from learning Green and Non Green concepts of "means" attribute of census dataset, whilst tables 4.7 and 4.8 are result from learning unMarried and Married concepts of "marst" attribute of IPUMS dataset. Each table between 4.1 and 4.8 is a ruleset either as ruleset R2 or R1and has 6 tuples (rules) include number of instances for each tuple (rule). Each table has four attributes (m in equation 3.1 at chapter 3) which are from five chosen attributes for each dataset as mentioned in section 4.2 minus one attribute learning.

Table 4.1. Ruleset R2 for learning government concept from "workclass" attribute of adult dataset

No	Education	Marital	Occupation	Country	Number of instances
0	Intermediate	ANY	ANY	ANY	3454
1	ANY	ANY	ANY	America	786
2	Advanced	ANY	ANY	Asia	30
3	Advanced	ANY	ANY	Europe	17
4	Basic	Married-spouse	Services	Europe	1
5	Advanced	Married-spouse	Services	Antartica	1

Table 4.2. Ruleset R1 for learning non government concept from "workclass" attribute of adult dataset

No	Education	Marital	Occupation	Country	Number of instances
0	7th-8th	Widowed	Tools	United-states	1
1	HS-grad	Never-married	ANY	United-states	4
2	HS-grad	Married-civ-spouse	ANY	ANY	5
3	Assoc-adm	Married-civ-spouse	Tools	United-states	1
4	Some-college	Married-civ-spouse	ANY	United-states	2
5	Some-college	Married-spouse-absent	Tools	United-states	1

Table 4.3. Ruleset R2 for learning AboutAverClump concept from "clump thickness" attribute of breast cancer dataset

No	Cell Size	Cell Shape	Bare Nuclei	Normal Nucleoli	Number of instances
0	ANY	ANY	ANY	ANY	496
1	mediumSize	smallShape	ANY	aboutAverNucleoli	3
2	VeryLargeSize	ANY	ANY	ANY	19
3	mediumSize	largeShape	aboveAverNuclei	ANY	7
4	VeryLargeSize	mediumShape	ANY	VeryLargeNucleoli	3
5	largeSize	VeryLargeShape	VeryLArgeNuclei	ANY	5

Table 4.4. Ruleset R1 for learning AboveAverClump concept from "clump thickness" attribute of breast cancer dataset

No	Cell Size	Cell Shape	Bare Nuclei	Normal Nucleoli	Number of instances
0	ANY	ANY	ANY	ANY	277
1	smallSize	largeShape	VeryLargeNuclei	VeryLargeNucleoli	1
2	mediumSize	VeryLargeShape	ANY	aboveAverNucleoli	5
3	largeSize	VeryLargeShape	ANY	ANY	4
4	VeryLargeSize	smallShape	mediumNuclei	VeryLargeNucleoli	1

No	Cell Size	Cell Shape	Bare Nuclei	Normal Nucleoli	Number of instances
5	largeSize	smallShape	mediumNuclei	largeNucleoli	1

Table 4.5. Ruleset R2 for learning Green concept from “means” attribute of census dataset

No	Class	Marital	Relat1	Yearsch	Number of instances
0	ANY	ANY	ANY	ANY	1929
1	ANY	2	ANY	Basic	29
2	ANY	ANY	Married	No Education	13
3	ANY	2	Married	Advanced	6
4	Child	1	Married	Advanced	2
5	Private	3	Married	Advanced	1

Table 4.6. Ruleset R1 for learning Non Green concept from “means” attribute of census dataset

No	Class	Marital	Relat1	Yearsch	Number of instances
0	ANY	ANY	ANY	ANY	592
1	Non Government	ANY	ANY	Basic	134
2	Private	ANY	ANY	ANY	43
3	Non County	ANY	Married	Advanced	9
4	ANY	0	Married	ANY	14
5	County	ANY	Family	ANY	17

Table 4.7. Ruleset R2 for learning unMarried concept from “marst” attribute of IPUMS dataset

No	Relateg	Educrec	Migrat5g	tranwork	Number of instances
0	ANY	ANY	ANY	ANY	108026
1	ANY	Secondary School	ANY	ANY	7632

No	Relateg	Educrec	Migrat5g	tranwork	Number of instances
2	ANY	Primary School	Other-state	ANY	10332
3	ANY	Reception School	Other-state	ANY	3175
4	ANY	Primary School	Not-known	ANY	6356
5	ANY	College	Not-known	ANY	4603

Table 4.8. Ruleset R1 for learning Married concept from “marst” attribute of IPUMS dataset

No	Relateg	Educrec	Migrat5g	tranwork	Number of instances
0	ANY	ANY	ANY	ANY	56087
1	ANY	Basic	Moved	ANY	6707
2	ANY	Academy	Not-known	ANY	5440
3	ANY	Primary School	Not-known	ANY	2296
4	ANY	College	Not-known	ANY	5706
5	ANY	Secondary School	Not-known	ANY	1217

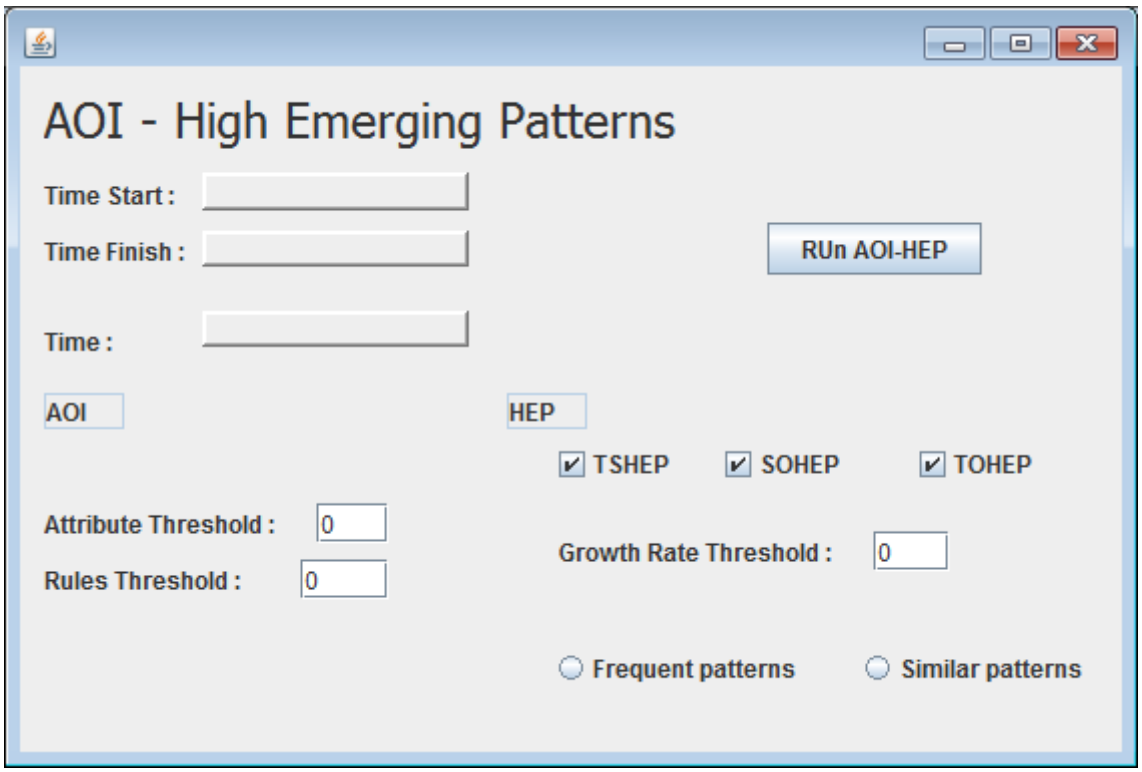


Figure 4.1. Screen display for AOI-HEP application

The results for running the AOI-HEP application for four experimental datasets can be seen between tables 4.9 and 4.15 where the adult dataset has two TSHEP, four SOHEP and no TOHEP, the breast cancer dataset has no TSHEP, two SOHEP and no TOHEP,whilst the census dataset has two TSHEP, six SOHEP and no TOHEP and the IPUMS dataset has no TSHEP, four SOHEP and two TOHEP. The results of running the AOI-HEP application are shown in table 4.16 which shows SLV values with equation 3.1 at chapter 3 and table 4.17 which shows growth rate values with equation 3.4 at chapter 3. Figure 4.2 shows the graph for table 4.16 and figure 4.3 shows the graph for table 4.17. Table 4.16 and 4.17 show that most datasets have SOHEP but not TSHEP and TOHEP (shown by 0), and the most rarely found were TOHEP. Tables 4.9 and 4.10 are two TSHEP and four SOHEP from adult dataset respectively, whilst table 4.11 is two SOHEP from breast cancer dataset. Meanwhile, tables 4.12 and 4.13 are two TSHEP and six SOHEP from census dataset respectively, whilst tables 4.14 and 4.15 are four SOHEP and two TOHEP from IPUMS dataset.

Tables 4.9 to 4.15 are outputs which are stated in line number 17 HEP algorithm in figure 3.3 at chapter 3. Each table has number of growth rates grouped either as TSHEP, SOHEP or TOHEP, where growth rate is discrimination between rulesets R2 and R1 as mentioned in equation 3.4 at chapter 3. Each table has position rulesets R2(X) and R1(Y), support D2(X), support D1(Y), Support D2(X)/Support D1(Y)=GR, HEP pattern and HEP%, where parameters X and Y, R2(X), R1(Y), support D2(X) and support D1(Y) refer to equation 3.4 at chapter 3. Columns R2(X) and R1(Y) in tables 4.9 to 4.15 refer to position tuple (rule) between tables 4.1 and 4.8. Columns Support D2(X) and Support D1(Y) are implementation equation 3.4 whilst column Support D2(X)/Support D1(Y)=GR is division between columns Support D2(X) and Support D1(Y) as implementation equation 3.4 as well. Moreover, column HEP pattern is implementation equation 3.1as comparation rule between rulesets R2(X) and R1(Y). Furthermore, column HEP% is percentage parameter LV in equation 3.1 between LV=0.4 or LV=0.5 and LV=2.0 and LV=2.1 as subsumption and overlapping respectively.

Columns R2(X) and R1(Y) in table 4.9 and 4.10 refer to position tuple (rule) in table 4.1 and 4.2 respectively since they are from the same dataset (adult dataset), where R2(X) and R1(Y) for learning government and non government concepts respectively from the same “workclass” attribute of adult dataset. While columns R2(X) and R1(Y) in table 4.11 refer to position tuple (rule) in table 4.3 and 4.4 respectively since they are from the same dataset (breast cancer dataset), where R2(X) and R1(Y) for learning AboutAverClump and AboveAverClump concepts respectively from the same “clump thickness” attribute of breast

cancer dataset. Meanwhile, columns R2(X) and R1(Y) in table 4.12 and 4.13 refer to position tuple (rule) in table 4.5 and 4.6 respectively since they are from the same dataset (census dataset), where R2(X) and R1(Y) for learning Green and Non Green concepts respectively from the same "means" attribute of census dataset. Moreover, columns R2(X) and R1(Y) in table 4.14 and 4.15 refer to position tuple (rule) in table 4.7 and 4.8 respectively since they are from the same dataset (IPUMS dataset), where R2(X) and R1(Y) for learning unMarried and Married concepts respectively from the same "marst" attribute of IPUMS dataset.

In tables 4.9 and 4.10, divisor 4289 and 14 in columns Support D2(X) and Support D1(Y) are number of instances for learning government and non government concepts respectively of "workclass" attribute of adult dataset.

Table 4.9. TSHEP from adult dataset

No	R2(X)	R1(Y)	Support D2(X)	Support D1(Y)
1	0	3	3454/4289=0.80532	1/14=0.07143
2	0	5	3454/4289=0.80532	1/14=0.07143

No	Support D2(X)/ Support D1(Y)=GR	HEP Pattern	HEP %
1	(3454/4289) / (1/14) =11.27442	0.5+0.5+0.5+0.5=2	100%
2	(3454/4289) / (1/14) =11.27442	0.5+0.5+0.5+0.5=2	100%

TSHEP number 1 in table 4.9 has R2(X)=0 and R1(Y)=3 which refer to rulesets number 0 in table 4.1 and number 3 in table 4.2 which have number of instances 3454 and 1 as numerator in column Support D2(X) and Support D1(Y) respectively. Likewise, TSHEP number 2 in table 4.9 has R2(X)=0 and R1(Y)=5 which refer to rulesets number 0 in table 4.1 and number 5 in table 4.2 which have number of instances 3454 and 1 as numerator in columns Support D2(X) and Support D1(Y) respectively. Column HEP pattern for TSHEP number 1 and 2 in table 4.9 havethe same comparation rule with all parameter LV=0.5 and SLV=2 as implementation equation 3.1. Therefore, columns HEP% have value 100% since all parameter LV have the same subsumption value 0.5.

Table 4.10. SOHEP from adult dataset

No	R2(X)	R1(Y)	Support D2(X)	Support D1(Y)
1	0	1	3454/4289=0.80532	4/14=0.28571
2	0	2	3454/4289=0.80532	5/14=0.35714
3	0	4	3454/4289=0.80532	2/14=0.14286
4	1	2	786/4289=0.18326	5/14=0.35714

No	Support D2(X)/ Support D1(Y)=GR	HEP Pattern	HEP %
1	(3454/4289) / (4/14) = 2.81861	0.5+0.5+2.1+0.5=3.6	75%+25%
2	(3454/4289) / (5/14) = 2.25488	0.5+0.5+2.1+2.1=5.2	50%+50%
3	(3454/4289) / (2/14) =5.63721	0.5+0.5+2.1+0.5=3.6	75%+25%
4	(786/4289) / (5/14) == 0.51313	0.5+0.5+2.1+0.4=3.5	75%+25%

SOHEP number 1 in table 4.10 has R2(X)=0 and R1(Y)=1 which refer to rulesets number 0 in table 4.1 and number 1 in table 4.2 which have number of instances 3454 and 4 as numerator in column Support D2(X) and Support D1(Y) respectively. Moreover, SOHEP number 2 in table 4.10 has R2(X)=0 and R1(Y)=2 which refer to rulesets number 0 in table 4.1 and number 2 in table 4.2 which have number of instances 3454 and 5 as numerator in columns Support D2(X) and Support D1(Y) respectively. Furthermore, SOHEP number 3 in table 4.10 has R2(X)=0 and R1(Y)=4 which refer to rulesets number 0 in table 4.1 and number 4 in table 4.2 which have number of instances 3454 and 2 as numerator in columns Support D2(X) and Support D1(Y) respectively. Likewise, SOHEP number 4 in table 4.10 has R2(X)=1 and R1(Y)=2 which refer to rulesets number 1 in table 4.1 and number 2 in table 4.2 which have number of instances 786 and 5 as numerator in columns Support D2(X) and Support D1(Y) respectively.

Meanwhile, column HEP pattern for SOHEP number 1 in table 4.10 has comparation rule with parameters LV=0.5,LV=0.5,LV=2.1,LV=0.5 and SLV=3.6 as implementation equation 3.1. Then, column HEP% has value 75%+25% since there are 75% with three subsumption parameters LV with all value 0.5 and 25%with one overlapping parameter LV=2.1. Moreover, column HEP pattern for SOHEP number 2 in table 4.10 has comparation rule with parameters LV=0.5,LV=0.5,LV=2.1,LV=2.1 and SLV=5.2 as implementation

equation 3.1. Then, column HEP% has value 50%+50% since there are 50% with two subsumption parameters LV with all value 0.5 and 50% with two overlapping parameters LV with all value 2.1. Furthermore,column HEP pattern for SOHEP number 3 in table 4.10 which similar with column HEP pattern for SOHEP number 1 has comparation rule with parameters LV=0.5,LV=0.5,LV=2.1,LV=0.5 and SLV=3.6 as implementation equation 3.1. Then, column HEP% has value 75%+25% since there are 75% with three subsumption parameters LV with all value 0.5 and 25% with one overlapping parameter LV=2.1. Likewise, column HEP pattern for SOHEP number 4 in table 4.10 has comparation rule with parameters LV=0.5,LV=0.5,LV=2.1,LV=0.4 and SLV=3.5 as implementation equation 3.1. Then, column HEP% has value 75%+25% since there are 75% with three subsumption parameters LV with value 0.5, 0.5, 0.4 and 25% with one overlapping parameter LV=2.1.

In table 4.11, divisor 533 and 289 in columns Support D2(X) and Support D1(Y) are number of instances for learning AboutAverClump and AboveAverClump concepts respectively of "clump thickness" attribute of breast cancer dataset.

Table 4.11. SOHEP from breast cancer dataset

No	R2(X)	R1(Y)	Support D2(X)	Support D1(Y)
1	2	4	19/533=0.03565	1/289=0.00346
2	5	3	5/533=0.00938	4/289=0.01384

No	Support D2(X)/ Support D1(Y)=GR	HEP Pattern	HEP %
1	(19/533) / (1/289) = 10.30206	2.0+0.5+0.5+0.5=3.5	75%+25%
2	(5/533) / (4/289) = 0.67777	2.0+2.0+0.4+2.1=6.5	25%+75%

SOHEP number 1 in table 4.11 has R2(X)=2 and R1(Y)=4 which refer to rulesets number 2 in table 4.3 and number 4 in table 4.4 which have number of instances 19 and 1 as numerator in column Support D2(X) and Support D1(Y) respectively. Likewise, SOHEP number 2 in table 4.11 has R2(X)=5 and R1(Y)=3 which refer to rulesets number 5 in table 4.3 and number 3 in table 4.4 which have number of instances 5 and 4 as numerator in columns Support D2(X) and Support D1(Y) respectively. Meanwhile, column HEP pattern for SOHEP number 1 in table 4.11 has comparation rule with parameters LV=2.0,LV=0.5,LV=0.5,LV=0.5 and SLV=3.5 as implementation equation 3.1. Then,

column HEP% has value 75%+25% since there are 75% with three subsumption parameters LV with all value 0.5 and 25% with one overlapping parameter LV=2.0. Likewise, column HEP pattern for SOHEP number 2 in table 4.11 has comparation rule with parameters LV=2.0,LV=2.0,LV=0.4,LV=2.1 and SLV=6.5 as implementation equation 3.1. Then, column HEP% has value 25%+75% since there is 25% with one subsumption parameters LV=0.4 and 75% with three overlapping parameter LV with value 2.0, 2.0, 2.1.

In tables 4.12 and 4.13, divisor 1980 and 809 in columns Support D2(X) and Support D1(Y) are number of instances for learning Green and No Green concepts respectively of "means" attribute of census dataset.

Table 4.12. TSHEP from census dataset

No	R2(X)	R1(Y)	Support D2(X)	Support D1(Y)
1	1	5	29/1980=0.01465	17/809=0.02101
2	3	2	6/1980=0.00303	43/809=0.05315

No	Support D2(X)/ Support D1(Y)=GR	HEP Pattern	HEP %
1	(29/1980) / (17/809) = 0.69700	0.5+0.4+0.5+0.4=1.8	100%
2	(6/1980) / (43/809) = 0.05701	0.5+0.4+0.4+0.4=1.7	100%

TSHEP number 1 in table 4.12 has R2(X)=1 and R1(Y)=5 which refer to rulesets number 1 in table 4.5 and number 4 in table 4.6 which have number of instances 29 and 17 as numerator in column Support D2(X) and Support D1(Y) respectively. Likewise, TSHEP number 2 in table 4.12 has R2(X)=3 and R1(Y)=2 which refer to rulesets number 3 in table 4.5 and number 2 in table 4.6 which have number of instances 6 and 43 as numerator in columns Support D2(X) and Support D1(Y) respectively. Meanwhile, column HEP pattern for TSHEP number 1 in table 4.12 has comparation rule with parameters LV=0.5,LV=0.4,LV=0.5,LV=0.4 and SLV=1.8 as implementation equation 3.1. Then, column HEP% has value 100% since all parameter LV have subsumption value 0.5, 0.4, 0.5, 0.4. Likewise, column HEP pattern for TSHEP number 2 in table 4.12 has comparation rule with parameters LV=0.5,LV=0.4,LV=0.4,LV=0.4 and SLV=1.7 as implementation equation 3.1. Then, column HEP% has value 100% since all parameter LV have subsumption value 0.5, 0.4, 0.4, 0.4.

Table 4.13. SOHEP from census dataset

No	R2(X)	R1(Y)	Support D2(X)	Support D1(Y)
1	1	1	29/1980=0.01465	134/809=0.16564
2	1	2	29/1980=0.01465	43/809=0.05315
3	2	2	13/1980=0.00657	43/809=0.05315
4	2	4	13/1980=0.00657	14/809=0.01731
5	3	3	6/1980=0.00303	9/809=0.01112
6	5	2	1/1980=0.00051	43/809=0.05315

No	Support D2(X)/ Support D1(Y)=GR	HEP Pattern	HEP %
1	(29/1980) / (134/809) = 0.08843	0.5+0.4+2.1+2.0=5.0	50%+50%
2	(29/1980) / (43/809) = 0.27556	0.5+0.4+2.1+0.4=3.4	75%+25%
3	(13/1980) / (43/809) = 0.12353	0.5+2.1+0.4+0.4=3.4	75%+25%
4	(13/1980) / (14/809) = 0.37940	2.1+0.5+2.0+0.4=5.0	50%+50%
5	(6/1980) / (9/809) = 0.27239	0.5+0.4+2.0+2.0=4.9	50%+50%
6	(1/1980) / (43/809) = 0.00950	2.0+0.4+0.4+0.4=3.2	75%+25%

SOHEP number 1 in table 4.13 has R2(X)=1 and R1(Y)=1 which refer to rulesets number 1 in table 4.5 and number 1 in table 4.6 which have number of instances 29 and 134 as numerator in column Support D2(X) and Support D1(Y) respectively. SOHEP number 2 in table 4.13 has R2(X)=1 and R1(Y)=2 which refer to rulesets number 1 in table 4.5 and number 2 in table 4.6 which have number of instances 29 and 43 as numerator in columns Support D2(X) and Support D1(Y) respectively. SOHEP number 3 in table 4.13 has R2(X)=2 and R1(Y)=2 which refer to rulesets number 2 in table 4.5 and number 2 in table 4.6 which have number of instances 13 and 43 as numerator in columns Support D2(X) and Support D1(Y) respectively. Moreover, SOHEP number 4 in table 4.13 has R2(X)=2 and R1(Y)=4 which refer to rulesets number 2 in table 4.5 and number 4 in table 4.6 which have number of instances 13 and 14 as numerator in columns Support D2(X) and Support D1(Y) respectively. Furthermore, SOHEP number 5 in table 4.13 has R2(X)=3 and R1(Y)=3 which

refer to rulesets number 3 in table 4.5 and number 3 in table 4.6 which have number of instances 6 and 9 as numerator in columns Support D2(X) and Support D1(Y) respectively.Likewise, SOHEP number 6 in table 4.13 has R2(X)=5 and R1(Y)=2 which refer to rulesets number 5 in table 4.5 and number 2 in table 4.6 which have number of instances 1 and 43 as numerator in columns Support D2(X) and Support D1(Y) respectively.

Meanwhile, column HEP pattern for SOHEP number 1 in table 4.13 has comparation rule with parameters LV=0.5,LV=0.4,LV=2.1,LV=2.0 and SLV=5.0 as implementation equation 3.1. Then, column HEP% has value 50%+50% since there are 50% with two subsumption parameters LV with value 0.5, 0.4 and 50% with two overlapping parameters LV with value 2.1, 2.0. Column HEP pattern for SOHEP number 2 in table 4.13 has comparation rule with parameters LV=0.5,LV=0.4,LV=2.1,LV=0.4 and SLV=3.4 as implementation equation 3.1. Then, column HEP% has value 75%+25% since there are 75% with three subsumption parameters LV with value 0.5, 0.4, 0.4 and 25% with one overlapping parameters LV=2.1. Column HEP pattern for SOHEP number 3 in table 4.13 has comparation rule with parameters LV=0.5,LV=2.1,LV=0.4,LV=0.4 and SLV=3.4 as implementation equation 3.1. Then, column HEP% has value 75%+25% since there are 75% with three subsumption parameters LV values 0.5, 0.4, 0.4 and 25% with one overlapping parameter LV=2.1. Moreover, column HEP pattern for SOHEP number 4 in table 4.13 has comparation rule with parameters LV=2.1,LV=0.5,LV=2.0,LV=0.4 and SLV=5.0 as implementation equation 3.1. Then, column HEP% has value 50%+50% since there are 50% with two subsumption parameters LV with value 0.5, 0.4 and 50% with two overlapping parameter LV with value 2.1, 2.0. Furthermore, column HEP pattern for SOHEP number 5 in table 4.13 has comparation rule with parameters LV=0.5,LV=0.4,LV=2.0,LV=2.0 and SLV=4.9 as implementation equation 3.1. Then, column HEP% has value 50%+50% since there are 50% with two subsumption parameters LV with value 0.5, 0.4 and 50% with two overlapping parameter LV with all value 2.0. Likewise, column HEP pattern for SOHEP number 6 in table 4.13 has comparation rule with parameters LV=2.0,LV=0.4,LV=0.4,LV=0.4 and SLV=3.2 as implementation equation 3.1. Then, column HEP% has value 75%+25% since there are 75% with three subsumption parameters LV with all value 0.4 and 25% with one overlapping parameter LV=2.0.

In tables 4.14 and 4.15, divisor 140124 and 77453 in columns Support D2(X) and Support D1(Y) are number of instances for learning unMarried and Married concepts respectively of “marst” attribute of IPUMS dataset.

Table 4.14. SOHEP from IPUMS dataset

No	R2(x)	R1(Y)	Support D2(X)	Support D1(Y)
1	1	1	7632/140124=0.05447	6707/77453=0.08659
2	1	5	7632/140124=0.05447	1217/77453=0.01571
3	2	1	10332/140124=0.07373	6707/77453=0.08659
4	3	1	3175/140124=0.02266	6707/77453=0.08659

No	Support D2(X)/ Support D1(Y)=GR	HEP Pattern	HEP %
1	(7632/140124) / (6707/77453) = 0.62898	2.1+0.4+0.5+2.1=5.1	50%+50%
2	(7632/140124) / (1217/77453) = 3.46636	2.1+2.0+0.5+2.1=6.7	25%+75%
3	(10332/140124) / (6707/77453) = 0.85149	2.1+0.4+0.4+2.1=5.0	50%+50%
4	(3175/140124) / (6707/77453) = 0.26166	2.1+0.4+0.4+2.1=5.0	50%+50%

SOHEP number 1 in table 4.14 has R2(X)=1 and R1(Y)=1 which refer to rulesets number 1 in table 4.7 and number 1 in table 4.8 which have number of instances 7632 and 6707 as numerator in column Support D2(X) and Support D1(Y) respectively. Moreover, SOHEP number 2 in table 4.14 has R2(X)=1 and R1(Y)=5 which refer to rulesets number 1 in table 4.7 and number 5 in table 4.8 which have number of instances 7632 and 1217 as numerator in columns Support D2(X) and Support D1(Y) respectively. Furthermore, SOHEP number 3 in table 4.14 has R2(X)=2 and R1(Y)=1 which refer to rulesets number 2 in table 4.7 and number 1 in table 4.8 which have number of instances 10332 and 6707 as numerator in column Support D2(X) and Support D1(Y) respectively. Likewise, SOHEP number 4 in table 4.14 has R2(X)=3 and R1(Y)=1 which refer to rulesets number 3 in table 4.7 and number 1 in table 4.8 which have number of instances 3175 and 6707 as numerator in columns Support D2(X) and Support D1(Y) respectively.

Meanwhile, column HEP pattern for SOHEP number 1 in table 4.14 has comparation rule with parameters LV=2.1,LV=0.4,LV=0.5,LV=2.1 and SLV=5.1 as implementation equation 3.1. Then, column HEP% has value 50%+50% since there are 50% with two subsumption parameters LV with value 0.4, 0.5 and 50% with two overlapping parameters LV with all value 2.1. Moreover, column HEP pattern for SOHEP number 2 in table 4.14 has comparation rule with parameters LV=2.1,LV=2.0,LV=0.5,LV=2.1 and SLV=6.7 as

implementation equation 3.1. Then, column HEP% has value 25%+75% since there are 25% with one subsumption parameters LV=0.5 and 75% with three overlapping parameters LV with value 2.1, 2.0,2.1. Furthermore, column HEP pattern for SOHEP number 3 and 4 in table 4.14 have the same comparation rule with parameters LV=2.1,LV=0.4,LV=0.4,LV=2.1 and SLV=5.0 as implementation equation 3.1. Then, column HEP% has the same value 50%+50% since there are 50% with two subsumption parameters LV with all value 0.4 and 50% with two overlapping parameters LV with all value 2.1.

Table 4.15. TOHEP from IPUMS dataset

No	R2(x)	R1(Y)	Support D2(X)	Support D1(Y)
1	4	3	6356/140124=0.045	2296/77453=0.029
2	5	4	4603/140124=0.033	5706/77453=0.074

No	Support D2(X)/ Support D1(Y)=GR	HEP Pattern	HEP %
1	(6356/140124) / (2296/77453) = 1.530	2.1+2.0+2.0+2.1=8.2	100%
2	(4603/140124) / (5706/77453) = 0.446	2.1+2.0+2.0+2.1=8.2	100%

TOHEP number 1 in table 4.15 has R2(X)=4 and R1(Y)=3 which refer to rulesets number 4 in table 4.7 and number 3 in table 4.8 which have number of instances 6356 and 2296 as numerator in column Support D2(X) and Support D1(Y) respectively. Likewise, TOHEP number 2 in table 4.15 has R2(X)=5 and R1(Y)=4 which refer to rulesets number 5 in table 4.7 and number 4 in table 4.8 which have number of instances 4603 and 5706 as numerator in column Support D2(X) and Support D1(Y) respectively. Meanwhile, column HEP pattern for TOHEP number 1 and 2 in table 4.15 have the same comparation rule with parameters LV=2.1,LV=2.0,LV=2.0,LV=2.1 and SLV=8.2 as implementation equation 3.1. Then, columns HEP% have value 100% where all parameters LV have overlapping value 2.1 and 2.0.

4.3.1. Composition SLV values for mining TSHEP, SOHEP and TOHEP

Table 4.16. Composition SLV values for four experimental datasets

Adult		Breast Cancer	Census		IPUMS	
TSHEP	**SOHEP**	**SOHEP**	**TSHEP**	**SOHEP**	**SOHEP**	**TOHEP**
2	3.6	3.5	1.8	5	5.1	8.2
2	5.2	6.5	1.7	3.4	6.7	8.2
0	3.6	0	0	3.4	5	0
0	3.5	0	0	5	5	0
0	0	0	0	4.9	0	0
0	0	0	0	3.2	0	0

The graph in figure 4.2 shows the consistency between minimum and maximum SLV values for TSHEP, SOHEP and TOHEP in figure 3.5 at chapter 3, where TSHEP, SOHEP and TOHEP have small, medium and high SLV values respectively. The graph in figure 4.2 shows the position TSHEP at the bottom of graph (below SLV=2) which indicates that TSHEP have small SLV values. The SOHEP position in the middle of the graph (between SLV=3 and SLV=7) indicates that SOHEP have medium SLV values and TOHEP position at the upper part of the graph (above SLV=8) indicates that TOHEP have high SLV values. The graph in figure 4.2 shows that TSHEP for adult and census datasets are consistent with mining TSHEP in section 3.5.1 at chapter 3 where TSHEP has SLV value with all subsumption LV values between minimum and maximum values 0.4 and 0.5.As mentioned in section 3.5.1 at chapter 3, TSHEP has minimum and maximum SLV values of m*c with equation 3.2 at chapter 3 where c=0.4 and c=0.5 then m*0.4 and m*0.5 respectively. Thus, graph in figure 4.2 shows that TSHEP for adult and census datasets have minimum and maximum values between SLV=m*0.4=4*0.4=1.6 and SLV=m*0.5=4*0.5=2 respectively with m=4 where for all experimental datasets m=4 in equation 3.1 at chapter 3.

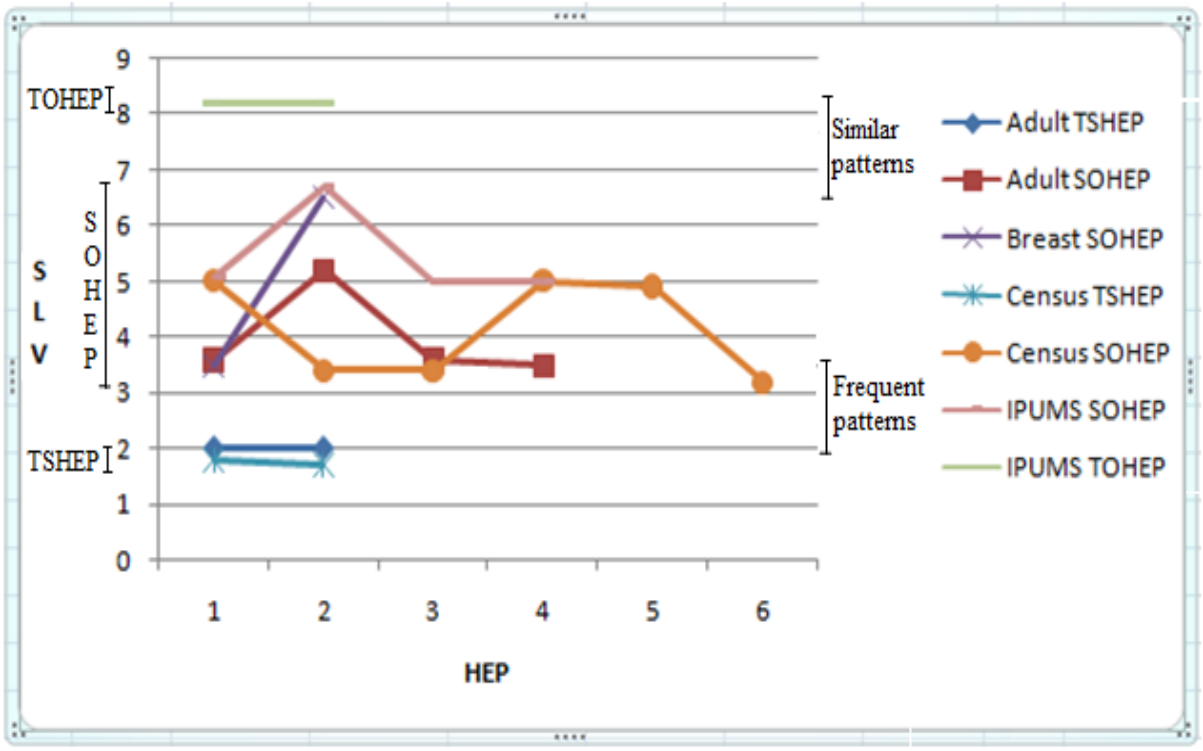


Figure 4.2. Composition SLV values for four experimental datasets

Moreover, the graph in figure 4.2 shows that SOHEP for all four experimental datasets are consistent with mining SOHEP in section 3.5.1 at chapter 3 where SOHEP has SLV value which combination between subsumption (LV=0.4 or LV=0.5) and overlapping (LV=2.0 or LV=2.1) categorizations. As mentioned in section 3.5.1 at chapter 3, SOHEP has minimum and maximum SLV values of (m-1)*c+c1 with equation 3.3 at chapter 3 where c=0.4,c1=2 and c=2.1,c1=0.5 then (m-1)*0.4+2 and (m-1)*2.1+0.5 respectively. Thus, graph in figure 4.2 shows that SOHEP for all four experimental datasets have minimum and maximum values between SLV=(m-1)*0.4+2=(4-1)*0.4+2=3.2 and SLV=(m-1)*2.1+0.5=(4-1)*2.1+0.5=6.8 respectively with m=4 where for all experimental datasets m=4 in equation 3.1 at chapter 3. Furthermore, the graph in figure 4.2 shows that TOHEP for IPUMS dataset is consistent with mining TOHEP in section 3.5.1 at chapter 3 where TOHEP has SLV value with all overlapping LV values between minimum and maximum values 2.0 and 2.1.As mentioned in section 3.5.1 at chapter 3, TOHEP has minimum and maximum SLV values of m*c with equation 3.2 at chapter 3 where c=2.0 and c=2.1 then m*2.0 and m*2.1 respectively. Thus, graph in figure 4.2 shows that TOHEP for IPUMS dataset has minimum and maximum values betweenSLV=m*2=4*2=8 and SLV=m*2.1=4*2.1=8.4 respectively with m=4 where for all experimental datasets m=4 in equation 3.1 at chapter 3.

4.3.2. Composition SLV values for mining frequent patterns

Meanwhile, the graph in figure4.2 shows the consistency between minimum and maximum SLV values for frequent and similar patterns in figure 3.5 at chapter 3. The graph in figure 4.2 shows that frequent pattern is consistent with mining frequent pattern in section 3.5.2 at chapter 3 where frequent pattern can be mined from TSHEP with SLV value full similarity subsumption LV=0.5 or from TSHEP or SOHEP with SLV value frequent similarity subsumption LV=0.5 at percentage value of (m-1)/m*100. As mentioned in section 3.5.2 at chapter 3, frequent pattern has minimum and maximum SLV values of (m-1)*c+c1 with equation 3.3 at chapter 3 where c=0.5,c1=0.4 and c=0.5,c1=2.1 then (m-1)*0.5+0.4 and (m-1)*0.5+2.1 respectively as shown in figure 3.5 chapter 3. Thus, graph in figure 4.2 shows that frequent pattern can be mined from TSHEP and SOHEP with minimum and maximum values betweenSLV=(m-1)*0.5+0.4=(4-1)*0.5+0.4=1.5+0.4=1.9 and SLV=(m-1)*0.5+2.1=(4-1)*0.5+2.1=1.5+2.1=3.6 respectively with m=4 where for all experimental datasets m=4 in equation 3.1 at chapter 3.

4.3.3. Composition SLV values for mining similar patterns

Moreover, the graph in figure 4.2 shows that similar patterns are consistent with mining similar patterns in section 3.5.3 at chapter 3 where similar patterns can be mined from TOHEP with full similarity overlapping LV=2.0 or TOHEP with combination overlapping LV=2.0 and LV=2.1, but not for TOHEP with full similarity overlapping LV=2.1. Furthermore, similar patterns can be mined from SOHEP with frequent similarity overlapping LV=2.0 or SOHEP with frequent combination overlapping LV=2.0 and LV=2.1 at percentage value of (m-1)/m*100, but not for SOHEP with frequent similarity overlapping LV=2.1 at percentage value of (m-1)/m*100. As mentioned in section 3.5.3 at chapter 3, similar patterns have minimum and maximum SLV values of (m-1)*c+c1 with equation 3.3 at chapter 3 where c=2.0,c1=0.4 and c=2.1,c1=2.0 then (m-1)*2.0+0.4 and (m-1)*2.1+2.0 respectively as shown in figure 3.5 at chapter 3. Thus, graph in figure 4.2 shows that similar patterns can be mined from SOHEP and TOHEP with minimum and maximum values between SLV=(m-1)*2.0+0.4=(4-1)*2.0+0.4=6+0.4=6.4 and SLV=(m-1)*2.1+2.0=(4-1)*2.1+2.0=6.3+2.0=8.3 respectively with m=4 where for all experimental datasets m=4 in equation 3.1 at chapter 3.

4.3.4. Composition Growth rate values

As mention in the beginning of section 4.3, table 4.17 shows growth rate values with equation 3.4 at chapter 3 as the results of running the AOI-HEP application for four experimental datasets from UCI machine learning, and figure 4.3 shows the graph for table 4.17. The results running the AOI-HEP application can be seen between tables 4.9 and 4.15 where the adult dataset has two TSHEP, four SOHEP and no TOHEP, the breast cancer dataset has no TSHEP, two SOHEP and no TOHEP, whilst the census dataset has two TSHEP, six SOHEP and no TOHEP and the IPUMS dataset has no TSHEP, four SOHEP and two TOHEP. Two TSHEP and four SOHEP in adult dataset have the same growth rates 11.27442 and growth rates 2.81861, 2.25488, 5.63721, 0.51313 respectively, whilst two SOHEP in breast cancer dataset have growth rates 10.30206, 0.6777. Meanwhile, two TSHEP and six SOHEP in census dataset have growth rates 0.69700, 0.05701 and growth rates 0.08843, 0.27556, 0.12353, 0.37940, 0.27239, 0.00950 respectively, whilst four SOHEP and two TOHEP in IPUMS dataset have growth rates 0.62898, 3.46636, 0.85149, 0.26166 and growth rates 1.530, 0.446 respectively.Table 4.17 shows that most datasets have SOHEP but not TSHEP and TOHEP (shown by 0), and the most rarely found were TOHEP.

Table 4.17. Composition Growth rate values for four experimental datasets

Adult		**Breast Cancer**	**Census**		**IPUMS**	
TSHEP	**SOHEP**	**SOHEP**	**TSHEP**	**SOHEP**	**SOHEP**	**TOHEP**
11.27442	2.81861	10.30206	0.69700	0.08843	0.62898	1.530
11.27442	2.25488	0.6777	0.05701	0.27556	3.46636	0.446
0	5.63721	0	0	0.12353	0.85149	0
0	0.51313	0	0	0.37940	0.26166	0
0	0	0	0	0.27239	0	0
0	0	0	0	0.00950	0	0

Meanwhile, the graph in figure 4.3 shows that types of HEP such as TSHEP, SOHEP or TOHEP in table 4.17 cannot be used as categorization of large and small growth rates. For example, SOHEP in adult, breast cancer and IPUMS datasets have large and small growth

rates, whilst SOHEP in census dataset has small growth rates only. Table 4.17 and the graph in figure 4.3 shows there are TSHEP, SOHEP or TOHEP with large and small growth rates. However, for HEP with a large growth rate has the possibility of becoming a frequent pattern with strong discrimination power. As mentioned in section 3.5.2 at chapter 3, TSHEP or SOHEP with a large growth rate can be categorized as a frequent pattern with strong discrimination power when they have a similar subsumption LV=0.5. For instance, two TSHEP from the adult dataset have the same large growth rates 11.2744, but two TSHEP from the census dataset have small growth rates between 0.697 and 0.05701. Another instance, from the breast cancer dataset has two SOHEP where one SOHEP has large growth rate 10.30206 but the other SOHEP has a small growth rate 0.6777. The next section shows two TSHEP from the adult dataset with a large growth rate 11.27744 and one SOHEP from the census dataset with a large growth rate 10.30206 which becomes a frequent pattern with strong discrimination power.

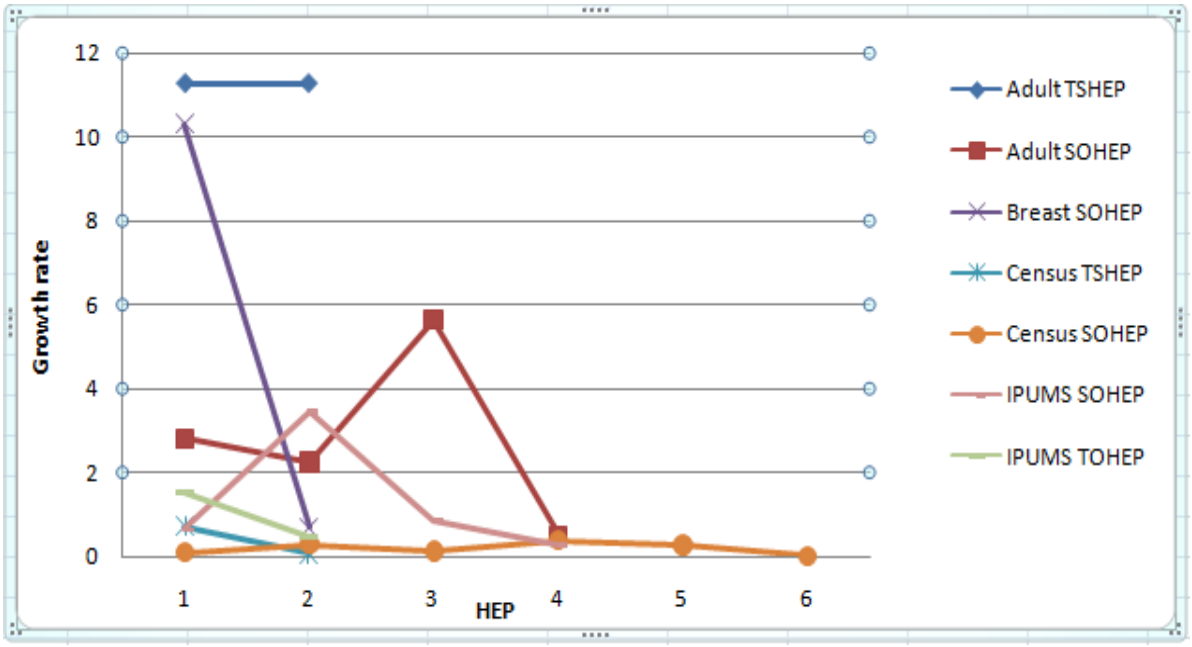


Figure 4.3. Composition growth rate values for four experimental datasets

4.4. Mining Frequent Patterns

As mentioned in section 3.5.2 at chapter 3, there are two conditions for mining frequent patterns with strong discrimination power and these are:

1. TSHEP with full similarsubsumption LV=0.5 or TSHEP with frequent similar subsumption LV=0.5 at percentage value of (m-1)/m*100.
2. SOHEP with frequent similar subsumption LV=0.5 at percentage value of (m-1)/m*100.

Process to mine frequent patterns use HEP algorithm in figure 3.3 at chapter 3 and as mention in section 3.5.2 at chapter 3, mining frequent patterns is executed by giving condition true to input frequent variable. In line number 8 HEP algorithm in figure 3.3 at chapter 3, variable counter F will be incremented when have subsumption LV=0.5 and in line number 16 if input Frequent variable is true and variable F=x or F=x-1 then the output will be categorized as frequent pattern with strong discrimination power, where x is m in equation 3.1 at chapter 3. F=x represents to TSHEP with full similarity subsumption LV=0.5, while F=x-1 represents to TSHEP or SOHEP with frequent similarity subsumption LV=0.5.

Next, a frequent pattern will be mined from TSHEP and SOHEP based on two conditions and m is m in equation 3.1 at chapter 3. From the experiments upon these four experimental datasets, m=4 for all four experimental datasets since they have the same number of attributes.

4.4.1. Mining frequent patterns from TSHEP

Frequent patterns can be mined from TSHEP with full similar subsumption LV=0.5 or frequent similar subsumption LV=0.5 at percentage value of (m-1)/m*100 and table 4.16 shows that there are two TSHEP in the adult and census datasets respectively.

1. **Mining frequent patterns from two TSHEP in adult dataset.**

   Two TSHEP in the adult dataset have the same SLV value 0.5+0.5+0.5+0.5=2 as shown in table 4.9. These two TSHEP are frequent patterns since all attributes have a full similar subsumption LV=0.5 and show that they have strong discriminating power since they have the same large growth rates 11.27442. Moreover, supports in target (D2) dataset (3454/4289=0.80532) are large than supports in contrasting (D1) dataset (1/14=0.07143).Table 4.18 and 4.19 show TSHEP number 1 and 2 in table 4.9 respectively, where these two TSHEP have the similar rule number 0 in their ruleset R2(X) and a different ruleset R1(Y). Ruleset R2(X)=0 refers to ruleset number 0 in table 4.1, whilst rulesets R1(Y)=3 and R1(Y)=5 for TSHEP number 1 and 2 refer to rulesets number 3 and 5 in table 4.2 respectively.

2. **Mining frequent patterns from two TSHEP in census dataset.**

Meanwhile, two TSHEP from the census dataset as shown in table 4.12 are infrequent patterns since they do not have full similar subsumption LV=0.5 and are not frequent similar subsumption LV=0.5 or less than percentage value of (m-1)/m*100.Firstly, TSHEP with SLV value 0.5+0.4+0.5+0.4=1.8 is an infrequent pattern since the number subsumption LV=0.5 is less than percentage value of (m-1)/m*100=(4-1)/4*100=3/4*100=75. Identical to the second TSHEP with SLV value 0.5+0.4+0.4+0.4=1.7 where the number subsumption LV=0.5 is less than percentage value of (m-1)/m*100=(4-1)/4*100=3/4*100=75. Moreover, both of TSHEP from the census dataset have small growth rates 0.697 and 0.05701 which indicates a weak discriminating power.

Table 4.18. TSHEP in adult dataset for rulesets R_3^1 to R_6^2 with GR=(3454/4289)/(1/14)=0.80532/0.07143=11.27442

Rulesets	Education	Marital	Occupation	Country	Instances
R_6^2	Intermediate	ANY	ANY	ANY	3454
R_3^1	Assoc-adm	Married-civ-spouse	Tools	United-states	1
LV	0.5	0.5	0.5	0.5	

Table 4.19. TSHEP in adult dataset for rulesets R_5^1 to R_6^2 with GR=(3454/4289)/(1/14)=0.80532/0.07143=11.27442

Rulesets	Education	Marital	Occupation	Country	Instances
R_6^2	Intermediate	ANY	ANY	ANY	3454
R_5^1	Some-college	Married-spouse-absent	Tools	United-states	1
LV	0.5	0.5	0.5	0.5	

4.4.2. Mining frequent patterns from SOHEP

Meawhile, frequent patterns can be mined from SOHEP with frequent attributes subsumption LV=0.5 at percentage value of (m-1)/m*100. Table 4.16 shows that all four

experimental datasets have SOHEP, however we are just interested in SOHEP with a strong discrimination power where there is frequent similar subsumption LV=0.5 at percentage value of (m-1)/m*100.

1. **Mining frequent patterns from SOHEP in the adult dataset.**

   Table 4.16shows there are four SOHEP from the adult dataset as shown in table 4.10 where SOHEP numbers 1 and 3 are frequent patterns whilst SOHEP numbers 2 and 4 are infrequent patterns. Thetwo frequent SOHEP numbers 1 and 3 have the same SLV value 0.5+0.5+2.1+0.5=3.6 and show as strong discriminating power where they have frequent similar subsumption LV=0.5 at percentage value of (m-1)/m*100=(4-1)/4*100=3/4*100=75. Moreover they have large growth rates 2.81861 and 5.63721 respectively. Meanwhile, the two infrequent SOHEP numbers 2 and 4 with SLV values 0.5+0.5+2.1+2.1=5.2 and 0.5+0.5+2.1+0.4=3.5 have weak discriminating power where frequent similar subsumption LV=0.5 are less than percentage value of (m-1)/m*100. Moreover SOHEP number 4 has a small growth rate 0.51313 as indication of a weak discriminating power. However, the infrequent SOHEP number 2 with growth rate 2.25488 has nearly equal growth rate with frequent SOHEP number 1 with a growth rate 2.81861. Since infrequent SOHEP number 2 has frequent similar subsumption LV=0.5 less than percentage value of (m-1)/m*100 then it is categorized as an infrequent SOHEP which has weak discriminating power.Tables4.20 and 4.21 show SOHEP numbers 1 and 3 in table 4.10 respectively where they have the same ruleset R2(X)=0 and different ruleset R1(Y). RulesetR2(X)=0 refers to ruleset number 0 in table 4.1, whilst rulesets R1(Y)=1 and R1(Y)=4 refer to rulesets number 1 and 4 in table 4.2.

2. **Mining frequent patterns from SOHEP in the breast cancer dataset.**

   Table 4.16 shows there are two SOHEP from the breast cancer dataset as shown in table 4.11 where SOHEP number 1 is a frequent pattern whilst SOHEP number 2 is an infrequent pattern. The frequent SOHEP number 1 has SLV value 2.0+0.5+0.5+0.5=3.5 which shows as a strong discriminating power which has a frequent similar subsumption LV=0.5 at percentage value of (m-1)/m*100=(4-1)/4*100=3/4*100=75 and moreover it has a large growth rate 10.30206. Whilst the infrequent SOHEP number 2 has SLV value 2.0+2.0+0.4+2.1=6.5 which shows as a weak discriminating power which has no frequent similar subsumption LV=0.5 which is less than percentage value of (m-1)/m*100=(4-1)/4*100=3/4*100=75 and moreover it has a small growth rate 0.6777.Table 4.22 shows

the frequent SOHEP number 1 in table 4.11 where ruleset R2(X)=2 refers to ruleset number 2 in table 4.3, whilst ruleset R1(Y)=4 refers to ruleset number 4 in table 4.4.

3. **Mining frequent patterns from SOHEP in census datasets.**

   Meanwhile, there are no frequent SOHEP with frequent subsumption LV=0.5 at percentage value of (m-1)/m*100in the census dataset. Table 4.16 shows six SOHEP from the census dataset which are shown detail in table 4.13 are infrequent patterns which have no frequent similar subsumption LV=0.5 or less than percentage value of (m-1)/m*100=(4-1)/4*100=3/4*100=75.Moreover, they have small growth rates between 0.0095 and 0.3794 which show as weak discriminating power.

4. **Mining frequent patterns from SOHEP in the IPUMS dataset.**

   Similar to the census dataset, there are no frequent SOHEP with frequent subsumption LV=0.5 at percentage value of (m-1)/m*100 in IPUMS dataset. Table 4.16 shows four SOHEP from the IPUMS dataset which are shown detail in table 4.14 are infrequent patterns which have no frequent similar subsumption LV=0.5 or less than percentage value of (m-1)/m*100=(4-1)/4*100=3/4*100=75. Moreover, three infrequent SOHEP have small growth rates between 0.26166 and 0.85149 which show as weak discriminating power. However, one of the infrequent SOHEP has a large growth rate 3.46636 but since frequent attributes subsumption LV=0.5 are less than percentage value of (m-1)/m*100=(4-1)/4*100=3/4*100=75, then cannot be categorized as frequent SOHEP.

Table 4.20.Frequent subsumptionSOHEP in adult dataset for rulesets R_1^1 to R_0^2 with GR=(3454/4289)/(4/14)=0.80532/0.28571=2.81861

Rulesets	**Education**	**Marital**	**Occupation**	**Country**	**Instances**
R_0^2	Intermediate	ANY	ANY	ANY	3454
R_1^1	HS-Grad	Never-married	ANY	United-states	4
LV	0.5	0.5	2.1	0.5	

Table 4.21.Frequent subsumptionSOHEP in adult dataset for rulesets R_4^1 to R_0^2 with GR=(3454/4289)/(2/14)=0.80532/0.14286=5.63721

Rulesets	Education	Marital	Occupation	Country	Instances
R_0^2	Intermediate	ANY	ANY	ANY	3454
R_4^1	Some-college	Married-civ-spouse	ANY	United-states	2
LV	0.5	0.5	2.1	0.5	

Table 4.22.Frequent subsumptionSOHEP in breast cancer dataset for rulesets R_4^1 to R_2^2 with GR=(19/533)/(1/289)=0.03565/0.00346=10.30206

Rulesets	Cell Size	Cell Shape	Bare Nuclei	Normal Nucleoli	Instances
R_2^2	VeryLargeSize	ANY	ANY	ANY	19
R_4^1	VeryLargeSize	smallShape	MediumNuclei	VeryLargeNucleoli	1
LV	2.0	0.5	0.5	0.5	

4.5. Strong discriminant rules from frequent patterns

As mention in section 3.5.2 at chapter 3, in AOI-HEP, the strong of discriminant rules are expressed by subsumption LV=0.5 where R2 in target (D2) dataset is superset and R1 in contrasting (D1) dataset is subset. The strong of discrimination rules with subsumption LV=0.5 show that have large support in target (D2) dataset and low support in contrasting (D1) dataset, where by the end will create large growth rate. In EP, the strong of discrimination rulesare expressed by its large growth rate and support in target (D2) dataset [69,71,79]. Frequent patterns between table 4.18 and 4.22 have SLV values between 2.0 and 3.6 and show consistency between minimum and maximum SLV value for frequent patterns in figure 3.5 at chapter 3 and graph in figure 4.2 shows that frequent patterns can be mined from TSHEP or SOHEP as mention in section 4.3. Table 4.23 shows the strong discrimination rules from frequent patterns from tables 4.18 to 4.22 and frequent patterns in table 4.23 are consistent with condition AOI-HEP strong discriminant rules where target dataset is larger than contrasting dataset where by the end will create large growth rate. Moreover, SLV values for all frequent patterns in table 4.23 are consistent as mentioned in section 4.4 where frequent patterns can be mined from TSHEP with SLV value full similarity

subsumption LV=0.5 or from TSHEP or SOHEP with SLV value frequent similarity subsumption LV=0.5 at percentage value of (m-1)/m*100.

Table 4.23. Frequent patterns for creating strong discrimination rules

No	HEP	Dataset	Target dataset	Contrasting dataset
1	TSHEP	Adult	3454/4289=0.80532	1/14=0.07143
2	TSHEP	Adult	3454/4289=0.80532	1/14=0.07143
3	SOHEP	Adult	3454/4289=0.80532	4/14=0.28571
4	SOHEP	Adult	3454/4289=0.80532	2/14=0.14286
5	SOHEP	Breast cancer	19/533=0.03565	1/289=0.00346

No	Target dataset/Contrasting dataset = Growth rate	SLV
1	(3454/4289) / (1/14) = 11.2744	0.5+0.5+0.5+0.5=2
2	(3454/4289) / (1/14) = 11.2744	0.5+0.5+0.5+0.5=2
3	(3454/4289) / (4/14) = 2.81861	0.5+0.5+2.1+0.5=3.6
4	(3454/4289) / (2/14) = 5.63721	0.5+0.5+2.1+0.5=3.6
5	(19/533) / (1/289) = 10.30286	2.0+0.5+0.5+0.5=3.5

Furthermore, SLV values for all frequent patterns in table 4.23 are consistent as mentioned in section 4.3.2 where frequent pattern can be mined from TSHEP and SOHEP with minimum and maximum values between SLV=(m-1)*0.5+0.4=(4-1)*0.5+0.4=1.5+0.4=1.9 and SLV=(m-1)*0.5+2.1=(4-1)*0.5+2.1=1.5+2.1=3.6 respectively with m=4 where for all experimental datasets m=4 in equation 3.1 at chapter 3. Frequent pattern has minimum and maximum SLV values of (m-1)*c+c1 with equation 3.3 at chapter 3 where c=0.5,c1=0.4 and c=0.5,c1=2.1 then (m-1)*0.5+0.4 and (m-1)*0.5+2.1 respectively as shown in figure 3.5 chapter 3. The graph in figure 4.3 shows that frequent patterns from tables 4.18 to 4.22 and as shown in table 4.23 appear at upper part of the graph which indicate they have large growth rates and as strong discriminant rules. However, table 4.17 and graph in figure 4.3 show that one SOHEP in IPUMS dataset with growth rate 3.46636 is not frequent pattern although larger than the smallest growth rate in number 3 table 4.23

(2.81861) . This is because SOHEP in IPUMS dataset with growth rate 3.46636 as shown at number 2 in table 4.14 has SLV value 2.1+2.0+0.5+2.1=6.7 where pattern SLV value has no frequent similarity subsumption LV=0.5 at percentage value of (m-1)/m*100.

From each of the frequent patterns from tables 4.18 to 4.22 and as shown in table 4.23, the strong discriminant rules can be described in the next list respectively.

1. There are 80.53% adults in government workclass with an intermediate education and 7.14% adults in non governmentworkclass with assoc-adm education, married-civ-spouse marital status, tools occupation and from the United States.

   or

   There are 11.2744 times more adults in government workclass with intermediate education than adults in non governmentworkclass with assoc-adm education, married-civ-spouse marital status, tools occupation and from the United States.

   or

   There are 11.2744growth rates for TSHEP adult dataset with 80.53% frequent pattern in government workclass and 7.14% infrequent pattern in non government workclass.
2. There are 80.53%adults in government workclass with an intermediate education and 7.14% adults in non government workclass with somecollege education, married- spouse-absent marital status, tools occupation and from the United States.

   or

   There are 11.2744 times more adults in government workclass with an intermediate education than adults in non government workclass with somecollege education, married-spouse-absent marital status, tools occupation and from the United States.

   or

   There are 11.2744growth rates for TSHEP adult dataset with 80.53% frequent pattern in government workclass and 7.14% infrequent pattern in non government workclass.
3. There are 80.53%adults in government workclass with an intermediate education and 28.57%adults in non government workclass with HS-Grad education, Never-married marital status and from the United States.

   or

   There are 2.81861 times more adults in government workclass with an intermediate education than adults in non government workclass with HS-Grad education, Never-married marital status and from the United States.

   or

There are 2.81861growth rates for SOHEP adult dataset with 80.53% frequent pattern in government workclass and 28.57% infrequent pattern in non government workclass.

4. There are 80.53%adults in government workclass with intermediate education and 14.28%adults in non government workclass with somecollege education, married-civ-spouse marital status and from the United States.

   or

   There are 5.63721 times more adults in government workclass with intermediate education than adults in non government workclass with somecollege education, married-civ-spouse marital status and from the United States.

   or

   There are 5.63721growth rates for SOHEP adult dataset with 80.53% frequent pattern in government workclass and 14.28% infrequent pattern in non government workclass.

5. There are 3.56%breast cancer in AboutAverClump “clump thickness” with VeryLargeSize “Cell Size” and 0.34%breast cancer in AboveAverClump “clump thickness” with VeryLargeSize “Cell Size”, SmallShape “Cell shape”, mediumNuclei “Bare Nuclei” and VeryLargeNucleoli “Normal Nucleoli”.

   or

   There are 10.30206 times more breast cancer in AboutAverClump “clump thickness” with VeryLargeSize “Cell Size” than breast cancer in AboveAverClump “clump thickness” with VeryLargeSize “Cell Size”, SmallShape “Cell shape”, mediumNuclei “Bare Nuclei” and VeryLargeNucleoli “Normal Nucleoli”.

   or

   There are 10.30206growth rates for SOHEP breast cancer dataset with 3.56% frequent pattern in AboutAverClump “clump thickness” and 0.34% infrequent pattern in AboveAverClump “clump thickness”.

Discriminating rules for table 4.18 to 4.22 show as strong discriminating power where they have large growth rates and supports in target (D2) datasets. They have large growth rates between 2.81861 and 11.2774 and large supports in target (D2) datasets between 3.56 and 80.53. Moreover, discriminating rules for table 4.18 to 4.22 have small supports in contrasting (D1) datasets where each of the support in contrasting (D1) dataset are less than the support in target(D2) dataset. They have small supports in contrasting (D1) dataset between 0.34 and 28.57 where each of the support in contrasting (D1) dataset is less than the support in target (D2) dataset.

4.6. Mining Similar Patterns

As mentioned in section 3.5.3 at chapter 3, there are 2 conditions for mining similar patterns and they are:

1. TOHEP with full similarity overlapping LV=2.0 or TOHEP with combination overlapping LV=2.0 and LV=2.1, but not for TOHEP with full similarity overlapping LV=2.1.

2. SOHEP with frequent similarity overlapping LV=2.0 or SOHEP with a frequent combination overlapping LV=2.0 and LV=2.1 at percentage value of (m-1)/m*100, but not for SOHEP with frequent similarity overlapping LV=2.1 at percentage value of (m-1)/m*100.

Process to mine similar patterns use HEP algorithm in figure 3.3 at chapter 3 and as mention in section 3.5.3 at chapter 3, mining similar patterns is executed by giving condition true to input similar variable. In line number 5 HEP algorithm in figure 3.3 at chapter 3, variable counter S will be incremented when have overlapping LV=2.1 and in line number 16 if input Similar variable is true and variable S<x-1 then the output will be categorized as similar pattern, where x is m in equation 3.1 at chapter 3. S<x-1 represents to SOHEP with frequent similarity overlapping LV=2.1 < x-1 where SOHEP with frequent similarity overlapping LV=2.1 at percentage value of (m-1)/m*100 is not interesting (for instance SOHEP with SLV=2.1+2.1+2.1+0.5).

Next, a similar pattern will be mined from TOHEP and SOHEP, and m in condition 2 is m in equation 3.1 at chapter 3. From the experiments in these four experimental datasets, m=4 for all four experimental datasets since they have the same number of attributes.

4.6.1. Mining similar patterns from TOHEP

Table 4.16 shows there are two TOHEP with same SLV value 2.1+2.0+2.0+2.1=8.2 as shown in table 4.15, where in these experiments, TOHEP is the most rarely mined are found in the IPUMS dataset only. Tables4.24 and 4.25 show TOHEP numbers 1 and 2 in table 4.15 respectively where ruleset R2(X)=4 and R2(X)=5 in each TOHEP refer to ruleset numbers 4 and 5 in table 4.7. Whilst rulesets R1(Y)=3 and R1(Y)=4 in each TOHEP refer to

ruleset numbers 3 and 4 in table 4.8. Tables 4.24 and 4.25 show both rules between rulesets R2 and R1 are full similar.

Table 4.24. TOHEP in IPUMS dataset for rulesetsR_3^1 to $R_4^2$with GR=(6356/140124)/(2296/77453)=0.045/0.029=1.530

Rulesets	Relateg	Educrec	Migrat5g	Tranwork	Instances
R_4^2	ANY	Primary School	Not-known	ANY	6356
R_3^1	ANY	Primary School	Not-known	ANY	2296
LV	2.1	2.0	2.0	2.1	

Table 4.25. TOHEP in IPUMS dataset for rulesetsR_4^1 to $R_5^2$with GR=(4603/140124)/(5706/77453)=0.033/0.074=0.446

Rulesets	Relateg	Educrec	Migrat5g	Tranwork	Instances
R_5^2	ANY	College	Not-known	ANY	4603
R_4^1	ANY	College	Not-known	ANY	5706
LV	2.1	2.0	2.0	2.1	

4.6.2. Mining similar patterns from SOHEP

As shown in table 4.16, all four experimental datasets have SOHEP, however we are just interested in SOHEP with frequent similarity overlapping LV=2.0 or SOHEP with a frequent combination overlapping LV=2.0 and LV=2.1 at percentage value of (m-1)/m*100, but not for SOHEP with frequent similarity overlapping LV=2.1 at percentage value of (m-1)/m*100.

1. **Mining similar patterns from SOHEP in the adult dataset.**
   Table 4.16shows there are four SOHEP from the adult dataset as shown in table 4.10and there is no similar patterns where all four SOHEP have frequent overlapping LV=2.0 and LV=2.1 less than percentage value of (m-1)/m*100=(4-1)/4*100=3/4*100=75.
2. **Mining similar patterns from SOHEP in the breast cancer dataset.**

Table 4.16 shows there are two SOHEP from the breast cancer dataset as shown in table 4.11. SOHEP number 1 with SLV value 2.0+0.5+0.5+0.5=3.5 is not a similar pattern since it is frequent overlapping LV=2.0 and LV=2.1 is less than percentage value of (m-1)/m*100=(4-1)/4*100=3/4*100=75. Whilst SOHEP number 2 is similar a pattern with SLV value 2.0+2.0+0.4+2.1=6.5 and has frequent overlapping LV=2.0 and LV=2.1 with percentage value of (m-1)/m*100=(4-1)/4*100=3/4*100=75.Table 4.26 shows the frequent similarity pattern SOHEP number 2 where rulesets R2 and R1 have a frequent similarity with percentage value of (m-1)/m*100 =(4-1)/4*100 =3/4*100=75. SOHEP number 2 has ruleset R2(X)=5 which refers to ruleset number 5 in table 4.3, whilst ruleset R1(Y)=3 refers to ruleset number 3 in table 4.4.

3. **Mining similar patterns from SOHEP in the census dataset.**

   Table 4.16 shows six SOHEP from the census dataset which are shown detail in table 4.13 and there is no similar pattern where all six SOHEP have frequent overlapping LV=2.0 and LV=2.1 less than percentage value of (m-1)/m*100=(4-1)/4*100=3/4*100=75.

4. **Mining similar patterns from SOHEP in the IPUMS dataset.**

   Table 4.16 shows four SOHEP from the IPUMS dataset which are shown in detail in table 4.14. SOHEP number 2 is a similar pattern, with SLV value 2.1+2.0+0.5+2.1=6.7 and has frequent overlapping LV=2.0 and LV=2.1 with percentage value of (m-1)/m*100=(4-1)/4*100=3/4*100=75. SOHEP number 2 has ruleset R2(X)=1 which refers to ruleset number 1 in table 4.7 and ruleset R1(Y)=5 refers to ruleset number 5 in table 4.8. Table 4.27 shows the frequent similarity pattern SOHEP number 2 which has frequent similarity with percentage value of (m-1)/m*100=(4-1)/4*100=3/4*100=75. For other SOHEP in table 4.14, such as SOHEP numbers 1,3 and 4 are not similar patterns since their frequent overlapping LV=2.0 and LV=2.1 are less than percentage value of (m-1)/m*100=(4-1)/4*100=3/4*100=75.

Table 4.26. Frequent overlapping SOHEP in breast cancer dataset for rulesetsR_3^1 to $R_5^2$with GR=(5/533)/(4/289)=0.00938/0.01384=0.67777

Rulesets	**Cell Size**	**Cell Shape**	**Bare Nuclei**	**Normal Nucleoli**	**Instances**
R_5^2	largeSize	VeryLargeShape	VeryLargeNuclei	ANY	5
R_3^1	largeSize	VeryLargeShape	ANY	ANY	4
LV	2.0	2.0	0.4	2.1	

Table 4.27. Frequent overlapping SOHEP in IPUMS dataset for rulesetsR_5^1 to $R_1^2$with GR=(7632/140124)/(1217/77453)=0.05447/0.01571=3.46636

Rulesets	Relateg	Educrec	Migrat5g	Tranwork	Instances
R_1^2	ANY	Secondary School	ANY	ANY	7632
R_5^1	ANY	Secondary School	Not-known	ANY	1217
LV	2.1	2.0	0.5	2.1	

4.7. Discriminant rules from similar patterns

As mention in section 3.5.3 at chapter 3, in AOI-HEP, the similar patterns are shown by overlapping LV=2.0 or LV=2.1. As shown in equation 3.1 chapter 3 where LV=2.0 when hierarchy level and values are the same and the attributes values are not ANY, whilst LV=2.1 when hierarchy level and values are the same and the attributes values are ANY. Similar patterns from tables 4.24 to 4.27 and as shown in table 4.28 have SLV values between 6.5 and 8.2 and show consistency between minimum and maximum SLV values for similar patterns in figure 3.5 at chapter 3 and graph in figure 4.2 shows that similar patterns can be mined from SOHEP or TOHEP as mention in section 4.3.

Table 4.28. Similar patterns for creating discrimination rules

No	HEP	Dataset	Target dataset	Contrasting dataset
1	TOHEP	IPUMS	6356/140124=0.045	2296/77453=0.029
2	TOHEP	IPUMS	4603/140124=0.033	5706/77453=0.074
3	SOHEP	Breast cancer	5/533=0.00938	4/289=0.01384
4	SOHEP	IPUMS	7632/140124=0.05447	1217/77453=0.01571

No	Target dataset/Contrasting dataset = Growth rate	SLV
1	(6356/140124) / (2296/77453) = 1.530	2.1+2.0+2.0+2.1=8.2
2	(4603/140124) / (5706/77453) = 0.446	2.1+2.0+2.0+2.1=8.2
3	(5/533) / (4/289) = 0.67777	2.0+2.0+0.4+2.1=6.5
4	(7632/140124) / (1217/77453) = 3.46636	2.1+2.0+0.5+2.1=6.7

Moreover, SLV values for all similar patterns in table 4.28 are consistent as mentioned in section 4.6 where similar patterns can be mined from TOHEP with full similarity overlapping LV=2.0 or TOHEP with combination overlapping LV=2.0 and LV=2.1, but not for TOHEP with full similarity overlapping LV=2.1. Still, similar patterns can be mined from SOHEP with frequent similarity overlapping LV=2.0 or SOHEP with frequent combination overlapping LV=2.0 and LV=2.1 at percentage value of (m-1)/m*100, but not for SOHEP with frequent similarity overlapping LV=2.1 at percentage value of (m-1)/m*100.Furthermore, SLV values for all similar patterns in table 4.28 are consistent as mentioned in section 4.3.3 where similar patterns can be mined from SOHEP and TOHEP with minimum and maximum values betweenSLV=(m-1)*2.0+0.4=(4-1)*2.0+0.4=6+0.4=6.4 and SLV=(m-1)*2.1+2.0=(4-1)*2.1+2.0=6.3+2.0=8.3 respectively with m=4 where for all experimental datasets m=4 in equation 3.1 at chapter 3. Similar patterns have minimum and maximum SLV values of (m-1)*c+c1 with equation 3.3 at chapter 3 where c=2.0,c1=0.4 and c=2.1,c1=2.0 then (m-1)*2.0+0.4 and (m-1)*2.1+2.0 respectively as shown in figure 3.5 at chapter 3.

Meanwhile, not like frequent pattern which are appeared at upper part of the graph in figure 4.3 which have large growth rates and indicate as strong discriminant rules, the graph in figure 4.3 cannot show the similar pattern interest position in the graph, but some of them at bottom of the graph indicate as weak discrimination rules since they have small growth rates.

From similar patterns between from tables 4.24 to 4.27 and as Shown in table 4.28, the discrimant rules can be described in the next list respectively.

1. There are 4.5% individuals in unmarried “marital status” and 2.9% individuals in Married “marital status” with a similar pattern in the Primary School education and Not-known “Migration status”

   or

   There are 1.53 times more individuals in unmarried “marital status” than individuals in Married “marital status” with a similar pattern in the Primary School education and Not-known “Migration status”

   or

   There are 1.53growth rates similar patterns for TOHEP IPUMS dataset with 4.5% unmarried “marital status”pattern and 2.9% Married “marital status” pattern.

2. There are 3.3% individuals in unmarried “marital status” and 7.4% individuals in Married “marital status” with a similar pattern in College education and Not-known “Migration status”

   or

   There are 0.446 times more individuals in unmarried “marital status” than individuals in Married “marital status” with a similar pattern in College education and Not-known “Migration status”

   or

   There are 0.446growth rates similar patterns for TOHEP IPUMS dataset with 3.3% unmarried “marital status”pattern and 7.4% Married “marital status” pattern.

3. There is 0.938%breast cancer in AboutAverClump “clump thickness” with VeryLargeNuclei “Bare Nuclei” and 1.384% breast cancer in AboveAverClump “clump thickness” with similar pattern largeSize “Cell Size” and VeryLargeShape “Cell shape”.

   or

   There is 0.6777 times more breast cancer in AboutAverClump “clump thickness” with VeryLargeNuclei “Bare Nuclei” than breast cancer in AboveAverClump “clump thickness” with similar pattern largeSize “Cell Size” and VeryLargeShape “Cell shape”.

   or

   There are 0.6777growth rates similar patterns for SOHEP breast cancer dataset with 0.938% AboutAverClump “clump thickness” pattern and 1.384% AboveAverClump “clump thickness” pattern.

4. There are 5.447%individuals in unMarried “marital status” and 1.571%individuals in Married “marital status” with Not-known “Migration status” and a similar pattern in Secondary School education.

   or

   There are 3.46636 times more individuals in unMarried “marital status” than individuals in Married “marital status” with Not-known “Migration status” and there is a similar pattern in Secondary School education.

   or

   There are 3.46636growth rates similar patterns for SOHEP IPUMS dataset with 5.447% unmarried “marital status”pattern and 1.571% Married “marital status” pattern.

Discriminant rules number 1 and 4 for tables 4.24 and 4.27 are strong discriminant rules since they have large growth rates (1.53 and 3.46636) and supports in target (D2)

datasets (4.5 and 5.447). Moreover, they have small supports in contrasting (D1) datasets (2.9 and 1.571) where each of the supports in the contrasting (D1) dataset is less than support in the target (D2) dataset. Meanwhile, discriminant rules numbers 2 and 3 for table 4.25 and 4.26 are weak discriminant rules since have small growth rates (0.446 and 0.6777) and supports in target (D2) datasets (3.3 and 0.938). Moreover, they have large supports in contrasting (D1) datasets (7.4 and 1.384) where each of supports in contrasting (D1) dataset is greater than the support in target(D2) dataset.

4.8. Experiment's analysis

Equation 4.1 and 4.2 are confidence equations for finding probability AOI-HEP mining interest between frequent or similar pattern in dataset by finding number of HEP types (TSHEP, SOHEP or TOHEP) with frequent or similar pattern in each of dataset. Equation 4.1 and 4.2 are applied on each of dataset from four UCI machine learning experimental datasets to find out the AOI-HEP mining interest for each dataset with learning high level concept on their chosen attribute. Equation 4.1 is used to count probability (confidence) AOI-HEP mining interest between frequent or similar pattern against HEP types such as TSHEP, SOHEP or TOHEP, whilst equation 4.2 is used to count average probability (confidence) AOI-HEP mining interest between frequent or similar pattern.

$$\frac{x1}{x2} * 100 \tag{4.1}$$

$$(\sum_{i=1}^{n} \frac{x1_i}{x2_i})/n * 100 \tag{4.2}$$

Where:

X1= number of HEP or TSHEP or SOHEP or TOHEP with frequent or similar pattern.

X2= number of HEP or TSHEP or SOHEP or TOHEP in table 4.16 or 4.17.

n= number of HEP types (TSHEP, SOHEP or TOHEP) plus 1.

4.8.1. AOI-HEP mining in adult dataset.

In AOI-HEP experiments, adult dataset learn on high level concept in "workclass" attribute and table 4.16 or 4.17 shows there are two mining types of HEP (TSHEP and SOHEP) and they are two TSHEP and four SOHEP. Since there are two mining types of HEP (TSHEP and SOHEP) in adult dataset, then n=3 in equation 4.2 where probability (confidence) AOI-HEP mining interest between frequent or similar pattern for adult dataset will be tested on two types of HEP (TSHEP and SOHEP) plus HEP itself as combination TSHEP and SOHEP. X2 in equation 4.1 and 4.2 are number of mining result for each n mining HEP types and because n=3 then firstly, X2=2 since there are two TSHEP. Secondly, X2=4 since there are four SOHEP and thirdly, X2=6 where two TSHEP plus four SOHEP.

X1 in equation 4.1 and 4.2 are number of frequent or similar pattern which come from mining result (X2) for each n mining HEP types. From section 4.4 there are two TSHEP and two SOHEP frequent patterns with strong discriminating power from adult dataset and because n=3 then firstly, X1=2 since there are two TSHEP frequent patterns. Secondly, X1=2 since there are two SOHEP frequent patterns and thirdly, X1=4 where two TSHEP frequent patterns plus two SOHEP frequent patterns. Moreover, with equation 4.1 and because n=3 then there are three types of AOI-HEP mining frequent patterns interest and they are :

1. TSHEP frequent patterns with confidence = $\frac{X1}{X2}$*100 = $\frac{2}{2}$*100= 100%
2. SOHEP frequent patternswith confidence = $\frac{X1}{X2}$*100 = $\frac{2}{4}$*100=50%
3. HEP (TSHEP and SOHEP) frequent patterns with confidence = $\frac{\mathrm{X1}}{\mathrm{X2}}$*100 = $\frac{4}{6}$*100=66.67%

Meanwhile, from section 4.6 there is no similar pattern in adult dataset and all X1=0 to show there is no similar pattern in adult dataset. The implementation equation 4.1 show that adult dataset which learn on high level concept in "workclass" attribute have AOI-HEP mining interest 100%, 50% and 66.67% probability (confidence) for TSHEP, SOHEP and HEP frequent patterns respectively and 0% probability (confidence) for similar patterns. Since adult dataset have AOI-HEP mining interest for frequent patterns (100%,50%,66.67%) rather than similar patterns (0%) and n=3 then equation 4.2 is used to decide average probability (confidence) for frequent patterns and they are :

$$(\sum_{i=1}^{n}\frac{x1_i}{x2_i})/n * 100 = (\sum_{i=1}^{3}\frac{x1_i}{x2_i})/3 * 100 = (\frac{2}{2}+\frac{2}{4}+\frac{4}{6})/3 * 100$$
$$= (1 + 0.5 + 0.67)/3 * 100 = 2.167/3 * 100 = 0.72233 * 100 = 72.23\%.$$

Thus, adult dataset which learn on high level concept in "workclass" attribute have AOI-HEP mining interest for frequent patterns rather than similar patterns with average probability (confidence) 72.23%.

4.8.2. AOI-HEP mining in breast cancer dataset.

In AOI-HEP experiments, breast cancer dataset learn on high level concept in "clump thickness" attribute and table 4.16 or 4.17 shows there is just only one mining type of HEP and they are two SOHEP. Since there is only one mining type of HEP (SOHEP) in breast cancer dataset, then n=1 in equation 4.2 where probability (confidence) AOI-HEP mining interest between frequent or similar pattern for breast cancer dataset will be tested on one type of HEP (SOHEP) only.

X2 in equation 4.1 and 4.2 are number of mining result for each n mining HEP types and because n=1 then just only one X2 and X2=2 since there are two SOHEP. Whilst, X1 in equation 4.1 and 4.2 are number of frequent or similar pattern which come from mining result (X2) for each n mining HEP types. From section 4.4 there is only one SOHEP frequent pattern with strong discriminating power from breast cancer dataset and because n=1 then just only one X1 and X1=1 since there is only one SOHEP frequent pattern. Moreover, with equation 4.1 and because n=1 then there is one type of AOI-HEP mining frequent pattern interest and it is :

SOHEP frequent pattern with confidence = $\frac{X1}{X2}$*100 = $\frac{1}{2}$*100=50%

Meanwhile, from section 4.6 there is only one SOHEP similar pattern in breast cancer dataset and because n=1 then just only one X1 and X1=1 since there is only one SOHEP similar pattern. Moreover, with equation 4.1 and because n=1 then there is one type of AOI-HEP mining similar patterns interest and it is :

SOHEP similar pattern with confidence = $\frac{X1}{X2}$*100 = $\frac{1}{2}$*100=50%

Since n=1 then equation 4.2 was not implemented. Thus, the implementation equation 4.1 show that breast cancer dataset which learn on high level concept in “clump thickness” attribute have AOI-HEP mining interest for both frequent and similar pattern with probability (confidence) 50% respectively.

4.8.3. AOI-HEP mining in census dataset.

In AOI-HEP experiments, census dataset learn on high level concept in “means” attribute and table 4.16 or 4.17 shows there are two mining types of HEP (TSHEP and SOHEP) and they are two TSHEP and six SOHEP. Since there are two mining types of HEP (TSHEP and SOHEP) in census dataset, then n=3 in equation 4.2 where probability (confidence) AOI-HEP mining interest between frequent or similar pattern for census dataset will be tested on two types of HEP (TSHEP and SOHEP) plus HEP itself as combination TSHEP and SOHEP. X2 in equation 4.1 and 4.2 are number of mining result for each n mining HEP types and because n=3 then firstly, X2=2 since there are two TSHEP. Secondly, X2=6 since there are six SOHEP and thirdly, X2=8 where two TSHEP plus six SOHEP.

X1 in equation 4.1 and 4.2 are number of frequent or similar pattern which come from mining result (X2) for each n mining HEP types. From section 4.4 there is no frequent pattern with strong discriminating power from census dataset and all X1=0 to show there is no frequent pattern with strong discriminating power from census dataset. Meanwhile, from section 4.6 there is no similar pattern in census dataset and all X1=0 to show there is no similar pattern. Since AOI-HEP mining interests for both frequent and similar patterns have 0% probability (confidence) then equation 4.2 was not implemented. Thus, the implementation equation 4.1 show that census dataset which learn on high level concept in “means” attribute have no AOI-HEP mining interest for both frequent and similar patterns.

4.8.4. AOI-HEP mining in IPUMS dataset.

In AOI-HEP experiments, IPUMS dataset learn on high level concept in “marst” attribute and table 4.16 or 4.17 shows there are two mining types of HEP (SOHEP and

TOHEP) and they are four SOHEP and two TOHEP. Since there are two mining types of HEP (SOHEP and TOHEP) in IPUMS dataset, then n=3 in equation 4.2 where probability (confidence) AOI-HEP mining interest between frequent or similar pattern for IPUMS dataset will be tested on two types of HEP (SOHEP and TOHEP) plus HEP itself as combination SOHEP and TOHEP. X2 in equation 4.1 and 4.2 are number of mining result for each n mining HEP types and because n=3 then firstly, X2=4 since there are four SOHEP. Secondly, X2=2 since there are two TOHEP and thirdly, X2=6 where four SOHEP plus two TOHEP.

X1 in equation 4.1 and 4.2 are number of frequent or similar pattern which come from mining result (X2) for each n mining HEP types. From section 4.4 there is no frequent pattern with strong discriminating power from IPUMS dataset and all X1=0 to show 4 there is no frequent pattern with strong discriminating power from IPUMS dataset. Meanwhile, from section 4.6 there are one SOHEP and two TOHEP similar patterns from IPUMS dataset and because n=3 then firstly, X1=1 since there is one SOHEP similar pattern. Secondly, X1=2 since there are two TOHEP similar patterns and thirdly, X1=3 where one SOHEP similar pattern plus two TOHEP similar patterns. Moreover, with equation 4.1 and because n=3 then there are three types of AOI-HEP mining similar patterns interest and they are :

1. SOHEP similar patterns with confidence = $\frac{X1}{X2}$*100 = $\frac{1}{4}$*100= 25%
2. TOHEP similar patterns with confidence = $\frac{X1}{X2}$*100 = $\frac{2}{2}$*100= 100%
3. HEP (SOHEP and TOHEP) similar patterns with confidence = $\frac{X1}{X2}$*100 = $\frac{3}{6}$*100= 50%

The implementation equation 4.1 show that IPUMS dataset which learn on high level concept in “marst” attribute have AOI-HEP mining interest 0% probability (confidence) for frequent pattern and 25%, 100% and 50% probability (confidence) for SOHEP, TOHEP and HEP similar patterns respectively. Since IPUMS dataset have AOI-HEP mining interest for similar patterns (25%, 100%,50%) rather than frequent pattern (0%) and n=3 then equation 4.2 is used to decide average probability (confidence) for similar patterns and they are :

$$(\sum_{i=1}^{n}\frac{x1_i}{x2_i})/n * 100 = (\sum_{i=1}^{3}\frac{x1_i}{x2_i})/3 * 100 = (\frac{1}{4}+\frac{2}{2}+\frac{3}{6})/3 * 100$$

$$= (0.25 + 1 + 0.5)/3 * 100 = 1.75/3 * 100 = 0.58333 * 100 = 58.33\%.$$

Thus, IPUMS dataset which learn on high level concept in “marst” attribute have AOI-HEP mining interest for similar patterns rather than frequent patterns with average probability (confidence) 58.33%.

4.8.5. Experiment’s analysis conclusion.

The experimental upon four UCI machine learning repository show that adult, breast cancer and IPUMS datasets are interested to be mined since there are AOI-HEP mining interest for frequent and/or similar patterns and not for census dataset since there is no AOI-HEP mining interest for both frequent and similar patterns. Adult dataset have AOI-HEP mining interest for frequent pattern with average probability (confidence) 72.23% while IPUMS dataset have AOI-HEP mining interest for similar pattern with average (confidence) 58.33%. Meanwhile, breast cancer dataset have AOI-HEP mining interest for both frequent and similar patterns with probability (confidence) 50% respectively.

Matrix in figure 4.4 shows AOI-HEP mining interest between frequent and similar patterns in four experimental datasets by finding type of HEP such as TSHEP, SOHEP and TOHEP based on explanation sub section 4.8.The matrix in figure 4.4 shows that frequent patterns can be mined from type of HEP such as TSHEP or SOHEP in adult or breast cancer datasets and similar patterns can be mined from type of HEP such as SOHEP or TOHEP in breast cancer or IPUMS datasets. Matrix in figure 4.4 shows that the more frequent the patterns, the more subsumption the type of HEP and the more similar the patterns, the more overlapping the type of HEP. This is accordance as mentioned in sub section 3.5.2 and 3.5.3 at chapter 3 where AOI-HEP can mine frequent and similar patterns in type of HEP such as TSHEP or SOHEP and SOHEP or TOHEP respectively. TSHEP and TOHEP in the matrix in figure 4.4 shows trend to frequent and similar patterns since have full or frequent similarity subsumption and overlapping as mentioned in sub section 3.5.2 and 3.5.3 at chapter 3 respectively. Meanwhile, differ from TSHEP and TOHEP, SOHEP show trend to frequent and similar patterns with frequent similarity subsumption and overlapping as mentioned in sub section 3.5.2 and 3.5.3 at chapter 3 respectively. Thus, there is no SOHEP with full similarity subsumption and overlapping for frequent and similar patterns respectively.

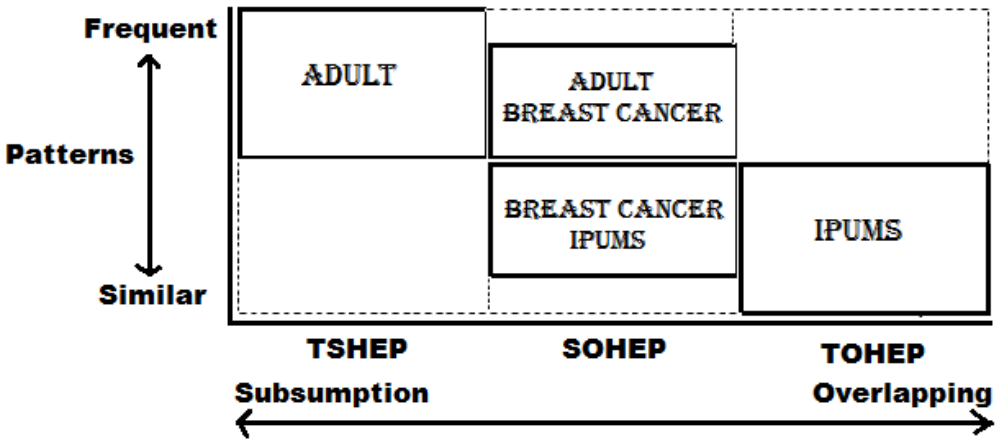


Figure 4.4. AOI-HEP mining interest matrix

Graph in figure 4.5 shows consistency with frequent pattern mining interest in matrix in figure 4.4 where TSHEP adult dataset, SOHEP adult dataset and SOHEP breast cancer dataset have 100%, 50% and 50% frequent pattern mining interest respectively. Meanwhile, graph in figure 4.6 shows consistency with similar pattern mining interest in matrix in figure 4.4 where SOHEP breast cancer dataset, SOHEP IPUMS dataset and TOHEP IPUMS dataset have 50%, 25% and 100% similar pattern mining interest respectively. The graphs in figure 4.5 and 4.6 show implementation equation 4.1 which is probability AOI-HEP mining interest between frequent and similar patterns against HEP types such as TSHEP, SOHEP or TOHEP, as mentioned in sub section 4.8.

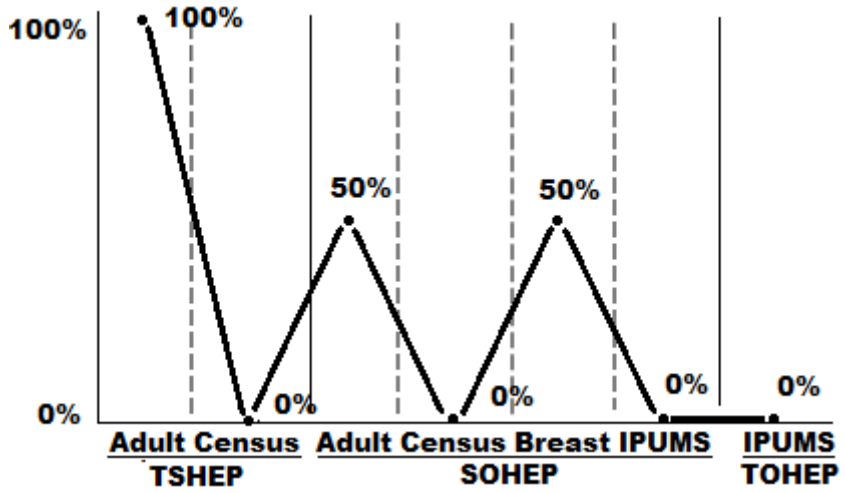


Figure 4.5. AOI-HEP Frequent pattern mining interest

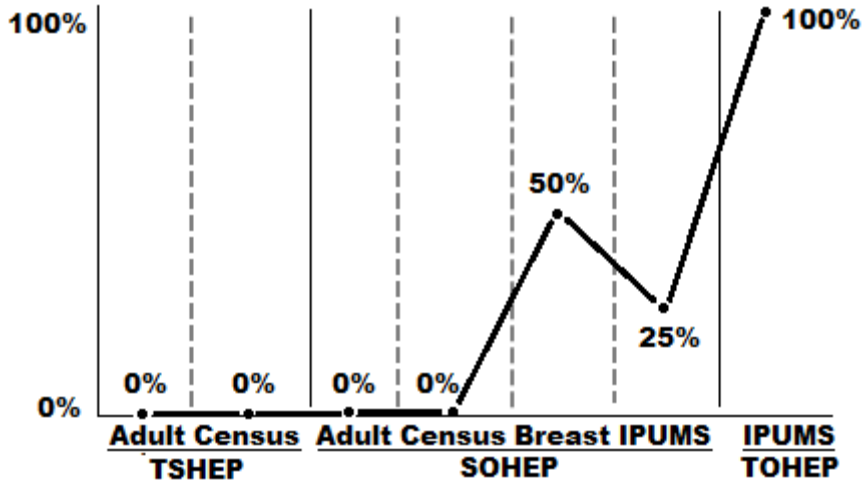


Figure 4.6. AOI-HEP Similar pattern mining interest

In accordance with above explanation, graph in figure 4.7 shows consistency with frequent and similar patterns mining interest in matrix in figure 4.4 where adult and breast cancer datasets have 72.23% and 50% frequent pattern mining interest respectively. Whilst breast cancer and IPUMS datasets have 50% and 58.33% similar pattern mining interest respectively. The graph in figure 4.7 shows implementation equation 4.2 which is average probability AOI-HEP mining interest between frequent and similar patterns, as mentioned in sub section 4.8.

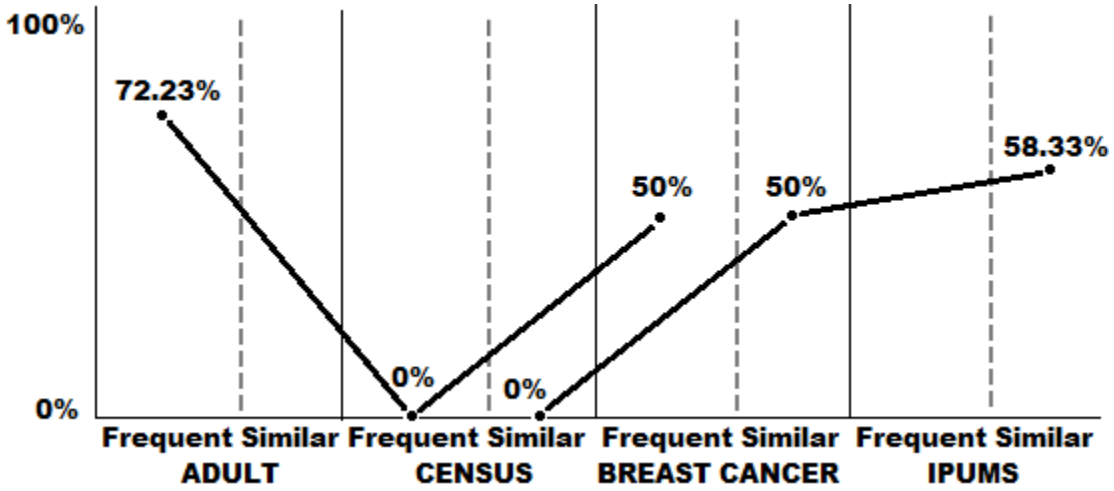


Figure 4.7. AOI-HEP Frequent and Similar patterns mining interest

The AOI-HEP mining interest between frequent and similar pattern for each dataset is influenced by learning on high level concept in one of chosen attribute. Adult, breast,cancer, census and IPUMS datasets learn on high level concept in “workclass”, ”clump thickness”,

“means” and “marst” attributes respectively. Extended experiment upon census dataset which have no AOI-HEP mining interest for both frequent and similar patterns shows that census dataset have AOI-HEP mining interest for similar patterns when learn on high level concept in “marital” attribute (appendix 12).

Extended experiment upon adult dataset which have AOI-HEP mining interest for frequent patterns shows that adult dataset have no AOI-HEP mining interest for both frequent and similar patterns when learn on high level concept in “marital-status” attribute (appendix 3). Moreover, extended experiment upon breast cancer dataset which have AOI-HEP mining interest for both frequent and similar patterns shows that breast cancer dataset have no AOI-HEP mining interest for both frequent and similar patterns when learn on high level concept in “cell size” or “cell shape” or “bare nuclei” attribute in appendix 7 or 8 or 9 respectively.

4.9. AOI-HEP justification

Attribute Oriented Induction High level Emerging Pattern (AOI-HEP) is proposed in order to explore a new data mining technique based on current approved data mining techniques, particularly Attribute Oriented Induction (AOI) and Emerging Pattern (EP). Since AOI-HEP is combination between two data mining techniques such as AOI and EP, then AOI-HEP is better than these two data mining techniques. Obviously, AOI-HEP is perfect since its mixture of strength of these two data mining techniques. Moreover, AOI-HEP has many possible features to be explored as will be listed in next chapter. Table 4.29 shows the performance metric with number of rules resulted and time to process among proposed data mining technique AOI-HEP and current two data mining techniques such as AOI and EP. Meanwhile table 4.30 shows the performance metric among data mining techniques AOI, AOI-HEP and EP based on list features in current data mining techniques such as AOI and EP.

Table 4.29. Performance metric for number of rules resulted and time to process

	AOI	**AOI-HEP**	**EP**
Number of rules resulted	**Intermediate**	**Few**	**Many**
Time to process	**Fastest**	**Medium**	**Slow**

In number of rules resulted, table 4.29 shows AOI-HEP has superiority rather than AOI and EP where AOI-HEP has a few number of rules resulted whilst AOI and EP have intermediate and many number of rules resulted respectively. AOI-HEP has superiority with a few number of rules resulted because AOI-HEP applies cartesian product between rulesets output from AOI characteristic rule algorithm, as mentioned in section 3.2 at chapter 3. Moreover, the cartesian product are eliminated by determining type of High level Emerging Pattern (HEP) such as TSHEP, SOHEP or TOHEP. Meanwhile, EP has weakness in many number of rules resulted since EP deals with low level data which have many low level rules. AOI-HEP and AOI use concept hierarchy to generalize from low level data into high level data, and as a result AOI-HEP and AOI mining high level rules which are less than low level rules. Thus, AOI-HEP has a few number of rules resulted because AOI-HEP mining high level rules which are less than low level rules, applies cartesian product and eliminates it by determining type of HEP.

For example, the experiments in section 4.3 shows that AOI has twelve rules for each dataset or forty eight rules from four experiments dataset from UCI machine learning repository as shown between tables 4.1 to 4.8. Whilst, AOI-HEP as shown in tables 4.16 and 4.17 has six rules for adult dataset, two rules for breast cancer dataset, eight rules for census dataset and six rules for IPUMS dataset.It means AOI-HEP has twenty two rules for four experiments dataset from UCI machine learning repository as shown between tables 4.9 to 4.15. Moreover, mining frequent and similar patterns with AOI-HEP has more less rules with only nine rules where five and four rules as shown in table 4.23 and 4.28 respectively. Meanwhile, since we do not have any experiments with EP then we can not measure the number of rules can be created. Also, since AOI and AOI-HEP mining high level rules whilst EP mining low level rules then AOI and AOI-HEP are not suitable to be compared with EP. However, since EP mining low level rules then the number of rules will be more than the number rules which created either by AOI or AOI-HEP since they mine high level rules. Thus, in term of number of rules resulted as shown in table 4.29, mining rules with AOI will have intermediate rules with forty eight rules and mining rules with AOI-HEP will have few rules with twenty two or nine rules. Moreover, mining rules with EP will have many rules since EP is running on low level data.

However, in time to process as shown in table 4.29, AOI-HEP has medium classification since AOI-HEP applies cartesian product between rulesets output from AOI characteristic rule algorithm as mentioned in previous paragraph. Performance metric in table

4.29 shows AOI-HEP and AOI have better performance in time to process against EP, since both of them deal with high level data. Similar as mentioned in previous paragraph since EP deals with low level data which have many low level rules then EP has weakness with slow performance in time to process. Similar as mentioned in previous paragraph, AOI-HEP and AOI use concept hierarchy to generalize from low level data into high level data where high level data have less data rather than low level data. Obviously, time to process high level data will have better performance since deal with less data and the other hand, time to process low level data will have slow performance since deal with huge data. Rather than AOI, AOI-HEP has lower performance in time to process, since AOI-HEP applies cartesian product between rulesets output from AOI characteristic rule algorithm, and cartesian product are eliminated by determining type of HEP such as TSHEP, SOHEP or TOHEP. Applying cartesian product and its elimination in AOI-HEP can be seen in section 3.2 at chapter 3.

For example as shown in section 4.3, the AOI-HEP experiments upon four datasets from UCI machine learning such as adult, breast cancer, census and IPUMS datasets have running time approximately three, three, four and thirteen seconds respectively. As mention in section 3.2 at chapter 3, AOI-HEP framework as shown in figure 3.1 at chapter 3 is combination between AOI characteristic rule algorithm and HEP algorithm where AOI characteristic rule algorithm will be run firstly. Since AOI-HEP framework is combination between AOI characteristic rule algorithm and HEP algorithm, then the running time with AOI upon four experiment datasets from UCI machine learning repository such as adult, breast cancer, census and IPUMS should have less running time. Meanwhile, since we do not have any experiments with EP then we can not measure the time to process in mining rules process. Since AOI and AOI-HEP mining high level rules whilst EP mining low level rules then AOI and AOI-HEP are not suitable to be compared with EP. However, since EP mining low level rules, then there will be many processes time upon low level data rather than AOI or AOI-HEP which mining high level rules. Thus, in term of time to process as shown in table 4.29, mining rules with AOI will have the fastest time while mining rules with AOI-HEP will have medium time with running time approximately three, three, four and thirteen seconds upon adult, breast cancer, census and IPUMS datasets respectively. Moreover, mining rules with EP will have slow time to process since EP process many low level data.

In table 4.30, there are 4 lists features belong to current data mining techniques AOI and EP. In table 4.30, sign ✓shows as applicable whilst sign X shows as not applicable and sign ✓* shows as future development particularly for AOI-HEP. First, growth rates is typical strength of EP and become justification for powerful discrimination in EP has been adopted in AOI-HEP for ratio of the supports at the same or different high level itemsets instead of the same low level itemsets as used in EP. However, AOI has not implemented growth rates. The next feature is Jumping Emerging Pattern (JEP) where JEP is EP with support is 0 in one dataset and more than 0 in the other dataset or EP as special type of EP which is having infinite growth rates (∞). Although AOI-HEP implements improvement growth rates, nevertheless as mentioned in section 3.6 at chapter 3, AOI-HEP does not have JEP since rule in ruleset has $|r_n^x|$ as the number of tuples. Meanwhile, since AOI has not implemented growth rates, then AOI has not had JEP.

Table 4.30. Performance metric for features from current data mining techniques AOI and EP

NO		AOI	AOI-HEP	EP
1	**Growth rates**	**X**	**✓**	**✓**
2	**JEP (Jumping Emerging Patterns)**	**X**	**X**	**✓**
3	**Concept hierarchy**	**✓**	**✓**	**X**
4	**Datawarehouse techniques (Roll up and drill down)**	**✓**	**✓***	**X**

The third feature is concept hierarchy where applicable for AOI and AOI-HEP and concept hierarchy is used to generalize from low level data into high level data. Meanwhile, EP does not use concept hierarchy since EP deals with low level data. The next features is datawarehouse technique such as roll up and drill down, and AOI has been recognized applicable for this datawarehouse technique using concept hierarchy which access either low level or high level concepts. Meanwhile, AOI-HEP is not yet applicable with this datawarehouse technique, but since AOI-HEP uses concept hierarchy then there is possibility to implement this datawarehouse technique in the future. Furthermore, EP is not applicable with datawarehouse technique since EP does not have access to high level data and just only deals with low level data and moreover EP does not implement concept hierarchy. Performance metric in table 4.30 shows that AOI-HEP is better than other two data mining techniques AOI and EP, and adoption these two data mining techniques will increase the

ability AOI-HEP as proposed data mining technique. Some features such as datawarehouse technique need to be explored in future will increase ability AOI-HEP as new proposed data mining technique.

4.10. Conclusion

Experimental evaluation from AOI-HEP mining framework on four datasets from the UCI machine learning repository is shown in this chapter. Each of the datasets has their own concept hierarchies built from five chosen attributes and one of the attributes with its concept hierarchy was chosen as learning to discriminate between datasets D1 and D2. The AOI-HEP application which is a combination between AOI and HEP algorithms, was run and shows the results where each of datasets has SOHEP, two TSHEP in adult and census datasets and only one TOHEP in the IPUMS dataset. From HEP patterns results, frequent patterns with strong discriminating power and similar pattern were mined. The experimental evaluation discovery showed that there are five frequent patterns which are two TSHEP, two SOHEP from the adult dataset and one SOHEP from the breast cancer dataset. Moreover, there are four similar patterns with two TOHEP and one SOHEP from IPUMS dataset and one SOHEP from the breast cancer dataset.

The discovery showed that a strong discrimination rule can be mined from frequent patterns since they have large growth rates and supports in target (D2) dataset, small supports in the contrasting (D1) dataset where support in the contrasting (D1) dataset is less than the support in the target (D2) dataset. All the frequent patterns mining have strong discriminant rules since support in contrasting (D1) dataset is less than the support in target (D2) dataset. Meanwhile from similar patterns there is possibility of having strong discriminant rules. The discovery showed that there are two similar patterns with strong discriminant rules when their supports in the contrasting (D1) dataset are less than the supports in target (D2) dataset. On the other hand, there are two similar patterns as non strong discriminant rules when their supports in contrasting (D1) dataset are greater than the supports in target (D2) dataset.

The experimental upon four UCI machine learning repository show that adult, breast cancer and IPUMS datasets are interested to be mined and not for census dataset. Adult dataset which learn on high level concept in “workclass” attribute have AOI-HEP mining

interest for frequent patterns with average probability (confidence) 72.23%. Whilst breast cancer dataset which learn on high level concept in “clump thickness” attribute have AOI-HEP mining interest for both frequent and similar pattern with probability (confidence) 50% respectively. Moreover, IPUMS dataset which learn on high level concept in “marst” attribute have AOI-HEP mining interest for similar patterns with average probability (confidence) 58.33%. Meanwhile, census dataset which learn on high level concept in “means” attribute have no AOI-HEP mining interest for both frequent and similar patterns. Next, chapter 5 will show the conclusion and future research.

Chapter 5: Conclusion

5.1. Introduction

This chapter presents a summary of the work in this thesis and further future research. Section 5.2 gives a summary of AOI-HEP as a proposed high-level emerging pattern mining framework and experimental evaluations from real datasets given in section 4.8 of chapter 4. Section 5.3 presents the possible future research in order to extend the AOI-HEP mining framework.

5.2. Summary

This thesis proposed Attribute Oriented Induction High level Emerging Pattern (AOI-HEP) as a new data mining framework combining two data mining techniques i.e. Attribute Oriented Induction (AOI) [30-47] and Emerging Patterns (EP) [65-72,75-84,90,92-104]. The AOI-HEP application was implemented as a hybrid between AOI characteristic rule mining and HEP algorithms. AOI-HEP combine the powerful features of AOI and EP by using concept hierarchy in AOI to generalize into high level data and applying growth rates in EP respectively. This approach produced powerful discrimination for high level data. AOI characteristic rule algorithm uses concept hierarchy as background knowledge for data generalization and uses attribute and rule thresholds to eliminate distinct attributes and tuples respectively. Meanwhile, HEP algorithm applies two functions i.e. similarity function C{ R_i^1, R_j^2 } and growth rate function GR{ R_i^1, R_j^2 }, (introduced in section 3.4 at chapter 3), where rulesets R_i^1 and R_j^2 are AOI's outputs from datasets D1 and D2 respectively.

The C{ R_i^1, R_j^2 } function is a metric similarity function which applies cartesian product between rulesets R_i^1 and R_j^2, and eliminate the cartesian product by determining between the type of HEP (TSHEP, SOHEP or TOHEP) and frequent or similar pattern. Determining between types of HEP and frequent or similar pattern, we applied a summing similarity value (SLV value) function to categorize attributes by comparing their values and hierarchy level between rulesets R_i^1 and R_j^2 (LV value). Threshold LV values were taken as

(LV=0.4 or LV=0.5) based on subsumption and (LV=2.0 or LV=2.1) based on overlapping combinations. TSHEP are rules that are completely subsumed, SOHEP are rules that overlap and are subsumed, whilst TOHEP are rules that are completely overlapping.

Meanwhile, frequent pattern were mined from TSHEP or SOHEP while similar pattern were mined from SOHEP or TOHEP. Frequent patterns were mainly TSHEP completely subsumed with full similar subsumption LV=0.5 or TSHEP (or SOHEP) with frequent similar subsumption LV=0.5 with value percentage (m-1)/m*100. Moreover, similar patterns are TOHEP that are completely overlapping with full similar overlapping LV=2.0 or TOHEP with combination overlapping LV=2.0 and LV=2.1. Furthermore, similar patterns are SOHEP with frequent similarity overlapping LV=2.0 or SOHEP with combination overlapping LV=2.0 and LV=2.1 with value percentage (m-1)/m*100. However, similar patterns are not TOHEP with full similarity overlapping LV=2.1 or SOHEP with frequent similarity overlapping LV=2.1 at percentage value of (m-1)/m*100. From frequent and similar patterns we can build discriminant rules, where from frequent pattern we can discover strong discrimination rules and from similar pattern we can discover strong and weak discrimination rules. The strong discrimination rules have large growth rates and supports in target (D2) dataset, small supports in contrasting (D1) dataset where support in contrasting (D1) dataset is less than support in target (D2) dataset.

The GR{ R_i^1, R_j^2 } function is ratio of the supports between rulesets R_i^1 and R_j^2 and eliminates the type of HEP as the outputs from C{ R_i^1, R_j^2 } function with growth rate less equal than given GrowthRate threshold. Thegrowth rate function GR{ R_i^1, R_j^2 } is influenced with a standard function used in Emerging Patterns (EP) and the difference is discovering high level emerging pattern with the same or different itemset instead of low level pattern with the same itemset in EP.

The experiments were carried out with four real datasets : adult, breast cancer, census and IPUMS datasets from UCI machine learning repository, where each of dataset has its own concept hierarchies built from five chosen attributes. One of the attribute was chosen to learn high level concepts in its concept hierarchy to discriminate between datasets D1 and D2. The AOI-HEP application was run with attribute and rule thresholds 6 and the experiments show that most of datasets have SOHEP, but not TSHEP and TOHEP and the most rarely to find is TOHEP. The experiments discovered that there are five frequent patterns with strong discrimination rules as shown in table 4.23 at chapter 4 which are two TSHEP and two SOHEP from adult dataset and one SOHEP from breast cancer dataset.

Moreover, the experiments discovered that there are four discriminant rules from similar patterns as shown in table 4.28 at chapter 4 which are two TOHEP and one SOHEP from IPUMS dataset and one SOHEP from breast cancer dataset. From these four discriminant rules as shown in table 4.28 at chapter 4, one TOHEP and SOHEP from IPUMS dataset are strong discriminant rules since they have large growth rates (1.53 and 3.46636) and supports in target (D2) datasets (4.5 and 5.447), small supports in contrasting (D1) datasets (2.9 and 1.571) where support in contrasting (D1) dataset is less than support in target (D2) dataset.

The AOI-HEP mining framework was presented to discover frequent and similar patterns for each dataset. This was influenced by learning high level concept in one of chosen attribute. Adult dataset which was learned on high level concept "workclass" attribute had frequent patterns with average probability (confidence) 72.23%. Whilst breast cancer dataset learned on high level concept "clump thickness" attribute had frequent and similar patterns with probability (confidence) 50% respectively. Moreover, IPUMS on "marst" attribute had similar patterns with average probability (confidence) 58.33%. In addition, census dataset which learn on high level concept "means" attribute did not have both frequent and similar patterns. However, extended experiment upon census dataset which have no frequent and similar patterns show that census dataset have AOI-HEP mining interest for similar patterns when learned on high level concept in "marital" attribute (appendix 12).

The major contributions from AOI-HEP are:

- A Hybrid approach between AOI characteristic rule mining and emerging patterns (EP).
- Powerful discrimination for high level data.
- Application of two functions for the HEP algorithm part: i.e. similarity function C{ R_i^1, R_j^2 } and growth rate function GR{ R_i^1, R_j^2 }. The C{ R_i^1, R_j^2 } function is a metric similarity function which applies cartesian product between rulesets R_i^1 and R_j^2, and eliminate the cartesian product by determining between the type of HEP (TSHEP, SOHEP or TOHEP) and frequent or similar pattern. The GR{ R_i^1, R_j^2 } function is ratio of the supports between rulesets R_i^1 and R_j^2 and eliminates the type of HEP as the outputs from C{ R_i^1, R_j^2 } function with growth rate less equal than given GrowthRate threshold. The growth rate function GR{ R_i^1, R_j^2 } is influenced with a standard function used in Emerging Patterns (EP) and the difference is discovering high level emerging pattern with the same or different itemset instead of low level pattern with the same itemset in EP.
- Mining different types of HEP patterns such as frequent and similar patterns.

- Frequent patterns can be mined from totally subsumed HEP (TSHEP) or subsumed overlapping (SOHEP). Frequent patterns were mainly TSHEP completely subsumed with full similar subsumption LV=0.5 or TSHEP (or SOHEP) with frequent similar subsumption LV=0.5 with value percentage (m-1)/m*100.
- Similar patterns can be mined from SOHEP or TOHEP.
- Finding frequent patterns that can build strong discriminant rules.
- Finding similarity patterns that can build strong or weak discriminant rules.

5.3. Future research

AOI-HEP as a hybrid between Attribute-Oriented Induction (AOI) and Emerging Patterns (EP) can be extended in many ways, including finding data irregularities and associations between two datasets. [39]. Whilst EP is recognized as a powerful mining technique to discriminate datasets [75,79]. Growth rate as ratio of the supports in one dataset to another dataset is justification for powerful discrimination, become the typical strength of EP.

We now show the extent to which AOI-HEP mining technique can be extended :

1. Inverse the discovery learning.

   In discovering interesting EP, the discovery can be done in both datasets where not just only from contrasting (D1) to target (D2) datasets ($\frac{D2}{D1}$), but can be extended from target (D2) to contrasting (D1) datasets ($\frac{D1}{D2}$). DeEP classifier uses three procedures to discover border representation of EPs with the third procedure to discover JEP and EP from both datasets. More specifically, one procedure uses INTERSECTOPERATION algorihtm to discover EP from both datasets namely commonT=[{0},Rp] $\cap$ [{0},Rn] [65, 72, 81]. AOI-HEP mining framework which is influenced by EP has been proved to learn HEP only from contrasting (D1) to target (D2) datasets ($\frac{D2}{D1}$). Since DeEP which is influenced by EP can discover interesting EPs in both datasets, than the discovery HEP can be extended in order to find many interesting HEP. The AOI-HEP ability can be extended not only to learn HEP from contrasting (D1) to target (D2) datasets ($\frac{D2}{D1}$), but can be extended to learn HEP from target (D2) to contrasting (D1) datasets ($\frac{D1}{D2}$). For instance, extended experiment upon census dataset which did not have both frequent and similar patterns shows that inverse discovery learning upon census dataset will have frequent

patterns from one SOHEP and one TSHEP. The inverse discovery learning on high level concept "means" attribute of census dataset will reverse from HEP learning $\frac{D2}{D1} = \frac{Green\ concept}{No\ Green\ concept} = \frac{1980\ instances}{809\ instances}$ to $\frac{D1}{D2} = \frac{No\ Green\ concept}{Green\ concept} = \frac{809\ instances}{1980\ instances}$.

2. Learning more than two datasets.

   EP with border-based algorithm influenced EP-based classifier[103] algorithms such as CAEP[84], CAEEP [100], DeEP[65], BCEP[78], CEP[77], JEP-C[83], JEP space[82] which can do classification task by learning more than two datasets. AOI-HEP mining framework which is influenced by EP has been proved to learn from only two datasets. Since EP with EP-based classifier has ability to do the classification by learning more than two datasets, then AOI-HEP can be extended to learn classification by learning more than two datasets. Moreover, the extended AOI-HEP has the ability to learn more than two datasets, learning classification rules can be extended to learn other knowledge rules as we explain in the next list.

3. Learning other knowledge rules.

   AOI is recognized can learn different kinds of knowledge rules such as characteristic rules, discriminant rules, classification rules, data evolution regularities, association rules and cluster description rules. AOI-HEP mining framework which is influenced by AOI has been proved to learn discrimination rules. Since AOI can learn different kinds of knowledge rules [39], then AOI-HEP ability can be extended to learn other knowledge rules other than discriminant rule such as characteristic rules, classification rules, data evolution regularities, association rules and cluster description rules.

4. Experiment's extension with other AOI algorithms apart from AOI characteristic rule algorithm.

   AOI-HEP uses AOI characteristic rule algorithm which is combined with HEP algorithm. AOI is recognized can learn other different kinds of knowledge rules [39] apart from characteristic rule such as discriminant rules, classification rules, data evolution regularities, association rules and cluster description rules. Thus, future research can be extended where not only using AOI characteristic rule algorithm, but using other AOI algorithms or to build new algorithm which replace AOI algorithms in order to access high level data by generalizing from low level to high level data.

5. Learning multidimensional view.

   AOI can perform datawarehouse techniques such as roll up (progressive generalization) or drill down (progressive specialization) operations, where data can be seen in

multidimensional view. AOI performs datawarehouse technique with concept hierarchies as AOI background knowledge when roll up is achieved by generalization low level with high level concepts, whilst drill down is achieved by specialization high level with low level concepts [44]. AOI-HEP mining framework which is influenced by AOI has been proved to uses concept hierarchies to access high level by generalizing from low level to high level concepts (roll up). Since AOI uses concept hierarchies to perform datawarehouse technique (roll up and drill down) then AOI-HEP ability can be extended from performing datawarehouse technique by using concept hierarchies to generalize from low level to high level concepts (roll up) to specialize from high level to low level concepts (drill down). The ability AOI-HEP can be extended as Online Analytical Mining (OLAM) (also called OLAP Mining) which integrates OLAP with data mining in order to create data cube for multidimensional view

6. Prediction from similar pattern.

   Searching similar patterns are important and can be used for segmentation or prediction. For example in banking system, banking segmentation and banking prediction with similar banking transaction could help to show banking transaction prediction, while similar customer behaviour pattern could help to uncover fraud, and loan prediction [109]. Prediction can be made based on past data which can be modelled by statistical techniques of regression. For instance, we can develop a model to predict the customer behaviour pattern to prevent uncover fraud and loan prediction[109] or develop model to predict the salary of graduate with 5 years of work experiences.

7. Extended experiment with the type of HEP from overlapping and subsumption into disjoint as dissimilar patterns.

   HEP algorithm which is part of AOI-HEP uses C$\{ R_i^1, R_j^2 \}$ function to determine the type of HEP (TSHEP, SOHEP or TOHEP) with categorization of attribute comparison value and hierarchy level between rulesets R_i^1 and R_j^2 (LV value) based on subsumption (LV=0.4 or LV=0.5) and overlapping (LV=2.0 or LV=2.1) combination. Subsumption represents for searching frequent and infrequent patterns while overlapping represents for searching similar patterns. The subsumption is subsumption between attributes comparison value while overlapping is similarity between attributes comparison value. The categorization of attribute comparison value and hierarchy level between rulesets R_i^1 and R_j^2 (LV value) can be extended from subsumption and overlapping into disjoint where

attribute comparison value is dissimilar. Thus, disjoint will represent for searching dissimilar patterns. Dissimilar patterns are interested to be mined since as reverse of similar patterns and will extend mining the type of HEP (TSHEP, SOHEP or TOHEP) into other types of HEP. Those other types of HEP are Total Disjoint HEP (TDHEP), Subsumption Disjoint HEP (SDHEP), Overlapping Disjoint HEP (ODHEP), Subsumption Overlapping Disjoint HEP (SODHEP). TDHEP are rules that are completely disjoint, SDHEP are rules that are subsumed and disjoint, ODHEP are rules that overlap and disjoint and SODHEP are rules that are subsumed, overlap and disjoint.

8. Comparing with other data mining techniques.

   In order to get public achievement, AOI-HEP needs to be compared with current data mining techniques by finding strengths and weaknesses. Comparation can be done by performance measurement in condition with the same dataset and concept hierarchies, run in the same computer to get transparent comparation.

9. Extended experiment for learning concept hierarchies.

   AOI-HEP used four real dataset from UCI machine learning repository [56] where each of real dataset has concept hierarchies built from five chosen attributes as shown between appendices 1 and 20. Each of real dataset was discriminate between two high level concepts in one of their chosen concept hierarchies attribute for learning purposes. Since the AOI-HEP experimental only learned from one of their chosen concept hierarchies attribute, then the learning for each of real dataset can be extended to other concept hierarchies attributes in order to find other HEP (TSHEP, SOHEP or TOHEP) and frequent or similar pattern. By the end, more knowledge can be explored from each of dataset.

10. Extended experiment for input dataset from flat files into relational databases and its combination.

    AOI-HEP which is influenced by AOI has been proven to be implemented with input real flat files dataset from UCI machine learning repository [56]. However, some business industries such as retail, banking, health care and education use relational databases for their daily transactions. Since AOI has input dataset from relational databases [30-32,39,45,47] then AOI-HEP ability can be extended to learn for input dataset in the form of relational databases and combination between flat files and relational databases. Performance issues will occur since flat files and relational databases have different architecture where relational databases have multiple tables which relate to each other through special key attributes and obviously using multiple tables are slower than using

one table. Somehow, SQL query language [49-51] can be used to increase relational databases performance.

Further work listed in 1,2,5,7 and 9 are easy to implement but not 3,4,6,8 and 10. In (1), it is easily implemented since there will be little changes in the HEP algorithm by reversing HEP learning from contrasting (D1) to target (D2) datasets ($\frac{D2}{D1}$) become HEP learning from target (D2) to contrasting (D1) datasets ($\frac{D1}{D2}$). With (2), it is easy to implement since we can use pair-wise feature concept [83] which is used by JEP-Classifier to deal for datasets with more than two classes. Moreover, list number 5 is easy to be implemented since AOI-HEP is able to generalize from low level to high level concepts (roll up) then specialize from high level to low level concepts (drill down) is easy to be implemented as well. Furthermore, line number 7 is easy to be implemented by extending categorization of attribute comparison value and hierarchy level between rulesets R_i^1 and R_j^2 (LV value) into disjoint. Finally, List number 9 is easy to be implemented since each of dataset has their concept hierarchies then the learning each concept hierarchy for datasets can be easily extended.

Meanwhile, number 3 is hard to be implemented since need to implement other knowledge rules other than discriminant rule such as characteristic rules, classification rules, data evolution regularities, association rules and cluster description rules. While list number 4 is hard to be implemented since AOI algorithms apart from AOI characteristic rule algorithm such as discriminant rules, classification rules, data evolution regularities, association rules and cluster description rules need to be implemented. Moreover, list number 6 is hard to be implemented since we need accurate past data and to develop statistical technique for modelling purposes. Furthermore, list number 8 is hard to be implemented since need to implement the current data mining techniques for comparing purposes. Finally, list number 10 is hard to be implemented since flat files and relational databases have different architecture. Some changing and improvement need to be done upon HEP algorithm regarding with changing input dataset from flat files into relational databases or combination between flat files and relational databases.

Finally, AOI-HEP has been used effectively with four large and real datasets from UCI machine learning repository [56] to discover strong discriminant rules from mining type of HEP (TSHEP, SOHEP, TOHEP), frequent and similar patterns. Since AOI-HEP is a new

data mining framework and proven successful, it presents itself as a viable interesting future work and useful in decision process making.

Publication List

1. Muyeba,M.K., Khan, M.S., Warnars, S. and Keane,J.2011. A Framework to Mine High-Level Emerging Patterns by Attribute-Oriented Induction. In proceedings of the 12th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL 2011), University of East Anglia, Norwich, United Kingdom, 7-9 September 2011,170-177.

2. Warnars, S.2012. Attribute Oriented Induction of High-level Emerging Patterns.In proceedings of the IEEE International Conference on Granular Computing (IEEE GrC2012), Hangzhou, China, 11-13 August 2012, 628-633.

A Framework to Mine High-level Emerging Patterns by Attribute-oriented Induction

[1]Maybin K. Muyeba [2]Muhammad S. Khan [3]Spits Warnars [4]John Keane
[1, 3]Sch. of Computing, Maths and Digital Techn., Manchester Metropolitan University, UK
{m.muyeba, s.warnars}@mmu.ac.uk
[2]Department of Computer Science, School of Electrical Engineering and Computer Science, University of Liverpool, UK
mskhan@liverpool.ac.uk
[4]School of Computer Science, The University of Manchester, UK
john.keane@cs.manchester.ac.uk

Abstract. This paper presents a framework to mine summary emerging patterns in contrast to the familiar low-level patterns. Generally, growth rate based on low-level data and simple supports are used to measure emerging patterns (EP) from one dataset to another. This consequently leads to numerous EPs because of the large numbers of items. We propose an approach that uses high-level data: high-level data captures the data semantics of a collection of attributes values by using taxonomies, and always has larger support than low-level data. We apply a well known algorithm, attribute-oriented induction (AOI), that generalises attributes using taxonomies and investigate properties of the rule sets obtained by generalisation algorithms.

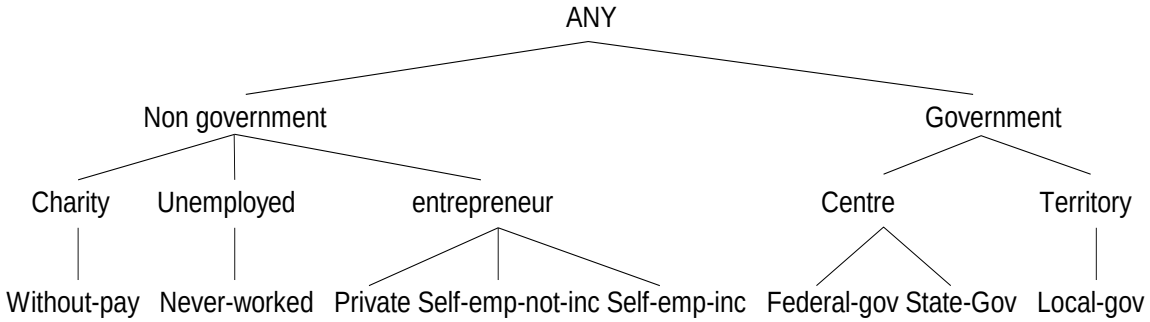
Attribute Oriented Induction of High-level Emerging Patterns

Spits Warnars
School of Computing, Maths and Digital Technology
Manchester Metropolitan University
Manchester, United Kingdom
08975791@stu.mmu.ac.uk

Abstract—Attribute Oriented Induction (AOI) produces high level characteristic summary data but does not discover new emerging patterns. Emerging Pattern (EP) algorithms discover emerging patterns between datasets but mostly consider low-level data. This paper introduces an algorithm, AOI-HEP, derived from both AOI and High-level Emerging Patterns (HEP), where HEP discriminates the high level data from AOI. The main objective is to discover characteristic HEP patterns using AOI. To filter out the large overlapping and subsuming attribute values in the output, a Cartesian product of attribute values, a similarity metric based on attribute values and attribute hierarchy level are applied. Experiments used four datasets from the UCI machine learning repository. Results show that various interesting HEP patterns can be generated by using the AOI-HEP algorithm.

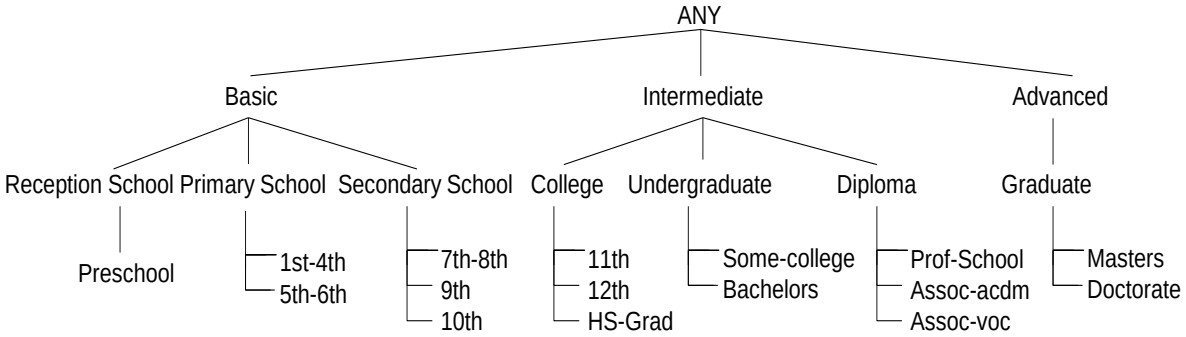
Appendices

Appendix 1: Concept hierarchy for workclass attribute of adult dataset.



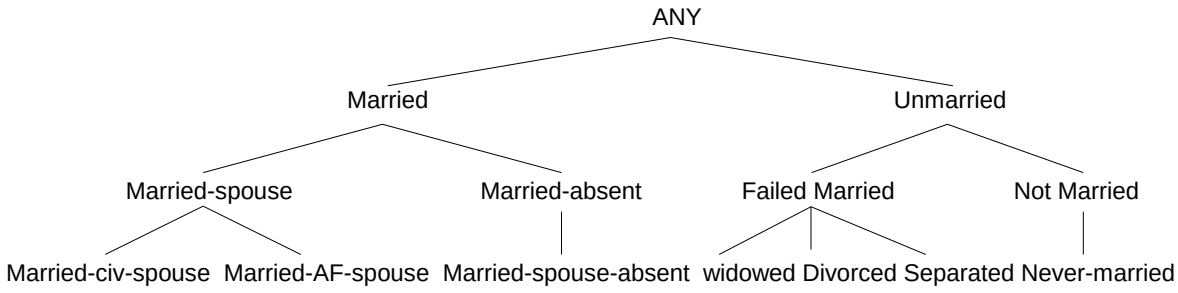
Without-pay	⊂	Charity
Never-worked	⊂	Unemployed
Private,Self-emp-not-inc,Self-emp-inc	⊂	entrepreneur
Federal-gov, State-gov	⊂	Centre
Local-gov	⊂	Territory
Charity, Unemployed, Entrepreneur	⊂	Non Government
Centre, Territory	⊂	Government
Non Government,Government	⊂	ANY

Appendix 2: Concept hierarchy for education attribute of adult dataset.



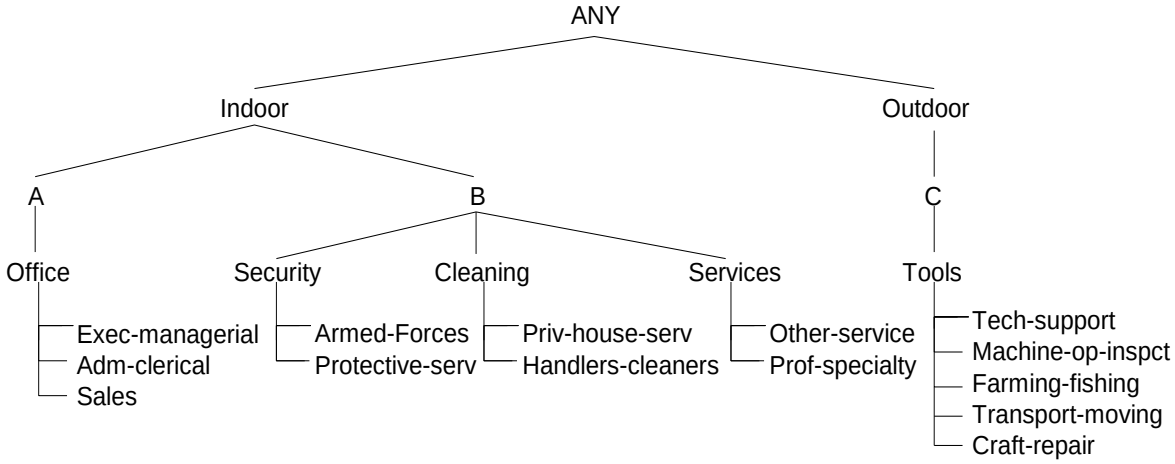
Preschool	⊂	Reception School
1st-4th,5th-6th	⊂	Primary School
7th-8th,9th,10th	⊂	Secondary School
11th,12th,HS-grad	⊂	College
Some-college,Bachelors	⊂	Undergraduate
Masters,Doctorate	⊂	Graduate
Prof-school,Assoc-acdm,Assoc-voc	⊂	Diploma
Reception School,Primary School,Secondary School	⊂	Basic
College,Undergraduate,Diploma	⊂	Intermediate
Graduate	⊂	Advanced
Basic,Intermediate,Advanced	⊂	ANY

Appendix 3: Concept hierarchy for marital-status attribute of adult dataset.



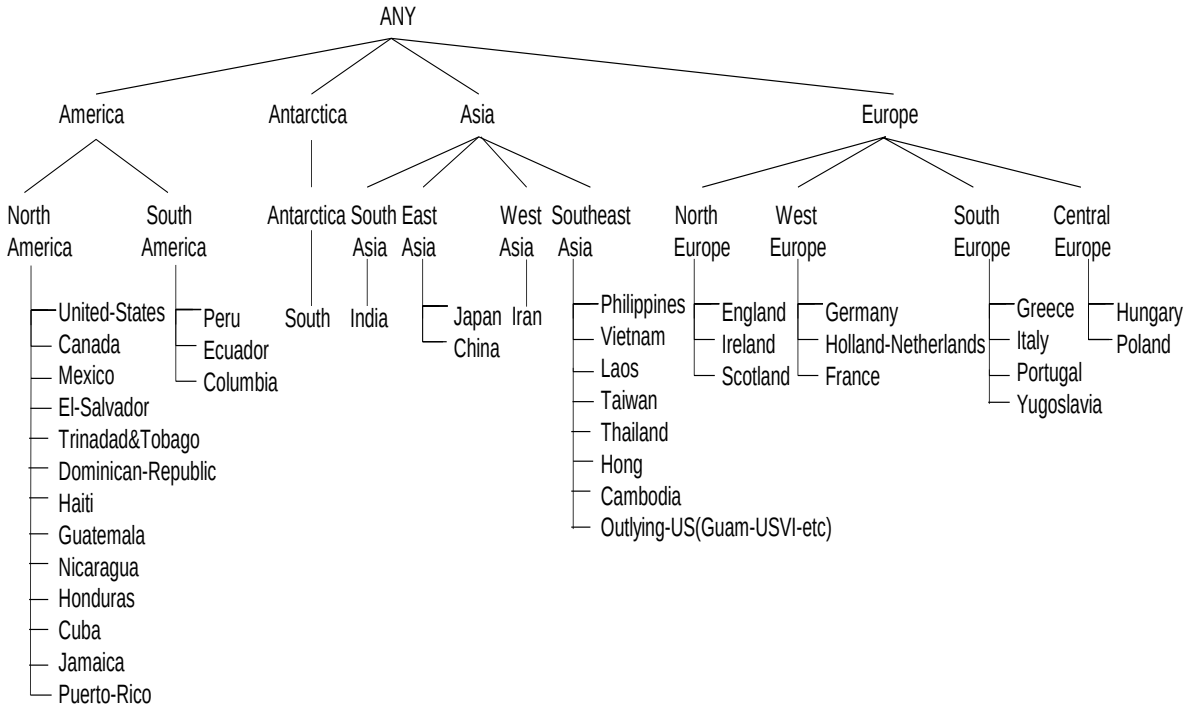
Married-civ-spouse, Married-AF-spouse	⊂	Married-spouse
Married-spouse-absent	⊂	Married-absent
Widowed,Divorced,Separated	⊂	Failed Married
Never-married	⊂	Not Married
Married-spouse, Married-absent	⊂	Married
Failed Married, Not Married	⊂	Unmarried
Married,Unmarried	⊂	ANY

Appendix 4: Concept hierarchy for Occupation attribute of adult dataset.



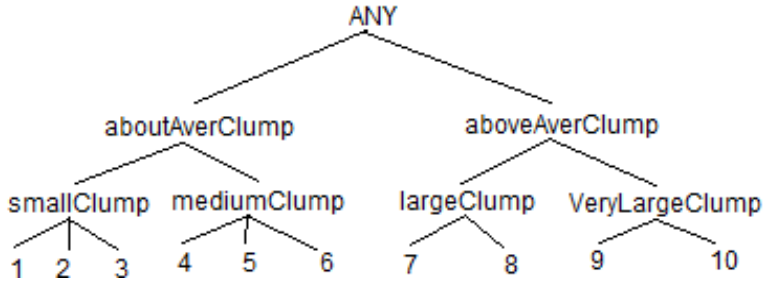
Tech-support,Machine-op-inspct,Farming-fishing, Transport-moving, Craft-repair	⊂	Tools
Armed-Forces, Protective-serv	⊂	Security
Exec-managerial,Adm-clerical,Sales	⊂	Office
Priv-house-serv, Handlers-cleaners	⊂	Cleaning
Other-service, Prof-specialty	⊂	Services
Office	⊂	A
Security,Cleaning, Services	⊂	B
Tools	⊂	C
A, B	⊂	Indoor
C	⊂	Outdoor
Indoor, Outdoor	⊂	ANY

Appendix 5: Concept hierarchy for native-country attribute of adult dataset.



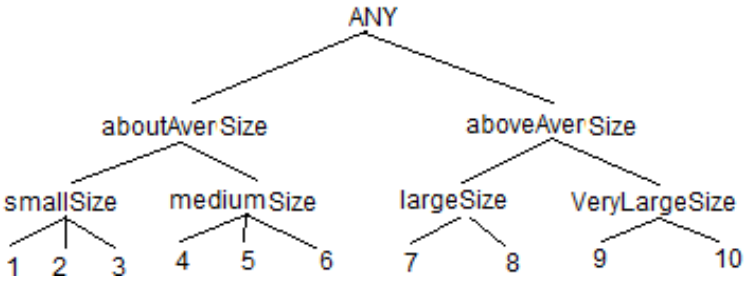
United-States,Canada,Mexico,El-Salvador,Trinadad&Tobago, Dominican-Republic,Haiti,Guatemala,Nicaragua,Honduras, Cuba,Jamaica,Puerto-Rico	⊂	North America
Peru,Ecuador,Columbia	⊂	South America
South	⊂	South Pole
India	⊂	South Asia
Japan,China	⊂	East Asia
Iran	⊂	West Asia
Philippines,Vietnam,Laos,Taiwan,Thailand,Hong,Cambodia, Outlying-US(Guam-USVI-etc)	⊂	Southeast Asia
England,Ireland,Scotland	⊂	North Europe
Germany,Holand-Netherlands,France	⊂	West Europe
Greece,Italy,Portugal,Yugoslavia	⊂	South Europe
Hungary,Poland	⊂	Central Europe
North America,South America	⊂	America
South Pole	⊂	Antarctica
Southeast Asia,South Asia,East Asia,West Asia	⊂	Asia
North Europe,West Europe,South Europe,Central Europe	⊂	Europe
America,Antarctica,Asia,Europe	⊂	ANY

Appendix 6: Concept hierarchy for clump thickness attribute of breast cancer dataset.



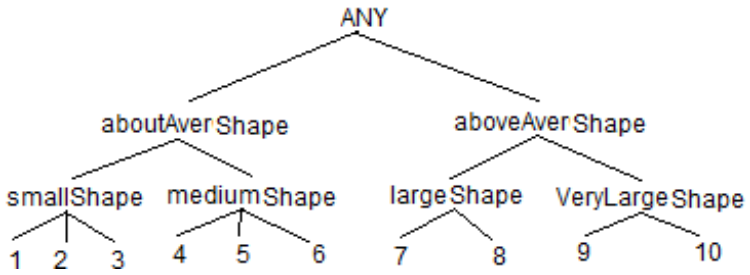
1,2,3	⊂	smallClump
4,5,6	⊂	mediumClump
7,8	⊂	largeClump
9,10	⊂	VeryLargeClump
smallClump, mediumClump	⊂	aboutAverClump
largeClump, VeryLargeClump	⊂	aboveAverClump
aboutAverClump,aboveAverClump	⊂	ANY

Appendix 7: Concept hierarchy for cell size attribute of breast cancer dataset.



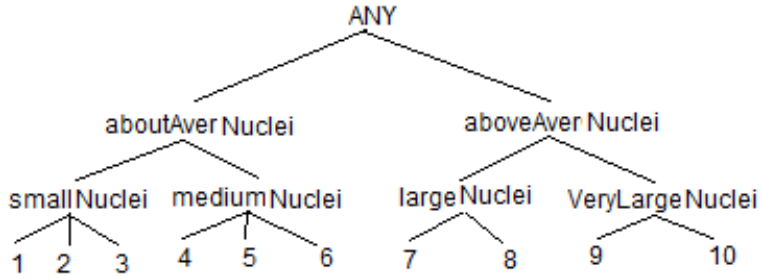
1,2,3	⊂	smallSize
4,5,6	⊂	mediumSize
7,8	⊂	largeSize
9,10	⊂	VeryLargeSize
smallSize, mediumSize	⊂	aboutAverSize
largeSize, VeryLargeSize	⊂	aboveAverSize
aboutAverSize,aboveAverSize	⊂	ANY

Appendix 8: Concept hierarchy for cell shape attribute of breast cancer dataset.



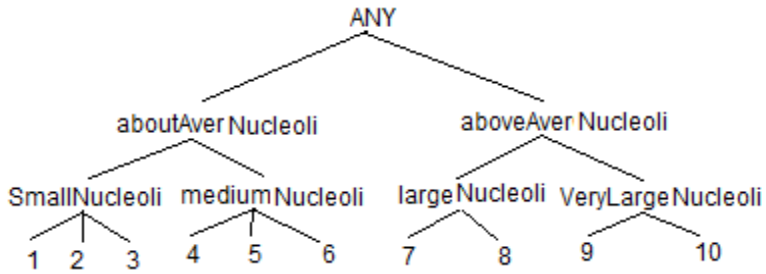
1,2,3	⊂	smallShape
4,5,6	⊂	mediumShape
7,8	⊂	largeShape
9,10	⊂	VeryLargeShape
smallShape, mediumShape	⊂	aboutAverShape
largeShape, VeryLargeShape	⊂	aboveAverShape
aboutAverShape,aboveAverShape	⊂	ANY

Appendix 9: Concept hierarchy for bare nuclei attribute of breast cancer dataset.



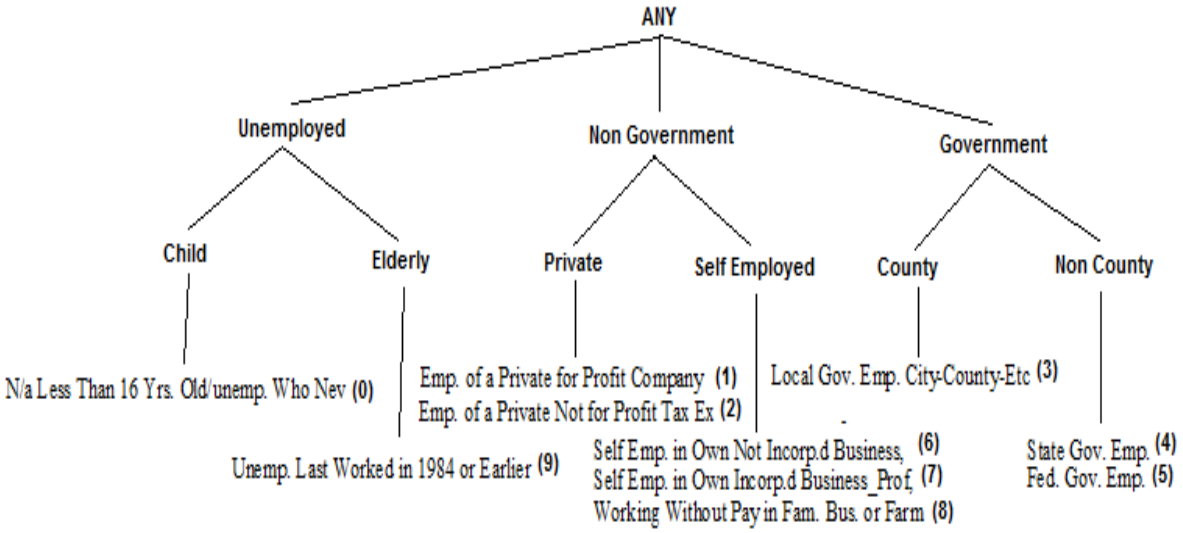
1,2,3	⊂	smallNuclei
4,5,6	⊂	mediumNuclei
7,8	⊂	largeNuclei
9,10	⊂	VeryLargeNuclei
smallNuclei, mediumNuclei	⊂	aboutAverNuclei
largeNuclei, VeryLargeNuclei	⊂	aboveAverNuclei
aboutAverNuclei,aboveAverNuclei	⊂	ANY

Appendix 10: Concept hierarchy for normal nucleoli attribute of breast cancer dataset.



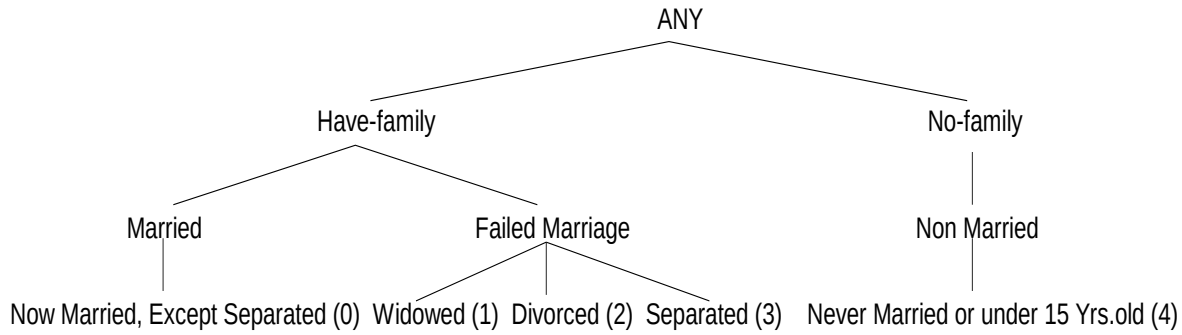
1,2,3	⊂	smallNucleoli
4,5,6	⊂	mediumNucleoli
7,8	⊂	largeNucleoli
9,10	⊂	VeryLargeNucleoli
smallNucleoli, mediumNucleoli	⊂	aboutAverNucleoli
largeNucleoli, VeryLargeNucleoli	⊂	aboveAverNucleoli
aboutAverNucleoli,aboveAverNucleoli	⊂	ANY

Appendix 11: Concept hierarchy for class attribute of census dataset.



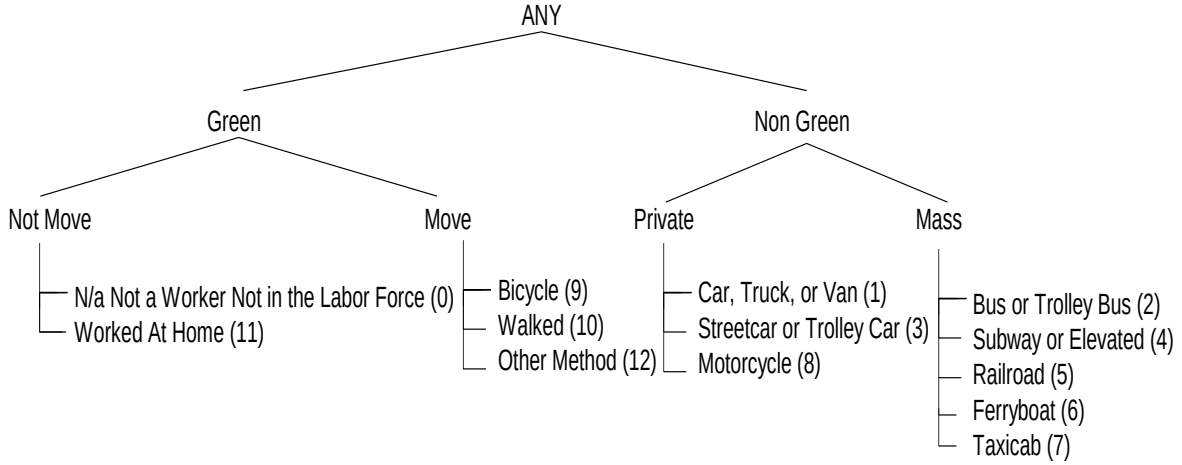
0	⊂	Child
1,2	⊂	Private
3	⊂	County
4,5	⊂	Non County
6,7,8	⊂	Self Employed
9	⊂	Elderly
Child, Elderly	⊂	Unemployed
County, Non County	⊂	Government
Private, Self Employed	⊂	Non Government
Unemployed, Government, Non Government	⊂	ANY

Appendix 12: Concept hierarchy for marital attribute of census dataset.



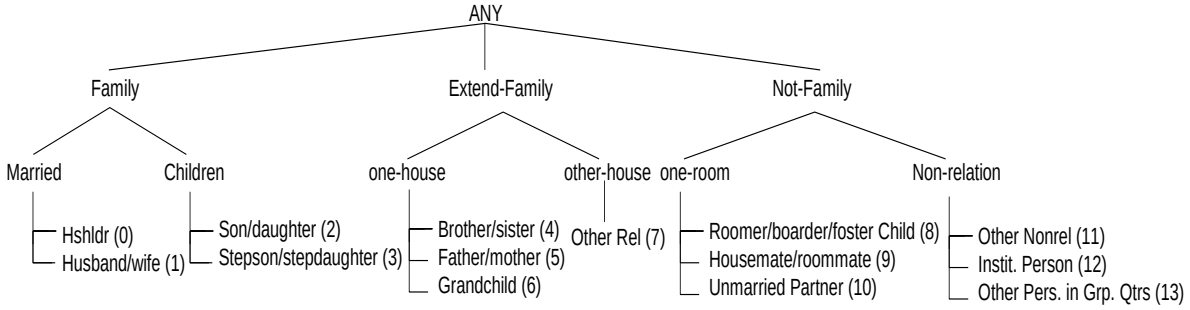
0	⊂	Married
1,2,3	⊂	Failed Marriage
4	⊂	Non Married
Married, Failed Marriage	⊂	Have-family
Non Married	⊂	No-family
Have-family, No-family	⊂	ANY

Appendix 13: Concept hierarchy for means attribute of census dataset.



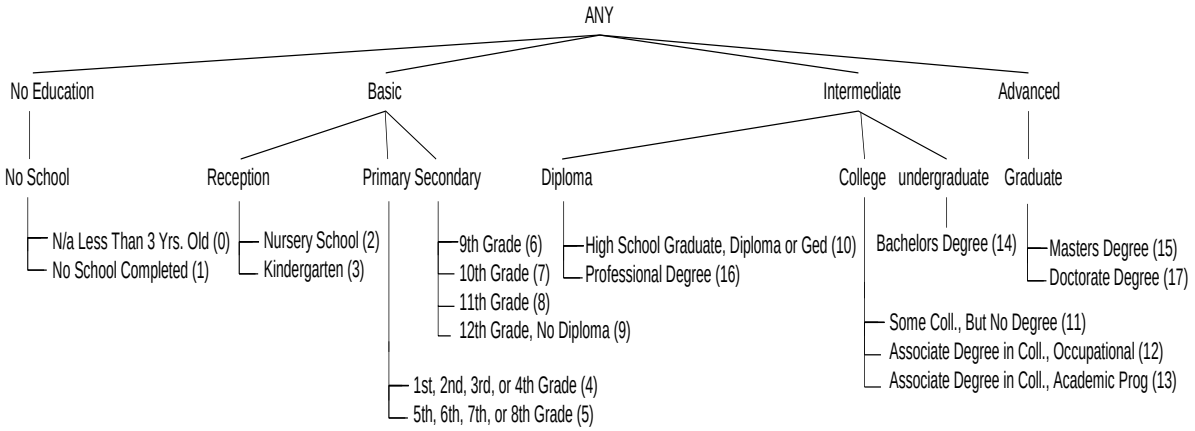
9,10,12	⊂	Move
0,11	⊂	Not Move
1,3,8	⊂	Private
2,4,5,6,7	⊂	Mass
Move, Not Move	⊂	Green
Private, Mass	⊂	Non Green
Green, Non Green	⊂	ANY

Appendix 14: Concept hierarchy for relat1 attribute of census dataset.



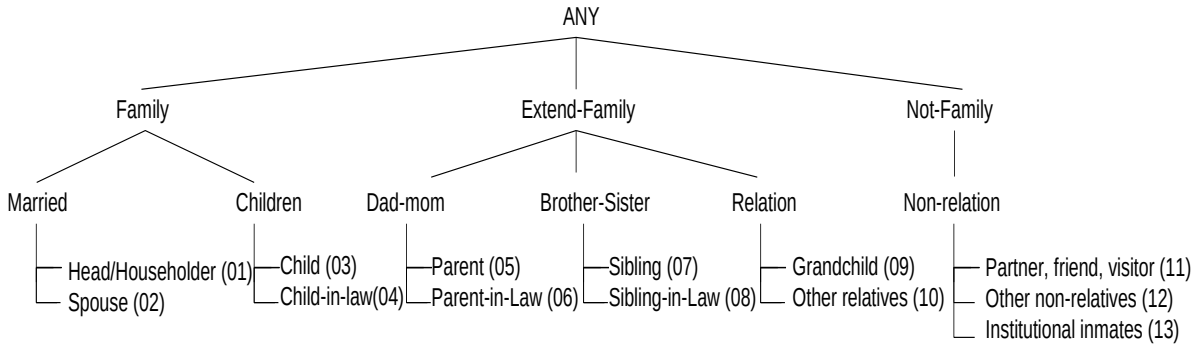
0,1	⊂	Married
2,3	⊂	Children
4,5,6	⊂	one-house
7	⊂	other-house
8,9,10	⊂	one-room
11,12,13	⊂	Non-relation
Married, Children	⊂	Family
One-house, other-house	⊂	Extend-Family
One-room, Non-relation	⊂	Not-Family
Family, Extend-Family, Not-Family	⊂	ANY

Appendix 15: Concept hierarchy for yearsch attribute of census dataset.



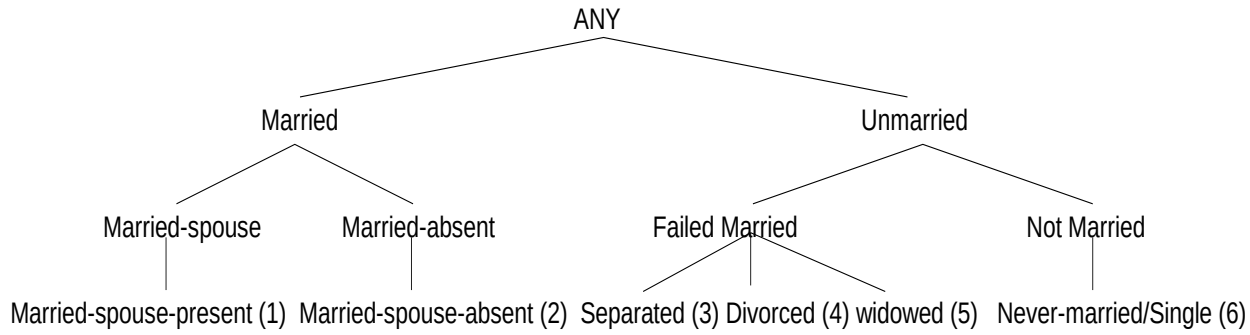
0,1	⊂	No School
2,3	⊂	Reception
4,5	⊂	Primary
6,7,8,9	⊂	Secondary
10,16	⊂	Diploma
11,12,13	⊂	College
14	⊂	undergraduate
15,17	⊂	Graduate
No School	⊂	No Education
Reception, Primary, Secondary	⊂	Basic
Diploma, College, undergraduate	⊂	Intermediate
Graduate	⊂	Advanced
No Education, Basic, Intermediate, Advanced	⊂	ANY

Appendix 16: Concept hierarchy for relateg attribute of IPUMS dataset.



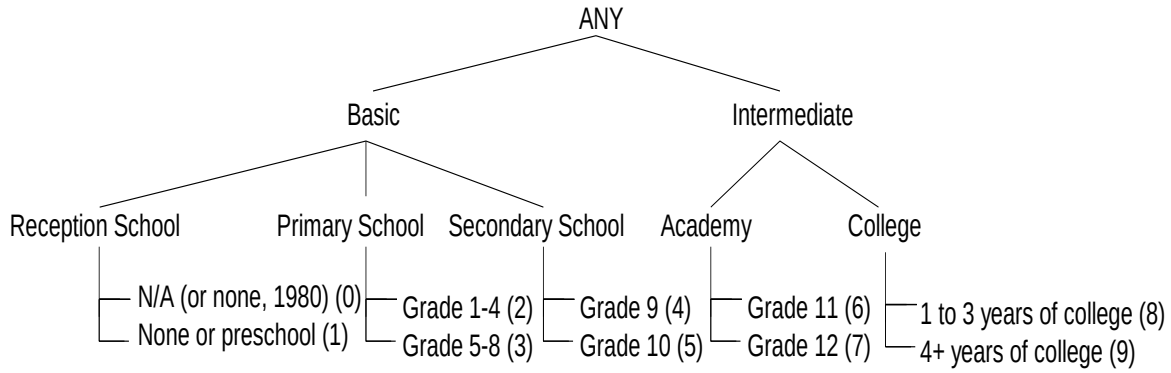
01,02	⊂	Married
03,04	⊂	Children
05,06	⊂	Dad-mom
07,08	⊂	Brother-Sister
09,10	⊂	Relation
11,12,13	⊂	Non-relation
Married, Children	⊂	Family
Dad-mom, Brother-Sister, Relation	⊂	Extend-Family
Non-relation	⊂	Not-Family
Family, Extend-Family, Not-Family	⊂	ANY

Appendix 17: Concept hierarchy for marst attribute of IPUMS dataset.



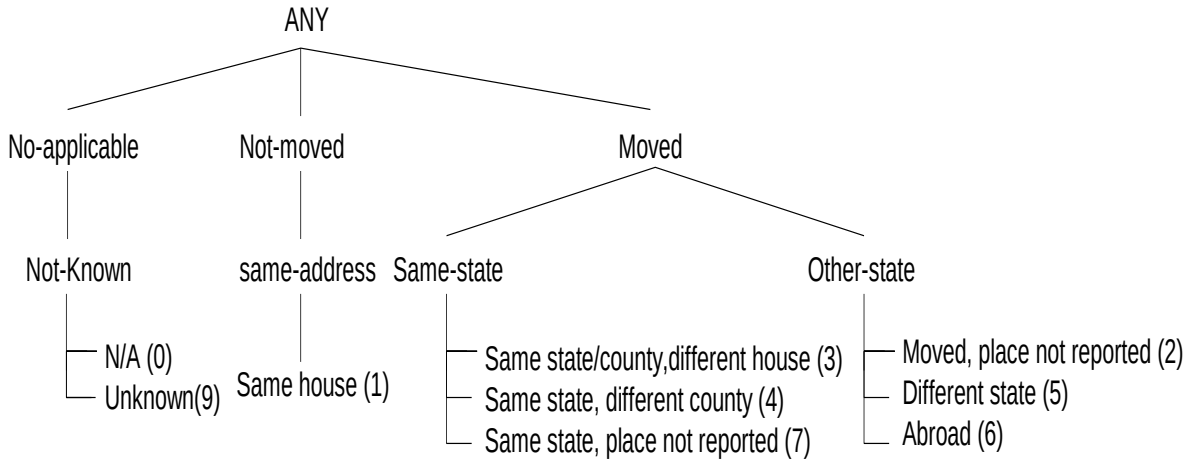
1	⊂	Married-spouse
2	⊂	Married-absent
3,4,5	⊂	Failed Married
6	⊂	Not Married
Married-spouse, Married-absent	⊂	Married
Failed Married, Not Married	⊂	Unmarried
Married, Unmarried	⊂	ANY

Appendix 18: Concept hierarchy for educrec attribute of IPUMS dataset.



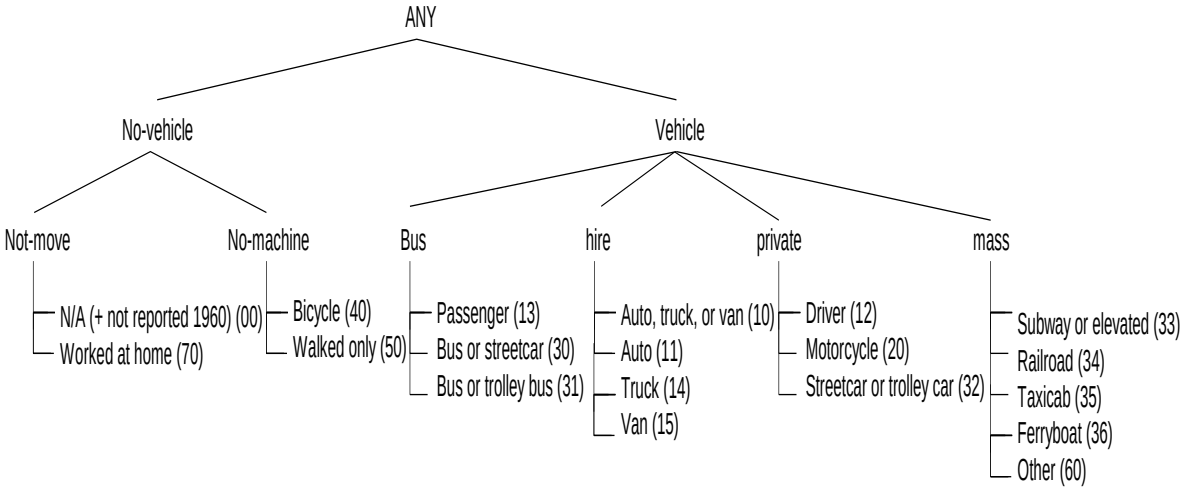
0,1	⊂	Reception School
2,3	⊂	Primary School
4,5	⊂	Secondary School
6,7	⊂	Academy
8,9	⊂	College
Reception School, Primary School, Secondary School	⊂	Basic
Academy, College	⊂	Intermediate
Basic, Intermediate	⊂	ANY

Appendix 19: Concept hierarchy for migrat5g attribute of IPUMS dataset.



0,9	⊂	Not-Known
1	⊂	same-address
3,4,7	⊂	Same-state
2,5,6	⊂	Other-state
Not-Known	⊂	No-applicable
Same-address	⊂	Not-moved
Same-state, Other-state	⊂	Moved
No-applicable, Not-moved, Moved	⊂	ANY

Appendix 20: Concept hierarchy for tranwork attribute of IPUMS dataset.



00,70	⊂	Not-move
40,50	⊂	No-machine
13,30,31	⊂	Bus
10,11,14,15	⊂	hire
12,20,32	⊂	private
33,34,35,36,60	⊂	mass
Not-move, No-machine	⊂	No-vehicle
Bus, hire, private, mass	⊂	Vehicle
No-vehicle, Vehicle	⊂	ANY

References

[1] Fayyad, U., Piatetsky-shapiro, G. and Smyth, P., 1996. From Data Mining to Knowledge Discovery. *AI Magazine*, 17(3), 37-54.

[2] Chakrabarti, S. et al., 2006. Data mining curriculum: A proposal (Version 1.0). Intensive Working Group of ACM SIGKDD Curriculum Committee,1-10, retrieved March 1, 2012, from http://www.sigkdd.org/curriculum/CURMay06.pdf.

[3] Chen, M.S., Han, J. and Yu, P.S. 1996. Data Mining: An Overview from a Database Perspective. IEEE Trans. on Knowl. and Data Eng, 8(6), 866-883.

[4] Fayyah U., Piatetsky-Shapiro G. and Smyth P. 1996. Knowledge discovery and data mining: Toward a unifying framework. In Proceeding of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD'96), 82-88.

[5] Keim, D.A. 2002. Information Visualization and Visual Data Mining. IEEE Transactions on Visualization and Computer Graphics 8(1), 1-8.

[6] Agrawal, R. and Srikant, R. 2000. Privacy-preserving data mining. In Proceedings of the 2000 ACM SIGMOD international conference on Management of data (SIGMOD '00), 439-450.

[7] Lindell, Y. and Pinkas, B. 2000. Privacy preserving data mining. Advances in Cryptology,36-54.

[8] Lindell, Y. and Pinkas, B. 2002. Privacy preserving data mining. Journal of cryptology, 15(3), 177-206.

[9] Agrawal, D. and Aggarwal, C.C. 2001. On the design and quantification of privacy preserving data mining algorithms. In Proceedings of the 20thACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (PODS '01), 247-255.

[10] Hearst, M.A. 1999. Untangling text data mining.In Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics (ACL '99), 3-10.

[11] Mooney, R.J. and Bunescu, R. 2005. Mining knowledge from text using information extraction. SIGKDD Explorations (special issue on Text Mining and Natural Language Processing), 7(1), 3-10.

[12] Mooney,R.J. and Nahm,U.Y. 2003. Text mining with information extraction. In Proceedings of the 4th International MIDP (Multilingualism and Electronic Language Management) Colloquium , 141-160.

[13] Piatetsky-Shapiro, G. 1991. Knowledge Discovery in Real Databases: A Report on the IJCAI-89 Workshop. AI Magazine, 11(5), 68–70.

[14] Piatetsky-shapiro, G. 2011. Industries/Fields where you applied Analytics/data mining in 2011, retrieved March 3, 2012, from http://www.kdnuggets.com/polls/2011/industries-applied-anaytics-data-mining.html.

[15] Piatetsky-shapiro, G. 2011. Algorithms for data analysis/data mining in 2011, retrieved March 6, 2012, from http://www.kdnuggets.com/polls/2011/algorithms-analytics-data-mining.html.

[16] Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. 1996. The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM, 39(11), 27-34.

[17] Piatetsky-shapiro, G. 2011. What data types you analyzed/mined in 2011, retrieved March 6, 2012, from http://www.kdnuggets.com/polls/2011/data-types-analyzed-mined.html .

[18] Witten, I.H., Frank, E. and Hall, M.A. 2011. Implementations: real machine learning schemes. In I.H.Witten, E. Frank and M.A. Hall (Ed.), Data mining : practical machine learning tools and techniques. Morgan Kaufmann, Burlington, MA, 3 edn,191-304.

[19] Piatetsky-shapiro, G. 2011. Largest dataset analyzed / data mined in 2011, retrieved March 6, 2012, from http://www.kdnuggets.com/polls/2011/largest-dataset-analyzed-data-mined.html .

[20] Wu, X., Kumar, V., Quinlan,J.R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Yu, P.S., Zhou, Z.H., Steinbach, M., Hand, D.J. and Steinberg,D. 2007. Top 10 algorithms in data mining. Knowl. Inf. Syst, 14(1), 1-37.

[21] Coenen, F. 2011. Data mining: past, present and future. Knowledge Eng. Review, 26(1),25-29.

[22] Piatetsky-shapiro, G. 2007. Data mining methodology in 2007, retrieved March 9, 2012, from http://www.kdnuggets.com/polls/2007/data_mining_methodology.htm .

[23] Azevedo, A. and Santos, M.F. 2008. KDD, SEMMA and CRISP-DM: a parallel overview. In Proceeding of the European Conference on Data Mining (IADIS), 182-185.

[24] Pachet, F., Westermann, G. and Laigre, D. 2001. Musical Data Mining for Electronic Music Distribution. In Proceedings of the 1st International Conference on WEB Delivering of Music (WEDELMUSIC'01),101-106.

[25] Liu, C.C., Hsu, J.L. and Chen A.L.P. 1999. Efficient Theme and Non-Trivial Repeating Pattern Discovering in Music Databases. InProceedings of the 15th International Conference on Data Engineering (ICDE '99), 14-21.

[26] Mikut, R. and Reischl, M. 2011. Data mining tools. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1(5), 431–445.

[27] Mitra, S., Pal, S.K. and Mitra, P. 2002. Data mining in soft computing framework: A survey. In IEEE Transactions on Neural Networks, 13(1),3-14.

[28] Freitas, A.A. 2003. A survey of evolutionary algorithms for data mining and knowledge discovery. In Advances in evolutionary computing,819-845.

[29] Koperski, K., Adhikary, J. And Han, J. 1996. Spatial data mining: progress and challenges survey paper. In SIGMOD Workshop on Research Issues on data Mining and Knowledge Discovery (DMKD), 1-10.

[30] Cai, Y., Cercone, N. and Han, J. 1990. An attribute-oriented approach for learning classification rules from relational databases. In Proceedings of 6th International Conference on Data Engineering, 281-288.

[31] Cai, Y., Cercone, N. and Han, J. 1991. Learning in relational databases: an attribute-oriented approach. Comput. Intell, 7(3),119-132.

[32] Cercone, N., Han, J., McFetridge, P., Popowich, F., Cai,Y., Fass, D., Groeneboer, C., Hall, G. and Huang, Y. 1994. System X and DBLearn: How to Get More from Your Relational Database, Easily. Integrated Computer-Aided Engineering, 1(4),311-339.

[33] Han, J., Cai, O., Cercone, N. and Huang, Y. 1995. Discovery of Data Evolution Regularities in Large Databases. Journal of Computer and Software Engineering,3(1),41-69.

[34] Han, J. 1998. Towards on-line analytical mining in large databases. SIGMOD Rec. 27(1), 97-107.

[35] Han,J. and Fu, Y. 1995. Discovery of Multiple-Level Association Rules from Large Databases. In Proceedings of the 21th International Conference on Very Large Data Bases (VLDB '95), 420-431.

[36] Han, J. 1997. OLAP Mining: An Integration of OLAP with Data Mining. In Proceedings of the 7th IFIP Working Conference on Database Semantics (DS-7),1-9.

[37] Han, J., Fu,Y., Koperski, K., Melli, G., Wang, W. And Zaïane, O.R. 1996. Knowledge Mining in Databases: An Integration of Machine Learning Methodologies with Database Technologies, Canadian Artificial Intelligence,(38),4-8.

[38] Fudger, D. and Hamilton, H.J. 1993. A Heuristic for Evaluating Databases for Knowledge Discovery with DBLEARN. In Proceedings of the International Workshop on Rough Sets and Knowledge Discovery: Rough Sets, Fuzzy Sets and Knowledge Discovery (RSKD '93), 44-51.

[39] Han,J., Cai, Y., and Cercone, N. 1993. Data-drivendiscovery of quantitative rules in relational databases. IEEE Trans on Knowl and Data Engin, 5(1),29-40.

[40] Han,J. and Fu, Y. 1995. Exploration of the power of attribute-oriented induction in data mining. in U. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy, eds. Advances in Knowledge Discovery and Data Mining, 399-421.

[41] Han, J., Cai, Y. and Cercone, N. 1992. Knowledge discovery in databases: An attribute-oriented approach. In Proceedings of the 18th International Conference on Very Large Data Bases, 547-559.

[42] Han,J. 1994. Towards efficient induction mechanisms in database systems. Theoretical Computer Science, 133(2), 361-385.

[43] Cheung, D.W., Fu, A.W. and Han, J. 1994. Knowledge discovery in databases: A rule-based attribute-oriented approach. In Proceedings of Intl Symp on Methodologies for Intelligent Systems, 164-173.

[44] Cheung, D.W., Hwang, H.Y., Fu, A.W. and Han, J. 2000. Efficient rule-based attribute-oriented induction for data mining. Journal of Intelligent Information Systems, 15(2), 175-200.

[45] Han,J., Fu, Y.,Wang, W., Chiang, J., Gong, W., Koperski, K., Li,D., Lu, Y., Rajan,A., Stefanovic,N., Xia,B. and Zaiane,O.R.1996. DBMiner:A system for mining knowledge in large relational databases. In Proceedings of International Conference on Data Mining and Knowledge Discovery, 250-255.

[46] Han,J., Lakshmanan, L.V.S. and Ng, R.T. 1999. Constraint-based,multidimensional data mining. IEEE Computer, 32(5), 46-50.

[47] Cai, Y. 1989. Attribute-oriented induction in relational databases. Master thesis, Simon Fraser University.

[48] Wu, Y., Chen, Y. and Chang, R. 2009. Generalized Knowledge Discovery from Relational Databases. International Journal of Computer Science and Network, 9(6),148-153.

[49] Imielinski, T. and Virmani, A. 1999. MSQL: A Query Language for Database Mining. in Proceedings of Data Mining and Knowledge Discovery, 3, 373-408.

[50] Muyeba, M. 2005. On Post-Rule Mining of Inductive Rules using a Query Operator. In Proceedings of Artificial Intelligence and Soft Computing.

[51] Meo, R., Psaila,G. and Ceri,S. 1998. An Extension to SQL for Mining Association Rules. In Proceedings of Data Mining and Knowledge Discovery,2,195-224.

[52] Muyeba,M.K. and Keane,J.A. 1999. Extending attribute-oriented induction as a key-preserving data mining method. In Proceedings 3rd European Conference on Principles of Data Mining and Knowledge Discovery, Lecture Notes in Computer science, 1704, 448-455.

[53] Muyeba, M. and Marnadapali, R. 2005. A framework for Post-Rule Mining of Distributed Rules Bases. In Proceeding of Intelligent Systems and Control.

[54] Zaiane, O.R. 2001. Building Virtual Web Views. Data and Knowledge Engineering, 39, 143-163.

[55] Han, J., Chiang, J. Y., Chee, S., Chen, J., Chen, Q., Cheng, S., Gong, W., Kamber, M.,Koperski, K., Liu, G., Lu, Y., Stefanovic, N., Winstone, L., Xia, B. B., Zaiane, O. R., Zhang, S., and Zhu, H. 1997. DBMiner: a system for data mining in relational databases and data warehouses. In Proceedings of the 1997 Conference of the Centre For Advanced Studies on Collaborative Research, 8-.

[56] Frank, A. and Asuncion, A. 2010. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[57] Elfeky, M.G., Saad, A.A. and Fouad, S.A. 2000. ODMQL: Object Data Mining Query Language. In Proceedings of the International Symposium on Objects and Databases, 128-140.

[58] Han, J. and Fu, Y. 1994. Dynamic Generation and Refinement of Concept Hierarchies for Knowledge Discovery in Databases. In Proceedings of AAAI Workshop on Knowledge Discovery in Databases, 157-168.

[59] Huang, Y. and Lin, S. 1996. An Efficient Inductive Learning Method for Object-Oriented Database Using Attribute Entropy. IEEE Transactions on Knowledge and Data Engineering, 8(6),946-951.

[60] Hu, X. 2003. DB-HReduction: A Data Preprocessing Algorithm for Data Mining Applications. Applied Mathematics Letters,16(6),889-895.

[61] Hsu, C. 2004. Extending attribute-oriented induction algorithm for major values and numeric values. Expert Systems with Applications, 27, 187-202.

[62] Han, J., Fu, Y., Huang, Y., Cai, Y., and Cercone, N. 1994. DBLearn: a system prototype for knowledge discovery in relational databases. ACM SIGMOD Record, 23(2), 516.

[63] Han, J., Fu, Y., and Tang, S. 1995. Advances of the DBLearn system for knowledge discovery in large databases. In Proceedings of the 14th international Joint Conference on Artificial intelligence, 2049-2050.

[64] Beneditto, M.E.M.D. and Barros, L.N.D. 2004. Using Concept Hierarchies in Knowledge Discovery. Lecture Notes in Computer Science, 3171,255–265.

[65] Li, J., Dong, G., Ramamohanarao, K. and Wong, L. 2004. DeEPs: A new instance-based discovery and classification system. Machine Learning, 54(2), 99-124.

[66] Zhang,X., Dong, G. and Ramamohanarao, K. 2001. Building Behaviour Knowledge Space to Make Classification Decision. In Proceedings of the 5th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD '01), 488-494.

[67] Zhang, X., Dong, G. and Ramamohanarao, K. 2000. Information-based classification by aggregating emerging patterns. In Proceedings of the 2nd International Conference on Intelligent Data Engineering and Automated Learning (IDEAL'00), 175-188.

[68] Zhang, X., Dong, G. and Ramamohanarao, K. 2000. Exploring constraints to efficiently mine emerging patterns from large high-dimensional datasets. In Proceedings of the 6th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD'00), 310-314.

[69] Ramamohanarao, K., Bailey, J. and Fan, H. 2005. Efficient Mining of Contrast Patterns and Their Applications to Classification. In Proceedings of the 3rd International Conference on Intelligent Sensing and Information Processing (ICISIP '05), IEEE Computer Society, 39-47.

[70] Fan, H. and Ramamohanarao, K. 2002. An Efficient Single-Scan Algorithm for Mining Essential Jumping Emerging Patterns for Classification. In Proceedings of the 6th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD '02), 456-462.

[71] Fan, H. and Ramamohanarao, K. 2003. A Bayesian approach to use emerging patterns for classification. In Proceedings of the 14th Australasian database conference (ADC '03), 17, 39-48.

[72] Li, J., Ramamohanarao, K., and Dong, G. 2000. Emerging Patterns and Classification. In Proceedings of the 6th Asian Computing Science Conference on Advances in Computing Science, 15-32.

[73] Han, J., Pei, J., Yin, Y., and Mao, R. 2004.Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach. Data Min. Knowl. Discov.,8(1), 53-87.

[74] Han, J., Pei, J., and Yin, Y. 2000. Mining frequent patterns without candidate generation. In Proceedings of the Inter. Conference on Management of Data, 1-12.

[75] Bailey, J., Manoukian, T. and Ramamohanarao. K. 2002. Fast Algorithms For Mining Emerging Patterns.In Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery, 39‑50.

[76] Bailey, J., Manoukian, T. and Ramamohanarao, K. 2003. Classification using constrained emerging patterns.In Proceedings of the 4th International Conferenceon Web Age Information Management, 226-237.

[77] Ramamohanarao, K., and Bailey, J. 2003. Discovery of Emerging Patterns and Their Use in Classification. In Proceedings of the 16th Australian Conference on Artificial Intelligence, 1-12.

[78] Fan, H. 2004. Efficient mining of interesting emerging patterns and their effective use in classification, Ph.D. thesis, University of Melbourne.

[79] Dong, G. and Li, J. 1999. Efficient mining of emerging patterns: discovering trends and differences. In Proceedings of the 5th ACM SIGKDD international conference on Knowledge discovery and data mining, 43-52.

[80] Dong, G. and Li, J. 2005. Mining border description of emerging patterns from dataset pairs. Journal of Knowledge and Information Systems, 8(2),178-202.

[81] Li,J., Dong,G. and Ramamohanarao, K. 2000. Instance-Based Classification by Emerging Patterns. In Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD '00), 191-200.

[82] Li, J., Ramamohanarao, K. and Dong, G. 2000. The Space of Jumping Emerging Patterns and Its Incremental Maintenance Algorithms. In Proceedings of the 17th International Conference on Machine Learning (ICML '00), 551-558.

[83] Li, J., Dong, G., and Ramamohanarao, K. 2000. Making Use of the Most Expressive Jumping Emerging Patterns for Classification. In Proceedings of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Current Issues and New Applications. Lecture Notes In Computer Science, 1805, 220-232.

[84] Dong, G., Zhang, X., Wong, L., and Li, J. 1999. CAEP: Classification by Aggregating Emerging Patterns. In Proceedings of the 2nd international Conference on Discovery Science. Lecture Notes in Computer Science, 1721, 30-42.

[85] Muyeba, M.K., Khan, M.S., Warnars,S. and Keane, J.A. 2011. A Framework to Mine High-Level Emerging Patterns by Attribute-Oriented Induction. In Proceedings of the 12th international conference on Intelligent data engineering and automated learning (IDEAL),170-177.

[86] Han, J., Cheng, H., Xin, D. and Yan, X.2007. Frequent pattern mining: current status and future directions. Data Min Knowl Disc, 15(1), 55-86.

[87] Danger, R., Ruiz-Shulcloper, J. and Llavori,R.B. 2004. Objectminer: A new approach for Mining Complex objects. In Proceedings of the 6th international conference on Enterprise Information Systems (ICEIS ’04), 42-47.

[88] Rodriguez-Gonzalez,A. Y. , Martinez-Trinidad, J.F., Carrasco-Ochoa, J.A. and Ruiz-Shulcloper, J. 2008. Mining Frequent Similar Patterns on Mixed Data.In Proceedings of the 13th Iberoamerican congress on Pattern Recognition: Progress in Pattern Recognition, Image Analysis and Applications(CIARP '08), 136-144.

[89] Podraza, R. and Tomaszewski, K. 2005. KTDA: Emerging Patterns Based Data Analysis System. In Proceedings of XXI Fall Meeting of Polish Information Processing Society, 213—221.

[90] Qian, X., Bailey, J. and Leckie, C. 2006. Mining generalised emerging patterns. In Proceedings of the 19th Australian joint conference on Artificial Intelligence: advances in Artificial Intelligence (AI'06), 295-304.

[91] Boulesteix, AL., Tutz, G, and Strimmer K. 2003. A CART-based approach to discover emerging patterns in microarray data. Bioinformatics. 19(18),2465-2472.

[92] Fan,H. andRamamohanarao, K. 2005. A weighting scheme based on emerging patterns for weighted support vector machines. In Proceedings of IEEE International Conference on Granular Computing (GrC), 435-440.

[93] Zhang, S., Ramamohanarao, K. and Bezdek, J.C. 2010. EP-based robust weighting scheme for fuzzy SVMs. In Proceedings of the 21st Australasian Conference on Database Technologies, 104, 123-132.

[94] Alhammady, H. And Ramamohanarao, K.. 2006. Using Emerging Patterns to Construct WeightedDecision Trees. IEEE Trans. Knowl. Data Eng, 18(7), 865-876.

[95] Ding, G., Wang, J. and Qin, K.. 2010. A visual word weighting scheme based on emerging itemsets for video annotation. Information Proces. Letters, 110(16), 692-696.

[96] Alhammady, H. And Ramamohanarao, K. 2004. Using Emerging Patterns and Decision Trees inRare-Class Classification. In Proceedings ofthe 4th IEEE International Conference on Data Mining (ICDM), 315-318.

[97] Alhammady, H. And Ramamohanarao, K. 2004.The application of emerging patterns for improving the quality of rare-class classification.In Proceeding of 8th Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD), 207–211.

[98] Alhammady, H. And Ramamohanarao, K. 2005. Expanding the Training Data Space Using Emerging Patterns and Genetic Methods. In proceeding of SIAM International Data Mining Conference (SDM), 481-485.

[99] Ramamohanarao, K., Bailey,J. and Fan, H. 2005. Efficient Mining of Contrast Patterns and Their Applications to Classification. In Proceedings of the 3rd International Conference on Intelligent Sensing and Information Processing (ICISIP '05). IEEE Computer Society, , 39-47.

[100] Sun, Q., Zhang, X. And Ramamohanarao, K. 2003. Noise Tolerance of EP-Based Classifiers. In Proceedings of the 16th Australian Conference on Artificial Intelligence Advances in Artificial Intelligence (AI), 2903,796-806.

[101] Fan, H. and Ramamohanarao, K. 2006. Fast Discovery and the Generalization of Strong Jumping Emerging Patterns for Building Compact and Accurate Classifiers. IEEE Trans. on Knowl. and Data Eng, 18(6), 721-737.

[102] Li, J., Liu, H., Downing, J.R., Yeoh, A. and Wong, L. 2003. Simple rules underlying gene expression profiles of more than six subtypes of acute lymphoblastic leukemia (ALL) patients. Bioinformatics, 19(1),71–78.

[103] Ramamohanarao, K. and Fan, H. 2007. Patterns Based Classifiers. World Wide Web, 10(1), 71-83.

[104] Cheng, M.W.K., Choi, B.K.K. and Cheung, W.K.W. 2010. Hiding emerging patterns with local recoding generalization. In Proceedings of the 14th Pacific-Asia conference on Advances in Knowledge Discovery and Data Mining (PAKDD'10), 158-170.

[105] Agrawal, R., Imielinski, T. and Swami, A. 1993. Mining association rules between sets of items in large databases. ACM SIGMOD Rec, 22(2), 207-216.

[106] Schubert, S. and Lee, T., 2011. Timer series data mining with SAS® Enterprise minerTM , in Proceedings of the SAS® Global forum.