On the Suitability of Type-1 Fuzzy Regression Tree Forests for Complex Datasets

Fathi Gasir, Keeley Crockett

Computer Science Department, Faculty of Information Technology, Misurata University, Misurata – Libya

F.Gasir@it.misuratau.edu.ly

The Intelligent Systems Group, School of Computing, Mathematics and Digital Technology, Manchester Metropolitan University, Chester Street, Manchester, M1 5GD, UK,

K.Crockett@mmu.ac.uk

Abstract.

One of the challenges in data mining practices is that the datasets vary in complexity and often have different characteristics such as number of attributes, dependent variables characteristics etc. In terms of regression problems, the features that describe the dataset will vary in their complexity, sparseness verses coverage in relation to the decision space, and the number of outcome classes. Fuzzy Decision trees are well-established classifiers in terms of building robust, representative models of the domain. In order to represent different perspectives of the same domain, fuzzy trees can be used to construct fuzzy decision forests to enhance the predictive ability of singular trees. This paper describes an empirical study which examines the applicability of fuzzy tree regression forests to seven different datasets which have complex properties. The relationship between dataset characteristics and the performance of fuzzy regression tree forests is debated.

Keywords: Fuzzy Decision trees, Fuzzy Regression Forests

1 Introduction

It is a known problem that the complexity of data is becoming increasing challenging for traditional machine learning algorithms to deal with, especially in the Big Data arena where data variety, veracity and volume have to be taken into consideration. However, the debate continues on whether the focus should be on developing better algorithms or to generate models using more data [1]. In the context of Big Data, Kwona and Simb [2] performed a comprehensive study on the performance of classification algorithms in relation to a datasets features. The experimental study found that legacy classification algorithms performed differently depending on how the data was

adfa, p. 1, 2011.

structured, its content and context in which it was applied [2]. For example, the number of features in any data set not only affects the time to produce an optimal model, but also influences the performance when using classification algorithms [3..5].

Fuzzy decision trees allow data instances to simultaneously fire multiple branches of a node with different degrees of membership whereby allowing all information to contribute towards the final classification [6,7]. More specifically, fuzzy regression trees are used where there is a non-linear relationship between input and output variables. Fuzzy decision tree forests have been shown to improve the predictive power of a singular fuzzy trees by allowing numerous insights and interpretations of the datasets that are being modelled [8..13]. Fuzzy Forests designed for classification problems have also be shown to be tolerant to noisy data [9].

The study presented in this paper investigates the relationship between the datasets characteristics, number of features, and datasets sizes and the performance of fuzzy regression tree forests in the context of regression tree problems. An empirical study on seven known datasets and generates for each fuzzy decision forests comprising each of five fuzzy trees using the Elgasir algorithm [13]. Fuzzification is optimised in each case by using the adapted version of an artificial immune network model (opt-aiNet [14]). A series of experiments is conducted to determine whether the characteristics of the data affects the performance of the fuzzy regression forests. This is determined through a comparison with singular crisp regression trees. This paper is structured as follows: Section 2 provides an overview of related work in the field of fuzzy regression trees and forests. Section 3 describes the algorithm for constructing type-1 fuzzy forests using the Elgasir algorithm. The characteristics of the dataset are described in 4, with the experimental methodology and results in section 5. Finally, conclusions are presented in section 6.

2 Related Work

Regression tree induction algorithms [15] are a technical approach which are used to construct a set of rules that will predict events in a given domain. Regression tree induction algorithms induce rules from the knowledge of a set of examples, known as a training dataset, whose predicted outcome is already known. The process of regression tree induction involves selecting [15]. CHAID provides a set of rules that can be applied to a new (unseen) dataset to predict the target or outcome. The CHAID algorithm stops growing a tree before overfitting occurs, as a result of using its unique dynamic branching strategy for determining the optimal number of branches. This strategy merges together attribute values that are shown to be statistically homogenous (similar), retaining the values that are heterogeneous (dissimilar). Trees generated from traditional tree induction algorithms are often referred to as “crisp” and suffer from sharp decision boundaries which results of using the strict partitioning for regression trees induction [7] and values are restricted to a limited number of discrete values as a result of using a discrete function to generate the tree output.

Fuzzy decision tree rule induction algorithms overcome such problems by allowing gradual transitions to exist between continuous attributes at tree nodes and utilizing fuzzy inference to combine information throughout the tree rather than following a single root to leaf node path. Early methods of fuzzy decision tree development replied on experts in the domain to pre-fuzzify the data prior to induction – a task that introduced a further uncertainty through subjectivity. Specific to this paper is attempts to fuzzily the CHAID algorithm. First achieved by Fowdar et al [6], the Fuzzy CHAID Induction Algorithm produced robust fuzzy trees with significantly higher accuracies than its crisp counterpart. The fuzzy regression tree algorithm known as Elgasir, also based on CHAID, incorporated degrees of uncertainty typical in data through the use of trapezoidal membership functions and Takagi-Sugeno fuzzy inference is applied to aggregate a final continuous output value. Elgasir alleviates Fowdar's defuzzification problem as a result of using Takagi-Sugeno fuzzy inference to aggregate fuzzy regression tree output as a single numeric value [16].

Fuzzy decision tree forests or assembles allow the concepts of fuzzy decision trees to be applied to allow different models of the same domain to be combined. Of significance in the field was Bonissone et al, [12] approach which used a fuzzy learning algorithm to create singular fuzzy trees using Breiman's metholdogy and then applied different configurations of combining leaf information. A different approach described in Crockett et al [7] involved the use of creating multiple fuzzy C4.5 decision trees from non-fuzzy really world data by selecting as the root attributes with high to low information content. Cadenas et al. [11] showed used a fuzzy random forest assemble method to select features for classification problems thus reducing dimensionality and improving classification accuracy. Work in this field has focused on classification and little work has reported on regression problems.

3 An Algorirthm for Constructing Type-1 Fuzzy decision tree forests

This section outlines the Elgasir fuzzy regression tree rule induction algorithm and describes how it is used to generate fuzzy regression tree forests.

3.1 The Elgasir Algorithm

The aim of the fuzzy regression tree algorithm Elgasir [13] was to apply appropriate membership functions to all branch split points in order to master the weakness of crisp decision trees, by allowing all the information used throughout the tree to contribute towards the outcome. Elgasir's foundations were based on Kass's CHAID Algorithm [15]. CHAID is a highly efficient statistical technique used to induce standard regression trees that are easy for humans to interpret. In order to reduce the strict partitioning at nodes and represent uncertainty, Elgasir combined principles of fuzzy theory and Takagi-Sugeno fuzzy inference technique to produce type-1 fuzzy regression trees. [16]. In order to optimise fuzzy set boundaries throughout the tree, an adapted version

of an artificial immune network model (opt-aiNet [13]) was applied. A brief overview of the algorithm is provided below and a full description can be found in [13].

1. Randomization and Partition the dataset into training and test data using multi-fold cross validation.
2. Apply CHAID Crisp Regression Trees rule induction algorithm for the first subset data.
    a. Generate optimal crisp CHAID regression tree from the training dataset by empirically applying various values to CHAID regression parameters.
    b. Evaluate performance of crisp tree using test dataset.
3. Convert near-optimal crisp tree into a set of IF-THEN rules.
4. Fuzzification of crisp CHAID regression tree.
    a. Repeat until the near-optimal performance of the fuzzy regression tree is reached.
        i. Apply adapted opt-aiNet to determine optimal amount of fuzzification to membership functions in all branch split points within the antecedent rules and Repeat
        ii. Parameterize consequent part of IF-THEN rules, where *n* is the total number of IF-THEN rules converted from the near-optimal crisp tree (step 2.a).
        iii. Identification of consequent parameters using the training dataset.
        iv. Evaluation of grades of membership.
5. Repeat steps 2 to 4, until each subset has been used once as a test dataset.

Step 6. Report on overall average error rate.

3.2 Constructing Forests

The Elgasir algorithm described in section 3.1 can be used to create fuzzy decision forests comprising of *n* fuzzy regression trees from one training sample where each tree represents a different perspective of the training sample. This allows better coverage of the domain which is less sensitive to noise in the data. The methodology reported in [13] comprises of three stages. Stage 1 generates *n* crisp regression trees using the CHAID algorithm and converts in to a fuzzy rule base; Stage 2 involves determination of fuzzy sets around each tree node and associated membership functions; Stage 3 requires optimization of fuzzy membership functions using the immune network opt-aiNet. In this work optimal forests are conducted for all datasets in this study.

4 Characteristics of Data

Seven known datasets are used in this study. Based in three criterion: the number of instances, number of attributes and number of unique values. They were selected the Boston Housing dataset is used to predict the median value of owner occupied homes, in $1,000's, as collected by the U.S Census Service concerning housing in the area of Boston, Massachusetts [17]. The Abalone dataset is concerned with predicting the age of abalone from physical measurements and has 28 unique outcome values [17]. The Compactiv dataset [18] is a collection of computer systems activity measures where the prediction task is to predict the variable *usr*, the portion of time that CPUs run in user mode. The Elevators dataset [17] is obtained from the task of controlling an F16 aircraft, and the goal variable is related to an action taken on the elevators of the aircraft. The Stock prices dataset [17] contains daily stock prices, from January 1988 through to October 1991, for ten aerospace companies. The task is to approximate the price of the 10th company, given the price of the others. The Concrete Compressive Strength Dataset comprising of 938 attributes is used to predict the concrete compressive strength 18]. Finally, the Communities and Crime dataset (120 attributes) is used to predict the per capita violent crime, 121 instances were left after the instances with missing attributes were remove [18]. This dataset has 120 attributes describing various social, economic and criminal factors. Table 1 presents a summary of the characteristics of these datasets.

Table 1. Dataset Characteristics

Name	Number of Instances	Number of attributes	Unique Values
Boston housing	506	14	229
Abalone	4177	9	28
Compactiv	8192	21	56
Elevators	16599	18	61
Stock prices	950	9	203
Crime	121	120	115
Concrete	1030	8	938

5 Experimental Methodology

For each dataset in Table 1, stratified 10-fold cross validation was applied. The training cases were partitioned into 10 equal-sized blocks with similar class distributions. Each block in turn is used to evaluate singular CHAID decision trees and the optimised fuzzy trees which were incrementally added the fuzzy forest. The singular CHAID trees were first optimised through parameter tuning to prevent any bias occurring. To create the second and subsequent trees in the forest, the attribute having the lowest p-value (the highest ranking) was constrained from formulating the root of the second tree. Five fuzzy trees were induced and compiled into each forest as it has been shown

that increasing the number of trees further would result in an increase in the error rate [13]. Fuzzification was optimised across each forest using opt-aiNet.

6 Results and Discussion

Table 2 present the result the average error rate of five Crisp CHAID regression trees for seven datasets and table 3 shows the results the average of each of five fuzzy regression tree forests for all datasets. The best result was obtained from the Concrete dataset where the error rate of fuzzy regression tree forests was reduced by 42% compared to Crisp CHAID regression trees which obtained a P-Value 0.0203. The Abalone dataset results show that fuzzy regression tree forests reduced the error rate by 41 % compared to Crisp CHAID regression trees with P-Value 0.0213. The reduction of the error rate was 34% on the Stock Price dataset by fuzzy regression tree forests compared to Crisp CHAID regression trees obtaining a P-Value of 0.0265. For the Crime dataset, the error was reduced by 27% by fuzzy regression tree forests compared to Crisp CHAID regression trees (P-Value 0.0422). The reduction of the error rate was 27% on the Compactiv dataset by fuzzy regression tree forests compared to Crisp CHAID regression trees with P-Value 0.0393. Whilst a 26% reduction in the error rate was achieved for the Boston housing dataset by fuzzy regression tree forests compared to Crisp CHAID regression trees which obtained a P-Value 0.0395. The Elevators dataset results show that fuzzy regression tree forests reduced the error rate by 24 % compared to Crisp CHAID regression trees obtaining a P-Value of 0.0412. Results of applying a paired t-test between results obtained from the singular crisp CHAID tree and Fuzzy regression tree forests can be found in Table 4. These results of all datasets show a statistically significant ($P<0.05$) in performance of fuzzy regression tree forests comparing with Crisp CHAID regression trees.

According to Tables 1 and 4, the number of attributes of dataset have been found to be significantly correlated to the performance of fuzzy regression tree forests. The biggest improvement in performance was obtained on the Concrete dataset, Abalone dataset and Stock Price dataset which have the smallest number of attributes 8,9 and 9 respectively compared with the rest of datasets. Based on these results, the number of attributes have inverse proportional relationship with the performance accuracy of the proposed method. On the other hand, the other dataset characteristics such as dataset size and unique outcome value been found not to be significantly correlated to the performance of fuzzy regression tree forests.

Table 2. Result the average of the five Crisp CHAID regression trees

Dataset	Training dataset (error value)	Test dataset (error value)
Boston housing	21.0576	21.4086
Abalone	4.4833	4.4982
Compactiv	25.4945	25.7666
Elevators	0.0000140389	0.0000140794
Stock Prices	7.6938	7.849
Crime	0.3419	0.3451
Concrete	0.1401	0.1492

Table 3. Result the average of the five fuzzy regression tree forests

Dataset	Training dataset (error value)	Test dataset (error value)
Boston housing	13.4618	15.8973
Abalone	2.41916	2.6545
Compactiv	17.9268	18.84
Elevators	0.0000106117	0.0000106593
Stock Prices	5.1245	5.1959
Crime	0.2489	0.2515
Concrete	0.0802	0.0863

Table 4. Results of paired t-test and Test Dataset of Crisp regression tree and Fuzzy regression tree forests

Dataset	Crisp regression tree (error value)	FRTF (error value)	P-Value
Boston housing	21.4086	15.8973	0.0395
Abalone	4.4982	2.6545	0.0213
Compactiv	25.7666	18.84	0.0393
Elevators	0.0000140794	0.0000106593	0.0412
Stock Prices	7.849	5.1959	0.0265
Crime	0.3451	0.2515	0.0422
Concrete	0.1492	0.0863	0.0203

7 Conclusion

This empirical study has shown that fuzzy regression tree forests, once optimized, can outperform singular crisp regression trees regardless of the number of instances, number of attributes and number of unique values. Optimization of each individual forest was domain dependent. As Elgasir is based on CHAID, the Chi-Square test of significance is used to evaluate all values of the predictor variable to select at each tree node the most significant attribute based on significance. Therefore, insignificant attributes are removed prior to the crisp trees fuzzification which typically reduces the number of attributes in the dataset that are modelled. In this study the relationship between dataset characteristics and the performance of fuzzy regression tree forests have been highlighted. The empirical results of seven datasets have shown that the number of attributes in a dataset have been found to be significantly correlated to the performance of fuzzy regression tree forests.

8 References

1. Amatriain, X. In Machine Learning, What is Better: More Data or better Algorithms. Available: https://www.quora.com/In-machine-learning-is-more-data-always-better-than-better-algorithms/answer/Xavier-Amatriain Date Accessed: 21/2/2016.
2. Kwona, O. Jae Mun Simb, Effects of data set features on the performances of classification algorithms, Expert Systems with Applications, Volume 40, Issue 5, April 2013, Pages 1847–1857.
3. Cadenas, J.M.; Garrido, M.C.; Martinez, R.; Bonissone, P.P., "Towards the learning from low quality data in a Fuzzy Random Forest ensemble," in Fuzzy Systems (FUZZ), 2011 IEEE International Conference on , vol., no., pp.2897-2904, 27-30 June 2011.

4. Cadenas. J. Garrido, M. Martínez, R. Selecting Features from Low Quality Datasets by a Fuzzy Ensemble, Computational Intelligence Volume 577 of the series Studies in Computational Intelligence pp 229-243, 2015

5. Bhatt, N. Thakkar, A. Ganatra, A. Bhatt, N. Ranking of Classifiers based on Dataset Characteristics using Active Meta Learning, International Journal of Computer Applications (0975 – 8887), International Journal of Computer Applications (0975 – 8887) Volume 69– No.20, May 2013 Volume 69– No.20, May 2013

6. Crockett, K. Bandar, Z. O'Shea, J. Fowdar, J. A Fuzzy Numeric Inference Strategy for Classification and Regression Problems, International Journal of Knowledge-based and Intelligent Engineering Systems, Vol12 ,No 4, pp. 255-269, 2008

7. Crockett, K., Bandar, Z., McLean, D. and O'Shea, J., (2006b). On Constructing a Fuzzy Inference Framework using Crisp Decision Trees. Fuzzy Sets and Systems, 157 (21), 2,809-2,832

8. Marsala, C.; Rifqi, M., "Summarizing Fuzzy Decision Forest by subclass discovery," in *Fuzzy Systems (FUZZ), 2013 IEEE International Conference on* , vol., no., pp.1-6, 7-10 July 2013

9. De Matteis, A.D.; Marcelloni, F.; Segatori, A., "A new approach to fuzzy random forest generation," in Fuzzy Systems (FUZZ-IEEE), 2015 IEEE International Conference on , vol., no., pp.1-8, 2-5 Aug. 2015

10. Jiang, R.; Bouridane, A.; Crookes, D.; Celebi, M.E.; Wei, Hua-Liang, "Privacy-Protected Facial Biometric Verification via Fuzzy Forest Learning," in Fuzzy Systems, IEEE Transactions on , vol.PP, no.99, pp.1-1

11. Cadenas, J.M.; Garrido, M.C.; Martinez, R., "Learning in a Fuzzy Random Forest ensemble from imperfect data," in Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on , vol., no., pp.277-282, 9-12 Oct. 2011 doi: 10.1109/ICSMC.2011.6083678

12. Bonissone, P., Cadenas, J. M., Garrido, M. C., & Díaz-Valladares, R. A. (2010). A fuzzy random forest. International Journal of Approximate Reasoning, 51(7), 729-747.

13. Gasir, F. Crockett, K, Bandar, Z. Inducing Fuzzy Regression Tree Forests Using Artificial Immune Systems. International Journal of Uncertainty, Fuzziness and Knowledge-Based SystemsInt. J. Unc. Fuzz. Knowl. Based Syst. 20, 133 (2012). DOI: 10.1142/S0218488512400181

14. De Castro, L.N. Timmis, J., An Artificial Immune Network for Multimodal Function Optimisation. Proc. Of IEEE World Congress on Evolutionary Computation. Pp. 669-674, 2002.

15. Kass, G. V., An Exploratory Technique For Investigating Large Quantities Of Categorical Data, Applied Statistics, 29(2) pp 119-127, 1979.

16. Takagi, T. Sugeno, M., Fuzzy identification of a system and its application to modeling and control, IEEE Transactions Systems, Man and Cybernetics, 15, (1985), pp.116-132.

17. Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

18. KEEL repository " Knowledge Extraction based on Evolutionary Learning" 2016. [Online]. Available: http://sci2s.ugr.es/keel/datasets.php, 2016.