OPEN ACCESS INTERNATIONAL JOURNAL OF SCIENCE & ENGINEERING

ENHANCING FUZZY BASED CLASSIFICATION SYSTEMS UNDER THE MAPREDUCE FRAMEWORK SYSTEM

K. SEETHALAKSHMI[1] J. JOSEPHINE SAHAYA VERGIN[2]
Asst. Professor, Department of Computer Applications, Idhaya College for Women, Sarugani, India[1]
Asst. Professor, Department of Computer Applications, Idhaya College for Women, Sarugani, India[2]

Abstract: **The inductive learning of fuzzy rule based classification systems affects from exponential growth of the fuzzy rule search space when the number of patterns and/or variables becomes high. This growth types the learning process more problematic and, in most cases, it indications to problems of scalability (in relationships of the time and memory consumed) and/or complexity (with respect to the number of rules obtained and the number of variables included in each rule). In this work, we propose a fuzzy association rule-based classification technique for high-dimensional problems based on three stages to find an accurate and compact fuzzy rule based classifier with a low computational cost. This technique parameters the order of the associations in the association rule extraction and considers the use of subgroup discovery based on an Improved Weighted Relative Accuracy portion to preselect the most interesting rules earlier a genetic post-processing process for rule selection and parameter change. The results found done twenty-six real-world datasets of altered sizes and with different numbers of variables establish the effectiveness of the proposed method.**

Keywords: *Data Mining, Associative Classification, Classification, Fuzzy Association Rules, Genetic Algorithms, Genetic Fuzzy Rule Selection, High-Dimensional Problems.*

I INTRODUCTION

Fuzzy Rule Based Classification Systems (FRBCSs) [1], [2] are valuable and well-known tools in the machine learning framework, since they can deliver an interpretable typical for the end user [3]–[6]. There are many real applications in which FRBCSs have been employed, with anomaly intrusion detection [7], image processing [8], among others. In most of these areas the accessible or useful data consists of a high number of patterns (instances or examples) and/or variables. In this situation, the inductive learning of FRBCSs suffers from exponential growth of the fuzzy rule search space. This development makes the learning procedure more problematic and, in most cases, it leads to problems of scalability (in terms of the time and memory consumed) and/or complexity (with respect to the amount of rules achieved and the number of variables included in every rule) [9], [10].

Association discovery is one of the most common Data Mining techniques used to mine stimulating knowledge from large datasets [11]. Many efforts have been made to use its benefits for classification under the name of associative classification [12]–[19]. Association discovery aims to find interesting relationships between the different items in a database [20], while classification aims to discover a model from training data that can be used to predict the class of test patterns [21]. Both association discovery and classification rules mining are essential in practical Data Mining applications [11], [22] and their integration could result in greater savings and suitability for the user.

A characteristic associative classification system is constructed in two stages:

1) Discovering the association rules inherent in a database;

2) Selecting a small set of relevant association rules to construct a classifier.

In order to improve the interpretability of the obtained classification rules and to avoid unnatural limitations in the partitioning of the attributes, different studies have been presented to get classification systems based on fuzzy association rules [23]–[25]. For instance, in [24] the authors have made use of a Genetic Algorithm (GA) [19], [20] to automatically determine minimum support and confidence thresholds, mining for each chromosome a fuzzy rule set for classification by means of an algorithm based on the Apriori algorithm [18] and adjusting the fuzzy confidence of these rules with the approach proposed by Nozaki et al in [12]. Consequently, this approach can only be used for small problems since its computational cost is very high when we consider problems that consist of a high number of patterns and/or variables. On the other hand, in [25] the authors used an algorithm based on the Apriori algorithm to mine association rules only up to a certain level and to select the K most confident ones for each class among them, in order to finally employ a genetic rule selection method that obtains a classifier from them. However, many patterns may be uncovered if we only consider the confidence measure to select the candidate rules.

In this paper we present a Fuzzy Association Rule-based Classification method for High-Dimensional problems (FARCHD) to obtain an accurate and compact fuzzy rule-based classifier with a low computational cost. This method is based on three stages:

1) Fuzzy association rule extraction for classification: A search tree is employed to list all possible frequent fuzzy item sets and to generate fuzzy association rules for classification, limiting the depth of the branches in order to find a small number of short (i.e., simple) fuzzy rules.

2) Candidate rule prescreening: Even though the order of the associations is limited in the association rule extraction, the number of rules generated can be very large. In order to decrease the computational cost of the genetic post-processing stage we consider the use of subgroup discovery based on an improved Weighted Relative Accuracy measure (wWRAcc’) to preselect the most interesting rules by means of a pattern weighting scheme [16].

3) Genetic rule selection and lateral tuning: Finally, we make use of GAs to select and tune a compact set of fuzzy association rules with high classification accuracy in order to consider the known positive synergy that both techniques present (selection and tuning). Several works have successfully combined the selection of rules with the tuning of membership functions (MFs) within the same process [14], [15], taking advantage of the possibility of different coding schemes that GAs provide. The successful application of GAs to identify fuzzy systems has led to the so-called Genetic Fuzzy Systems (GFSs) [16]–[18].

In order to assess the performance of the proposed approach, we have used twenty-six real-world datasets with a number of variables ranging from 4 to 90 and a number of patterns ranging from 150 to 19020. We have developed the following studies; first, we have shown the results obtained from comparison with three other GFSs [18]. Second, we have compared the performance of our approach with two approaches to obtain fuzzy associative classifiers. Third, we have shown the results obtained from the comparison with four other classical approaches for associative classification and with the C4.5 decision tree [19]. Furthermore, in these studies we have made use of some non-parametric statistical tests for pair-wise and multiple comparison of the performance of these classifiers. Then, we have shown a study on the influence of the depth of the trees and the number of evaluations in the genetic selection and tuning process. Finally, we have analyzed the scalability of the proposed approach.

A. Stage 1. Fuzzy Association Rule Extraction for Classification

To generate the RB we employ a search tree to list all the possible fuzzy item sets of a class. The root or level 0 of a search tree is an empty set. All attributes are assumed to have an order (in our case, the order of appearance in the training data), and the one-item sets corresponding to the attributes are listed in the first level of the search tree according to their order. If an attribute has j possible outcomes (qj linguistic terms for each quantitative attribute), it will have j one-item sets listed in the first level. The children of a one-item node for an attribute A are the two-item sets that include the one item set of attribute A and a one-item set for another attribute behind attribute A in the order, and so on. If an attribute has j > 2 possible outcomes, it can be replaced by j binary variables to ensure that no more than one of these j binary attributes can appear in the same node in a search tree. An example with two attributes (V1 and V2) with two linguistic terms (L and H) is detailed in Figure 1

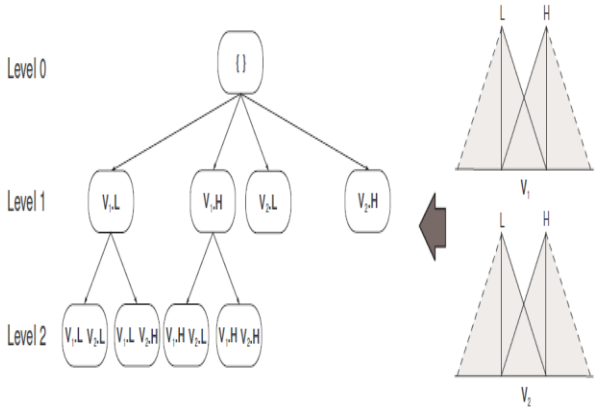

Figure 1. The search tree for two quantitative attributes V1 and V2 with two linguistic terms L and H

An item set with a support higher than the minimum support is a frequent item set. If the support of an n-item set in a node J is less than the minimum support, it does not need to be extended more because the support of any item set in a node in the subtree led by node J will also be less than the minimum support. Likewise, if a candidate item set generates a classification rule with confidence higher than the maximum confidence, this rule has reached the quality level demanded by the user and it is again unnecessary to extend it further. These properties greatly reduce the number of nodes needed for searching.

B. Stage 2. Candidate Rule Prescreening

In the previous stage, we can generate a large number of candidate rules. In order to decrease the computational costs of stage 3, we consider the use of subgroup discovery to preselect the most interesting rules from RB obtained in the previous stage by means of a pattern weighting scheme [23]. This scheme treats the patterns in such a way that covered positive patterns are not deleted when the current best rule is selected. Instead, each time a rule is selected, the algorithm stores a count i for each pattern of how many times (with how many of the selected rules) the pattern has been covered.

In the first iteration all target class patterns are assigned the same weight w(ej , 0) = 1, while in the following iterations the contributions of patterns are inversely proportional to their coverage by previously selected rules. In this way the patterns already covered by one or more selected rules decrease their weights while uncovered target class patterns whose weights have not been decreased will have a greater chance of being covered in the following iterations. Covered patterns are completely eliminated when they have been covered more than kt times.

Table 1: Five Patterns Example

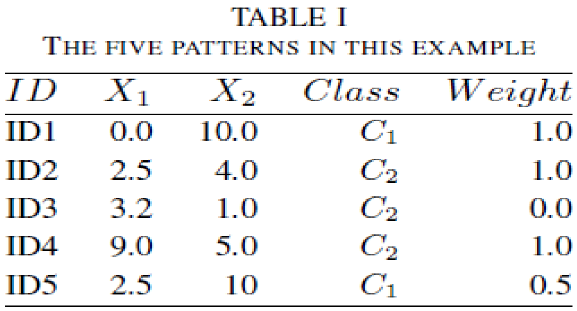

TABLE I
THE FIVE PATTERNS IN THIS EXAMPLE

ID	X_1	X_2	$Class$	$Weight$
ID1	0.0	10.0	C_1	1.0
ID2	2.5	4.0	C_2	1.0
ID3	3.2	1.0	C_2	0.0
ID4	9.0	5.0	C_2	1.0
ID5	2.5	10	C_1	0.5

Thus, in each iteration of the process the rules are ordered according to a rule evaluation criteria from best to worst. The best rule is selected, covered patterns are reweighted, and the procedure repeats these steps until one of the stopping criteria is satisfied: either all patterns have been covered more than kt times, or there are no more rules in the RB. This process is to be repeated for each

C. Stage 3. Rule Selection and Lateral Tuning

We consider the use of GAs to select and tune a compact set of fuzzy association rules with high classification accuracy from the RB obtained in the previous stage. We consider the approach proposed in [25] where rules are based on the linguistic 2-tuples representation [19]. This representation allows the lateral displacement of the labels considering only one parameter (symbolic translation parameter), which involves a simplification of the tuning search space that eases the derivation of optimal models, particularly when it is combined with a rule selection within the same process enabling it to take advantage of the positive synergy that both techniques present. In this way, this process for contextualizing the MFs enables them to achieve a better covering degree while maintaining the original shapes, which results in accuracy improvements without a significant loss in the interpretability of the fuzzy labels. The symbolic translation parameter of a linguistic term is a number within the interval [-0.5, 0.5) that expresses the domain of a label when it is moving between its two lateral labels. Let us consider a set of labels S representing a fuzzy partition. Formally, we have the pair, (Si , αi), Si ∈ S, αi ∈ [-0.5, 0.5). An example is illustrated in Fig. 3 where we show the symbolic translation of a label represented by the pair (S2, -0.3).

II RESULT AND DISCUSSION

Taking into account the results obtained, we can conclude that our model is a solid approach to deal with high dimensional datasets, as it obtains the best accuracy in the experimental study. Moreover, FARC-HD obtains models with a reduced number of rules (39,2 rules on average) and particularly with few attributes in the antecedent. Finally, the limit in the depth of the trees, along with candidate rule prescreening using the fuzzy measure wWRACC", allow us to reduce the search space considerably. Thus, the genetic process for selection and tuning does not introduce an excessive computational cost in to the whole process.

III CONCLUSION

In this paper we have future a new fuzzy associative classification method for high-dimensional datasets, named FARC-HD. Our aim is to get accurate and compact fuzzy associative classifiers with a low computational cost. To do this, we mine fuzzy association rules limiting the order of the associations in order to obtain a summary set of candidate rules with less attributes in the antecedent. We have made use of a pattern weighting scheme in order to reduce the number of candidate rules, preselecting the rules with the best quality. A genetic rule selection and lateral tuning is applied to select a small set of fuzzy association rules with a high classification accuracy.

REFERENCES

[1] H. Ishibuchi, T. Nakashima, and M. Nii, Classification and Modeling with Linguistic Information Granules: Advanced Approaches to Linguistic Data Mining. Berlin: Springer-Verlag, 2004.

[2] L. Kuncheva, Fuzzy Classifier Design. Berlin: Springer-Verlag, 2000.

[3] Y. Jin, W. Seelen, and B. Sendhoff, "Generating fc(3) fuzzy rule systems from data using evolution strategies,"

IEEE Transactions on Systems, Man and Cybernetics - Part B: Cybernetics, vol. 29, pp. 829–845, 1999.

[4] S. Ho, H. Chen, S. Ho, and T. Chen, “Design of accurate classifiers with a compact fuzzy-rule base using an evolutionary scatter partition of feature space,” IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, vol. 34, no. 2, pp. 1031–1044, 2004.

[5] H. Wang, S. Kwong, Y. Jin, W. Wei, and K. F. Man, “Multi-objective hierarchical genetic algorithm for interpretable fuzzy rule-based knowledge extraction,” Fuzzy Sets and Systems, vol. 149, pp. 149–186, 2005.

[6] Y. Zhang, X.-b. Wu, Z.-y. Xing, and W.-L. Hu, “On generating interpretable and precise fuzzy systems based on pareto multi-objective cooperative co-evolutionary algorithm,” Applied Soft Computing, vol. 11, pp. 1284–1294, 2011.

[7] C. Tsang, S. Kwong, and H. Wang, “Genetic-fuzzy rule mining approach and evaluation of feature selection techniques for anomaly intrusion detection,” Pattern Recognition, vol. 40, no. 9, pp. 2373–2391, 2007.

[8] T. Nakashima, G. Schaefer, and Y. Yokota, “A weighted fuzzy classifier and its application to image processing tasks,” Fuzzy Sets and Systems, vol. 158, no. 3, pp. 284–294, 2007.

[9] W. Combs and J. Andrews, “Combinatorial rule explosion eliminated by a fuzzy rule configuration,” IEEE Transactions on Fuzzy Systems, vol. 6, no. 1, pp. 1–11, 1998.

[10] Y. Jin, “Fuzzy modeling of high-dimensional systems: complexity reduction and interpretability improvement,” IEEE Transactions on Fuzzy Systems, vol. 8, pp. 212–221, 2000.

[11] J. Han and M. Kamber, Data Mining: Concepts and Techniques. Second Edition. San Fransisco: Morgan Kaufmann, 2006.

[12] B. Liu, W. Hsu, and Y. Ma, “Integrating classification and association rule mining,” in Proceedings of the International Conference on Knowledge Discovery and Data Mining (SIGKDD), New York, USA, 1998, pp. 80–86.

[13] B. Liu, Y. Ma, and C. Wong, “Classification using association rules: Weaknesses and enhancements,” in Data Mining for Scientific and Engineering Applications, R. Grossman, C. Kamath, and V. Kumar, Eds. Kluwer Academic Publishers, 2001, pp. 591–601.

[14] W. Li, J. Han, and J. Pei, “Cmar: accurate and efficient classification based on multiple class-association rules,” in Proceedings IEEE International Conference on Data Mining (ICDM), California, USA, 2001, pp. 369–376.

[15] X. Yin and J. Han, “Cpar: Classification based on predictive association rules,” in Proceedings of 3rd SIAM International Conference on Data Mining (SDM), San Francisco, CA, USA, 2003, pp. 331–335.

[16] J. Li, G. Dong, K. Ramamohanarao, and L. Wong, “Deeps: A new instance-based lazy discovery and classification system,” Machine Learning, vol. 54, no. 2, pp. 99–124, 2004.

[17] Y. Wang and A. Wong, “Boosting an associative classifier,” IEEE Transactions on Knowledge and Data Engineering, vol. 18, no. 7, pp. 988–992, 2006.

[18] F. Thabtah, “A review of associative classification mining,” Knowledge Engineering Review, vol. 22, no. 1, pp. 37–65, 2007.

[19] R. Rak and L. K. M. Reformat, “A tree-projection-based algorithm for multi-label recurrent-item associative-classification rule generation,” Data and Knowledge Engineering, vol. 64, no. 1, pp. 171–197, 2008.

[20] C. Zhang and S. Zhang, Association Rule Mining: Models and Algorithms Series. Berlin: Lecture Notes in Computer Science, LNAI 2307, Springer-Verlag, 2002.

[21] V. Cherkasski and F. Mulier, Learning from Data: Concepts, Theory, and Methods. Wiley-Interscience, 1998.

[22] P. Tan, M. Steinbach, and V. Kumar, Introduction to data mining. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2005.

[23] Y.-C. Hua and G.-H. Tzeng, “Elicitation of classification rules by fuzzy data mining,” Engineering Applications of Artificial Intelligence, vol. 16, no. 7-8, pp. 709–716, 2003.

[24] Y. Hu, R. Chen, and G. Tzeng, “Finding fuzzy classification rules using data mining techniques,” Pattern Recognition Letters, vol. 24, no. 1-3, pp. 509–519, 2003.

[25] Y. Yi and E. Hullermeier, “Learning complexity-bounded rule-based classifiers by combining association analysis and genetic algorithms,” in Proceedings of 4th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT 2005), Barcelona, Spain, 2005, pp. 47–52.

[26] Y.-C. Hua, “Determining membership functions and minimum fuzzy support in finding fuzzy association rules for classification problems,” Knowledge-Based Systems, vol. 19, no. 1, pp. 57–66