Extracting fuzzy classifier rules from mixed data

Roman Ostapenko
Laboratory of Intelligent Systems
Tomsk State University of Control Systems and Radioelectronis
Tomsk, Russia
ro99man2@mail.ru

Ilya Hodashinsky
Laboratory of Intelligent Systems
Tomsk State University of Control Systems and Radioelectronis
Tomsk, Russia
hodashn@rambler.ru

***Abstract* — Real world data cannot exist without mixing continuous and nominal data. This paper presents fuzzy classifier modification for mixed data classification. Basic fuzzy classifiers can only be applied to continuous attribute values and cannot handle mixed data. This problem can be solved by counting nominal attribute frequencies in each cluster. The K-means algorithm is used for clusterization. The proposed method was tested on data sets from the KEEL database and demonstrated higher classification accuracy than the method that does not consider nominal data.**

Keywords — Fuzzy classifier, classification, KEEL, K-means, mixed data.

I. Introduction

Classification is an important component of the scientific direction, called "machine learning". However, the meaning of "classification" is ambiguous, it contains several interpretations. In this article, classification is the division of a set of objects (observations) into groups (classes), based on the analysis of their indicative description. Classifier is defined as a function of the set of objects X and the set of class labels C: X → C, capable of specifying a class label for an arbitrary object from the original set; c = c(**a**; **θ**) is the label corresponding to the feature vector **a**, and **θ** is the vector of classifier parameters. Among the many different types of classifiers, the fuzzy classifier (FC) is distinguished by the possibility of interpreting both the actual result obtained and the process of obtaining it [1]. The process of building an FC includes three stages: feature selection, rule base generation, and optimization of FC parameters [2].

Many real-world datasets contain objects with both continuous and categorical features. To apply fuzzy classifiers to such problems, it is necessary to extend their current learning algorithms so that they can deal effectively with data with mixed attributes. In [3] proposed a mixed rule learning approach, which involves a multi-stage fuzzy rule learning part for handling continuous attributes in more depth and a traditional rule learning part for handling both discrete attributes and continuous attributes in diverse ways. In [4] presents a classification system called AFS classifiers that is applicable to mixed attribute data sets. The main disadvantage of this approach is that it generates a large number of rules. The most common way to classify data sets with mixed features is to use coding methods to convert categorical values into numeric values. One of the disadvantages of original learning algorithms is the inability to process and learn directly from data with mixed attributes without using encoding methods, which, when used, lose the interpretability of nominal features.

This article proposes an original algorithm for generating fuzzy rules and fuzzy terms. The algorithm takes into account mixed data and improves interpretability and classification accuracy.

II. Existing approaches

One of the most common approaches to handle mixed data is feature replacement when mixed data is replaced by continuous data [5]. This method is easy to use and does not require any special algorithms. However, the meaning of the nominal data is lost. This leads to a loss of interpretability, which contradicts the main advantages of the fuzzy classifier. There is also accuracy loss due to term generation specific.

Another approach is based on the mixed data processing method [6]:

$$b_i = \frac{1}{n}\sum_{i=1}^{n} g\left(d_{hi}, d_{ji}\right), \quad (1)$$

where n is the number of nominal features, d_{hi} is the value of the discrete attribute i of input, d_{ji} is the value of ith discrete attribute associated with current rule, and $g(d_{hi}, d_{ji})$ is given by:

$$g\left(d_{hi}, d_{ji}\right) = \begin{cases} 1, \ if \ \ d_{hi} \,\&\, d_{ji} > 0 \\ 0, \ if \ \ d_{hi} \,\&\, d_{ji} = 0 \end{cases}. \quad (2)$$

Based on this function, a special membership function is defined. For continuous features, a triangular membership function is used. Equation (1) is used to handle nominal features. This approach cannot work without data preprocessing. The nominal data must be represented as a binary data string. A binary data string is obtained by replacing the value of a nominal feature with a number that is a power of 2. If different values of a nominal feature occur in one cluster, then these values are combined into a new binary string. This binary string is formed using a bitwise logical OR operator. In this form, binary strings are included in hybrid fuzzy rules.

Ideas of replacement method (RM) [5] and hybrid algorithm (HA) [6] were adapted by us to work with a fuzzy classifier. Both methods were used with triangular membership function to handle continuous features.

III. Frequency-based method for rule generation

In contrast to the early described approaches, the proposed method has the following additional properties:

- The method provides higher interpretability;
- There is no classification accuracy loss due to the specifics of nominal term generation. Traditional fuzzy rules with the replacement method can include redundant nominal feature values that can lead to classification accuracy loss.

The proposed algorithm for generating a fuzzy classifier structure is as follows.

First, training data is clustered by a clustering algorithm selected for the task. Then each cluster is considered separately. The unique values of each nominal feature are selected in cluster. The number of occurrences of each unique value is counted in the given cluster. The membership degree is calculated according to Eq. (3):

$$f_i = \frac{fr_i}{\max(FR)}, \quad (3)$$

where fr_i is the number of occurrences of element i in nominal feature and FR is the list of occurrences of current nominal feature every unique element and defined as $\{fr_1, \ldots, fr_n\}$.

Membership degrees for continuous features are calculated using triangular membership functions. In contrast, membership degrees are defined as number of occurrences for nominal features. Both resulting membership degrees are used to form fuzzy rules.

IV. Experiment

To test the performance of the proposed approach, we selected mixed data sets from the KEEL repository. These data were also used to subsequently compare the described approaches. The experiment was carried out according to the scheme of 10-fold cross-validation. Table I shows the characteristics of the used data sets.

TABLE I. Data sets description

Data set	Features (Real/Integer/Nominal)	Classes	Number of instances
adult	(6/0/8)	28	4174
abalone	(7/0/1)	2	45222
australian	(3/5/6)	2	690
automobile	(15/0/10)	6	205
crx	(3/3/9)	2	653
german	(0/7/13)	2	1000
lymphography	(0/3/15)	4	148
saheart	(5/3/1)	2	462

In this experiment, K-means algorithm forms clusters. Fuzzy rules are formed based on each cluster. The K-means algorithm takes into account all continuous features with the replacement method.

V. Results

The proposed frequency-based method (FM), a hybrid algorithm (HA), and a replacement method (RM) were used to generate fuzzy rules. The data were classified based on the obtained rules. Classification accuracy was measured for each data set. The obtained experiment results are presented in Table II.

The proposed method showed the highest classification accuracy on most data sets. The described methods were compared to each other using the Friedman test. The null hypothesis states that there is no relationship between two observed phenomena. The alternative hypothesis states that the results of different algorithms have non-random differences. The significance level 0,05. Ranks and Friedman test values are presented in Table III.

TABLE II. Classification results

Data set	Training set			Test set		
	FM	***HA***	***RM***	***FM***	***HA***	***RM***
adult	24,73	23,35	**74,63**	23,23	23,8	**74,17**
abalone	**25,05**	25,03	21,89	**25,32**	25,25	19,17
australian	**57,68**	55,27	49,23	**57,10**	54,49	50,14
automobile	47,72	**47,86**	42,62	**44,94**	43,66	34,49
crx	**58,01**	55,05	54,16	**58,03**	54,99	53,64
german	**70,00**	69,79	**70,00**	**70,00**	69,70	**70,00**
lymphography	**76,87**	70,94	53,99	68,58	**70,45**	46,75
saheart	**67,32**	65,90	61,88	**66,69**	66,03	62,79
Mean	53,42	51,65	**53,55**	**51,74**	51,05	51,39

TABLE III. Ranks and p-values of Friedman test

Data set	FM	HA	RM	p-value
Train set	2,69	1,88	1,43	0,036
Test set	2,56	2,00	1,44	0,073

Statistical analysis showed that there are non-random differences in the training set results The results in the test set have random differences.

In the frequency-based method we developed, nominal features are processed according to the frequencies of their occurrence in the cluster. Higher frequencies mean higher impact on class assumption. That provides solid degree of interpretability. Replacement method reduces meaning of feature which leads to the difficulty of interpretation. That makes frequencies based method suitable for tasks that require high interpretability such as credit risk assessment, medical diagnosis, and fault detection [1].

VI. Conclusion

The purpose of this work was to develop a method that takes into account mixed data when using a fuzzy classifier. Eight data sets were used to evaluate the performance of the classifier. The obtained classification accuracy was compared with the accuracy obtained using the replacement method, and the hybrid membership function method. The classification accuracy of frequency based method was higher on most data sets. In the future, it is planned to use other methods of grouping and clustering data.

References

[1] Porebski, S. Evaluation of fuzzy membership functions for linguistic rule-based classifier focused on explainability, interpretability and reliability / S. Porebski // Expert Syst. Appl. – 2022. – Vol. 199. – P. 117116.

[2] Bardamova, M.B. Formation of fuzzy classifier structure by a combination of the class extremum algorithm and the shuffled frog leaping algorithm for imbalanced data with two classes / M.B. Bardamova, I.A. Hodashinsky // Optoelectron. Instrum. Data Process. – 2021. – Vol. 57. – P. 378–387.

[3] Liu, X. Extraction of fuzzy rules from fuzzy decision trees: An axiomatic fuzzy sets (AFS) approach / X. Liu, X. Feng, W. Pedrycz // Data Knowl. Eng. – 2013. – Vol. 84. – P. 1–25.

[4] Liu, H. Multi-stage mixed rule learning approach for advancing performance of rule-based classification / H. Liu, S.-M. Chen // Inf. Sci. – 2019. – Vol. 495. – P. 65–77.

[5] von Eye, A. Categorical variables in developmental research: Methods of analysis / A. von Eye, C. Clifford. – San Diego: Academic Press, 1996. – 286 p.

[6] Shinde, S. Extracting classification rules from modified fuzzy min–max neural network for data with mixed attributes / S. Shinde, U. Kulkarni // Appl. Soft Comput. – 2016. – Vol. 40. – P. 364–378.