"Use of BEXA family algorithms in bioinformatics data classification"

Madara Gasparovica (*Riga Technical University*), Ludmila Aleksejeva, Valdis Gersons

***Keywords* – Classification algorithms, bioinformatics data, BEXA, UCI data.**

I. Introduction

This article studies the possibilities of BEXA family classification algorithms – BEXA, FuzzyBexa and FuzzyBexa II in data, especially bioinformatics data, classification. Three different types of data sets were used in the study – data sets often used in the literature (like Iris data set), UCI data repository real life data sets (like breast cancer data set) and real bioinformatics data sets that have the specific character – a large number of attributes (several thousands) and a small number of records. For the comparison of classification results experiments were carried out using all data sets and other classification algorithms. As a result, conclusions were drawn and recommendations given about the use of each algorithm of BEXA family for classification of various real data, as well as an answer is given to the question, whether the use of these algorithms is recommended for bioinformatics data.

II. Used algorithms and Data sets

A. Used algorithms

BEXA – cover procedure – iteratively examines each concept (class). The concept is examined until all positive records (instances) of the concept are covered. After adding each rule to the set all positive records covered by the rule are deleted from the set of positive instances. The best rule is found using Laplace evaluation function. Creation of specializations creates specializations for conjunctions. They are returned to the best rule searching procedure for assessment [1].

FuzzyBexa – in this algorithm the BEXA basis algorithm is fitted for fuzzy data. The highest (Cover) and middle level (searching for the best rule) procedures include membership variables (alpha). The lowest level procedure examines linguistic attribute values (the possible linguistic values of each attribute in the final fuzzy data instead of attribute values or intervals as it is in the case of BEXA) [2].

FuzzyBexa II – in this algorithm each class is not examined individually; instead it generates rules for all classes. The highest level (Cover) does not split the training set into positive and negative sets, it transfers the whole training set and the set of concepts to the middle level procedure. The middle level procedure - find best conjunction – in its turn, finds both the conditional (antedecant, IF) and the resulting (consequent, THEN) part for each rule. Respectively, the lowest level procedure that generates specializations also processes the whole training set (or its part) instead of positive and negative instances of a split data set [3].

B. Used data sets

This study uses 16 data sets that can be conditionally divided into three groups. Initially the classification algorithms are tested using popular UCI data sets like Iris data set to evaluate the result of these algorithms comparing it to other algorithms. Then a series of experiments is carried out using real natural data available in the UCI repository to assess the accuracy of the algorithms using real medium-sized data sets. The section of practical experiments is concluded with experiments that use real bioinformatics data sets. The description of the data sets is given in Table 1.

TABLE I
USED DATA SETS

Name	Number of samples	Number of attributes	Number of classes
Iris data set (UCI)	150	4	3
Auto MPG Data Set (UCI)	398	8	2
Ionosphere Data Set (UCI)	351	34	2
Nursery Data Set (UCI)	12960	8	3
…			
GSE89 (bladder cancer)	40	5724	3
GSE1987 (lung cancer)	34	10541	3

III. Experiments

All experiments using algorithms of BEXA family were carried out in an application created using Java programming language (using Weka libraries). All experiments include evaluation using cross-validation. The experiment plan includes the data sets that were described in the previous section. To compare the results of BEXA family classification algorithms experiments were conducted using other popular algorithms and the same data sets. It is done with the aim to ascertain the competitiveness of the classification algorithms and draw the necessary conclusions, as well as answer the raised question – is the use of BEXA family classification algorithms recommended for bioinformatics data classification and whether it has potential.

IV. Conclusions

The algorithms of BEXA family can be used in bioinformatics data classification but the obtained results are not competitive when compared to other popular data mining algorithms. Additional experiments are necessary to improve classification results and assess the impact of various membership functions on the classification accuracy of BEXA family algorithms. Recommendations are given about the use of the most successful algorithm of BEXA family based on the used data set.

V. References

[1] H. Theron, I. Cloete, "BEXA: A Covering Algorithm for Learning Propositional Concept Descriptions," in Machine Learning, Vol. 24, Boston: Kluwer Academic Publishers, 1996, pp.5-40.

[2] J. van Zyl, I.Cloete, "FuzzConRi – A Fuzzy Conjunctive Rule Inducer," in Proc. Workshop on Advances in Inductive Rule Learning, ECML, 2004, pp.194-203.

[3] J. van Zyl, I.Cloete, "Simultaneous Concept Learning of Fuzzy Rules," in Proc. Workshop on Advances in Inductive Rule Learning, CCML, 2004, pp.194-203.