EMCSR: pp-pp, year
ISSN xxxx-xxxx
http://conference.emcsr.net

European Meetings on Cybernetics and Systems Research

Rule weight use in bioinformatics data classification

Madara Gasparovica, Ludmila Aleksejeva

Riga Technical University, 1 Kalku Street, Riga, LV-1658, Latvia, madara.gasparovica@rtu.lv, ludmila.aleksejeva_1@rtu.lv

Abstract: *This article studies the possible application of fuzzy classification methods that use rule weights in classification of bioinformatics data, in particular, it uses algorithms Fuzzy Rule Learning Model, whose results are compared to those of Fuzzy Unordered Rule Induction Algorithm. To assess the benefits of approach specifics, the experiments were carried out using eleven real bioinformatics data sets provided by University of Ljubljana, Faculty of computer and information science Bioinformatics Laboratory on their home page. The article provides conclusions about the efficacy of rule weight assessment methods and outlines the course of future research.*

Keywords: bioinformatics, classification task, fuzzy rule induction, rule weighing

Acknowledgement: This work has been supported by the European Social Fund within the project «Support for the implementation of doctoral studies at Riga Technical University». Thanks to Dr.habil.sc.comp. professor Arkady Borisov for help and support.

Introduction. Often in data mining there are specific algorithms created and applied to characteristic tasks that match them the best instead of using universal algorithms; but frequently they can be successfully implemented in other fields of data mining. This study examines the application of a pattern recognition algorithm in a classical data classification task using specific bioinformatics data that ask for definite algorithms due to their complex structure (large number of attributes and very few records). The article examines classification results of seven different data sets using Fuzzy Rule Learning Model algorithm (implemented in the data mining software KEEL (Alcalá-Fdez,J. et al.(2011)) and Fuzzy Unordered Rule Induction Algorithm (implemented in the data mining software Weka).

1. Experimental Study

The study uses an algorithm Fuzzy Rule Learning Model by the Chi et al. (Chi et al., 1995) approach with rule weights to determine its suitability to the specifics of bioinformatics data (thousands of attributes and few records). To assess this algorithm, its gain in other less complex data sets was compared to its accuracy gain in bioinformatics data, using FURIA algorithm (Hühn et al., 2009) as a benchmark for classification accuracy in bioinformatics data (Gasparovica, 2012). The datasets used for algorithm evaluation were taken from University of Ljubljana, Faculty of computer and information science Bioinformatics Laboratory on their home page (University of Ljubljana, 2012) and they contained gene expression data and other typical bioinformatics data with the previously described specifics. To evaluate the accuracy of the algorithms, the experiments were carried out using 10-fold cross-validation.

2. Results and Discussion

The experimental results show that when applying algorithm Fuzzy Rule Learning Model available in Keel software and using 10-fold cross-validation the results do not exceed the performance of FURIA algorithm in all cases. It can be explained by the specific nature of the data

sets – the large number of attributes and comparatively small number of records (e.g., 12625 attributes and 24 records in the Breast cancer data set), therefore the rules induced in the training phase have a complex structure and cannot be interpreted as easily as rules obtained using FURIA. The small training set cannot provide enough information about class structures and the induced rules do not cover/satisfy the previously unknown instances from the test sets. Another aspect, why FURIA has more successful rules, is its ability to stretch rules to cover previously unknown new records that differ from the training set instances and classify them accordingly.

During this study some experiments were carried out to determine, whether the number of labels per variable has an influence on classification accuracy in tasks that are solved using rule weights. It was concluded that its influence is relevant – when the number of labels per variable is increased by one unit, the change in the results was minimal, but increasing it by several units showed a significant change in classification accuracy. The change mostly was negative – the increase in the number of labels per variable decreased the classification accuracy proportionally.

The future research could involve the use of rule weight technology but it should be adapted to bioinformatics data specifics and tasks, that would be a narrower solution that could not be used in all fields but, instead, fitted for bioinformatics data, where it would show comparatively high and robust results. Nevertheless, to evaluate and approve this theory, there should be more research carried out using other bioinformatics data sets. Future research also holds more studies of Fuzzy Rule Learning Model algorithm, its parameters and options.

References

Alcalá-Fdez,J. et al.(2011). Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework. Journal of Multiple-Valued Logic and Soft Computing 17:2-3, pp. 255-287

Chi, Z., Yan, H. Pham. T. (1996). Fuzzy Algorithms: With Applications To Image Processing and Pattern Recognition. World Scientific.

Gasparoviča, M., Aleksejeva, L., (2012). Feature selection for bioinformatics data sets – recommended?: Proceedings of 5 th conference Applied information and communication technology, Jelgava, Latvia, April 26-27.(Submitted)

Hühn, J., Hüllermeier, E.,(2009). FURIA: An Algorithm for Unordered Fuzzy Rule Induction, Data Mining and Knowledge Discovery, Vol.19, No.3, pp.293-319.

University of Ljubljana, Faculty of computer and information science Bioinformatics Laboratory home page. Retrieved January 2, 2012, from http://www.fri.uni-lj.si/en/laboratories/biolab/

About the Author

Madara Gasparoviča

Received her diploma of Mg. sc. ing. in Information Technology from Riga Technical University in 2010. Now she is a PhD student of Information Technology program at Riga Technical University. Previous publications: Gasparovica M., Aleksejeva L. Using Fuzzy Unordered Rule Induction Algorithm for Cancer Data Classification Proceedings of 17th International Conference on Soft Computing, MENDEL 2011, Czech Republic, Brno, June 15-17, 2011, pp. 141-147. Her interests include decision support systems, data mining tasks and modular rules. She is a member of IEEE.

Ludmila Aleksejeva

Received her Dr. sc. ing. degree from Riga Technical University in 1998. She is associate professor in the Department of Modelling and Simulation of Riga Technical University. Her research interests include decision making techniques and decision support systems design principles as well as data mining methods and tasks, and especially mentioned techniques collaboration and cooperation. Previous publications: Gasparoviča M., Aleksejeva L., Tuleiko I. Finding Membership Functions for Bioinformatics Data // Proceedings of 17th International Conference on Soft Computing, MENDEL 2011, Czech, Brno, June 15-17, 2011.pp. 133-140.