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

EMCSR
European Meetings on Cybernetics and Systems Research

Robust dimensionality reduction in bioinformatics data

Inese Polaka, Arkady Borisov

Riga Technical University, 1 Kalku Street, Riga, LV-1658, Latvia, Inese.Polaka@rtu.lv, +37126418323

Abstract: *This article proposes a technique to reduce data dimensionality in bioinformatics tasks. In the cases where there are thousands of attributes and only few hundred instances even the scalable classification methods benefit from diminishing the number of features. But at the moment there is no way of finding the best feature selection and evaluation method for a particular data set and the performance of these methods varies a lot depending on the specific nature of a data set. Therefore it is necessary to find a robust technique that would perform the process of feature selection without the impact of a specific method. The system proposed in this article minimizes this effect while still producing highly informative feature subsets that are easier to comprehend, interpret and use in classification.*

Keywords: feature selection, data mining, classification, bioinformatics.

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».

One of the solutions to the problem of high dimensionality in biomedical and omics studies is data mining methods that depict the information and knowledge in such ways and structures that are familiar to all fields of research. An approach to this issue can be found in data mining techniques of dimensionality reduction. Such techniques are already used in other fields of research and can be successfully applied to high-dimensional data like those of bioinformatics (gene or protein microarrays).But there are no guidelines to select the most appropriate feature selection or evaluation technique. A use of data-mining based techniques is proposed in this study and blended into a system that allows processing bioinformatics data with high dimensionality without losing indicative information or interpretability. The system selects the most informative features based on several different evaluation techniques and then selects the features that are recognized as significant by most methods. Furthermore, the experiments show that the array does not lose significant information and the discriminative powers of classification models are similar or higher (due to avoidance of overfitting in large attribute sets).

1. Classification

Classifiers used in this research are the most popular classification methods in bioinformatics so far (according to the available literature) – Random Forests (Breiman, 2001), that create several tree-based structures that depict hierarchical relationships between features, and SVM (Vapnik, 1995) that bases the classification on finding functional relations in class discrimination. Another method that was included in the experimental study is C4.5 algorithm (Quinlan, 1993) that builds tree-based models of classification that are easily interpretable by experts, who are not related to data mining, and present the inter-feature connections.

2. Feature selection

There are several feature selection techniques that analyze the ability of a feature or a group of features to discriminate between target conditions. These techniques allow selecting a small

subgroup of features that hold the most information about the specific condition. But the problem with these techniques is their unstable nature because the best selection technique in one data set may perform poorly (select less informative features than other methods) in another data set and also change its performance as the data set increases, which is a common case in bioinformatics because new experiments are often performed using the same gene/protein libraries for microarrays (Saeys, 2007).

3. The proposed system

The system proposed in the study includes the preprocessing of data, by imputing missing values (the data sets are normalized after scanning by BMC), finding subsets of autoantibodies with significant discrimination abilities and computing a disease classification model in the terms of autoantibody panel and interactions between autoantibodies in a form of a decision tree. The scheme of the system is given in the Figure 1.

The selection of the feature subset is implemented using various feature subset selection methods and evaluation techniques to eliminate the impact of the method – our previous studies have shown that different methods give changing results based on the specifics of data sets. The robustness of the feature subset is ensured by using different evaluation metrics (Chi-squared, OneR classifier, Information gain etc.) to select the preliminary groups of autoantibodies and then the system filters the features that are present in more than *n*% of the subsets (*n* is changed between sets of experiments to explore the changes in classification accuracy). This ensures that only the truly significant features are present in the feature subset – a feature can be selected because of the specific nature of data by one method but the used methods differ enough to avoid inclusion of such random into the ultimate feature subset.

This final autoantibody panel is then used to build a decision tree classification model using the standard C4.5 algorithm that holds the most significant autoantibodies and explains their inter-relations.

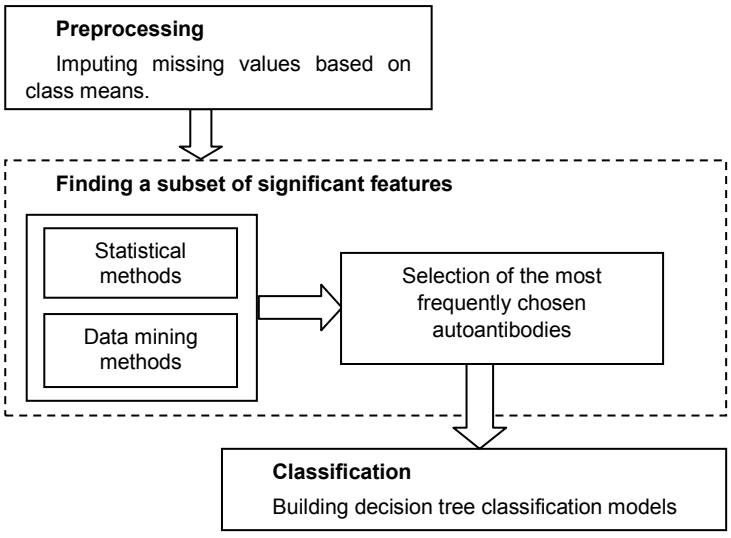

Figure 1: Autoantibody array data processing system

The results of the previous studies (Polaka, 2011) suggest that the accuracy of discrimination between affected patients and healthy individuals does not significantly decrease and in some cases it improves due to the overfitted models built on full data. The experimental results of the most frequent feature subsets were compared to the average results of other feature selection methods that rank features separately (Chi-squared, Information Gain, Gain Ratio, OneR, ReliefF) with the corresponding number of features per set. The results show that the proposed method

beats the average result of other methods in 19 cases out of 45, one result was tied and 25 were slightly worse than the average (in seven cases the loss in accuracy was less than half percent). The average change in accuracy of C4.5 was -0,95 percent, for Random Forests it increased by 0,54 percent and for SVM it increased by 0,41 percent. When the classification results of most frequent feature sets were compared with the worst results of other single feature selection method, the proposed method presented better results in 38 out of 45 cases. The average accuracy increase for C4.5 was 1,3 percent, for Random Forests it was 5,63 percent and for SVM it was 4,82 percent.

This means that choosing the most frequent features over several evaluation methods leads to better feature subsets in most cases and it allows escaping the consequences of selecting the wrong feature selection method. It will not grant the best possible results but the selected feature subsets perform well in most cases and allow skipping the search for the best method that could turn out being worse as the data set grows in size and the number of records increases.

4. Conclusions

The experiments show that the Frequent feature data subsets perform at the average level of other methods but provides the necessary robustness for the large-scale and changing data of bioinformatics, without the risk of choosing the wrong feature selection method that would decrease the results of classification.

References

Breiman, L. (2001). Random Forests. In *Machine Learning Volume 45 Issue 1* (pp. 5-32). Hingham, MA, USA: Kluwer Academic Publishers.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Berlin: Springer-Verlag.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers.

Saeys, Y., Inza, I., Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. In A. Bateman & A Valencia (Eds.), *Bioinformatics Volume 23 Issue 19* (pp. 2507-2517). Oxford, UK: Oxford University Press.

Polaka, I., Borisov, A. (2011). Impact of Antibody Panel Size on Classification accuracy. In *Scientific Journal of RTU. 5. series., Datorzinātne Volume 45* (In press). Riga, Latvia: RTU Publishing.

About the Authors

Inese Polaka

Inese Polaka is a Reasearch Fellow at the Institute of Information Technology, Riga Technical University (Latvia). She received her Mg.sc.ing. degree in 2009 from Riga Technical University; and is now continuing her research and education for the second year of PhD at the Institute of Information Technology, Riga Technical University.

Inese has worked at the Research department of LLC "Riga East university hospital" as Research Fellow. The most significant research areas of interest include bioinformatics, data mining, statistics and their application in medicine and life sciences, mostly focusing on classification tasks and data processing for diagnostics and prognostics studies, also using genetic algorithms in system implementations.

Inese is a member of IEEE and the Chair of Student Branch of Latvian Section.

Arkady Borisov

Arkady Borisov is a Professor of Computer Science at the Institute of Information Technology, Riga Technical University (Latvia). He received his Doctor of Technical Sciences degree from Taganrog State Radio-Engineering University (Russia) in 1986 and Dr.habil.sci.comp. degree from the Latvian Council of Science in 1992.

The research areas include inductive learning, fuzzy logic, classification tasks artificial neural networks, genetic algorithms, bioinformatics.

Prof. Borisov is member of IFSA European Fuzzy System Working Group, Russian Fuzzy System and Soft Computing Association, Honorary member of the Scientific Board, member of the Scientific Advisory Board of the Fuzzy Initiative Nordrhein-Westfalen (Dortmund, Germany), member of the Latvian National Automation Organisation.