Poster: Analysis of gene ranking algorithms with extraction of relevant biomedical concepts from Pubmed publications

Simon Kocbek[1,*], Rune Sætre[2], Gregor Stiglic[1], Jin-Dong Kim[2], Igor Pernek[1], Yoshimasa Tsuruoka[4], Peter Kokol[1], Sophia Ananiadou[3], and Jun'ichi Tsujii[2,3,*]
[1]Faculty of Health Sciences, University of Maribor, Maribor, Slovenia
[2]Dep. of Computer Science, University of Tokyo, Tokyo, Japan
[3]National Centre for Text Mining (NaCTeM), Manchester UK.
[4]School of Information Science, JAIST Ishikawa, Japan
*simon.kocbek@uni-mb.si, tsujii@is.s.u-tokyo.ac.jp

DNA microarray is a technology that can simultaneously measure the expression levels of thousands of genes in a single experiment. Often, the most informative genes have to be selected from different gene expression level datasets. One of the possible ways to rank the genes is to use a feature selection (FS) method. There are many FS methods which can be used, but how do researches know which one is the best? Several different methods were proposed to estimate the "goodness" of the ranked gene lists [2-4]. However, these methods usually require computer experts which know how FS methods and learning algorithms work. Therefore we propose AGRA (Analysis of Gene Ranking Algorithms), a novel method where biologists and other experts with low or no previous computer knowledge can compare different FS methods with help of evidence mined from PubMed publications. To achieve this, AGRA uses the FACTA+ [1] system which is an online text search engine for MEDLINE abstracts and it helps users browse biomedical concepts (e.g. genes/proteins, diseases, symptoms, drugs, enzymes and chemical compounds) which co-occur in the documents retrieved by a search query.

With the AGRA method, we define a biomedical concept space (BCS) for each gene list. BCS is defined as a set of ranked biomedical concepts gathered through FACTA+ where they are grouped into the six categories. The quality of the uploaded ranked gene lists can be compared with two metrics. In the first method, the overlap between each pair of two gene list BCSs is calculated to compare the effectiveness/performance of the FS methods. Overlap is a simple method to measure similarity between two BCSs where biomedical concepts that appear in both BCS are divided by the number of concepts in the shorter BCS. Another option to compare the FS methods with the AGRA method is to search for the position of relevant biomedical concepts in the final gene list BCS. Position of a single biomedical concept is defined as its ranked number amongst all the concepts in one of the categories. This way, researchers can decide which FS method selects the most important concepts and ranks them higher compared to other methods.

We tested the system with seven different gene ranking algorithms. The AGRA method was compared to overlaps calculated from feature selection based rankings. The results demonstrate higher sensitivity of AGRA when multivariate and univariate algorithms are compared. Figure 1 shows some of the results calculated with the proposed method. Additional experiment where we searched for the specific concept ("breast cancer" in the Disease category) shows that SVM-RFE ranks the searched concept in top 25% of concepts despite low stability of SVM-RFE in overlap scores - i.e. much higher than in competitive feature selection methods.

Gene/Protein

	GAINRATIO	INFOGAIN	ONER	RELIEFF	SVM-RFE	SYMMETRICAL
ChiSquared	63.64	100.0	72.99	21.66	25.48	68.84
GainRatio		59.85	66.67	9.09	9.85	90.91
InfoGain			70.07	16.31	22.7	66.67
OneR				18.25	13.14	72.26
ReliefF					22.77	15.22
SVM-RFE						13.04

Disease

	GAINRATIO	INFOGAIN	ONER	RELIEFF	SVM-RFE	SYMMETRICAL
ChiSquared	70.83	100.0	82.98	49.06	32.08	74.0
GainRatio		68.75	80.85	29.17	29.17	93.75
InfoGain			80.85	49.06	32.08	72.0
OneR				40.43	25.53	82.98
ReliefF					24.0	38.0
SVM-RFE						28.0

Figure 1. An example of the AGRA overlap results with seven FS methods.

ACKNOWLEDGMENT

This research work is funded by Slovenian Research Agency, under grant BI-JP/09-11-00, Japan Society for the Promotion of Science Bilateral Joint Project grant and BBSRC BB/G013160/1 grant.

REFERENCES

[1] Y. Tsuruoka, J.Tsujii, S. Ananiadou., "FACTA: a text search engine for finding associated biomedical concepts," in Bioinformatics, 2008.

[2] Ma S. Empirical study of supervised gene screening. BMC Bioinformatics. 2006;7, article 537

[3] Qiu X, Xiao Y, Gordon A, Yakovlev A. Assessing stability of gene selection in microarray data analysis. BMC Bioinformatics. 2006;7(1, article 50)

[4] Ein-Dor L, Zuk O, Domany E. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. Proceedings of the National Academy of Sciences of the United States of America. 2006;103(15):5923–5928.