

Paper reference

E. Georggi, L. Richter, U. Ruckert, and S. Kramer. **Analyzing microarray data using quantitative association rules.** *Bioinformatics* 2005 Sep 1; 21 Suppl 2:ii123-ii129.

Abstract

The main goal of the paper is to find regularities in microarray data through association rule mining. The main contribution of the paper is their mining algorithm which finds quantitative association rules between genes. In contrast to the conventional discrete mining algorithms, this algorithm can be directly applied on the unmodified continuous expression data and hence prevent any loss of information due to discretization. Furthermore, this algorithm is able to account for cumulative effects of genes unlike its discrete counterpart. This association rule mining algorithm is based on half-spaces which are capable of finding regularities that are not axis-parallel.

The authors test their association algorithm on two set of microarray data and provide a biological interpretation to their results. Moreover, they perform a list of “sanity checks” on their algorithm to characterize its reliability, robustness, statistical significance and scalability. Finally they compare their method to their discrete counterpart and conclude that they mostly find very distinct set of rules among genes.

Discussion

In traditional association rule setting, the conditions on the right hand side and left hand side are based on discrete attributes. In order to extend these rules to continuous data as was done here, the authors choose a smooth separation function class, namely hyperplanes, in order to minimize the error caused by random noise and measurement errors in data. In the association rules case, hyperplanes are used as conditions and the way an association is detected is by testing a weighted sum of variables against a threshold. With this, for instance one can build rules relating the expression levels of three genes in arginine metabolism: ‘ $0.99 \text{ ARG1} - 0.11 \text{ CAR1} > 0.062 \rightarrow 1.00 \text{ ARG3} > -0.032$ ’. In order to force the algorithm to find interesting association rules, the authors make sure that the two conditions (hyperplanes) are not correlated by forcing two hyperplanes to be perpendicular.

In contrast with discrete association mining methods, this method is not able to exhaustively enumerate all possible association rules since downward closure property does not hold for such rules. Hence the idea of generating all possible solutions is abandoned and an optimization approach that finds locally optimal solutions with respect to a parameterized score function is adopted. In this way, the user specifies the sort of rules that he/she is looking for and the algorithm returns the locally optimal solutions (similar to K-means and EM).

The authors further constraint their solution set by defining criterions for interestingness of the association rules, 1) high confidence score, i.e. the fraction of instances in the data that fulfill both conditions on the left and right divided by the instances that fulfill only the left hand side condition should be as high as possible, 2) coverage, i.e. fraction of instances of the dataset that satisfy the left-hand side condition (this is given by the user), 3) sparseness, i.e. most users prefer finding sparse association rules. Finally, a post-processing threshold is applied in order to set the small coefficients to zero while keeping the two hyperplanes perpendicular.

In the results section, the utility of the algorithm to find quantitative association rules was shown on microarray data. The results showed that the association rules that were found were biologically interpretable. The authors also managed to provide empirical evidence for the reliability of their algorithm to find known association rules using synthetic data. They also showed the statistical significance of the results and their reproducibility as well as scalability of their approach to larger datasets. Finally, a comparison with discretization-based mining algorithm showed that the two different algorithms found mostly distinct set of rules. In conclusion, this paper presented a somewhat robust algorithm, to find hidden regulation rules in expression data.