BIOINFORMATICS

Vol. 21 Suppl. 2 2005, pages ii123–ii129
doi:10.1093/bioinformatics/bti1121

Microarrays

Analyzing microarray data using quantitative association rules

Elisabeth Georgii, Lothar Richter, Ulrich Rückert and Stefan Kramer*

Technische Universität München, Institut für Informatik/I12, Boltzmannstr. 3, 85748 Garching bei München, Germany

ABSTRACT

Motivation: We tackle the problem of finding regularities in microarray data. Various data mining tools, such as clustering, classification, Bayesian networks and association rules, have been applied so far to gain insight into gene-expression data. Association rule mining techniques used so far work on discretizations of the data and cannot account for cumulative effects. In this paper, we investigate the use of quantitative association rules that can operate directly on numeric data and represent cumulative effects of variables. Technically speaking, this type of quantitative association rules based on half-spaces can find non-axis-parallel regularities.

Results: We performed a variety of experiments testing the utility of quantitative association rules for microarray data. First of all, the results should be statistically significant and robust against fluctuations in the data. Next, the approach should be scalable in the number of variables, which is important for such high-dimensional data. Finally, the rules should make sense biologically and be sufficiently different from rules found in regular association rule mining working with discretizations. In all of these dimensions, the proposed approach performed satisfactorily. Therefore, quantitative association rules based on half-spaces should be considered as a tool for the analysis of microarray gene-expression data.

Availability: The code is available from the authors on request.

Contact: kramer@in.tum.de

1 INTRODUCTION

In recent years microarray technology has been gaining increasing popularity among molecular biologists. Today, it is an important tool to gain insights into many aspects of the cell's inner workings. One of its main advantages is, that it is—unlike many traditional methods in molecular biology—not limited to investigating 'one gene at a time'. Instead it allows to monitor the expression levels of many thousand genes at once. Thus, microarray experiments can better capture the 'big picture' and gain insight into the cell's gene interaction network. Also, it has been proven to be an important tool in medical research, e.g. to determine tumor subtypes or investigate the effects of chemical treatments on cells. As the technology matures, the size of typical microarray experiments increases from tens to hundreds of samples and thousands of genes. Analyzing such amounts of data is not a trivial undertaking.

Traditionally, researchers have been using clustering techniques to identify clusters of similarly expressed genes. Unfortunately, clustering gives mainly information about similar genes, not about the interactions between differently expressed genes. Current research in bioinformatics is, therefore, investigating methods that are able to model interactions between genes, such as Bayesian networks (Friedman *et al*., 2000) and association rules (Creighton and Hanash, 2003). To use association rules for gene-expression data, the expression values are usually discretized in a preprocessing step. High values are marked with a '↑' ('highly expressed'), low values with a '↓' ('highly repressed'). An association rule, such as '{gene F↑, gene D↓} → {gene B↓}', could then indicate, that gene B is very likely to be downregulated, whenever gene F is up and gene D is down. Of course, such a crude discretization leads to a certain loss of information.

Recently, new methods have been devised to deal with quantitative data directly without any need for discretization. For instance, Webb (2001) presented an approach that characterizes selected attributes using statistical moments. In previous work, we proposed a new approach based on half-spaces (Rückert *et al*., 2004b). In this paper, we investigate the use of quantitative association rules for analyzing gene-expression data. We present results from comprehensive experiments showing the potential of quantitative association rules for this type of data. Quantitative association rules based on half-spaced are rules of the form 'if the weighted sum of some variables is greater than a threshold, then, with high probability, a different weighted sum of variables is greater than a second threshold'. A typical example of such a rule is '$0.99\ \mathrm{ARG1} - 0.11\ \mathrm{CAR1} > 0.062 \rightarrow 1.00\ \mathrm{ARG3} > -0.032$', relating the expression levels of three genes in arginine metabolism. They are an interesting type of representation for the analysis of gene-expression data, because biochemical networks usually consist of main pathways as well as 'side roads' that can be used if the other ways are blocked. Weighted sums of variables are a suitable representation to model this kind of phenomenon in one rule. In contrast to most discrete association mining methods, algorithms for quantitative association rules based on half-spaces are not able to exhaustively enumerate all relevant association rules. Instead the presented framework uses an optimization approach. The aim is to find rules that are locally optimal with respect to a parameterized score function. Consequently, the user can adjust the parameters of the algorithm to obtain association rules that match his/her individual interests. For instance, it is possible to specify target values for certain parameters, such that the algorithm attempts to find rules near the target (penalizing rules that are too far off), while simultaneously optimizing the rules' confidence.

The paper is organized as follows: Sections 2 and 3 review the representation of quantitative association rules based on half-spaces and the optimization setting for finding such rules. In Section 4, we apply the approach to several gene-expression datasets to study the

*To whom correspondence should be addressed.

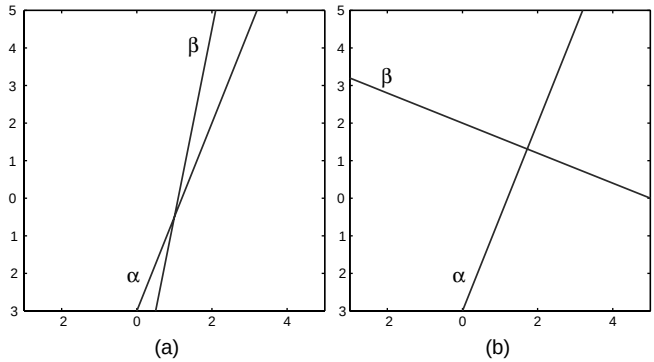

Fig. 1. Two non-perpendicular hyperplanes α and β (**a**), and two perpendicular hyperplanes α and β (**b**).

statistical significance of the rules as well as the algorithm's robustness, scalability and practicability. Finally we conclude in Section 5.

2 QUANTITATIVE ASSOCIATION RULES BASED ON HALF-SPACES

In the traditional association rule setting, the conditions on the right-hand side and left-hand side are based on hyperrectangles of discrete attributes. To extend association rules to continuous data, we therefore need to decide which kind of conditions the quantitative association rules should be based on. For numerical data, it makes sense to select a smooth separation function that minimizes the error caused by random noise or measurement errors in the data. A natural choice are hyperplanes, a particularly simple but powerful class of separation functions. From a geometrical perspective, a hyperplane α is given by a vector $\bar{\alpha}$ and an intercept α_0. An instance x is then assigned to one half-space, if the dot product $\bar{\alpha} \cdot x + \alpha_0$ is positive and to the other half-space, if it is negative. In Figure 1b, the 1D hyperplane α (i.e. a line) separates the 2D space into two half-spaces, one left of α, the other right of α.

In the case of association rules, the use of hyperplanes as conditions boils down to testing a weighted sum of variables against a threshold; i.e. an instance x in an n-dimensional space meets the condition $\alpha \in \mathbb{R}^{n+1}$, if $\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n \geq -\alpha_0$. With this, one could build an association rule such as $x_1 \geq 31 \rightarrow 0.9x_5 + 1.2x_6 \geq 250$. In a particular medical application this association rule might be interpreted as 'if the body mass index is $\geq$31, then the weighted sum of the systolic and diastolic blood pressure is $\geq$250'.

Of course, it is quite easy to generate a large number of uninteresting association rules with high confidence. In particular, situations like the one in Figure 1a are problematic: we have two hyperplanes α and β in a 2D space, that define an association rule $\alpha_1 x_1 + \alpha_2 x_2 \geq -\alpha_0 \rightarrow \beta_1 x_1 + \beta_2 x_2 \geq -\beta_0$. The problem is that α and β are highly correlated. If an instance is left of the α hyperplane, it is very likely to be left of the β hyperplane as well, simply because the space that is right of β, but left of α is much smaller than the space left of α and left of β. For our purposes it is, therefore, essential that α and β are uncorrelated, i.e. they have to be perpendicular as in Figure 1b.

It is remarkable that quantitative association rules based on half-spaces can be viewed as a direct generalization of classical Boolean association rules and disjunctive association rules (Nanavati *et al*., 2001): any conjunction or disjunction of items can be represented by a linear inequality. Another nice property of quantitative association rules is that we can visualize the distances between the instances and the two hyperplanes in a scatter plot. For example, Figure 2 suggests, that given the right-hand side of the hyperplane α, it is indeed the case that almost all of the cases are lying above the second hyperplane β. Such information can be valuable for assessing the relevance and usefulness of the pattern at hand.

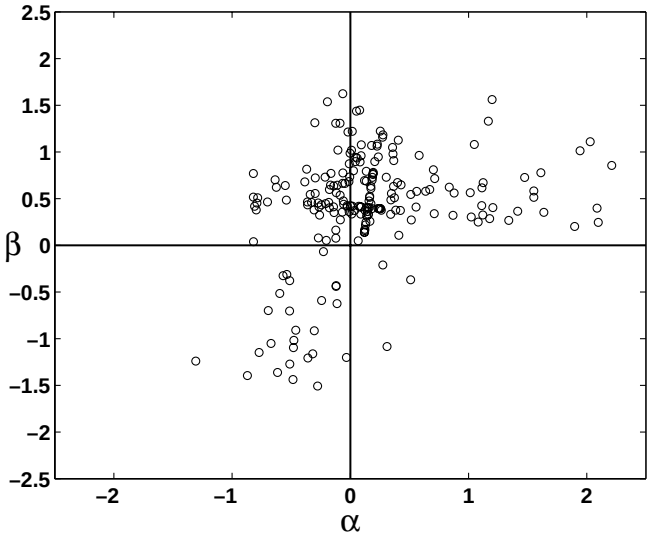

Fig. 2. Scatter plot of the data instance's distance to the α hyperplane (x-axis) and the β hyperplane (y-axis).

3 QUANTITATIVE ASSOCIATION RULE MINING

The main problem with finding good quantitative association rules is that the space of rules is uncountably infinite and, therefore, not suited to an enumeration strategy. In particular, the downward closure property does not hold for such rules, and thus we have to abandon the idea of generating the complete set of solutions. However, we can adopt an optimization approach, where the user can specify clearly the sort of rules he/she is looking for, and the algorithm returns locally optimal solutions. Although this may seem unusual for association rule mining, it is common practice in other areas, for instance, clustering (e.g. K-means clustering) and Bayesian learning (e.g. the EM algorithm). In the following we describe one particular algorithm for mining quantitative association rules in this setting. First, we define a score function to assess the 'interestingness' of an association rule. Then, we use standard techniques from numerical optimization theory to find association rules with a low score. Owing to space constraints we are not able to give the exact scoring function and the algorithm here and have to refer to Rückert *et al*. (2004a) for details.

For mining quantitative association rules we are given a dataset $X \subset \mathbb{R}^n$. We are now looking for association rules that are defined by two hyperplanes $\alpha, \beta \in \mathbb{R}^{n+1}$. The α hyperplane specifies the condition on the left-hand side of the association rule, the β hyperplane specifies the right-hand side. First of all, we are mainly interested in association rules with high confidence, i.e. the fraction of instances in X, that fulfill both conditions α and β divided by the fraction of

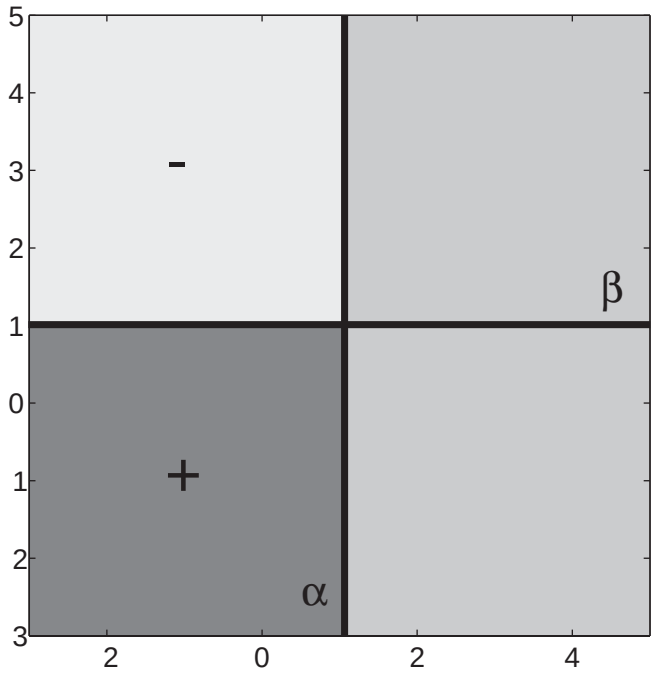

Fig. 3. Confidence is optimal if the distribution is uneven left of α, whereas contrast is optimal if it is even right of α.

instances that fulfill only the α condition should be as high as possible. Figure 3 illustrates this idea: if an instance x is located left of α and below β, it contributes to a high confidence score. If it is located left of α, but above β, it decreases the confidence measure. As outlined in Rückert *et al*. (2004a), one can construct a continuous function $l(\alpha, \beta, X)$ that is minimal for those high confidence settings.

A second criterion for the interestingness of an association rule is its coverage. The coverage is simply the fraction of instances in the dataset that satisfy the left-hand side condition. Unfortunately, the coverage of interesting rules is not clear a priori. In practice, it is often determined empirically. Similar to the confidence score, we design a function $c(\alpha, g, t, X)$ that is minimal if the coverage of α is close to a user-specified target value t. The g parameter determines how coverage should be weighted relative to confidence.

The confidence and coverage scores regulate what the optimization algorithm is looking for on the left side of α in Figure 3. For quantitative association rule mining, this is not enough: one can simply move the β hyperplane upward until it is located above all instances. Although this achieves maximal confidence, the resulting association rule is not very interesting. To overcome this problem we introduce the new criterion 'contrast'. The rationale is that the contrast between the instances above and below β should be as low as possible on the right side of α in Figure 3. Again, we formalize this criterion using a function $r(\alpha, \beta, X)$, that is minimal for low contrast settings.

To support the interpretability by human experts, we included one last criterion: the components of the α and β vectors usually contain n addends on both sides of the implication. However, most users prefer finding sparse association rules, i.e. rules where most coefficients are zero and only the relevant coefficients are given. To account for these pragmatic considerations, one can add a term $a(\alpha, h)$ to penalize non-sparse association rules. The user can adjust the importance of sparseness relative to the other scores using the parameter h. The final interestingness scoring function simply calculates the sum of scores given above.

A high score indicates that the association rule is uninteresting with regard to the selected parameter settings, a low score means we found an interesting rule. As the scoring function is continuous, there usually is a whole subspace of 'good' rules and it is easy to modify a rule with a low score to some small extent and obtain a rule with an even lower score. We are, therefore, aiming at finding association rules with optimally low score, i.e. the local optima of the scoring function, subject to the constraint that $\bar{\alpha}^{\mathrm{T}}\bar{\beta} = 0$. This constrained optimization problem can be tackled using established methods from optimization theory [see Rückert *et al*. (2004a) for details]. We take an approach that alternatingly keeps α fixed while optimizing β and vice versa.

As any other optimization procedure, this algorithm can get stuck in local optima with comparably high scores. For the sake of simplicity we use random restarts to obtain association rules with low score. A high sparseness parameter leads to rules that have a few large and many small (but non-zero) coefficients. A post-processing step is thus required to set the small coefficients to zero while retaining the perpendicularity of α and β. During this step all coefficients below a certain threshold are set to zero and a last optimization step is performed on the non-zero components of α and β.

4 EXPERIMENTAL RESULTS

To assess the applicability and feasibility of the described algorithm, we implemented a version in MATLAB. In the following we give a short biological interpretation of some rules we found in two microarray datasets. We also describe a range of experiments designed to assess reliability, robustness and scalability of the implementation on synthetic and real world data. Finally, we compare quantitative rules with their discrete counterparts generated by a traditional itemset mining algorithm.

4.1 Biological interpretation

For our first experiment, we chose the gene-expression dataset of Hughes *et al*. (2000). The dataset contains the expression levels of 6316 genes in the yeast genome measured for 300 diverse mutations and chemical treatments of yeast cells. It has been investigated qualitatively by Creighton and Hanash (2003).

For the experiment we manually selected 23 genes known to be involved in arginine metabolism. Please note that this resembles very much the envisaged real-world usage of such rules, where a few genes are selected to elicit the regulation mechanism among them. Besides, it would be easy to adapt the approach to consider only those rules that satisfy user-specified constraints. We applied the quantitative association mining algorithm with a target coverage of 0.5, a coverage weight of 1.0, a sparseness parameter of 0 and a post-processing threshold of 0.3. The best five rules are depicted in Table 1. The first rule states that ARG1 and ARG4 are downregulated, if CPA1, CPA2 and ARG3 are repressed. CPA1, CPA2 and ARG3 are responsible for the formation of carbamoyl phospate and citrulline which are essential intermediates in the biosynthesis of arginine. ARG1 and ARG4 perform the remaining steps to produce arginine from its precursor citrulline. The rule seems to be sensible, because all enzymes work on the same single path and it is clearly a waste of energy and

Table 1. The five best rules found in the Hughes *et al*. (2000) dataset (focus on arginine metabolism)

Quantitative association rule	Score	Coverage	Support	Contrast	Confidence
-0.66 CPA2-0.57 CPA1-0.49 ARG3 $> -0.06 \rightarrow -0.75$ ARG1-0.66 ARG4 > -0.2	-149.1	0.49	0.49	0.51	1.00
-0.72 HOM2-0.69 ARG5, 6 $> -0.042 \rightarrow -0.62$ MET22-0.55 MET13-0.55 ARG1 > -0.16	-148.6	0.50	0.50	0.49	1.00
-0.79 ARG3-0.62 CPA2 $> -0.042 \rightarrow -0.62$ ARG1-0.57 MET22-0.55 MAK31 > -0.12	-148.6	0.50	0.50	0.51	1.00
-0.57 ARG5, 6-0.55 ARG3-0.46 CPA1-0.41 ARG1 $> -0.054 \rightarrow$ -0.56 MET13-0.52 MET22-0.47 ARG1 $+ 0.43$ CPA1 > -0.098	-146.4	0.50	0.50	0.50	1.00
-0.75 CPA2-0.66 CPA1 $> -0.059 \rightarrow -0.68$ ARG4-0.53 CAR2-0.51 ARG5, 6 > -0.2	-145.6	0.50	0.49	0.50	0.99

Table 2. The five best rules found in the dataset taken from the NCBI Gene-Expression Omnibus database

Quantitative association rule	Score	Coverage	Support	Contrast	Confidence
-0.98 STE3-0.18 SAG1-0.10 LEU2 $> 2 \rightarrow -0.98$ SAG1 $+ 0.18$ STE3 > 0.91	-44.3	0.49	0.49	0.52	1.00
-1.00 LEU2 $> 2.4 \rightarrow -0.17$ STE3 $+ 0.11$ HO $+ 0.11$ URA1 $+ 0.97$ OAC1 > 0.052	-43.8	0.51	0.51	0.52	1.00
-0.99 STE3-0.11 SAG1 $> 1.5 \rightarrow -0.99$ SAG1 $+ 0.11$ STE3 > 1	-43.3	0.50	0.49	0.53	0.98
-0.98 STE3-0.14 LEU2-0.13 YER124C $> 1.9 \rightarrow -1.00$ SAG1 > 1	-43.3	0.49	0.49	0.52	1.00
-1.00 STE3 $> 1.2 \rightarrow -1.00$ SAG1 > 1.1	-43.0	0.50	0.49	0.51	0.98

nutrients to form only parts of a biosynthesis pathway. Note also, that the coefficients in the rule are approximately of the same magnitude, indicating that the individual genes are equally relevant for the pathway. The second rule states, that MET22, MET13 and ARG1 are downregulated, if HOM2 and ARG5,6 are repressed. ARG5,6 and ARG1 are enzymes for arginine synthesis, whereas MET13, MET22 and HOM2 take part in methionine synthesis. HOM2 is also involved in an early synthesis step in the threonine synthesis. Apparently, there is no need to produce arginine, if the demand for methionine and threonine synthesis is also low.

For the second experiment, we selected microarray experiment GDS464 for *Saccharomyces cerevisiae*, which is available in the Gene-Expression Omnibus database at NCBI (2005, http://www.ncbi.nlm.nih.gov/geo/gds) and consists of 90 instances referring to 1191 genes. We selected the 50 genes with the highest variance[1] and generated a range of quantitative association rules setting h to 0.2 and the post-processing threshold to 0.1. The five best rules are given in Table 2. Most of the rules involve STE3 and SAG1, two membrane-bound proteins involved in conjugation and responsible for cell communication aggregation. The fifth rule basically states that STE3 is down whenever SAG1 is down as well. From a biological point of view, both are required at the same time and both are involved in the same process. In particular, both are specifically expressed in alpha-type cells, whereas they are repressed in a-type cells.

As shown in the section on statistical significance (Section 4.3), rules with the above scores should be regarded as highly statistically significant. In fact, the p-value would be <0.01, since none of the rules from the randomly permuted dataset have a score in this range. This gives us some confidence that the algorithm, in fact, finds interesting and interpretable structure in the data.

[1]Feature selection for reducing the dimensionality of gene-expression data is an interesting topic in its own right, but in a sense orthogonal to approaches for finding dependencies among variables [see for instance Scholz *et al*. (2004) for an interesting new approach]. Related approaches like discrete association rules (Creighton and Hanash, 2003) or Bayesian networks (Friedman *et al*., 2000) have to reduce the dimensionality in one way or the other.

4.2 Reliability

In this experiment we investigate whether or not the algorithm is indeed able to find known quantitative rules in synthetic data. Even if the data are known to contain significant structure, it might be the case that the optimization algorithm gets stuck in local optima prematurely and never reaches the optimum representing the known structure. In order to investigate this question, we prepare synthetic data in the following way: first, we generate several quantitative association rules by randomly selecting suitable values for α and β. Then we sample new instances according to the uniform distribution within a 10D hypercube. A new instance is added to the dataset only if it is consistent with the conditions imposed by all the rules. We then run the described algorithm with a target coverage of 0.5, a coverage weight of 1.0, a sparseness parameter of 0.25 and 0, respectively, and a post-processing threshold of 0.3 for the first try. We perform 1000 restarts per experiment.

As the density of the data points in the considered space is clearly too low to uniquely identify a quantitative association rule, we cannot expect to find rules with exactly matching values for α and β. Instead we need some way to assess the 'similarity' between two rules. Given the rules (α_1, β_1) and (α_2, β_2), a simple similarity measure is to determine the fraction of instances in the dataset that are located in the same quadrants, i.e. that are located on the same sides of the α_1 and α_2 hyperplanes as well as the β_1 and β_2 hyperplanes. Table 3 gives the similarity measures between the randomly generated rules in the synthetic data and the most similar rules found by the algorithm. In both experiments we were able to find reasonably good approximations to most of the rules hidden in the dataset. All the rules with a similarity measure $>90\%$ were also ranked within the top 15 rules with the highest score. All in all, there is clear empirical evidence that the proposed algorithm is able to find structural properties in synthetic data.

Table 3. Reliability: the similarity measure between the generated rules hidden in the synthetic datasets and the rules found by the algorithm for two different settings

No. of instances	h	No. of LHS attributes	No. of RHS attributes	Similarity measure (%)
100	0.25	1	1	97
		2	1	98
		1	2	99
		2	2	82
		2	2	87
300	0	10	10	91
		10	10	86
		10	10	70

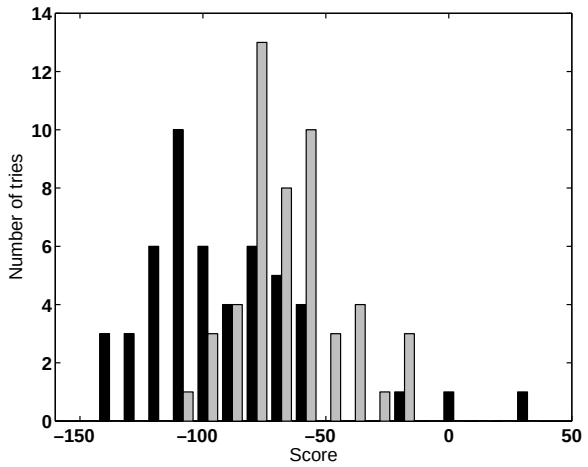

Fig. 4. Statistical significance: a histogram of the scores for the rules in the Hughes *et al*. dataset (black) and a permuted version of the same dataset.

4.3 Statistical significance

In the following paragraphs we describe experiments on three gene-expression datasets. In particular, we would like to investigate if the generated rules are indeed statistically significant. For the first experiment, we used the dataset of Hughes *et al*. (2000). We selected the 50 genes with the largest standard deviation for our experiment. The parameters were set as follows for all experiments in this section: $g = 1.0$, $t = 0.5$ and the sparseness parameter h was set to 0.5. To assess the robustness and statistical significance of the rules, we performed a randomization test, where the values of the columns are permuted randomly to generate a new dataset with the same distribution but no structural relations among the columns. We ran 50 tries on the original data, each try consisting of 10 restarts. For each try we kept the rule with the best score. Then, we repeated the same process on the permuted dataset. This yields the distribution of scores that can be expected on random, but similar data. Figure 4 gives the resulting histograms for the original and the permuted data. The scores for the permuted data are peaked around -70, whereas the original data feature a large number of association rules in the range from -140 to -60. Thus, we can be highly confident that the best induced rules describe indeed structural properties of the yeast dataset.

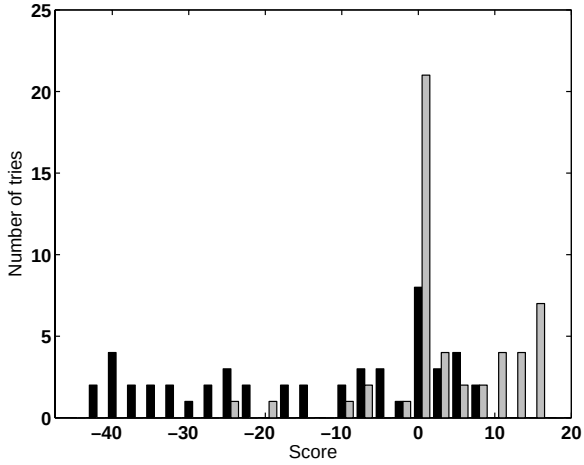

Fig. 5. Statistical significance: a histogram of the scores for the rules in the GDS464 dataset (black) and a permuted version of the same data.

Figure 5 gives the histogram for the second dataset taken from the Gene-Expression Omnibus database at NCBI (2005), selecting again the 50 genes with the largest standard deviation and using the same parameters as above. The third dataset (Sese *et al*., 2004)[2] is obtained by a different technique for gene-expression measurements called ATAC–PCR (Kato, 1997), which yields values with a much higher precision; after a preprocessing step the set contains 213 instances for a total of 1993 genes, originating from three types of human cells, namely tumor cells from liver cancer patients, non-tumors tissue cells from liver cancer patients and normal liver cells. In both cases the best rules on the original data have significantly lower scores than the rules on the permuted data. As the score function depends on the number of instances, we adapted the bin widths in the histograms; the best possible scores are about -150, -45 and -107 for the three datasets.

4.4 Robustness

The algorithm uses random restarts and gradient descent to sample a collection of rules as represented by the local minima of the scoring function. Its performance, therefore, depends to a large degree on the number and the shape of the local optima: if there are only a few 'deep' local optima, we can expect to repeatedly find relevant rules. If there are many 'shallow' optima with comparatively bad scores, the algorithm might come up with plenty of different mediocre rules, but only a few suited start configurations will lead to rules with globally optimal score. To investigate this issue further, we performed some experiments to assess how reliably the algorithm is able to induce rules with good scores on real world data.

The first experiment was performed on the 12 genes with the highest standard deviation from the Hughes *et al*. (2000) dataset. As before, the selection of genes only serves to illustrate our point. Experiments on a larger scale would be conceivable. We ran 10 tries with 500 restarts each, setting sparseness parameter h to 0.2 and the post-processing threshold to 0.2. For each try, we stored the 30 rules with the best scores after post-processing for sparseness. As a first result, we found that all tries contained a large fraction of rules that

[2]The data are not shown owing to lack of space.

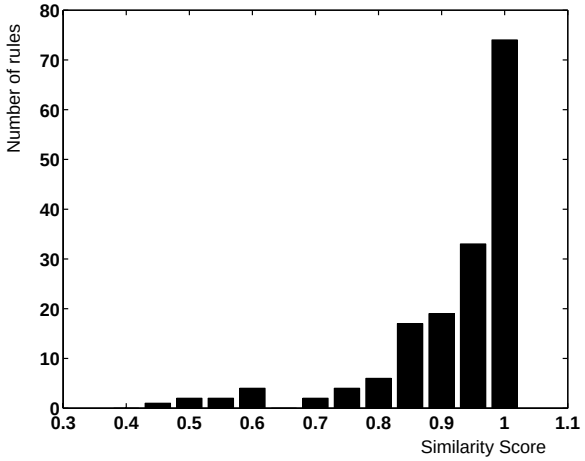

Fig. 6. Robustness: a histogram of the similarity measure for the rules and their most similar counterparts in the other tries.

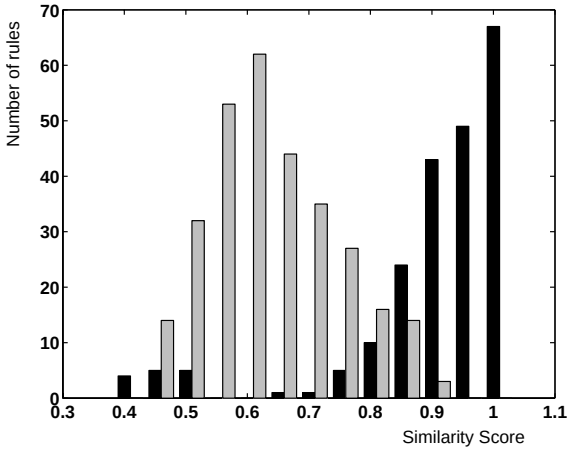

Fig. 7. Robustness: a histogram of the similarity measure between the rules and their most similar counterparts determined from a slightly modified dataset on the original data (black bars) and a permuted version (grey bars).

were slight variations of each other. It appears that the regions of the search space with comparatively good scores still contain a range of local optima. This is, of course, not too surprising given the fact that the dataset is a way too small to identify a rule uniquely. Consequently, most underlying quantitative regularities can be expressed by more than one rule. From the user's perspective, this is not a big problem: variations of the same rule can be easily identified. Much more important is the question whether the algorithm is able to find all of the dataset's clearly distinct rules in each try.

To answer this question, we use the same similarity measure as in Section 4.2. We consider a rule as redundant, if it has a similarity measure of 1.0 to another rule with a better score in the same try. Removing these redundant rules leaves 164 rules in 10 tries, each try containing only distinct rules. The main question is now: does each try contain a representative sample of distinct rules? If this is the case, each rule in each try should have at least one similar rule in every other try. Thus, we determine for each rule in each try the nine most similar rules in the other tries. Then, we take the average over the nine corresponding similarity measures. Figure 6 gives the histogram of the distribution of those averages. As can be seen, over two-thirds of the rules have equivalents in the other tries, whose average similarity is >0.9. It seems that the rules in every try are comparatively similar to each other and that each try gives a representative collection of the best rules in the dataset. Thus, the algorithm is robust in the sense that its output does not vary too much for different tries.

Although this result indicates that there is a certain robustness with respect to repeated runs, it is not clear whether the algorithm is also robust with regard to slight fluctuations in the dataset. After all, the algorithm might come up with completely different rules when running on slightly modified data or new data taken from the same source. To investigate this issue, we repeated the outlined experiment with a slight modification. We partitioned the dataset's instances randomly into 10 folds. We again ran 10 tries, but this time we gave only 9 of the 10 folds as input to the algorithm and set the tenth fold aside. As in the experiment above, we removed the redundant rules from each try and calculated for each rule in each try, the average similarity to the most similar rules in the other tries (we used all instances from all folds for the similarity calculation). The resulting histogram is given by the black bars in Figure 7. Apparently, the rule similarity between the 10 tries is still rather high, even though the individual runs used data that differed in about 11% of the instances. The grey bars in Figure 7 give the histogram of the same experiment performed on the dataset, after the columns were permuted to generate a new dataset with the same distribution, but no structural relations between the columns. This distribution is peaked at 0.6 rather than at 1.0, giving some empirical evidence that the algorithm is indeed robust even to random fluctuations in the dataset.

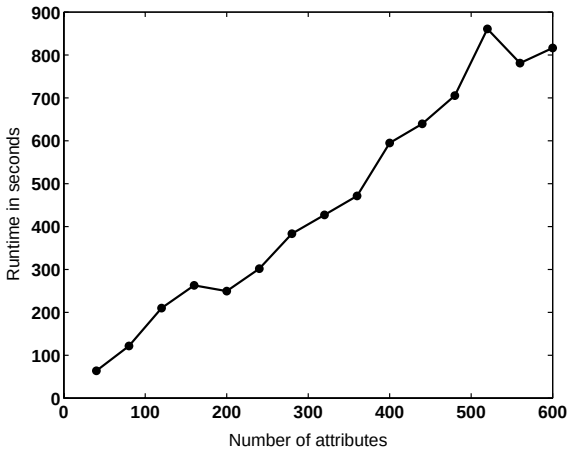

Fig. 8. Scalability: the run-time of the algorithm plotted against the number of the attributes in the dataset.

4.5 Scalability

Since microarray data often contain a large number of genes, another important question is whether the algorithm also scales well with the number of the attributes in the dataset. The following experiment investigates this issue. We ordered the genes in the Hughes *et al.* (2000) dataset by decreasing standard deviation. Then we generated 15 datasets containing between 40 and 600 attributes with increasing standard deviation in the steps of 40 attributes. We ran the algorithm with 50 restarts for each dataset, setting the parameters to the same values as in Section 4.3. In Figure 8 we plot the total run-time for

each dataset against its size. The run-time seems to scale nicely with the number of attributes. Even for a large number of attributes the run-time remains within a reasonable range.

4.6 Comparison with interval-based association rules

The strength of the presented algorithm lies in the fact that it can be applied on the unmodified continuous data, as an a priori discretization always goes along with loss of information. Moreover, quantitative association rules reveal non-axis-parallel and cumulative patterns, whereas discrete rules are not appropriate to discover such relationships. However, it might turn out that the most significant regularities in the microarray data are axis-parallel and non-sensitive to discretization. In that case, the computational overhead of using quantitative association rules might not be worth the gain in expressiveness. In the following, we compare the quantitative rules with the ones generated by a conventional algorithm.

We chose the 12 attributes with the highest standard deviation from the Hughes *et al*. (2000) dataset. For the traditional itemset approach, we used the same discretization step as in Creighton and Hanash (2003): an expression value >0.2 for the log base 10 of the fold change was binned as being up; a value <-0.2 as being down; and a value between -0.2 and 0.2 as being neither up nor down. We then applied the Magnum Opus itemset mining algorithm (G.I. Webb and associates 2005, http://www.giwebb.com/demoeula.html) to generate itemset rules with one attribute on the right side each. Magnum Opus includes a filter to remove insignificant rules. We set this threshold to 0.1. This yields 52 interval rules for a minimum coverage of 0.35 and a minimum confidence of 0.85. For the quantitative association rule algorithm we performed 1000 random restarts and set the parameters g and t as in the previous experiments, h to 0.1 and the post-processing threshold to 0.2. We stored the best 50 rules. Eliminating rules with similarity measure >0.9 to a better rule leaves us with 49 sufficiently distinct quantitative association rules.

Analyzing the resulting rules we could find only three discrete rules having a quantitative counterpart using the same attributes and seven quantitative rules with discrete counterparts. This indicates that discrete and quantitative rule mining algorithms seem to find mostly distinct patterns. We also took a further look at the rules that did use the same attributes. It turned out that for six of the seven quantitative rules, 75% of the instances fulfilling both conditions also fulfill both conditions of the corresponding discrete rule. So, even though the two approaches tend to find different patterns, there are some regularities that are detected by both methods.

5 CONCLUSION

As more and more microarray datasets become available, the need for advanced analysis tools for such data becomes ever more pressing. One important class of tools from the data mining community discovers associations between variables. Association rule mining is particularly appealing because it might help to shed some light on regulation mechanisms and also on the role of genes of yet unknown function. So far, only discrete variants of association rule mining have been applied to microarray data, requiring a discretization step in order to transform the numerical measurements into 0/1 representation. This obviously might lead to a loss of information that cannot be compensated in subsequent steps. Also note that association rule mining in 0/1 data usually cannot account for cumulative effects of variables. The items are just conjunctively connected, i.e. they all have to be present to make the left-hand side or the right-hand side of a rule true.

In this paper, we investigated the utility of quantitative assocation rules for the analysis of microarray data. Quantitative association rules do not require a discretization before the actual data mining step. We chose a recently developed variant of quantitative association rules that is based on half-spaces or linear combinations of variables thresholded against a constant. These half-space rules can account for cumulative effects of variables and discover regularities in the data that are not axis-parallel. To make quantitative association rules based on half-spaces a useful tool for analyzing micorarray data, they have to be sufficiently resistant against noise. Therefore, we included a variety of sanity checks in this paper. First we tested the reliability of the results in a 'hide-and-seek' manner using synthetic data. Next, we determined the statistical significance of the results and their reproducibility. Subsequently, we evaluated the scalability of the approach in the number of variables. Finally, a comparison with discretization-based association rules working on itemsets was performed. The experiments showed that the presented approach is a feasible alternative to regular association rule mining. Moreover, a biologist's interpretation of sample rules shows that quantitative association are, in fact, comprehensible patterns that help to gain insight into regularities that are hidden in microarray data. In conclusion, this paper showed that quantitative association rules, if sufficiently robust und noise-resistant, could be a valuable tool for the analysis of gene-expression data.

Conflict of Interest: none declared.

REFERENCES

Creighton,C. and Hanash,S. (2003) Mining gene expression databases for association rules. *Bioinformatics*, **19**, 79–86.

Friedman,N. *et al*. (2000) Using Bayesian networks to analyze expression data. *J. Comp. Biol.*, **7**, 601–620.

Hughes,T.R. *et al*. (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109–126.

Kato,K. (1997) Adaptor-tagged competitive PCR: a novel method for measuring relative gene expression. *Nucleic Acids Res.*, **25**, 4694–4696.

Nanavati,A.A., Chitrapura,K.P., Joshi,S. and Krishnapuram,R. (2001) Mining generalised disjunctive association rules. In Paques,H., Liu,L. and Grossman,D. (eds), *Proceedings of the Tenth International Conference on Information and Knowledge Management.* ACM Press, New York, pp. 482–489.

NCBI (2005) January 2005.

Rückert,U., Richter,L. and Kramer,S. (2004a). Quantitative association rules based on half-spaces: an optimization approach. *Technical Report TUM-I0412*, Institut für Informatik, TU München.

Rückert,U., Richter,L. and Kramer,S. (2004b) Quantitative association rules based on half-spaces: an optimization approach. In Morik,K., Rastogi,R. and Bramer,M. (eds), *Proceedings of the 4th International Conference on Data Mining.* IEEE Computer Society, Los Alamitos, pp. 507–510.

Scholz,M., Gibon,Y., Stitt,M. and Selbig,J. (2004) Independent component analysis of starch deficient pgm mutants. In Giegerich,R. and Stoye,J. (eds), *Proceedings of the German Conference on Bioinformatics 2004.* Gesellschaft für Informatik, Bonn, pp. 95–104.

Sese,J. *et al*. (2004) Constrained clusters of gene expression profiles with pathological features. *Bioinformatics*, **20**, 3137–3145.

Webb,G.I. (2001) Discovering associations with numeric variables. In *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM Press, New York, pp. 383–388.