Research Paper ■

Application of closed itemset mining for class labeled data in functional genomics

Petra Kralj, Ana Rotter, Nataša Toplak, Kristina Gruden, Nada Lavrač, Gemma C. Garriga

Abstract. This paper applies a recently introduced methodology of closed itemset mining for class labeled data to potato microarray data. The study shows the discovered rules that best distinguish between virus resistant and virus sensitive transgenic potato lines. The discovered rules are interpretable and meaningful to domain experts.

■ **Infor Med Slov**: 2006; 11(1): 40-45

Authors' institutions: Institut Jožef Stefan, Ljubljana, Slovenia (PK, NL), National Institute of Biology, Ljubljana, Slovenia (AR, NT,KG), Nova Gorica Polytechnic, Nova Gorica, Slovenia (NL), Universitat Politecnica de Catalunya, Barcelona, Spain (GCG).

Contact person: Petra Kralj, Department of Knowledge Technologies, Jozef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia. email: petra.kralj@gmail.com.

Introduction

Microarray technology offers researchers the ability to simultaneously examine expression levels of hundreds or thousands of genes in a single experiment. Knowledge about gene regulation and expression can be gained by dividing samples into control samples, in our case mock infected plants, and treatment samples, in our case virus infected plants. Studying the differences between gene expression of the two groups (control and treatment) can provide useful insights into complex patterns of host relationships between plants and pathogens. [1]

Since the dimensions of microarrays are typically very large, statistical and data mining methods have to be used in order to draw significant conclusions from the data. Careful data preprocessing has to be done before using statistics or data mining. Data preprocessing includes, but is not limited to, filtering of data, leaving out low intensity signals or high background (noisy) signals. At later preprocessing stages, irrelevancy filtering[2] can also be used.

The task of data mining on microarray data differs from traditional data mining tasks because microarray domains are characterized by very large numbers of attributes (genes) relative to the number of examples (observations, samples). Standard classification rule learning algorithms do not perform well on microarray data because of this dimensionality problem.

This work applies a recently developed approach named RelSets[3] to microarray data. The approach[3] uses closed itemset mining to detect relevant rules of the form

IF Conditions THEN Class

from class labeled data. First, all frequent closed itemsets are found on data instances labeled as positive. In the second phase, itemsets that would form irrelevant rules are removed. Only relevant rules are returned.

In our study, we aimed to find differences between two classes of resistance in four transgenic potato lines. For this purpose, 48 potato samples were used leading to 24 microarrays. We applied the algorithm RelSets[3] on these microarray data. The resulting rules are meaningful to biology experts.

The paper is organized as follows. First, the algorithm for mining closed itemsets from class labeled data is reviewed. Next, the biological experiment is outlined with the description of the data preparation steps. The data mining task is then outlined, followed by data mining results, their interpretation and conclusions.

Methodological background

The closed set for class labeled data technique[3] used in our experiment is based on the theory of closed itemset mining[4], upgraded by the recently developed theory of relevancy. [2]

Closed itemsets

Searching for descriptions from data has been addressed in descriptive data mining, in particular association rule learning. [5] An innovative insight was provided by closure systems, [4] aimed at compacting the whole dataset into a reduced system of relevant sets of items that formally conveys the same information as the complete dataset.

Let E denote a set of training examples, described by a fixed set of features $F = \{f1, \ldots, fn\}$. Features are logical variables representing attribute-value pairs (called *items* in association rule learning). Each example e is represented as a tuple of features f from F with an associated class label.

From the point of view of data mining algorithms, closed itemsets are maximal sets among any other itemsets occurring in the same examples. Formally, let $supp(X)$ denote the number of examples where the itemset X is contained. Then: Set $X \subseteq F$ is said to be a *closed itemset* when there is no other

set $Y \subseteq F$ such that $X \subset Y$ and $supp(X) = supp(Y)$. [4]

Feature and rule relevancy

The rule induction problem can be viewed as a process of searching the space of concept descriptions. In our case, the space of descriptions to be searched is the space of itemsets (conjunctions of features) that form rule conditions. Some descriptions in this hypothesis space may turn out to be more relevant than others for characterizing and/or discriminating the target class. [6]

A rule is said to *cover* an example if the condition part of the rule is satisfied for that example. A rule *correctly covers* an example if the rule covers the example and the predicted class of the rule matches the class label of the example. The rule *incorrectly covers* an example if the rule covers an example and the class of the rule differs from the class label of the example.

Quality of rule *R* is measured by *rule coverage*, determined by two quantities: the number of correctly covered examples *TP(R)* (*True Positives*) and the number of incorrectly covered examples *FP(R)* (*False Positives*). Good rules correctly cover many examples (many true positives) and incorrectly cover as few examples as possible (few false positives).

The notion of relative *relevancy of features*[2] can be generalized to apply to rules. [3]

Feature *f1* is *relatively irrelevant* with respect to feature *f2* if $TP(f1) \subseteq TP(f2)$ and if $FP(f2) \subseteq FP(f1)$. A feature is *relatively irrelevant* if there exists another feature in the dataset compared to which it is irrelevant.

The definition of feature relevancy can be generalized to rule relevancy as follows. Rule *R1* is *relatively irrelevant* with respect to rule *R2* if $TP(R1) \subseteq TP(R2)$ and if $FP(R2) \subseteq FP(R1)$. A rule is *relatively irrelevant* if there exists another rule in the ruleset compared to which it is irrelevant. [3]

The RelSets algorithm

A closed itemset mining algorithm and a rule relevancy filter are used in the approach applied in this paper. In this section we briefly recall the algorithm for closed itemset mining for labeled data, called RelSets. [3]

The input to RelSets is the dataset and one parameter: the minimum true positive count *(minTP)*. This is a constraint that implies that only rules that cover at least *minTP* positive examples should be constructed.

The dataset is first divided into two parts depending on the class label of the examples: the positive examples *P* and the negative examples *N*.

Closed itemsets in the positive examples are mined with a minimum support constraint *(minTP)*. These closed sets can be directly interpreted as rules:

IF Closed set THEN Positive

These rules have high true positives count since they were built with a *minTP* constraint. The theory[3] proves that these are all the most specific rules that have the potential to be relevant. Nothing is yet known about the coverage of false positives.

In the second phase RelSets confronts the rules found in the first phase with the negative data. It removes relatively irrelevant rules on the negative data. A maximum false positives count constraint can be implied.

The RelSets algorithm returns all the relatively relevant rules which fulfill the minimum true positive count constraint. The algorithm is complete in the sense that it finds all the most specific rules satisfying the constraints. This is very appropriate for microarray data since not many

examples are available and the complete search of the space is very adequate.

Biological experiment

The goal of our experiment is to investigate differences between virus sensitive and resistant transgenic potato lines. Since potato cultivation is economically important worldwide, infection pathway research is motivated not only by scientists but also by the industry.

Four transgenic potato lines (two of them resistant and two of them sensitive to a viral infection) were tested. Plants from each transgenic line were divided into four groups: one half was infected with potato virus PVY^{NTN} and the other half was mock inoculated. One PVY^{NTN} inoculated group and one mock inoculated group of each transgenic line were harvested 8 hours and the rest 12 hours after the infection. Every experiment was repeated 3 times, thus yielding 48 samples. Each microarray was hybridized with a virus inoculated sample and mock inoculated sample from the same transgenic line, yielding to 24 microarray experiments (Figure 1). Depending on the individual microarray design, red or green labeling for different samples was used.

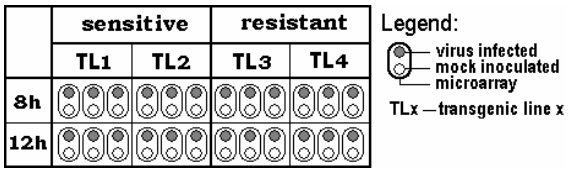

Figure 1 Schematic representation of the experiment. Each potato sample is represented by a circle: empty circles represent mock infected samples and full circles represent virus infected samples. An ellipse around two dots represents one microarray.

The dimensions of the initial data matrix after image scanning were 31200 x 24. As each gene is represented twice (as duplicate) on a microarray, the actual data dimensions are 15600 x 24. Data were filtered using the image analysis software ArrayPro Analyzer®. Spots that were unevenly distributed, had stained background (low signal-to-noise ratio) and low intensity signal on both channels (red and green) were left out of further analysis. After this first filtering, an average of 20,000 expression values per array remained for further analysis.

Two expression values for the same gene in a microarray were averaged. Second data filtering had two conditions: if 10 out of 24 microarray experiments for a given gene resulted in expression values in the interval (-0.3 , +0.3) or were missing values, the gene was discarded from further analysis. Both conditions for filtering were chosen arbitrarily to yield a suitable number of potentially regulated genes for further analysis. The dimensionality of data matrix was thus reduced to 6377 x 24.

Data mining task and results

The data mining task was to find differences in gene expression levels characteristic for virus sensitive potato transgenic lines, discriminating them from virus resistant potato transgenic lines and vice versa. For this purpose we used the RelSets[1] algorithm.

Our dataset contains 12 examples. Each example is a pair of microarrays (8 and 12 hours after infection) from the same transgenic line. All the data was discretized by using expert background knowledge. Features of the form $|$*gene expression value*$| > \pm 0.3$ were generated and enumerated.

Three groups of features were generated:

- the gene expression levels 8 hours after infection (feature numbers < 12493)
- the gene expression levels 12 hours after infection (feature numbers between 12494 and 24965)
- the difference between gene expression levels 12 and 8 hours after infection (feature numbers > 24966)

We ran our algorithm twice: once the sensitive examples were considered positive and once the resistant ones were considered positive. In both cases the constraint of minimal true positive count was set to 4, and in the first phase the algorithm returned 22 rules. The second part of the algorithm, which involves rule relevancy filtering, filtered the rules to just one relevant rule with true positive rate 100% and false positive rate of 0%. The results gained are shown below, where features are represented by numbers.

IF (13031 13066 19130 23462 24794 25509 29938 33795 33829 35003 35190 36266) THEN *sensitive* (TP=6)(FP=0)

Twelve features determined the potato sensitivity class for the samples used.

IF (16441 20474 20671 24030 25141 29777 30111 32459 33225 33248 33870 34108 34114 34388 37252 37484) THEN *resistant* (TP=6)(FP=0)

Sixteen features determined the potato *resistance* class for the samples used.

Biological interpretation

Based on the samples tested it seems that the response to the infection after 8 hours is not strong enough to distinguish between resistant transgenic lines and sensitive ones. None of the gene expression changes after 8 hours appeared significant for the data mining algorithm. However, gene expression levels after 12 hours and the comparison of gene expression difference (12-8) seem to determine the resistance to the infection with potato virus PVY^{NTN} for the transgenic lines tested.

According to the mechanism of virus plant interaction, genes that proved to be important for rule building, appearing in the output rules, have been classified into the following categories:

- *rec*: genes, whose products are responsible for sensing the infections by viral pathogen (receptors) and whose products are part of the cell membrane
- *sig*: genes, whose products are responsible for intracellular signaling transduction
- *TF*: genes, whose products are influencing transcription in cell nucleus
- *def*: genes, whose products are effectors for defense
- *hk*: housekeeping genes, whose expression was historically accepted as constant regardless of the physiological state of the plant
- *uf*: unknown function.

The distribution of the genes, appearing in class discriminating rules, determining whether a transgenic line is sensitive or resistant, can be viewed from Table 1. It can be argued that the downregulation of the first part of virus-plant interaction (reception and signaling) is an indicator of plant's sensitivity for infection, whereas upregulation of the second part (transcription and defense) of the interaction determines the resistance in plants tested.

Table 1 Functional distribution of genes, important to determine the sensitivity or resistance of a potato.

	sensitive	resistant
rec	1	0
sig	5	3
TF	2	4
def	7	6
hk	1	4
uf	1	3

The pattern can be visualized in Figure 2 and Figure 3. Housekeeping genes for which expression levels were argued to remain unchanged regardless of the treatment given to the plant seem to be important for determining the resistance of samples tested (Table 1).

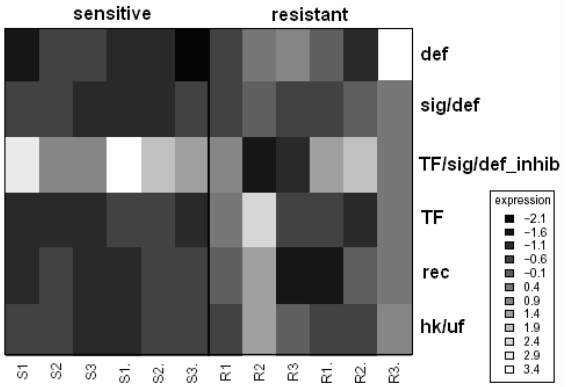

Figure 2 Heatmap for sensitive transgenic lines. Genes that have been found to be important for determining the sensitivity in samples tested were grouped into parts of plant-host interaction pathway: rec, sig, TF, def, hk, uf and their combinations if found. The heatmap shows that most of the genes of sensitive transgenic lines (marked with S, left side of the heatmap) were downregulated. Products of upregulated genes are inhibitors, important for signaling and defense pathways.

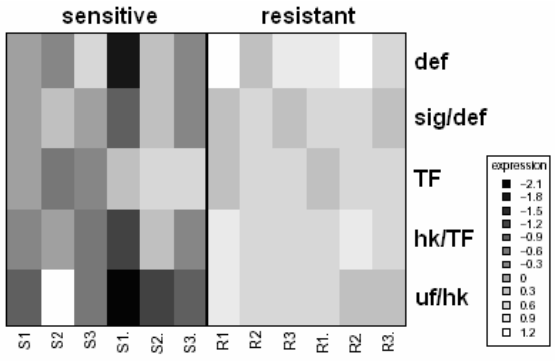

Figure 3 Heatmap for resistant transgenic lines. The heatmap shows that genes that have been found to be important of determining resistance in sample tested were upregulated (left side of the heatmap). Genes were grouped into parts of plant-host interaction pathway: rec, sig, TF, def, hk, uf and their combinations if found.

Conclusions

Using data mining on microarray data is a challenge because of the unusual dimensionality of the data specific for these domains. The recently developed method for closed itemset mining for labeled data shows no difficulties when applied to this kind of data. Furthermore, the results are in a form of comprehensible rules that are easy to be interpreted by domain experts. The approach was proven to perform well on microarray data, where the goal was to find differences between virus resistant and virus sensitive potato transgenic lines. The results proved to be meaningful to domain experts.

Literature

1. Taiz L , Zeiger E: Plant physiology, second edition 1998;(372:374). Sinauer Associates.
2. Lavrač N, Gamberger D: Relevancy in constraint-based subgroup discovery. In Boulicaut JF, De Raedt L, Mannila H (eds.) Constraint-Based Mining and Inductive databases 2006. In press.
3. Garriga GC, Kralj P, Lavrač N: Subgroup discovery by closed itemset mining from labeled data, Jozef Stefan Institute Technical Report, No. 9351, December 2005.
4. Carpineto C, Romano G: Concept data analysis. Theory and applications 2004. Wiley.
5. Agrawal R, Srikant R: Fast Algorithms for Mining Associa tion Rules. In Proceedings of the 20th International Conference on Very Large Databases 1994:207-216.
6. Gamberger D, Lavrač N, Železny F, et al.: Induction of comprehensible models for gene expression datasets by subgroup discovery methodology. Journal of biomedical informatics 2004 (37):269-284.