Subgroup Discovery Techniques and Applications

Nada Lavrač[1,2]

[1] Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
[2] Nova Gorica Polytechnic, Vipavska 13, 5000 Nova Gorica, Slovenia

Abstract. This paper presents the advances in subgroup discovery and the ways to use subgroup discovery to generate actionable knowledge for decision support. Actionable knowledge is explicit symbolic knowledge, typically presented in the form of rules, that allow the decision maker to recognize some important relations and to perform an appropriate action, such as planning a population screening campaign aimed at detecting individuals with high disease risk. Two case studies from medicine and functional genomics are used to present the lessons learned in solving problems requiring actionable knowledge generation for decision support.

1 Introduction

Rule learning is an important form of *predictive* machine learning, aimed at inducing classification and prediction rules from examples [2]. Developments in *descriptive induction* have recently also gained much attention of researchers interested in rule learning. These include mining of association rules [1], subgroup discovery [11, 4, 6] and other approaches to non-classificatory induction.

This paper discusses actionable knowledge generation by means of subgroup discovery. The term *actionability* is described in [10] as follows: "a pattern is interesting to the user if the user can *do something with it* to his or her advantage." As such, actionability is a subjective measure of interestingness.

The lessons in actionable knowledge generation, described in this paper, were learned from two applications that motivated our research in actionable knowledge generation for decision support. In an ideal case, the induced knowledge should enable the decision maker to perform an action to his or her advantage, for instance, by appropriately selecting individuals for population screening concerning high risk for coronary heart disease (CHD). Consider one rule from this application:

$$\text{CHD} \leftarrow \text{body mass index} > 25\ kgm^{-2}\ \&\ \text{age} > 63\ \text{years}$$

This rule is actionable as the general practitioner can select from his patients the overweight patients older than 63 years.

This paper provides arguments in favor of actionable knowledge generation through recently developed subgroup discovery approaches, where a subgroup

T.B. Ho, D. Cheung, and H. Liu (Eds.): PAKDD 2005, LNAI 3518, pp. 2–14, 2005.

discovery task is informally defined as follows [11, 4, 6]: Given a population of individuals and a specific property of individuals that we are interested in, find population subgroups that are statistically 'most interesting', e.g., are as large as possible and have the most unusual distributional characteristics with respect to the property of interest.

We restrict the subgroup discovery task to learning from class-labeled data, and induce individual rules (describing individual subgroups) from labeled training examples (labeled positive if the property of interest holds, and negative otherwise), thus targeting the process of subgroup discovery to uncovering properties of a selected target population of individuals with the given property of interest. Despite the fact that this form of rules suggests that standard supervised classification rule learning could be used for solving the task, the goal of subgroup discovery is to uncover individual rules/patterns, as opposed to the goal of standard supervised learning, aimed at discovering rulesets/models to be used as accurate classifiers of yet unlabeled instances [4].

In subgroup discovery, the induced patterns must be represented in explicit symbolic form and must be relatively simple in order to be recognized as actionable for guiding a decision maker in directing some targeted campaign. We provide arguments in favour of actionable knowledge generation through recently developed subgroup discovery algorithms, uncovering properties of individuals for actions like population screening and functional genomics data analysis. For such tasks, actionable rules are characterized by the experts' choice of the 'actionable' attributes to appear in induced subgroup descriptions, as well as by high coverage (support), high sensitivity and specificity[1], even if this can be achieved only at a price of lower classification accuracy, which is the quality to be optimized in classification and prediction tasks.

This paper is structured as follows. Two applications that have motivated our research in actionable knowledge generation are described in Section 2. Section 3 introduces the ROC and the TP/FP space needed for better understanding of the task and results of subgroup discovery. Section 6 introduces the functional genomics domain in more detail, where the task is to distinguish between different cancer types.

2 Two Case Studies

The motivation for this work comes from practical data mining problems in a medical and a functional genomics domain.

[1] *Sensitivity* measures the fraction of positive cases that are classified as positive, whereas *specificity* measures the fraction of negative cases classified as negative. If TP denotes true positives, TN true negatives, FP false positives, FN false negatives, Pos all positives, and Neg all negatives, then $Sensitivity = TPr = \frac{TP}{TP+FN} = \frac{TP}{Pos}$, and $Specificity = \frac{TN}{TN+FP} = \frac{TN}{Neg}$, and $FalseAlarm = FPr = 1 - Specificity = \frac{FP}{TN+FP} = \frac{FP}{Neg}$. Quality measures in association rule learning are *support* and *confidence*: $Support = \frac{TP}{Pos+Neg}$ and $Confidence = \frac{TP}{TP+FP}$.

The medical problem domain is first outlined: the problem of the detection and description of Coronary Heart Disease (CHD) risk groups [4]. Typical data collected in general screening include anamnestic information and physical examination results, laboratory tests, and ECG at-rest test results. In many cases with significantly pathological test values (especially, for example, left ventricular hypertrophy, increased LDL cholesterol, decreased HDL cholesterol, hypertension, and intolerance glucose) the decision is not difficult. However, the hard problem in CHD prevention is to find endangered individuals with slightly abnormal values of risk factors and in cases when combinations of different risk factors occur. The results in the form of risk group models should help general practitioners to recognize CHD and/or to detect the illness even before the first symptoms actually occur. Expert-guided subgroup discovery discovery is aimed at easier detection of important risk factors and risk groups in the population.

In functional genomics, gene expression monitoring by DNA microarrays (gene chips) provides an important source of information that can help in understanding many biological processes. The database we analyze consists of a set of gene expression measurements (examples), each corresponding to a large number of measured expression values of a predefined family of genes (attributes). Each measurement in the database was extracted from a tissue of a patient with a specific disease; this disease is the class for the given example. The domain, described in [9, 5] and used in our experiments, is a typical scientific discovery domain characterised by a large number of attributes compared to the number of available examples. As such, this domain is especially prone to overfitting, as it is a domain with 14 different cancer classes and only 144 training examples in total, where the examples are described by 16063 attributes presenting gene expression values. While the standard goal of machine learning is to start from the labeled examples and construct models/classifiers that can successfully classify new, previously unseen examples, our main goal is to uncover interesting patterns/rules that can help to better understand the dependencies between classes (diseases) and attributes (gene expressions values).

3 Background: The ROC and the TP/FP Space

A point in the ROC space (ROC: Receiver Operating Characteristic) [8] shows classifier performance in terms of false alarm or *false positive rate* $FPr = \frac{|FP|}{|TN|+|FP|} = \frac{|FP|}{|N|}$ (plotted on the X-axis), and sensitivity or *true positive rate* $TPr = \frac{|TP|}{|TP|+|FN|} = \frac{|TP|}{|P|}$ (plotted on the Y-axis).

A point (FPr, TPr) depicting rule R in the ROC space is determined by the covering properties of the rule. The ROC space is appropriate for measuring the success of subgroup discovery, since rules/subgroups whose TPr/FPr tradeoff is close to the diagonal can be discarded as insignificant; the reason is that the rules with TPr/FPr on the diagonal have the same distribution of covered positives and negatives as the distribution in the training set. Con-

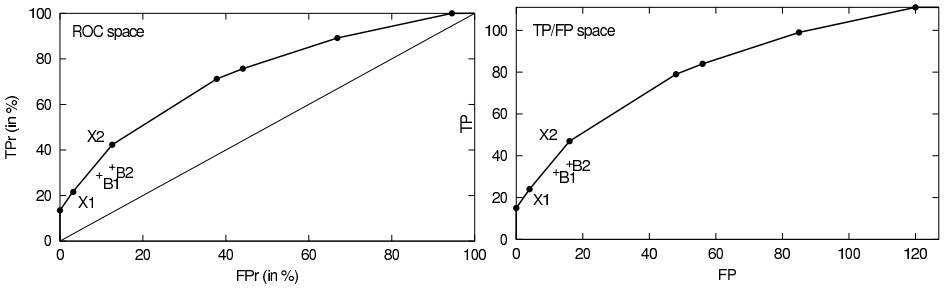

Fig. 1. The left-hand side figure shows the ROC space with a convex hull formed of seven rules that are optimal under varying TPr/FPr tradeoffs, and two suboptimal rules B1 and B2. The right-hand side presents the positions of the same rules in the corresponding TP/FP space

versely, significant rules/subgroups are those sufficiently distant from the diagonal. Subgroups that are optimal under varying TPr/FPr tradeoffs form a convex hull called the ROC curve. Figure 1 presents seven rules on the convex hull (marked by circles), including $X1$ and $X2$, while two rules $B1$ and $B2$ below the convex hull (marked by +) are of lower quality in terms of their TPr/FPr tradeoff.

It was shown in [6] that for rule R, the vertical distance from the (FPr, TPr) point to the ROC diagonal is proportional to the significance of the rule. Hence, the goal of a subgroup discovery algorithm is to find subgroups in the upper-left corner area of the ROC space, where the most significant rule would lie in point $(0, 1)$ representing a rule covering only positive and none of the negative examples ($FPr = 0$ and $TPr = 1$).

An alternative to the ROC space is the so-called TP/FP space (see the right-hand side of Figure 1), where FPr on the X-axis is replaced by $|FP|$ and TPr on the Y-axis by $|TP|$.[2] The TP/FP space is equivalent to the ROC space when comparing the quality of subgroups induced in a single domain. The reminder of this paper considers only this simpler TP/FP space representation.

4 Constraint-Based Subgroup Discovery

Subgroup discovery is a form of supervised inductive learning of subgroup descriptions of the target class. As in all inductive rule learning tasks, the language bias is determined by the syntactic restrictions of the pattern language and the vocabulary of terms in the language. In this work the hypothesis language is restricted to simple if-then rules of the form $Class \leftarrow Cond$, where $Class$ is the target class and $Cond$ is a conjunction of features. Features are logical condi-

[2] The TP/FP space can be turned into the ROC space by simply normalizing the TP and FP axes to the [0,1]x[0,1] scale.

tions that have values *true* or *false*, depending on the values of attributes which describe the examples in the problem domain: subgroup discovery rule learning is a form of two-class propositional inductive rule learning, where multi-class problems are solved through a series of two-class learning problems, so that each class is once selected as the target class while examples of all other classes are treated as non-target class examples.

This section briefly outlines a recently developed approach to subgroup discovery that can be applied to actionable knowledge generation.

4.1 Constraint-Based Subgroup Discovery with the SD Algorithm

In this paper, subgroup discovery is performed by SD, an iterative beam search rule learning algorithm [4]. The input to SD consists of a set of examples E and a set of features F that can be constructed for the given example set. The output of the SD algorithm is a set of rules with optimal covering properties on the given example set. The SD algorithm is implemented in the on-line Data Mining Server (DMS), publicly available at `http://dms.irb.hr`.[3]

In a constraint-based data mining framework [3], a formal definition of subgroup discovery involves a set of constraints that induced subgroup descriptions have to satisfy. The following constraints are used to formalize the SD constraint-based subgroup discovery task.

Language Constraints

- Individual subgroup descriptions have the form of rules $Class \leftarrow Cond$, where $Class$ is the property of interest (the target class), and $Cond$ is a conjunction of features (conditions based on attribute value pairs) defined by the language describing the training examples.
- For discrete (categorical) attributes, features have the form $Attribute = value$ or $Attribute \neq value$, for continuous (numerical) attributes they have the form $Attribute > value$ or $Attribute \leq value$. Note that features can have values *true* and *false* only, that every feature has its logical complement (for feature f_1 being $A_1 = v_1$ its logical complement $\overline{f_1}$ is $A_1 \neq v_1$, for $A_2 > v_2$ its logical complement is $A_2 \leq v_2$), and that features are different from binary valued attributes because for every attribute at least two different features are constructed.
- To simplify rule interpretation and increase rule actionability, subgroup discovery is aimed at finding short rules. This is formalized by a language constraint that every induced rule R has to satisfy: rule size (i.e., the number of features in $Cond$) has to be below a user-defined threshold: $size(R) \leq MaxRuleLength$.

[3] The publicly available Data Mining Server and its constituent subgroup discovery algorithm SD can be tested on user submitted domains with up to 250 examples and 50 attributes. The variant of the SD algorithm used in gene expression data analysis was not limited by these restrictions.

Evaluation/Optimization Constraints

- To ensure that induced subgroups are sufficiently large, each induced rule R must have high support, i.e., $sup(R) \geq MinSup$, where $MinSup$ is a user-defined threshold, and $sup(R)$ is the relative frequency of correctly covered examples of the target class in examples set E:

$$sup(R) = p(Class \cdot Cond) = \frac{n(Class \cdot Cond)}{|E|} = \frac{|TP|}{|E|}$$

- Other evaluation/optimization constraints have to ensure that the induced subgroups are highly significant (ensuring that the class distribution of examples covered by the subgroup description will be statistically significantly different from the distribution in the training set). This could be achieved in a straight-forward way by imposing a significance constraint on rules, e.g., by requiring that rule significance is above a user-defined threshold. Instead, in the SD subgroup discovery algorithm [4] the following rule quality measure assuring rule significance, implemented as a heuristic in rule construction, is used:

$$q_g(R) = \frac{|TP|}{|FP| + g} \tag{1}$$

  In Equation 1, TP are true positives (target class examples covered by rule R), FP are false positives (non-target class examples covered by rule R), and g is a user defined generalization parameter. High quality rules will cover many target class examples and a low number of non-target examples. The number of tolerated non-target class cases, relative to the number of covered target class cases, is determined by parameter g. It was shown in [4] that by using this optimization constraint (choose the rule with best $q_g(R)$ value in beam search of best rule conditions), rules with a significantly different distribution of covered positives and negatives, compared to the prior distribution in the training set, are induced.

5 Experiments in Patient CHD Risk Group Detection

Early detection of artherosclerotic coronary heart disease (CHD) is an important and difficult medical problem. CHD risk factors include artherosclerotic attributes, living habits, hemostatic factors, blood pressure, and metabolic factors. Their screening is performed in general practice by data collection in three different stages.

A Collecting anamnestic information and physical examination results, including risk factors like age, positive family history, weight, height, cigarette smoking, alcohol consumption, blood pressure, and previous heart and vascular diseases.

B Collecting results of laboratory tests, including information about risk factors like lipid profile, glucose tolerance, and thrombogenic factors.

C Collecting ECG at rest test results, including measurements of heart rate, left ventricular hypertrophy, ST segment depression, cardiac arrhythmias and conduction disturbances.

In this application, the goal was to construct at least one relevant and interesting CHD risk group for each of the stages A, B, and C, respectively.

A database with 238 patients representing typical medical practice in CHD diagnosis, collected at the Institute for Cardiovascular Prevention and Rehabilitation, Zagreb, Croatia, was used for subgroup discovery [4]. The database is in no respect a good epidemiological CHD database reflecting actual CHD occurrence in a general population, since about 50% of gathered patient records represent CHD patients. Nevertheless, the database is very valuable since it includes records of different types of the disease. Moreover, the included negative cases (patients who do not have CHD) are not randomly selected persons but individuals considered by general practitioners as potential CHD patients, and hence sent for further investigations to the Institute. This biased dataset is appropriate for CHD risk group discovery, but it is inappropriate for measuring the success of CHD risk detection and for subgroup performance estimation in general medical practice.

5.1 Results of Subgroup Discovery

The process of expert-guided subgroup discovery was performed as follows. For every data stage A, B and C, the SD algorithm was run for values g in the range 0.5 to 100 (values 0.5, 1, 2, 4, 6, ...), and a fixed number of selected output rules equal to 3. The rules induced in this iterative process were shown to the expert for selection and interpretation. The inspection of 15–20 rules for each data stage triggered further experiments, following the suggestions of the medical expert to limit the number of features in the rule body and avoid the generation of rules whose features would involve expensive and/or unreliable laboratory tests.

In the iterative process of rule generation and selection, the expert has selected five most interesting CHD risk groups. Table 1 shows the induced subgroup descriptions. The features appearing in the conditions of rules describing the subgroups are called the *principal factors*. The described iterative process was successful for data at stages B and C, but it turned out that medical history data on its own (stage A data) is not informative enough for inducing subgroups, i.e., it failed to fulfil the expert's subjective criteria of interestingness. Only after engineering the domain, by separating male and female patients, interesting subgroups $A1$ and $A2$ have actually been discovered.

Separately for each data stage A, B and C, we have investigated which of the induced rules are the best in terms of the TP/FP tradeoff, i.e., which of them are used to define the convex hull in the TP/FP space. At stage B, for instance, seven rules (marked by +) are on the convex hull of the TP/FP space shown in Figure 1. Notice that the expert-selected subgroups B1 and B2 are significant, but are not among those lying on the convex hull in Figure 1. The reason for selecting exactly those two rules at stage B are their simplicity (con-

Table 1. Induced subgroups in the form of rules. Rule conditions are conjunctions of principal factors. Subgroup A1 is for male patients, subgroup A2 for female patients, while subgroups B1, B2, and C1 are for both male and female patients. The subgroups are induced from different attribute subsets (A, B and C, respectively) with different g parameter values (14, 8, 10, 12 and 10, respectively)

	Expert Selected Subgroups
A1	CHD ← positive family history & age over 46 year
A2	CHD ← body mass index over 25 kgm^{-2} & age over 63 years
B1	CHD ← total cholesterol over 6.1 $mmolL^{-1}$ & age over 53 years & body mass index below 30 kgm^{-2}
B2	CHD ← total cholesterol over 5.6 $mmolL^{-1}$ & fibrinogen over 3.7 gL^{-1} & body mass index below 30 kgm^{-2}
C1	CHD ← left ventricular hypertrophy

sisting of three features only), their generality (covering relatively many positive cases) and the fact that the used features are, from the medical point of view, inexpensive laboratory tests. Additionally, rules B1 and B2 are interesting because of the feature *body mass index below 30* kgm^{-2}, which is intuitively in contradiction with the expert knowledge that both increased body weight as well as increased total cholesterol values are CHD risk factors. It is known that increased body weight typically results in increased total cholesterol values while subgroups B1 and B2 actually point out the importance of increased total cholesterol when it is not caused by obesity as a relevant disease risk factor.

5.2 Statistical Characterization of Subgroups

The next step in the proposed subgroup discovery process starts from the discovered subgroups. In this step, statistical differences in distributions are computed for two populations, the target and the reference population. The target population consists of true positive cases (CHD patients included into the analyzed subgroup), whereas the reference population are all available non-target class examples (all the healthy subjects). Statistical differences in distributions for all the descriptors (attributes) between these two populations are tested using the χ^2 test with 95% confidence level ($p = 0.05$).

To enable testing of statistical significance, numerical attributes have been partitioned in up to 30 intervals so that in every interval there are at least 5 instances. Among the attributes with significantly different value distributions there are always those that form the features describing the subgroups (the principal factors), but usually there are also other attributes with statistically significantly different value distributions. These attributes are called *supporting*

Table 2. Statistical characterizations of induced subgroup descriptions (supporting factors)

	Supporting Factors
A1	psychosocial stress, cigarette smoking, hypertension, overweight
A2	positive family history, hypertension, slightly increased LDL cholesterol, normal but decreased HDL cholesterol
B1	increased triglycerides value
B2	positive family history
C1	positive family history, hypertension, diabetes mellitus

attributes, and the features formed of their values that are characteristic for the discovered subgroups are called *supporting factors*.

Supporting factors are very important for subgroup descriptions to become more complete and acceptable for medical practice. Medical experts dislike long conjunctive rules which are difficult to interpret. On the other hand, they also dislike short rules providing insufficient supportive evidence. In this work, we found an appropriate tradeoff between rule simplicity and the amount of supportive evidence by enabling the expert to inspect all the statistically significant supporting factors, whereas the decision whether they indeed increase the user's confidence in the subgroup description is left to the expert. In the CHD application the expert has decided whether the proposed supporting factors are meaningful, interesting and actionable, how reliable they are and how easily they can be measured in practice. Table 2 lists the expert selected supporting factors.

6 Experiments in Functional Genomics

The gene expression domain, described in [9, 5] is a domain with 14 different cancer classes and 144 training examples in total. Eleven classes have 8 examples each, two classes have 16 examples and only one has 24 examples. The examples are described by 16063 attributes presenting gene expression values. In all the experiments we have used gene presence call values (A, P, and M) to describe the training examples. The domain can be downloaded from http://www-genome.wi.mit.edu/cgi-bin/cancer/datasets.cgi. There is also an independent test set with 54 examples. The standard goal of machine learning is to start from such labeled examples and construct classifiers that can successfully classify new, previously unseen examples. Such classifiers are important because they can be used for diagnostic purposes in medicine and because they can help to understand the dependencies between classes (diseases) and attributes (gene expressions values).

6.1 Choice of the Description Language of Features

Gene expression scanners measure signal intensity as continuous values which form an appropriate input for data analysis. The problem is that for continuous valued attributes there can be potentially many boundary values separating the classes, resulting in many different features for a single attribute. There is also a possibility to use presence call (signal specificity) values computed from measured signal intensity values by the Affymetrix GENECHIP software. The presence call has discrete values A (absent), P (present), and M (marginal). Subgroup discovery as well as filtering based on feature and rule relevancy are applicable both for signal intensity and/or the presence call attribute values. Typically, signal intensity values are used [7] because they impose less restrictions on the classifier construction process and because the results do not depend on the GENECHIP software presence call computation. For subgroup discovery we prefer the later approach based on presence call values. The reason is that features presented by conditions like $Gene = P$ is *true* (meaning that *Gene* is present, i.e., expressed) or $Gene = A$ is *true* (meaning that *Gene* is absent, i.e., not expressed) are very natural for human interpretation and that the approach can help in avoiding overfitting, as the feature space is very strongly restricted, especially if the marginal value M is encoded as value unknown.

In our approach, the M value is handled as an unknown value, as we do not want to increase the relevance of features generated from attributes with M values. The M values are therefore handled as unknown values as follows: unknown values in positive examples are replaced by value *false*, while unknown values in negative examples are replaced by value *true*. As for the other two values, A and P, it holds that two features for gene X, $X = A$ and $X \neq P$, are identical. Consequently, for every gene X there are only two distinct features $X = A$ and $X = P$.

6.2 The Experiments

The experiments were performed separately for each cancer class so that a two-class learning problem was formulated where the selected cancer class was the target class and the examples of all other classes formed non-target class examples. In this way the domain was transformed into 14 inductive learning problems, each with the total of 144 training examples and between 8 and 24 target class examples. For each of these tasks a complete procedure consisting of feature construction, elimination of irrelevant features, and induction of subgroup descriptions in the form of rules was repeated. Finally, using the SD subgroup discovery algorithm [4], for each class a single rule with maximal q_g value was selected, for $q_g = \frac{|TP|}{|FP|+g}$ being the heuristic of the SD algorithm and $g = 5$ the generalization parameter default value. The rules for all 14 tasks consisted of 2–4 features. The procedure was repeated for all 14 tasks with the same default parameter values. The induced rules were tested on the independent example set.

Table 3. Covering properties on the training and on the independent test set for rules induced for three classes with 16 and 24 examples. Sensitivity is $\frac{|TP|}{|P|}$, specificity is $\frac{|TN|}{|N|}$, while precision is defined as $\frac{|TP|}{|TP|+|FP|}$

Cancer	Training set			Test set		
	Sens.	Spec.	Prec.	Sens.	Spec.	Prec.
lymphoma	16/16	128/128	100%	5/6	48/48	100%
leukemia	23/24	120/120	100%	4/6	47/48	80%
CNS	16/16	128/128	100%	3/4	50/50	100%

There are very large differences among the results on the test sets for various classes (diseases) and the precision higher than 50% was obtained for only 5 out of 14 classes. There are only three classes (lymphoma, leukemia, and CNS) with more than 8 training cases and all of them are among those with high precision on the test set, while for only two out of eleven classes with 8 training cases (colorectal and mesothelioma) high precision was achieved. The classification properties of rules induced for classes with 16 and 24 target class examples (lymphoma, leukemia and CNS) are comparable to those reported in [9] (see Table 3), while the results on eight small example sets with 8 target examples were poor. An obvious conclusion is that the use of the subgroup discovery algorithm is not appropriate for problems with a very small number of examples because overfitting can not be avoided in spite of the heuristics used in the SD algorithm and the additional domain-specific techniques used to restrict the hypothesis search space. But for larger training sets the subgroup discovery methodology enabled effective construction of relevant rules.

6.3 Examples of Induced Rules

For three classes (lymphoma, leukemia, and CNS) with more than 8 training cases the following rules were induced by the constraint-based subgroup discovery approach involving relevancy filtering and handling of unknown values described in this chapter.

Lymphoma class:
(CD20_receptor EXPRESSED) AND
(phosphatidylinositol_3_kinase_regulatory_alpha_subunit NOT EXPRESSED)
Leukemia class:
(KIAA0128_gene EXPRESSED) AND
(prostaglandin_d2_synthase_gene NOT EXPRESSED)
CNS class:
(fetus_brain_mRNA_for_membrane_glycoprotein_M6 EXPRESSED) AND
(CRMP1_collapsin_response_mediator_protein_1 EXPRESSED)

The expert interpretation of the results yields several biological observations: two rules (for the lymphoma and leukemia classes) are judged as reassuring and one (the CNS class) has a plausible, albeit partially speculative explanation. Namely, the best-scoring rule for the lymphoma class in the multi-class cancer recognition problem contains a feature corresponding to a gene routinely used as a marker in diagnosis of lymphomas (CD20), while the other part of the conjunction (phosphatidylinositol, the PI3K gene) seems to be a plausible biological co-factor. The best-scoring rule for the leukemia class contains a gene whose relation to the disease is directly explicable (KIAA0128, Septin 6). Both M6 and CRMP1 appear to have multifunctional roles in shaping neuronal networks, and their function as survival (M6) and proliferation (CRMP1) signals may be relevant to growth promotion and CNS malignancy.

Both good prediction results on an independent test set as well as expert interpretation of induced rules prove the effectiveness of described methods for avoiding overfitting in scientific discovery tasks.

Acknowledgments

The paper describes joint work with Dragan Gamberger from Rudjer Bošković Institute, Zagreb, Croatia, supported by the Slovenian Ministry of Higher Education, Science and Technology.

References

1. R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A.I. Verkamo. Fast discovery of association rules. In U.M. Fayyad, G. Piatetski-Shapiro, P. Smyth and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, 307–328. AAAI Press, 1996.
2. P. Clark and T. Niblett. The CN2 induction algorithm. *Machine Learning*, 3(4):261–283, 1989.
3. R.J. Bayardo, editor. *Constraints in Data Mining. Special issue of* SIGKDD Explorations, 4(1), 2002.
4. D. Gamberger and N. Lavrač. Expert-guided subgroup discovery: Methodology and application. *Journal of Artificial Intelligence Research* 17: 501–527, 2002.
5. D. Gamberger, N. Lavrač, F. Železný, and J. Tolar. Induction of comprehensible models for gene expression datasets by the subgroup discovery methodology. *Journal of Biomedical Informatics* 37:269–284, 2004.
6. N. Lavrač, B. Kavšek, P. Flach and L. Todorovski. Subgroup discovery with CN2-SD. *Journal of Machine Learning Research*, 5: 153–188, 2004.
7. J. Li and L. Wong. Geography of differences between two classes of data. In *Proc. of 6th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD2002)*, Springer, 325–337, 2002.
8. F. Provost and T. Fawcett. Robust classification for imprecise environments. *Machine Learning*, 42(3): 203–231, 2001.
9. S. Ramaswamy et al. Multiclass cancer diagnosis using tumor gene expression signitures. In *Proc. Natl. Acad. Sci. USA*, 98(26): 15149–15154, 2001.

10. A. Silberschatz and A. Tuzhilin. On subjective measures of interestingness in knowledge discovery. In *Proc. First International Conference on Knowledge Discovery and Data Mining (KDD)*, 275–281, 1995.
11. S. Wrobel. An algorithm for multi-relational discovery of subgroups. In *Proceedings of the 1st European Symposium on Principles of Data Mining and Knowledge Discovery*, Springer, 78–87, 1997.