Research Paper ■

Subgroup discovery: An experiment in functional genomics

Nada Lavrač, Dragan Gamberger

Abstract. Functional genomics is a typical scientific discovery domain characterized by a very large number of attributes (genes) relative to the number of examples (observations). This work presents an approach to subgroup discovery in supervised inductive learning of short rules that are appropriate for human interpretation. The approach is based on the subgroup discovery rule learning framework, enhanced by methods of restricting the hypothesis search space by exploiting the relevancy of features that enter the rule construction process as well as their combinations that form the rules. A multi-class functional genomics problem of classifying fourteen cancer types based on more than 16000 gene expression values is used to illustrate the methodology.

■ **Infor Med Slov**: 2006; 11(1): 46-51

Authors' institutions: Jožef Stefan Institute, Ljubljana, Slovenia (NL), Rudjer Bošković Institute, Zagreb, Croatia (DG).

Contact person: Nada Lavrač, Department of Knowledge Technologies, Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia. email: nada.lavrac@ijs.si.

Introduction

Construction of understandable and explainable models is important for scientific discovery as well as for the generation of actionable knowledge. It is possible to extract the most informative features or attributes from complex classifiers (the attributes with this property are called disease markers) but logical connections among these features or attributes are missing. This disables the construction and expert interpretation of models describing the target class. In contrast, short rules, despite being potentially less accurate than the complex classifiers, are much more appropriate for scientific discovery tasks in which the interpretability of induced models is of ultimate importance.

Functional genomics is a typical scientific discovery domain characterized by a very large number of attributes (genes) relative to the number of examples (observations). The danger of data overfitting is crucial in such domains. This work presents an approach to subgroup discovery, complemented by an approach which can help in avoiding data overfitting in supervised inductive learning of short rules that are appropriate for human interpretation. The approach is based on the subgroup discovery rule learning framework, enhanced by methods of restricting the hypothesis search space by exploiting the relevancy of features that enter the rule construction process as well as their combinations that form the rules.

This paper presents an approach, based on the subgroup discovery rule learning framework, enhanced by a method for filtering of irrelevant features. The results of its application on a multi-class functional genomics problem, aimed at classifying fourteen cancer types based on more than 16000 gene expression values, illustrate the use of the proposed methodology.

Subgroup Discovery

Subgroup discovery is a form of supervised inductive learning of subgroup descriptions for the target class in a two class domain. The descriptions have the form of rules built as logical conjunctions of features. Features are logical conditions 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 therefore a form of two-class propositional inductive rule learning. Multi-class problems can be solved as 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.

Formally, the task of subgroup discovery is defined as follows: given a population of individuals and a specific property of the individuals that we are interested in, find population subgroups that are 'most interesting', e.g., are as large as possible and have the most unusual distributional characteristics with respect to the property of interest.

Standard classification rule learning algorithms can be adapted to perform subgroup discovery. For instance, subgroup discovery algorithms, CN2-SD[1] and Apriori-SD[2] are adaptations of classification rule learners: CN2-SD is an adaptation of CN2[3] and Apriori-SD is an adaptation of APRIORI-C[4] and APRIORI[5]. These algorithms take as input the training examples described by discrete attribute values.

Method

In this work, subgroup discovery is performed by the SD algorithm[6,7], implemented in the on-line Data Mining Server (DMS), publicly available at http://dms.irb.hr, a relatively simple iterative beam search rule learning algorithm.

The input to SD consists of a set of examples E (E=P U N, P is the set of target class examples, N the set of non-target class examples) and a set of features F constructed for the given example set. For discrete (categorical) attributes, features have the form Attribute = value, while for continuous (numerical) attributes they have the form Attribute > value or Attribute < value. The output of the SD algorithm is a set of rules with optimal covering properties on the given example set. As in classification rule learning, an induced rule (subgroup description) has the form of a (backwards) implication:

$$Class \leftarrow Cond.$$

In terms of rule learning, the property of interest for subgroup discovery is the target class (Class) that appears in the rule consequent, and the rule antecedent (Cond) is a conjunction of features (attribute-value pairs) selected from the features describing the training instances.

A rule with ideal covering properties is true for all target class examples and not true for all non-target class examples. Target class examples covered by a rule are also called true positives, TP, while non-target class examples covered by the rule are called false positives, FP. All remaining non-target class examples not covered by the rule are called true negatives, TN. An ideal rules has TP=P and TN=N. In the proposed subgroup discovery approach, the following rule quality measure q is used in heuristic search of rules:

$$q = |TP| / (|FP| + g)$$

where 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 negative examples, relative to the number of covered target class cases, is determined by parameter g.

The flexibility of subgroup discovery is due to its search of rules that satisfy groups of examples of the target class, not necessary excluding all of the non-target examples. Sizes of subgroups are not defined in advance but the algorithm tends to make them as large as possible. Due to this flexibility the algorithm is able to incorporate different rule relevancy methods with the goal to prevent the construction of target class subgroup descriptions which do not have sufficient supportive evidence for being significantly different from non-target samples. An equally important part of the methodology for avoiding overfitting is that each feature that enters the subgroup discovery algorithm should itself be a relevant target class descriptor.

Relevancy of features

The relevancy of features is determined by a combination of methods for restricting the hypothesis search space and for eliminating features with low covering properties. The later methods based on absolute and relative relevancy are universally applicable to any domain and their use is suggested in all feature based inductive learning tasks. The restrictions of the hypothesis search space are related to the form of rules and to the properties of the domain. In this section we present an effective approach that can strongly reduce the number of features and its application is suggested for descriptive induction tasks in gene expression domains.

The features are restricted to simple forms only, because their complex forms may enable that, despite testing feature covering properties, features with insufficient supportive evidence may enter the rule construction process. For example, for discrete attributes the simple features have the form $Ai=a$. No complex logical forms like ($Ai=a$ *&* $Aj=b$) or ($Ai=a$ *V* $Aj=b$) are acceptable. The first form is not needed as all potential conjunctions are tested by the beam search procedure of the subgroup discovery algorithm. The second form is dangerous because, for example, the feature $Ai=a$ may be relevant while the feature $Aj=b$ may be irrelevant. Their combination $Ai=a$ *V* $Aj=b$ may be even more relevant than $fi=a$ itself, which may cause that

condition *fj=b* may be included into the finally constructed rules while its inclusion is not justified by its covering properties on the training set. Notice that if both conditions *fi=a* and *fj=b* are relevant, it does not mean that by restricting the form of used features some important logical combinations of features will be ignored. In the subgroup discovery approach both features can build separate subgroup descriptions and - if they are relevant - they both have a chance to appear in the final set of induced rules.

Results

The gene expression domain, described by Ramaswamy et al.[8] and Gamberger et al.,[9] and used in our experiments, 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. 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.

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. Another possibility is to use presence call (signal specificity) values computed from measured signal intensity values by the Affymetrix GENECHIP software. In all experiments we used only the presence call values. The presence call has discrete values A (absent), P (present), and M (marginal). The M value can be interpreted as a *do not know* state, so for every attribute there are only two distinct features *Attribute = A* and *Attribute = P* generated for each attribute. The reason is that features presented by conditions like *Ai* is true (*Ai* is present) or *Aj* is false (*Aj* is absent) are very natural for human interpretation. A more important reason for using GENECHIP presence call values (instead of continuous signal intensity values) is that the approach can help in avoiding overfitting, as the feature space is very strongly restricted: instead of many features per attribute we have only two. Also, as the measured gene expression values are not completely reliable (which is reflected by the fact that for the same sample measured values may change from one measurement to another), some robustness of constructed rules is welcome, which is achieved by treating the marginal presence call attribute value M as a *do not know* state. The value can neither be used to support the relevancy of a feature or a rule, nor it can be used for prediction purposes. In this way it additionally restricts the hypothesis search space.

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 with 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, for each class a single rule with maximal q value has been selected, for q being the heuristic of the SD algorithm, and g being equal 5 in all experiments presented in this work. The rules for all 14 tasks consisted of 2-4 features. The induced rules were tested on the independent example set. The procedure was repeated for all 14 tasks with the same default parameter values and tested on an independent test set. The results are presented in Table 1.

The table presents measured covering properties both on the training set and on the test set. Although the obtained covering values on the training sets are very good, the measured prediction quality on the test sets is for many

classes very low, significantly lower than those reported by Ramaswamy et al.[8] For 7 out of 14 classes the measured precision on the test sets is 0%. But from the table an interesting and important relationship between prediction results on the test set and the number of target class examples in the training set can be noticed. There are very large differences among the results on the test sets for various classes (diseases) and the precision higher than 50% has been 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.

Table 1 Covering properties on the training and on the independent test set for rules induced for 14 classes. Sensitivity is |TP|/|P|, specificity is |TN|/|N|, while precision is defined as |TP|/(|TP|+|FP|).

Cancer	Training set			Test set		
	Sens.	Spec.	Prec.	Sens.	Spec.	Prec.
breast	5/8	136/136	100%	0/4	49/50	0%
prostate	7/8	136/136	100%	0/6	45/48	0%
lung	7/8	136/136	100%	1/4	47/50	25%
colorectal	7/8	136/136	100%	4/4	49/50	80%
lymphoma	16/16	128/128	100%	5/6	48/48	100%
bladder	7/8	136/136	100%	0/3	49/51	0%
melanoma	5/8	136/136	100%	0/2	50/52	0%
uterus_adeno	7/8	136/136	100%	1/2	49/52	25%
leukemia	23/24	120/120	100%	4/6	47/48	80%
renal	7/8	136/136	100%	0/3	48/51	0%
pancreas	7/8	136/136	100%	0/3	45/51	0%
ovary	7/8	136/136	100%	0/4	47/50	0%
mesothelioma	7/8	136/136	100%	3/3	51/51	100%
CNS	16/16	128/128	100%	3/4	50/50	100%

The classification properties of rules induced for classes with 16 and 24 target class examples (lymphoma, leukemia and CNS, presented below) are comparable to those reported by Ramaswamy et al.,[8] while the results on eight small example sets with 8 target examples were poor.

The following rule was found for the lymphoma class:

Lymphoma ← CD20_receptor EXPRESSED AND phosphatidylinositol_3_kinase_regulatory_alpha_subunit NOT_EXPRESSED.

For the leukemia class, we have the following rule:

Leukemia ← KIAA0128_gene EXPRESSED AND prostaglandin_d2_synthase_gene NOT EXPRESSED.

The best-scoring rule for the lymphoma class contains a feature corresponding to a gene routinely used as a marker in diagnosis of lymphomas (CD20), while the other part of the conjunction (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 (Septin 6).

Lastly, we address the rule found for the CNS class:

CNS ← fetus_brain_mRNA_for_membrane_glycoprotein_M6 EXPRESSED AND CRMP1_collapsin_response_mediator_protein_1 EXPRESSED.

Conclusion

For larger training sets the subgroup discovery methodology enabled effective construction of relevant knowledge. The result, illustrated in Figure 1, demonstrates that mean values of rule sensitivity and precision are significantly higher for three tasks with 16 and 24 target class examples than for eleven tasks with only 8 target class examples. The mean values for the specificity are also higher but they were over 95% already for small target class sets.

The induced rules for lymphoma, leukemia and CNS were evaluated by a domain expert and most of features used in them were recognized as known disease markers for the target class cancers.[9] Expert evaluation proved the relevancy of induced rules. Both good prediction results on an independent test set as well as expert interpretation of induced rules show the

effectiveness of described methods for avoiding overfitting in scientific discovery tasks. Mostly bad results for tasks with only eight target class examples demonstrate that the methods can not be successful in all situations, especially those with a very small number of examples.

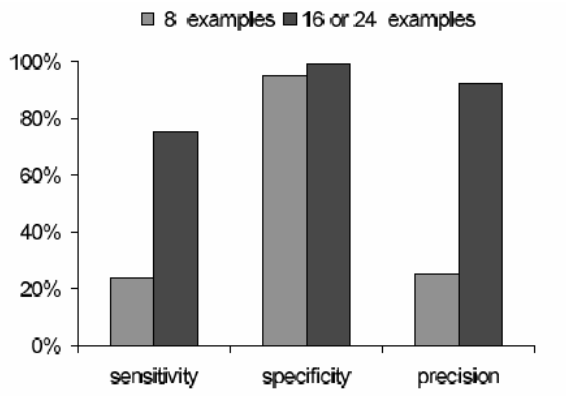

Figure 1 Mean values of sensitivity, specificity, and precision measured on the independent test set versus the number of target class cases in the training set.

In spite of the number of findings in agreement with the bio-medical state-of-the-art, discovery of known factors in the considered malignancies was not the ultimate goal of this study. The main goal of the methodology is the discovery of unknown and never thought-off relationships, in a form instantly understandable to an expert. The presented experiments have succeeded in discovering human understandable rules, some of which have uncovered interesting regularities in the data.

Literature

1. Lavrač N, Kavšek B, Flach P, et al.: Subgroup discovery with CN2-SD. Journal of Machine Learning Research 2004; 5: 153-188.
2. Kavšek B, Lavrač N: APRIORI-SD: Adapting association rule learning to subgroup discovery. In: Proceedings of the 5th International Symposium on Intelligent Data Analysis 2003; 230-241, Springer.
3. Clark P, Niblett T: The CN2 induction algorithm. Machine Learning 1989; 3(4):261-283.
4. Jovanovski V, Lavrač N: Classification rule learning with APRIORI-C. In Progress in Artificial Intelligence: Proceedings of the 10th Portuguese Conference on Artificial Intelligence 2001; 44-51, Springer.
5. Agrawal R, Srikant R: Fast algorithms for mining association rules. In: Proceedings of the 20th International Conference on Very Large Databases 1994; 207-216.
6. Gramberger D, Lavrač N: Expert-guided subgroup discovery: methodology and application. Journal of Artificial Intelligence Research 2002; 17:501-527.
7. Gamberger D, Krstačić A, Krstačić G, et al.: Data analysis based on subgroup discovery: experiments in brain ischaemia domain. In Proceedings of the 10th International Workshop on Intelligent Data Analysis in Medicine and Pharmacology 2005; 52-56, University of Aberdeen.
8. Ramaswamy S, et al.: Multiclass cancer diagnosis using tumor gene expression signitures. In Proc. Natl. Acad. Sci USA 2001; 98(26): 15149-15154.
9. Gamberger D, Lavrač N, Železny F, et al.: Induction of comprehensible models for gene expression datasets by subgroup discovery methodology. Journal of Bioinformatics 2004; 37: 269-284.