Symbolic approach to unsupervised learning

Enis Avdičaušević, Mitja Lenič, Bruno Stiglic
University of Maribor, FERI, Slovenia
aenis@uni-mb.si, mitja.lenic@uni-mb.si, stiglic@uni-mb.si

Abstract
As important as data mining techniques are, they still require too much technical knowledge from users. This is most evident in the process of interpretation of results, where domain experts are involved. One of the most challenging tasks in the area of knowledge discovery is to express discovered knowledge in a form, which can be understood by domain experts (e.g. medical experts). In the paper we present our approach to unsupervised learning using multivariate symbolic hybrid. Main advantage of multimethod symbolic hybrid is that learned knowledge is expressed in a form of symbolic rules. Learned knowledge is much more understandable to domain experts, which increases its value and makes it much easier to apply.

1. Introduction

Many real-world medical problems are nowadays being treated with tools for automatic intelligent data analysis. Various methods have been developed to improve the quality of analysis for specific domains. One of the most important fields in this area are machine learning and knowledge extraction. These research fields have long tradition in data mining, which is obviously needed in today's society that has relatively inexpensive and generally available means to collect and store the data from various environments. The aggressive rate of growth of disk storage and thus the ability to store enormous quantities of data has far outpaced our ability to process and utilize them. This challenge has produced a phenomenon called data tombs – data is deposited to merely rest in peace, never to be accessed again. But the growing appreciation that data tombs represent missed opportunities in for example supporting scientific discovering, business exploitation or complex decision making has awaken commercial interest in knowledge discovery and data mining techniques. That has also stimulated new interest in the automatic knowledge extraction from examples stored in large databases. Another important aspect of data mining is to be able to discover relations in patterns and represent it in a symbolic form so that domain experts can use and understand discovered knowledge. In the field of supervised learning many methods have been developed, but the main disadvantage of supervised learning is that target concept (decision class) has to be specified. In the area of unsupervised learning commonly clustering is used to group objects with no explanation why objects is in the group. Other symbolic approaches use simple attribute relations to discover knowledge. In this paper a new approach for unsupervised learning that is also able to discover complex relations with application of methods from supervised learning is presented. We constructed a hybrid approach to unsupervised learning with integration of complex object transformation and various supervised learning methods. Main contributions of the paper are:

- Complex rule extraction with symbolic knowledge representation
- Application of supervised learning in the field of unsupervised learning
- Introduction of interestingness measure for rule evaluation
- Application of presented method on the breastfeeding database[1]

2. Related work

In the area of unsupervised learning or more specific in the area of data clustering research there has been a considerable amount of effort already invested. There are several approaches and methods developed in this area: hierarchical algorithms, partitioning algorithms, density based algorithms, grid-based algorithms and also some newer approaches based on neural networks and genetic algorithms. All of these existing approaches can be considered as successful considering quality of partition they produce or for example time complexity. On the other hand results produced by these approaches are usually expressed in a form of sets of objects, which are members of cluster. Results expressed in such form give us membership relation between object and clusters but not the reasons for membership nor relationship between objects belonging to the same cluster. Therefore results

[1] Breastfeeding database was collected in various hospitals in Slovenia and sponsored by Slovenian Ministry of science and technology.

expressed in such way have poor explanatory power to humans. Finding a suitable way to increase explanatory power of learned knowledge and to present it in form understandable to humans is receiving much attention in research community in recent time [1][2][3][4][5]. For example neural networks have great ability to learn. They are able to learn very complex concepts but their expressive power is very low so different methods are developed to extract knowledge learned by neural networks by extracting learned features or combining with symbolic methods. Considering the complexity of supervised and especially unsupervised learning, genetic approaches are more and more used in this field. One of the most interesting genetic approaches is described in [6]. Authors present approach based on rules expressed in form of hypercubes. A set of cluster rules is subject to evolution which leads to partition of data space in form of disjunctive hypercubes representing clusters. However our method can express more complex knowledge and is also suitable for more complex data forms types like for example signals.

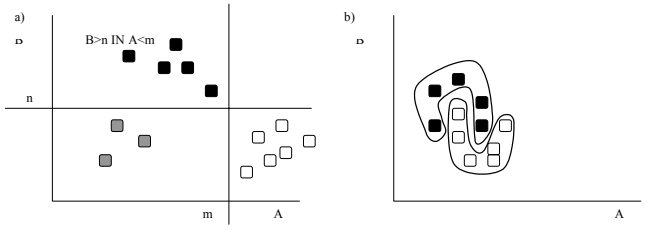

Figure 1 – Linear a) and nonlinear b) division

3. Complex Rule Mining

In order to be able to capture more complex concepts and knowledge hidden in data we first have to define a suitable form in which these concepts and learned knowledge will be expressed. This form has to be simple enough to be understandable to domain experts and on the other hand it has to be flexible enough to allow representation of more complex knowledge. Even more important is that this form has to be adaptive in order to start learning simple concepts and progress towards more complex knowledge.

Many of already existing approaches use rules to express learned knowledge. These rules are usually constructed as conjunction of attribute range tests in following form: $C_1<=a<=C_2$. (a is attribute and C_1 and C_2 are constants). Such rules are very simple to compute and understand but also knowledge captured in such rules is limited to linear relationships (Figure 1). In order to increase the power of such rules we extended these rules to combination of tests in following general form:

$$f(a_1,..,a_{j1}, k_1,..,k_{r1})\ \text{REL\_OP}\ C$$

Symbol f represents function constructed in the process of learning. Each of these functions can have different number of attributes and constant parameters. REL_OP is relational operator like for example <, > or =. C is a constant to which the value of function is compared using REL_OP. With this form we are able to transform initial attribute space and make decisions in transformed space, which gives us a possibility to make non-linear decisions based on multiple attributes at the same time. By selecting the appropriate complexity of these functions we can find optimal relationship between complexity of rules and their understandability.

Obviously this extension expands the search space and learning algorithm has to explore different combinations of functions and attributes. However this makes the learning task even more difficult. Because of the complexity of search space, classical algorithms are usually not appropriate approach in such cases. Instead we decided to use evolutionary algorithms, which have proved to give much better results on such complex problems.

3.1. Genetic programming

Genetic algorithms [8][9][10] (GA) are adaptive heuristic search methods, which may be used to solve all kinds of complex search and optimization problems. GA are based on evolutionary ideas of natural selection and genetic processes of biological organisms. As the natural populations evolve according to the principles of natural selection and “survival of the fittest”, first laid down by Charles Darwin. By simulating this process, genetic algorithms are able to evolve solutions to real-world problems, if they have been suitably encoded. They are often capable of finding optimal solutions even in the most complex of search spaces or at least they offer significant benefits over other search and optimization techniques.

A typical genetic algorithm operates on a set (a population) of solutions (chromosomes) within the search space. The search space represents all possible solutions that can be obtained for the given problem, and is usually very complex or even infinite. Every point of the search space is one of the possible solutions and therefore the aim of genetic algorithm is to find an optimal point or at least come as close to it as possible.

The genetic algorithm consists of three genetic operators: selection, crossover (recombination), and mutation. Selection is the survival of the fittest individuals within the genetic algorithm with the aim of giving the preference to the best ones. For this purpose all solutions have to be evaluated, which is done with the use of fitness function. Selection determines individuals to be used for the second genetic operator -

crossover or recombination, where from two good individuals a new, possibly even better one is constructed. The crossover process is repeated until the whole new population is completed with the offspring produced by the crossover. All constructed individuals have to preserve the feasibility regarding the given problem. In this manner it is important to co-ordinate internal representation of individuals with genetic operators. The last genetic operator is mutation, which is an occasional alteration of an individual that helps to find an optimal solution to the given problem faster and more reliably.

If the search space of genetic algorithm is consisted of programs we are talking about special kind of genetic algorithms named genetic programming [7]. In genetic programming population is a set of programs, which are possible candidates to solve the problem. Programs are usually represented as expression trees, which are subject to process of evolution in which programs are gradually adjusted to improve their capability to solve the original problem. In our approach we are using genetic programming to obtain rules, which best describe the knowledge hidden in data.

3.2. Multivariate symbolic hybrid

As we already described, in our approach rules have two parts, which form IF-THEN rule (Figure 2). First part of the rule (condition) is expression, which is a conjugation of extended range tests. This part of the rule is built using genetic programming. Condition expression is represented as tree, which consists of functions and symbols. Symbols can have different types, for example integer for discrete attributes and floating point for continuous attributes. Symbols, which usually appear in such expressions, are attribute value in a row, different constants and references on columns of attributes (@A in Figure 2). References to

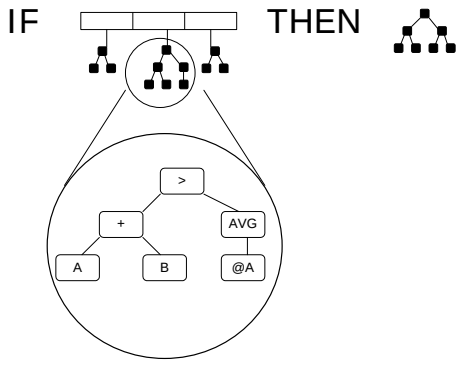

Figure 2 – Rule structure

columns give us the possibility to include average values, moments and other statistical elements to the expression. Since columns are not available at the time when learned knowledge is used on concrete example, final expression has to be partially evaluated in order to calculate subexpressions referencing complete columns. Expression can include different kinds of relationship functions (<,>, <=) and arithmetic functions (+, -, /, *, exp, …). Additional functions can be added to accommodate special needs of the application or some other data types.

For definition of then part of the rule we used novel approach to avoid additional search for most appropriate rule. When expression of if part is determined by genetic programming there is only a limited (but still infinite) set of possibilities for then part. If we would search for those then parts we would need a lot of computational power to find proper then part. This means that these two parts cannot be constructed independently. Considering the complexity of the problem and size of the search space there is a very small chance that we would be able to independently build two expressions, which would match together using evolutionary computation. Consequence is also an expression but can contain only those attributes, which were not used in condition to avoid senseless rules (for example if a>5 then a>5).

We took another approach by constructing multivariate symbolic hybrid, which is a hybrid that applies supervised learning algorithms on unsupervised learning problem with combination of evolutionary algorithms. That is achieved with the idea of transforming unsupervised learning problem to supervised learning problem by adding information from condition part to the problem.

Conversion of unsupervised learning problem to supervised is implemented by introduction of additional attribute to the learning set. Additional attribute is constructed from condition expression that resulted from evolutionary part and calculating its value for all learning objects in database. Additional attribute represents decision class (label) of learning object. That separates all objects in database in two classes, those for which condition expression is valid and those for which condition expression is not valid. In the next step multimethod learning system (using multiple learning methods to construct various classifiers) for supervised learning is used [15]. Only symbolic methods should be used to extract symbolic information [14]. Commonly we use decision trees [11][12][13] for knowledge representation. To avoid senseless rules system for supervised learning is instructed to use only those attributes, which were not used in condition of the rule. The Result is a decision tree, which is able to distinguish between those two classes of objects. Resulting decision tree has two purposes. First we can extract consequence expression from its structure and depending on our needs extracted expression can be simple or more complex.

Supervised learning algorithms estimate best attribute for division of objects at certain level. In our case we use simpler forms of expression consisted from one or two attributes. Secondly resulting tree is used to evaluate quality of complete IF-THEN rule. Considering quality, size and structure of decision tree we can evaluate how similar were the objects selected by condition. For example if objects, which were selected by condition are not similar this affects the structure of decision tree, which has more branches and also it is not able to accurately distinguish between objects.
We can see this as a kind of distance measure. Distance measure computed in this way is different from usual distance measures based on Euclidian space. It is based on relationship between attributes, which describe objects in database.

3.3 Evaluation of discovered knowledge

Considering the size of search space, evaluation of discovered knowledge is extremely important for its practical usage. Learning algorithms can produce very large number of rules. Browsing and searching through such large set of rules is very difficult and makes it very hard for user to find interesting rules. Automatic learning algorithms have to be able to evaluate different rules constructed in the learning process and present just those that could represent certain value to the user.
Obvious criteria for evaluation of learned rules are different objective properties of such rules such as rule support, rule confidence, size and attribute costs. For example, rules with higher support are more valuable to users than those with lower support. Lets take a few moments to think about that statement. By definition of support, rules with high support hold for most individuals in the observed population. If some rule holds for most of the individuals of the population we can expect that it is probably already known to the user and that user probably does not need some high tech automatic learning algorithm to tell him “old jokes”.
In general valuable rules represent either facts, which were previously unknown to the user or facts that contradict users knowledge of the problem domain. Such rules have real value to the users since they represent new knowledge and as such can be used to improve processes conducted by users in the problem domain. This value is usually referred as rule interestingness, since users usually have more interest in rules with greater value.
Naturally, our first impression is that rule interestingness is purely subjective measure and therefore very hard to describe in formal way. We certainly have to allow possibility that user will show interest in some rule based on pure subjective basis, but most of the users interest in certain rules is based on prior knowledge, which user has about the problem domain. If this knowledge can be modeled or at least approximated then also rule interestingness can be computed.
In our system we estimate rule interestingness based on two approaches:

- Attribute Correlation; Rule is interesting if attributes in IF and THEN part have low correlation, but rule still has high support and confidence. This means that by using different combinations of functions and attributes, original attribute space was transformed in a such way that correlation is evident. Since relation between attributes was not obvious we consider such rule to be interesting to the user.
- Rule templates; with rule templates user can describe possible candidates for rules that are in his interest. Rule templates are defied in a form of $A \rightarrow X$, where A is set of attributes, which should appear in IF part and X is attribute, which should appear in THEN part. Rules matching these patterns are more interesting to the user.

By combining these two approaches we are able to present to the user just those rules, which are considered as interesting. Off course resulting set of rules can be still quite large and can still require involvement of domain expert.

4. Results and discussion

Multivariate symbolic hybrid approach was tested on several synthetic and real world medical databases. As examples will show, described approach is very promising and it able to discover different concepts hidden in the date. These tests also showed real complexity of described problem. Especially challenging task is evaluation of rule interestingness and we are working on further improvements of the approach.

4.1 Synthetic case

On figure 3 we can see one of synthetic examples. Simple database was created artificially to test systems ability to find complex relationships in data.

Rule examples:

```
if (C=[1]) then D in [1,2]

if ((A+B)*(A+B) <= 30.0) and (C=[1]) then D in [1]
```

4.2 Breastfeeding database

A	B	C	D
1,00	1,00	1	1
2,00	1,00	1	1
1,00	2,30	1	1
3,00	2,00	1	1
1,00	2,10	1	1
2,00	1,00	1	1
3,00	1,10	1	1
2,00	1,10	1	1
1,20	4,00	1	1
2,20	2,93	1	1
2,40	1,50	1	1
5,00	5,00	1	2
4,00	6,00	1	2
11,00	6,00	1	2
2,00	5,00	1	2
3,00	4,00	1	2
1,50	4,00	1	2
1,00	1,00	2	3
2,00	1,00	2	3
1,00	2,30	2	3
3,00	2,00	2	3
5,00	5,00	2	3
4,00	6,00	2	3

Figure 3 – Synthetic case.

For the purpose of qualifying the influence of feeding the baby to his/her mental and physical development we have measured and collected data in form of precisely defined protocols. For 383 infants who were present at the preventive monthly check-ups basic and specific information about their habits and behavior were collected. Some data like birth weight and birth length is based on maternity's records while the rest was gained from the parents of checked children. The whole process was performed very carefully, in order to prevent mistakes. Nevertheless, there are some cases where some of the attribute values are missing or incorrect, which makes the processing of the data quite difficult, but we were able to produce several interesting rules. As an example we present following rules:

```
if(F_CIGARETTES  == 0.0) then M_SMOKING = no
```

System discovered correlation between number of cigarettes smoked by father (F_CIGARETTES) and the fact if mother is smoking. If father is not smoking then mother usually does not smoke.

```
if(M_CIGARETTES - APGAR2 - 1.0 > -9.0) then
APGAR1<=9.0
```

In this case relationship between number of cigarettes smoked by mother and apgar tests is presented. This rule has a little bit complicated form but in general it states that if mother is smoking then apgar tests show lower values. As expected this confirms that smoking has negative influence on infants health.

```
if(PEDIATRI - APGAR3 == -10.0)
       then HOSPITAL not in
[Ljubljana,Kranj,Slovenj
Gradec,Maribor,Sempeter]
```

This is one of the most interesting rules discovered on this database. It presents relationship between number of days in pediatric hospital (PEDIATRI), apgar and hospital. Interesting statement, which can be extracted from this rule is that children with APGAR3 and no days spent in hospital are not from following hospitals: Ljubljana, Kranj, Slovenj Gradec, Maribor, Sempeter. Interesting fact is that these are bigger and more equipped hospitals in Slovenia. Probable explanation for this rule is that physicians in these hospitals usually more strictly evaluate apgar tests.
Multivariate symbolic hybrid proved to be able to find complex knowledge in both cases but still has some disadvantages. Sometimes the rules are too complicated and complex and huge involvement of medical experts is necessary. But nevertheless the approach is very promising and we are currently working on further improvements.

5. Conclusion

Multivariate symbolic hybrid is a system for unsupervised learning. It can discover and learn complex concepts hidden in data. Its main feature is that learned knowledge is expressed in symbolic form understandable to humans. Approach is not limited to discrete and continuous data but it can support data mining also on alternative and more complex data types like for example signals. Multivariate symbolic hybrid is still in research phase and we are working on further improvements of the approach but it already shows main benefits of the approach: Learned knowledge is much more understandable to domain experts, which increases its value and makes it much easier to apply.
With proper combination of objective and subjective criteria we are able to present to the user those rules, which have potential interestingness. Evaluation of subjective criteria is very challenging especially in cases where initial database is not well constructed or has many missing attribute values. Our future research efforts will be concentrated mainly in this area in order to improve quality (interestingness) of presented results. One of the possible promising directions is research of knowledge contained in exceptions. Rules that describe general behavior are probably already known to users and therefore not very interesting, but exceptions from that rules can contain very interesting knowledge to the user since they contradict general behavior.
Evolutionary systems require a lot of computing power. This is also true for multivariate symbolic hybrid. We are already working on distributed architecture, which will

allow us to employ more computers to work together on common problem and therefore reduce computing time and improve quality of results.

6. References

[1] Coit DW, Smith AE, Solving the redundancy allocation problem using a combined neural network/genetic algorithm approach, Computers & Operations Research, Volume 23, Issue 6, June 1996, Pages 515-526.

[2] Furuya T, Satake T and Minami Y, Evolutionary programming for mix design, Computers and Electronics in Agriculture, Volume 18, Issues 2-3, August 1997, Pages 129-135.

[3] Andrews R, Diedrich J, Tickle AB: A Survey And Critique Of Techniques For Extracting Rules From Trained Artificial Neural Networks. Neurocomputing Research Centre, 1995.

[4] Sethi IK: Entropy Nets: From Decision Trees to Neural Networks. Proceedings if IEEE, Volume 78, 1990.

[5] Kohonen T, Kaski S, Lappalainen H. Self-Organized Formation of Various Invariant-Feature Filters in the Adaptive-Subspace SOM. Neural Computation. 1997.

[6] Sarafis I, Zalzala AMS, Trinder PW. A Genetic Rule-Based Data Clustering Toolkit. Congress on Evolutionary Computation (CEC), Honolulu, USA (May 2002).

[7] Koza JR. Genetic Programming: On the Programming of Computers by Means of Natural Selection. MIT Press. 1992.

[8] Banzhaf W, Nordin P, Keller RE and ED. Francone, *Genetic Programming* - An *Introduction,* Morgan Kaufmann Publishers Inc., 1998.

[9] Michalewicz Z. *Genetic Algorithms + Data Structures = Evolution Programs*, Springer Verlag, 1996.

[10] Goldberg DE: *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison Wesley, Reading MA, 1989.

[11] Quinlan JR, Induction of decision trees, Machine Learninig 1 (1986), 81-106.

[12] Quinlan JR, Decision trees and instance based classifiers, Artificial Intelligence and Robotics (I 997), 521-535.

[13] Lenič M, Kokol P. Combining methods with multimethod approach. V: European symposium on intelligent techniques, hybrid systems and their implementation of smart adaptive systems [also] Eunite 2002, September 19-21, 2002, Albufeira, Algarve, Portugal.

[14] Lenič M, Kokol P, Zorman M, Povalej P, Stiglic B, Yamamoto R. Improved knowledge mining with the multimethod approach. V: The foundation of data mining and knowledge discovery : IEEE ICDM 2002 Workshop proceedings, TERRSA, Maebashi City, Japan, 9-12 December, 2002.

[15] Lenič M, Kokol P. Combining classifiers with multimethod approach. V: ABRAHAM, Ajith (ur.), RUIZ-DEL-SOLAR, Javier (ur.), KÖPPEN, Mario (ur.). Second international conference on Hybrid Intelligent Systems, Santiago de Chile, December 1-4, 2002.