EXPERIMENTAL COMPARISON OF THREE SUBGROUP DISCOVERY ALGORITHMS: ANALYSING BRAIN ISCHAEMIA DATA

Petra Kralj(1), Nada Lavrač(1,2), Blaž Zupan (3,4), Dragan Gamberger (5)
(1) Jozef Stefan Institute, Jamova 39, Ljubljana, Slovenia
(2) Nova Gorica Polytechnic, Vipavska 13, Nova Gorica, Slovenia
(3) Faculty of Computer and Information Science, Tržaška 25, Ljubljana, Slovenia
(4) Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX
(5) Rudjer Bošković Institute, Zagreb, Croatia
e-mail: Petra.Kralj@campus.fri.uni-lj.si

ABSTRACT

This paper presents experimental results of subgroup discovery algorithms SD, CN2-SD and Apriori-SD implemented in the Orange data mining software. The experimental comparison shows that algorithms perform quite differently on data discretized in different ways. From the experiments, performed in the brain ischemia domain, it is impossible to conclude which discretization is the most adequate for subgroup discovery.

1 INTRODUCTION

This paper addresses the problem of subgroup discovery in a medical domain. Subgroup discovery is an appropriate method for analyzing medical data, since it provides short and understandable descriptions of subgroups regarding the property of interest.

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 statistically 'most interesting', e.g., are as large as possible and have the most unusual statistical (distributional) characteristics with respect to the property of interest.

Standard classification rule learning algorithms can be adapted to perform subgroup discovery. In this paper we discuss three subgroup discovery algorithms, SD [1], CN2-SD [2] and Apriori-SD [3], two of which are adaptations of classification rule learners: CN2-SD is an adaptation of CN2 [8] and Apriori-SD is an adaptation of APRIORI [6]. We compare the results of these three algorithms, implemented in the Orange data mining environment [5], on brain ischaemia data [4].

These algorithms take as their input the training examples described by discrete attribute values. Since some of the attributes in the brain ischemia domain are continuous, data discretization is needed in the pre-processing phase. Discretization is performed in two different ways.

This paper is organized as follows: Sections 2, 3 and 4 present the basic ideas of algorithms SD, CN2-SD and Apriori-SD, respectively. Section 5 explains how subgroups can be used for classification purposes. In Section 6 the data set and its pre-processing are presented. Experimental results are provided in Section 7. Finally, Section 8 provides conclusions and references.

2 THE SD ALGORITHM

The SD algorithm [1] is a variation of the beam search algorithm. At the beginning all the subgroup descriptions in the beam are initialized to empty. The algorithm builds subgroup descriptions in a general-to-specific fashion by adding conjunctions to subgroup descriptions. Discovered subgroups must satisfy criteria of minimal support and they must be relevant. The new subgroup is irrelevant if there exists a subgroup R such that true positives of the new subgroup are a subset of true positives of R and false positives of the new subgroup are a superset of false positives of R.

The algorithm keeps the best subgroup descriptions in a beam of fixed width (beam width is a parameter of the algorithm). In each iteration of the algorithm it adds a conjunction to every subgroup descriptions in the beam and replaces the worst subgroup in the beam if the new subgroup in better.

The goal of the subgroup discovery algorithm SD is to find subgroups that maximize the generalization quotient heuristic (Equation 1), where TP are the true positives, FP are the false positives, and g is a generalization parameter.

$$q_g = \frac{TP}{FP + g} \qquad (1)$$

High quality subgroups cover many target class examples and a low number of non-target examples. The number of tolerated non-target class examples, relative to the number of covered target class examples, is determined by parameter g. For low g, induced rules will have high specificity since the coverage of every single non-target class example is made relatively very 'expensive'. On the

other hand, by selecting a high g value, more general rules will be generated, covering also non-target class instances.

3 THE CN2-SD ALGORITHM

The CN2-SD algorithm [2] consists of two main procedures: the bottom-level search procedure that performs beam search in order to find a single rule, and a top-level control procedure that repeatedly executes the bottom-level search and performs the weighting of covered examples to induce a rule set.

The bottom-level procedure performs search in a general-to-specific fashion, specializing only the subgroup descriptions in the beam by iteratively adding features. This procedure stops when no specialized subgroup description can be added to the beam, because none of the specializations has a higher weighted relative accuracy (Equation 2).

$$WRAcc(X \to Y) = \frac{n'(X)}{N'}\left(\frac{n'(XY)}{n'(X)} - \frac{n(Y)}{N}\right) \qquad (2)$$

In this equation, N is the number of all examples, N' is the sum of the weights of all examples, n'(X) is the sum of weights of all covered examples, N(Y) is the number of examples within the target class, and n'(XY) is the sum of the weights of all correctly covered examples. The weights are calculated as follows:

$$w(e_j, i) = \frac{1}{i+1} \qquad (3)$$

In this equation, e_j is an example that is covered i times.

4 THE APRIORI-SD ALGORITHM

The APRIORI-C algorithm [7] uses techniques from the association learning algorithm APRIORI [8] to build classification rules. Some adaptations of the APRIORI algorithms are needed to perform the classification task, like building only the rules with the target variable on the right hand side and others, described in [7].

The main modification of the APRIORI-C algorithm, making it appropriate for subgroup discovery, involve the implementation of an example weighting scheme in rule post-processing, a modified rule quality function incorporating example weights and a probabilistic classification scheme.

Algorithm Apriori-SD [3] is very similar to the CN2-SD algorithm, since they have very similar top-level control procedures that repeatedly execute the bottom-level search and perform the weighting of covered examples to induce a rule set. In Apriori-SD, a set of potential subgroup descriptions is generated at the beginning of the control procedure by executing the Apriori-C algorithm. The condition parts of the generated rules can be interpreted as subgroup descriptions.

The bottom-level procedure in Apriori-SD finds the subgroup with the highest weighted relative accuracy (WRAcc, Equation 2) among the subgroup descriptions (rules) generated by algorithm Apriori-C. It removes the found subgroup description from the set and returns this rule.

5 SUBGROUP DESCRIPTIONS AS CLASSIFIERS

Even though subgroup discovery belongs to descriptive induction, using and testing it as a classifier enables us to better evaluate the general descriptive usefulness and generalization properties of the found subgroups.

Subgroup discovery can be used as predictive induction by building subgroups for every class within the target variable. When classifying a new example, this approach calculates the average of distributions of all the discovered subgroups that cover this example and classifies it into the class that has the highest probability estimation. In this way the votes of all the subgroups have the same weight when deciding in which class to classify, regardless how many examples they cover.

6 BRAIN ISCHAEMIA DATA

The brain ischemia dataset consists of records of patients who have been treated in the Intensive Care Unit of the Department of Neurology, University Hospital Center "Zagreb", in Zagreb, Croatia during the year 2003. 300 patients are included in the database: 209 with confirmed diagnosis of brain attack, and 91 patients who entered the same department with adequate neurological symptoms and disorders, but were diagnosed with other diagnosis. In this paper, the goal of subgroup discovery is to discover regularities that characterize brain attack patients.

Patients are described with 26 attributes; 14 of them are discrete, 12 continuous. They are described in [4].

Since the subgroup discovery algorithms we compare in this paper take as their input discrete attribute descriptions of data, we need to discretize the continuous attributes. We do it in two ways: one is by performing the entropy based discretization [10] as implemented in Orange, the other is by performing binarization in feature construction and feature subset selection [9] and using the results as binary attributes.

- The Orange implementation of the entropy based discretization transforms 14 continuous attributes into seven discrete attributes: six binary and one with three values. Five attributes are discarded as irrelevant.
- By performing feature generation and feature selection on the continuous attributes we obtain 509 features of the form attribute<value or attribute>value. We use them as binary attributes, therefore when they appear in a subgroup description, they look like "attribute>value =y" or "attribute>value =n", where y and n stand for logical values *true* and *false*, respectively.

7 EXPERIMENTAL RESULTS

We performed tests of all three algorithms on data discretized in both ways. Unfortunately we were unable to test the algorithms CN2-SD and Apriori-SD on the data,

discretized by feature generation and selection because the implementations of these algorithms are not capable of dealing with that many attributes.

We first ran the algorithms on the entire data set and calculated which subgroups are on the convex hull in the ROC space (marked by *) and calculated the corresponding area under the ROC convex curve (AUC). In another experiment we performed ten-fold cross validation and calculated the average classification accuracy (CA) of the algorithms. The results are shown in the following tables.

Tables 1 to 4 show the results of three subgroup discovery algorithms on the brain ischemia domain. The discovered subgroups show the importance of attributes Age and Fibr, since all the algorithms discovered subgroups containing these attributes in their subgroup descriptions.

The tables are formatted as follows: The first column contains subgroups names. The second column contains subgroup descriptions, where spaces between conjuncts denote logical and. Columns TPr and FPr show the rate of positive and negative examples covered by each individual subgroup. The asterisks in the last column denotes that the specific rule is on the ROC convex hull. The numbers in the last column show the classification accuracy obtained by performing ten-fold cross validation.

Apriori-SD (Table 1) has a high classification accuracy while its area under the ROC convex hull is not large. Algorithm CN2-SD (Table 2) produced only three very short and understandable rules.

Ref.	**Subgroup description**	**TPr** [%]	**FPr** [%]	**CA** **AUC**
a1	D_Fibr≥4.30	54	5	*
a2	D_Age≥66.0 D_RRsys≥158.0	48	8	
a3	D_Age≥66.0 D_Gluc≥5.90 AHyp=yes	39	4	
a4	D_au≥378.0 D_Gluc≥5.90	27	2	*
a5	D_Age≥66.0 D_Gluc≥5.90 D_RRdya≥89.0 Stat=no	37	3	*
a6	D_Gluc≥5.90 D_RRsys≥158.0 ASS=no Ahypo=no	29	4	
a7	FA=yes AHyp=yes	28	5	
a8	alcoh=yes stres=no	28	5	
a9	D_au≥378.0 AHyp=yes	28	4	
a10	D_Age≥66.0 Fhis=yes Stat=no	31	4	
a11	D_Gluc≥5.90 D_RRsys≥158.0 D_RRdya≥89.0 ASS=no Acoag=no Stat=no	28	3	
a12	D_Age≥66.0 Smok=no stres=no	28	5	
classification accuracy				**0.83**
area under ROC convex hull				**0.74**

Table 1: *Subgroup descriptions induced by algorithm **Apriori-SD** on discretized data.*

If we compare the classification accuracy of algorithm SD on differently discretized data, we can see that the simple entropy based discretization (Table 4) works better for small values of the generalization parameter g, while the feature based discretization (Table 3) is better for large values of the g parameter.

Ref.	**Subgroup description**	**TPr [%]**	**FPr** [%]	**CA** **AUC**
c1	D_Fibr≥4.30	54	5	*
c2	D_Age≥66.0 D_Fibr≥4.30	41	1	*
c3	D_Fibr≥4.30 Fhis=yes	34	1	*
classification accuracy				**0.77**
area under ROC convex hull				**0.75**

Table 2: *Subgroup descriptions induced by algorithm **CN2-SD** on discretized data. The algorithm induced 20 descriptions, but only three of those are different.*

Ref.	**Subgroup description**	**TPr** [%]	**FPr** [%]	**CA** **AUC**
generalization parameter value 5				0.75
g5a	Fibr>2.75=n Age>62.50=y	44	1	
g5b	Fibr>2.75=n Plat<145.50=n Age>70.50=y	34	0	
g5c	Fibr>2.75=n PT<0.99=n	33	0	
generalization parameter value 10				0.76
g10a	Acoag=yes=n Trig<1.48=n Trig>1.48=y Fibr>4.55=n	49	2	
g10b	Fibr>2.75=n Age>60.50=y	43	0	*
g10c	Acoag=yes=n Trig<1.48=n Trig>1.42=y Fibr>4.55=n	49	2	*
generalization parameter value 20				0.74
g20a	Trig<1.48=n Age>59.50=y Plat<133.0=n	66	10	*
g20b	Trig<1.48=n Age>55.50=y	60	7	
g20c	Trig>1.48 Age>59.50 Gluc>6.85 Plat>133.0	61	7	*
generalization parameter value 50				0.8
g50a	Trig<1.48=n Plat<145.50=n	78	21	
g50b	Trig<1.52=n Age>61.50=y	75	16	*
g50c	Trig<1.48=n Plat<133.0=n	77	19	*
generalization parameter value 100				0.85
g100a	Trig<1.52=n	84	31	*
g100b	Trig<1.52=n RRsys>169.0=n	82	29	
g100c	Trig<1.52=n Plat<145.50=n	82	27	*
average classification accuracy				**0.78**
area under the ROC convex hull				**0.85**

Table 3: *Subgroup descriptions induced by algorithm **SD** on feature data. The subgroup descriptions and the*

classification accuracy are induced for different values of generalization parameter g in the range [5, 100].

Ref.	Subgroup description	TPr [%]	FPr [%]	CA / AUC
generalization parameter value 5				0.8
d5a	D_Fibr≥4.30	54	5	*
d5b	D_Age≥66.0 D_Gluc≥5.90 D_RRdya≥89.0 D_RRsys≥158.0	34	2	*
d5c	D_Fibr≥4.30 Stat=no	47	4	
generalization parameter value 10				0.81
d10a	D_Fibr≥4.30	54	5	*
d10b	D_Age≥66.0 D_RRsys≥158.0 Stat=no	43	5	
d10c	D_Fibr≥4.30 Stat=no	47	4	
generalization parameter value 20				0.8
d20a	D_Age≥66.0	66	23	*
d20b	D_Fibr≥4.30	54	5	
d20c	D_Age≥66.0 D_RRdya≥89.0	54	14	
generalization parameter value 50				0.7
d50a	D_RRdya≥89.0	80	55	
d50b	D_Gluc≥5.90	74	48	
d50c	AHyp=yes	74	46	
generalization parameter value 100				0.81
d100a	D_Age≥66.00	66	23	
d100b	D_RRdya≥89.00	80	55	
d100c	-	-	-	*
average classification accuracy				**0.78**
area under the ROC convex hull				**0.76**

Table 4: *Subgroup descriptions induced by algorithm **SD** on discretized data. The subgroup descriptions and the classification accuracy are induced for different values of the generalization parameter in the range [5, 100].*

8 CONCLUSIONS

In this paper we confronted three subgroup discovery algorithms on the brain ischemia domain. We discovered that algorithms perform quite differently on data discretized in different ways. From the experiments we made it is impossible to conclude which discretization is the most adequate for subgroup discovery – this evaluation should be performed by the medical expert in future work. Additionally, comparison of algorithms on many other domains should be performed to get relevant statistical results from which one could conclude which algorithm and discretization perform the best.

Acknowledgements

The authors acknowledge the support of the Slovenian Ministry of Higher Education, Science and Technology and the 6FP EU project Inductive Quesries for Mining Patterns and Models.

References

[1] D. Gramberger, N. Lavrač. Expert-Guided Subgroup Discovery: Methodology and Application. Journal of Artificial Intelligence Research, 17:501–527, 2002.

[2] N. Lavrač, B. Kavšek, P. Flach, L. Todorovski. Subgroup Discovery with CN2-SD. Journal of Machine Learning Research, 5: 153–188, 2004.

[3] B. Kavšek, N. Lavrač. APRIORI-SD: Adapting Association rule Learning to Subgroup Discovery. In: Proceedings of the 5th International Symposium on Intelligent Data Analysis, pages 230–241, Springer, 2003.

[4] D. Gamberger, A. Krstačić, G. Krstačić, N. Lavrač, M. Sebag. 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, pages 52-56, University of Aberdeen 2005.

[5] J. Demšar, B. Zupan, G. Leban. (2004). Orange: From Experimental Machine Learning to Interactive Data Mining. White Paper (www.ailab.si/orange), Faculty of Computer and Information Science, University of Ljubljana.

[6] R. Agrawal, R. Srikant. Fast Algorithms for Mining Association Rules. In: Proceedings of the 20th International Conference on Very Large Databases, pages 207–216, 1994.

[7] V. Jovanovski, N. Lavrač. Classification Rule Learning with APRIORI-C. In: Progress in Artificial Intelligence: Proceedings of the 10th Portuguese Conference on Artificial Intelligence, pages 44–51, Springer, 2001.

[8] P. Clark, T. Niblett. The CN2 induction algorithm. Machine Learning, 3(4):261–283, 1989.

[9] N. Lavrač, D. Gamberger. Relevancy in constraint-based subgroup discovery. In J.F. Boulicaut, L. De Raedt, H. Mannila (eds.) Constraint-Based Mining and Inductive Databases, Springer, 2006 (to appear).

[10] U.M. Fayyad, F.B. Irani. Multiinterval discretization of continuous-valued attributes for classification learning. In: Proceeding of the 13th Int. Joint Conference on Attificial Intelligence, pages 1022-1027. Morgan Kaufman, 1993.