RULE INDUCTION FOR SUBGROUP DISCOVERY: A CASE STUDY IN MINING UK TRAFFIC ACCIDENT DATA

Branko Kavšek, Nada Lavrač
Department of Intelligent Systems
Jozef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
Tel: +386 1 4773343; fax: +386 1 1251038
e-mail: {branko.kavsek,nada.lavrac}@ijs.si

John C Bullas
AA Foundation for Road Safety Research / CSS
Hampshire County Council
Winchester, SO23 7RX, United Kingdom
Tel+44 1962 873759, Fax +44 1962 873745
e-mail: john.bullas@hants.gov.uk

ABSTRACT

Rule learning is typically used in solving classification and prediction tasks. However, learning of classification rules can be adapted also to subgroup discovery. Such an adaptation has already been done for the CN2 rule learning algorithm. In previous work this new algorithm, called CN2-SD, has been described in detail and applied to the well known UCI data sets showing its appropriateness for subgroup discovery. This paper summarizes the modifications needed for the adaptation of the CN2 rule learner to subgroup discovery and presents some results of its application to a real-life data set of UK traffic accidents, together with an initial evaluation of results by the traffic expert.

1 INTRODUCTION

Classical rule learning algorithms were designed to construct classification and prediction rules [3], [6]. In addition to this area of machine learning, referred to as *predictive induction*, developments in *descriptive induction* have recently gained much attention. These involve mining of association rules (e.g., the APRIORI association rule learning algorithm [1]), subgroup discovery (e.g., the MIDOS subgroup discovery algorithm [9]), and other approaches to non-classificatory induction.

This paper summarizes the methodology, presented in [4], which was used for upgrading the classical rule learning algorithm CN2 [3], [2] to a subgroup discovery algorithm CN2-SD. In contrast with the CN2-SD implementation described in [4], this paper uses a new implementation of CN2-SD in which we have modified the original Boswell's implementation of the CN2 algorithm [2] to accommodate the changes needed to make it suitable to a subgroup discovery task. The goal of this paper is not to analyze in detail the deficiencies and benefits of the CN2-SD algorithm - this has already been done in [4], where the algorithm was studied on the data sets from the UCI Repository of Machine Learning [7]. The purpose of this paper is to apply the new algorithm to a real-life problem and have an expert evaluate the results.

The paper is organized as follows. In Section 2 the background for this work is explained in short: the standard CN2 rule induction algorithm and the standard CN2 heuristics, as well as the weighted relative accuracy. Section 3 presents the modified CN2 algorithm, called CN2-SD, adapting the CN2 algorithm for subgroup discovery. Sections 2 and 3 are short versions of the same sections in [4]. They are given here for the sake of completeness and to make Section 4 more understandable. Section 4 presents the real-life data set together with the experimental settings, results of experiments and expert evaluation of results. Section 5 concludes by commenting the results and giving some directions for further work.

2 BACKGROUND

This section briefly presents the backgrounds: classical CN2 rule induction algorithm and standard CN2 heuristics, as well as the weighted relative accuracy heuristic.

The CN2 Rule Induction Algorithm

CN2 is an algorithm for inducing propositional classification rules [3]. CN2 consists of two main procedures: the *search procedure* that performs beam search in order to find a single rule and the *control procedure* that repeatedly executes the search.

The search procedure performs beam search using the Laplace estimate [2] of the rule as a heuristic function. We replaced the accuracy measure with the weighted relative accuracy measure [8], defined in Equation 1. Additionally, CN2 can apply a significance test to the induced rule. The rule is considered to be significant, if it locates regularity unlikely to have occurred by chance. To test significance, CN2 uses the likelihood ratio statistic [3].

Two different control procedures are used in CN2: one for inducing an *ordered list* of rules and the other for the *unordered* case. Both ordered and unordered control procedures induce rules in a similar fashion, running the search procedure that finds the best rule, removing the examples covered by that rule and iteratively repeating this step until all examples have been covered. Detailed description of the difference between the two control procedures is given in [2].

More important than how the rules are produced is how they are interpreted. In the ordered case each rule depends on the rules that precede it, while in the unordered case each rule is interpreted separately and thus each rule represents an independent "chunk" of knowledge.

The Weighted Relative Accuracy Heuristic

Weighted relative accuracy (WRAcc) can be meaningfully applied both in the descriptive and predictive induction framework; in this paper we apply this heuristic for subgroup discovery.

We use the following notation. Let $n(Cond)$ stand for the number of instances covered by a rule $Class \leftarrow Cond$, $n(Class)$ stand for the number of examples of class $Class$, and $n(Class \cdot Cond)$ stand for the number of correctly classified examples (true positives). We use $p(Class \cdot Cond)$ etc. for the corresponding probabilities. WRAcc [5], [8] is then defined as follows:

$$WRAcc(Class \leftarrow Cond) = p(Cond) \cdot \left(p(Class \mid Cond) - p(Class)\right)$$

Equation 1: *The Weighted Relative Accuracy heuristic.*

WRAcc consists of two components: *generality* $p(Cond)$, and *relative accuracy* $p(Class \mid Cond) - p(Class)$. The second term, relative accuracy, is the accuracy gain relative to the fixed (default) rule $Class \leftarrow true$. However, it is easy to obtain high relative accuracy with highly specific rules, i.e., rules with low generality $p(Cond)$. To this end, generality is used as a "weight", so that weighted relative accuracy trades off generality of the rule ($p(Cond)$, i.e., rule coverage) and relative accuracy .

3 SUBGROUP DISCOVERY ALGORITHM CN2-SD

The main modifications of the CN2 algorithm, making it appropriate for subgroup discovery, involve the implementation of the *weighted covering algorithm* and incorporation of example weights into the weighted relative accuracy heuristic. Both modifications are briefly described below. The complete description of the changes is given in [4].

The Weighted Covering Algorithm

In the classical covering algorithm only the first few induced rules may be of interest as subgroup descriptors with sufficient coverage, since subsequently induced rules are induced from biased example subsets, i.e., subsets including only positive examples not covered by previously induced rules. This bias constrains the population for subgroup discovery in a way that is unnatural for the subgroup discovery process which is, in general, aimed at discovering interesting properties of subgroups of the entire population. In contrast, the subsequent rules induced by the weighted covering algorithm allow for discovering interesting subgroup properties of the entire population.

The weighted covering algorithm is modified in such a way that covered positive examples are not deleted from the current training set. Instead, in each run of the covering loop, the algorithm stores with each example a count how many times (with how many rules induced so far) the example has been covered. Weights derived from these example counts then appear in the computation of *WRAcc*. We have implemented two approaches:

Multiplicative weights. In the first approach, weights decrease multiplicatively. For a given parameter $\gamma < 1$, weights of covered examples decrease as follows: $e(i) = \gamma^i$, where $e(i)$ is the weight of an example being covered i times.

Additive weights. In the second approach, weights of covered examples are modified as follows: $e(i) = 1/(i+1)$.

Modified WRAcc Heuristic with Example Weights

The modification of CN2 reported in [8] affected only the heuristic function: weighted relative accuracy was used as search heuristic, instead of the Laplace heuristic of the original CN2, while everything else stayed the same. In [4], the heuristic function was further modified to enable handling example weights, which provide the means to consider different parts of the instance space in each iteration of the weighted covering algorithm.

In the *WRAcc* computation (Equation 1) all probabilities are computed by relative frequencies. An example weight measures how important it is to cover this example in the next iteration. The initial example weight $e(0) = 1$ means that the example hasn't been covered by any rule, while lower weights mean that it has already been covered by previously generated rules. The modified *WRAcc* measure is then defined as follows:

$$WRAcc(Class \leftarrow Cond) = \frac{n'(Cond)}{N'}\left(\frac{n'(Class \cdot Cond)}{n'(Cond)} - \frac{n'(Class)}{N'}\right)$$

Equation 2: *The modified WRAcc heuristic.*

where N' is the sum of the weights of all examples, $n'(Cond)$ is the sum of the weights of all covered examples, and $n'(Class \cdot Cond)$ is the sum of the weights of all correctly covered examples.

4 EXPERIMENTAL EVALUATION

In contrast with the CN2-SD implementation described in [4], this paper uses a new implementation of CN2-SD in which we have modified the original Boswell's implementation of the CN2 algorithm [2] to accommodate the changes needed to make it suitable to a subgroup discovery task.

We evaluated the new CN2-SD approach on a real-life problem, namely the TRAFFIC data set described in Section 4.1. Due to large amounts of data, some preprocessing was needed before running the experiments. The data preprocessing step is described in Section 4.2. In Section 4.3 the results of experiments are given. These results were then shown to the domain expert whose comments are presented in Section 4.4.

4.1 The TRAFFIC data set

The TRAFFIC data set includes the records of all the accidents that happened on the roads of Great Britain between years 1979 and 1999. It is a relational data set consisting of 3 related sets of data: the *ACCIDENT* data, the *VEHICLE* data and the *CASUALTY* data. The ACCIDENT data consists of the records of all accidents happened over the given period of time; VEHICLE data includes data about all the vehicles involved in those accidents; CASUALTY data includes the data about all the casualties involved in the accidents. Consider the following example: “Two vehicles crashed in a traffic accident and three people were seriously injured in the crash”. In terms of the TRAFFIC data set this is recorded as one record in the ACCIDENT set, two records in the VEHICLE set and three records in the CASUALTY set. We can also see that the three sets are related one with the other. Every separate set is described by around 20 attributes and consists of more than 5 million records.

4.2 Preprocessing of the data

The enormous quantity of data in the TRAFFIC data set makes it practically impossible to run any data mining algorithm on the whole set.

Therefore we have decided to take samples of the data set and perform the experiments on these samples, rather than on the whole data set. We focused on the ACCIDENT set of data and decided to examine only the accidents that happened in 10 districts (called Local Authorities (LAs)) across Great Britain. We have chosen the 5 areas with the most increasing trend of accidents and 5 areas with the most decreasing trend according to the results of regression analysis of the number of accidents that happened in each LA over the years. In this way we selected 10 data sets (one for each LA) with some ten thousands of examples each. We further sampled this data taking only 10% of the examples from each of the 10 sets. The characteristics of these 10 data sets are given in Table 1. Since all 10 sets are subsets of the same data set, they all have the same number of attributes (26 including the class attribute). Therefore Table 1 only gives the number of examples in each set and the distribution of the class attribute. The sets 1 to 5 represent the 5 areas with the most decreasing trend of accidents (set 1 being the “best”) and sets 6 to 10 the ones with the most increasing trend (set 6 being the “worst”). The Code numbers 1 through 10 do not correspond to the codes 1 through 10 used for Local Authorities in the Database.

LA	NO of exs.	Class Dist. (%)
1	6039	0.64/15.35/84.01
2	3627	1.15/16.80/82.04
3	2916	0.95/17.37/81.67
4	3182	1.10/19.60/79.29
5	2684	0.88/16.87/82.25
6	5487	1.35/12.74/85.90
7	1477	1.64/14.81/83.54
8	6381	1.66/17.31/81.02
9	1645	2.05/18.18/79.77
10	4375	1.82/17.11/81.05

Table 1: *Characteristics of data sets.*

Among the 26 attributes describing each of the 10 data sets we chose the attribute “accident severity” to be the class attribute. The task that we have posed was therefore to find rules that predict the severity of an accident (slight, serious or fatal) from other attributes describing the accident, such as: “road class”, “speed limit”, “light condition”, etc.

4.3 Results of experiments

We further wanted to investigate if by running CN2-SD on the data sets described in Table 1, we are able to get some rules that are typical for the 5 areas with the most increasing trend of accidents as well as rules typical for the 5 areas with the most decreasing trend. In Table 2 CN2-SD (we used the additive weights approach) and standard CN2 are compared on the 10 LA sets in terms of: “number of induced rules” (*R*), ”relative average coverage” (*CVG*) and “accuracy of rules” (*Acc*).

We have used 10-fold cross-validation to compute the accuracies of induced rule sets, whereas the number of rules and the relative average coverage were computed on rules induced from all available data. The relative average coverage measures the percentage of examples covered on average by one rule from the induced rule set. It is computed as:

$$CVG = \sum_{i=1}^{n_R} \text{covered}(i) \,/\, (n_R \cdot n_{EXS}) \, ,$$

where n_R is the number of induced rules, n_{EXS} is the number of examples in the data set and $\text{covered}(i)$ is the number of examples covered by the i-th rule in the rule set.

LA	Standard CN2			CN2-SD (additive)		
	R	**CVG**	**Acc**	**R**	**CVG**	**Acc**
1	22	6.25	83.95	10	13.08	83.94
2	19	4.78	81.98	9	13.20	82.20
3	31	1.44	79.05	11	10.66	79.15
4	25	6.54	81.46	9	14.49	81.66
5	20	5.51	81.46	14	7.49	81.23
6	29	5.26	85.88	13	10.45	85.86
7	17	5.29	81.02	9	12.96	81.02
8	34	4.37	79.73	14	8.80	79.76
9	13	6.82	83.40	14	8.40	82.74
10	28	5.05	80.97	11	11.64	80.83
Avg	**23.8**	**5.13**	**81.89**	**11.4**	**11.12**	**81.84**

Table 2: *Experimental comparison of standard CN2 with CN2-SD.*

4.4 Comments of the domain expert

We examined further the rules induced by the CN2-SD algorithm (additive weights). We focused on rules with high coverage and rules that cover a high percentage of the predicted class as those are the rule that are likely to reflect some regularity in the data. We have found a very surprising fact. One might expect the more severe the accident the greater number of people hurt up to the total number of occupants in the vehicles. Also common sense would tell the more the vehicles involved in the accident the more severe the accident. Contrary to our expectations we found two types of rules:

- rules that classify an accident as “fatal” or “serious” when just one vehicle is involved in the accident;
- rules that classify an accident as “slight” when two or more vehicles are involved and there are few casualties

We have shown these results to the domain expert who pointed out an interesting fact about collecting the data for the ACCIDENT data set.

The severity code in the ACCIDENT data set relates to the most severe injury among those reported for that Accident. Therefore a multiple vehicle accident with 1 fatal and 20 slight injuries would be classified as fatal as one fatality occurred. Each individual CASUALTY injury severity is coded in the CASUALTY data set.

Some injuries may be unreported at the accident scene, if the policeman compiles/revises the report after the event, new casualty/injury details can be reported (injuries that came to light after the event or reported for reasons relating to injury/insurance claims). However this is a very surprising fact that needs to be further investigated. We agreed with the expert that examining the ACCIDENT data set was not enough. Further examination of the VEHICLE and CASUALTY set is needed.

5 CONCLUSIONS

The comparative results in Section 4.3 (Table 2) show that CN2-SD induced on average smaller rule sets that included rules that had on average a higher coverage that those induced by the standard CN2 algorithm. The latter fact makes CN2-SD more suitable for the subgroup discovery task as each rule with high coverage represents potentially an interesting subgroup in the data. On the other hand the average accuracy of the CN2-SD rule sets was more or less the same as the accuracy of standard CN2 rules, which is very good given that the CN2-SD algorithm does not optimize rule accuracy. The above findings are not new an reflect the findings in [4].

It is worth noticing that that both CN2-SD and standard CN2 performed “worse than default” in terms of accuracy of induced rules, meaning that if we predicted the majority class (Table 1) we would have got better classification accuracy than by applying the induced rules. This fact is not surprising due to a very unbalanced class distribution and the way how the experiments were performed. Since classification was not the task addressed, we were not really interested in the accuracy of the rule sets but more in detecting interesting subgroups that the rules represented.

The most interesting finding was the rule interpretation by the domain expert. What we found in our case study was that the result of a data mining process depends not only on the accuracy of the chosen method and the data that is at hand but also on how the data was collected.

ACKNOWLEDGEMENTS

Thanks to Dragan Gamberger and Peter Flach for joint work in subgroup discovery, and to Damjan Demšar and Peter Ljubič for the help in collecting the results. The work reported in this paper was supported by the Slovenian Ministry of Education, Science and Sport and the IST-1999-11495 project Data Mining and Decision Support for Business Competitiveness: A European Virtual Enterprise.

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 (Eds.) *Advances in Knowledge Discovery and Data Mining* (pp. 307-328). AAAI Press, 1996.

[2] P. Clark and R. Boswell. Rule induction with CN2: Some recent improvements. In Y. Kodratoff, editor, *Proceedings of the 5th European Working Session on Learning*, Springer-Verlag, pp. 151-163, 1989.

[3] P. Clark and T. Niblett. The CN2 induction algorithm. *Machine Learning*, Kluwer, 3(4):261-283, 1989.

[4] N. Lavrač, P. Flach, B. Kavšek and L. Todorovski. Rule induction for subgroup discovery with CN2-SD. *Technical Report*. 2002.

[5] N. Lavrač, P. Flach and B. Zupan. Rule Evaluation Measures: A Unifying View. *In Proceedings of the Ninth International Workshop on Inductive Logic Programming*, *volume 1634 of Lecture Notes in Artificial Intelligence*: pp. 74-185. Springer- Verlag, 1999.

[6] R.S. Michalski, I. Mozetič, J. Hong, and N. Lavrač. The multi-purpose incremental learning system AQ15 and its testing application on three medical domains. *In Proc. Fifth National Conference on Artificial Intelligence*, (pp. 1041-1045), Morgan Kaufmann, 1986.

[7] P.M. Murphy and D.W. Aha. *UCIrepository of machine learning databases* [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science, 1994.

[8] L. Todorovski, P. Flach and N. Lavrač. Predictive Performance of Weighted Relative Accuracy. In D.A Zighed, J. Komorowski and J. Zytkow, editors, *Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD2000)*, Springer-Verlag, pp. 255-264, 2000.

[9] S. Wrobel. An algorithm for multi-relational discovery of subgroups. *Proc. First European Symposium on Principles of Data Mining and Knowledge Discovery*, (pp. 78-87), Springer, 1997.