Discovering Subgroups Using Descriptive Models of Adverse Outcomes in Medical Care

Gregor Stiglic[1,2], Peter Kokol[1,2]

[1]Faculty of Health Sciences,
[2]Faculty of Electrical Engineering and Computer Science,
University of Maribor, Slovenia

Corresponding author:

Assist. Prof. Gregor Stiglic, PhD
Faculty of Health Sciences
University of Maribor
Zitna ulica 15
2000 Maribor
Slovenia
Tel.: +386 2 220 4750
Fax: +386 2 220 4747
E-mail: gregor.stiglic@uni-mb.si
Website: http://ri.fzv.uni-mb.si/gstiglic

Summary

Objectives:

Hospital discharge databases store hundreds of thousands of patients. These datasets are usually used by health insurance companies to process claims from hospitals, but they also represent a rich source of information about the patterns of medical care. The proposed subgroup discovery method aims to improve the efficiency of detecting interpretable subgroups in data.

Methods:

Supervised descriptive rule discovery techniques can prove inefficient in cases when target class samples represent only an extremely small amount of all available samples. Our approach aims to balance the number of samples in target and control groups prior to subgroup discovery process. Additionally, we introduce some improvements to an existing subgroup discovery algorithm enhancing the user experience and making the descriptive data mining process and visualization of rules more user friendly.

Results:

Instance-based subspace subgroup discovery introduced in this paper is demonstrated on hospital discharge data with focus on medical errors. In general, the number of patients with a recorded diagnosis related to a medical error is relatively small in comparison to patients where medical errors did not occur. The ability to produce comprehensible and simple models with high degree of confidence, support, and predictive power using the proposed method is demonstrated.

Conclusions:

This paper introduces a subspace subgroup discovery process that can be applied in all settings where a large number of samples with relatively small number of target class samples are present. The proposed method is implemented in Weka machine learning environment and is available at http://ri.fzv.uni-mb.si/ssd.

Keywords: Subgroup Discovery, Data Mining, Descriptive Trees.

1 Introduction

In contrast to past decades when a typical medical data set in data mining research contained at most hundreds of data samples, nowadays, data miners work with datasets containing data collected from millions of patients. Consequently, the data mining techniques follow the trend, and new data mining solutions are presented to cope with huge amounts of data. This paper presents a novel approach to subgroup discovery in data sets where the group of interest consists of an extremely low percentage of all available samples. Examples of such data sets are hospital discharge data sets where data from hundred thousands of patients, together with information on hundreds of different diagnosis and procedure codes, are stored. These datasets are usually used by health insurance companies to process claims from hospitals, but they also represent a rich source of information about the patterns of medical care. Therefore, hospital discharge data have been previously used to asses risk patterns for different diseases [1] or for proposal of new approaches to comorbidity factor calculations [2]. The same data also are interesting from the economic perspective to estimate the costs of specific diseases or processes in healthcare on the global level. Such studies often focus on medical errors and increased medical costs related to them.

Our approach to the analysis of medical errors and extraction of potentially surprising patterns is based on subgroup discovery. In contrast to decision trees, the subgroup discovery methods attempt to induce a model for a specific subgroup of the population instead of a general model suitable for classification of data. However, it has to be noted that classification trees also can be interpreted in a descriptive manner, and the subgroups also can have predictive abilities. In comparison to other hospital discharges, medical error-related diagnoses represent extremely small subgroups that are appropriate for application of subgroup discovery algorithms. Additionally, our approach allows using one or more diagnosis codes as a conjunctive or disjunctive target class for subgroup discovery algorithm. The resulting rules are represented in a very intuitive discovery tree form that is visually more appropriate for analysis, compared with classical rule representation. By visually tuning the descriptive tree to fit the predefined size of a single screen, we achieve simpler representation in comparison to a long list of rules that require more time to be studied.

The remainder of this paper is organized as follows: Section 2 introduces the methodological part including the novel approach to subgroup discovery. Experimental setup and more detailed description of database are presented in Section 3. In Section 4, we present the experimental results and include some examples of extracted models. We conclude with a discussion of findings and future work directions in Section 5.

2 Methods

Subgroup discovery [3, 4] is a supervised descriptive rule discovery technique [5] for extraction of rules describing interesting relationships with regard to a target class group. Subgroup discovery techniques have been used in some experiments from medical domain recently [6, 7, 8]. Multiple measures were proposed to evaluate the quality of extracted rules, most of them focusing on maximizing the coverage and unusual statistical characteristics of the samples in the target group. In our approach, we use HotSpot algorithm, originally implemented in Weka machine learning framework [9], with some

adaptations in visualization, rule compactness, and statistical significance testing. Original implementation of Weka HotSpot algorithm is an implementation of simple subgroup discovery with a minimal support constraint that aims to maximize the confidence of extracted rules with regard to the selected target class. In contrast to most subgroup discovery techniques, the HotSpot rules are presented in a node-link structure that is called a descriptive tree in this paper. One should note that there is a significant difference between decision and descriptive trees, although classification rule discovery often is employed in real-world data mining applications in which the objective is to find interesting rather than predictive rules [10]. Another important difference between both types of trees is related to interpretation of the trees. In case of decision tree, usually, only terminal nodes represent rules, whereas in descriptive trees, each node (except the root node) represents a single rule.

Heuristic beam search, usually with small size of beam that can be defined by the user, is used to build the rules. Consequently, the rules derived using HotSpot often lack diversity, especially in cases with low branching factor. On the contrary, by increasing branching factor, the discovery trees become incomprehensible and inappropriate for visualization because of their complexity. Additionally, the user can define the minimal improvement parameter, which defines the percentage of needed improvement to extend the subgroup discovery rule. HotSpot also is able to work with continuous target values where it aims to improve confidence in specific segment of the target class.

2.1 Subspace Subgroup Discovery

In this section, we describe a novel approach to visual presentation of subgroup discovery rules in datasets with high number of samples and relatively small target subgroups. To allow the effective mining of subgroup discovery rules presented in the form of descriptive tree, we propose a few modifications to the original Weka HotSpot algorithm. Our adapted HotSpot2 and the proposed Subspace Subgroup discovery (SSD) algorithms can be obtained as Weka packages (http://ri.fzv.uni-mb.si/ssd). HotSpot2 still allows user-defined branching parameter, but it is only applied in the initial branching phase to increase the diversity of generated rules. Binary branching is used in all of the following nodes to keep the width of the top-down representation as narrow as possible. Therefore, each node obtained at the initial branching phase can contain only a single split value (numeric attributes) or a single nominal value. Additionally, we introduce the automated visual tuning option that allows totally automated fitting of the descriptive tree in predefined visual boundaries. Default values for maximal dimensions of the decision tree are set to 1280 x 800 pixels corresponding to Widescreen eXtended Graphics Array (WXGA) video standard that can be displayed on most displays in use today. To fit the tree in the predefined boundaries, a variant of binary search-based tuning is applied to minimal improvement parameter. The user can still change the minimal support constraint and influence the minimal coverage of the produced rules. As proposed by Gamberger and Lavrač [11], the default minimal rule support value in SSD is set to $\sqrt{n_t}/n$, where n_t represents the number of target class samples and, n represents all available samples.

Another improvement to the original HotSpot algorithm includes testing of statistical significance for all rules in the descriptive tree, thus introducing the components of contrast-set mining. HotSpot2 uses χ^2 test to evaluate the significance of rules comparing differences in coverage (i.e., confidence) between selected target class and control groups. Adjustment of critical values to reduce the risk of false

discoveries was done using layered critical values as proposed by Webb [12]. Layered adjustment of critical values aims to reduce the level of strictness, especially on shorter rules that often are rejected by statistical significance tests using direct adjustment of critical value. The following adjustment is used in HotSpot2:

$$\alpha'_L = \alpha/(L \times H_L)$$

where L is the current level of search, and H_L is the number of all hypotheses that should be tested on level L. In case of numerical attributes, $H_L = \sum_i 2(|A_i| - 1)$ and $H_L = \sum_i |A_i|$ for discrete attributes, where $|A_i|$ represents all possible values of attribute A_i. Rules that do not pass the χ^2 test are not removed from the tree but are marked with red color and the user decides whether to use such rules. Additionally, the unadjusted p-value is displayed in each node, that is, for each rule.

HotSpot2 presented above is an integral part of the proposed SSD approach to rule discovery in large datasets. Similar to generalization of subgroup discovery called Exceptional Model Mining by Leman et al. [13], SSD defines three groups of attributes. The first group of attributes $\{A_1^D, \dots, A_k^D\}$ represents *k* descriptive attributes, the second group $\{A_1^M, \dots, A_l^M\}$ represents *l* model attributes, and finally, the user also should define a group of *m* target attributes $\{A_1^T, \dots, A_m^T\}$ and corresponding target values $\{V_1^T, \dots, V_m^T\}$. In case of the proposed SSD approach, we use descriptive attributes A^D to select the most appropriate group of samples that will be used in the model building based on independent model attributes A^M. Target attribute-value pairs can be combined in a conjunctive or disjunctive way to define a subgroup of target samples S^T from a group of all samples S. Each sample s_i^T is used to find r most similar samples from $S^{\bar{T}}$ in space of A^D, where r represents user specified ratio of $|S^{\bar{T}}|/|S^T|$. Nearest neighbors algorithm is used to find the most similar samples. This way we are guaranteed to compare samples that are similar (e.g., have similar diagnoses and belong to similar age groups) based on descriptive attributes. Furthermore, it is well known that large sample sizes can make a small difference statistically significant [14].

In the following phase of SSD, we keep instance-based subspace $S_S = S^T \cup S^{\bar{T}}$ and use A^M and A^T to build subgroup discovery tree. HotSpot2 is used in our implementation of SSD, but the SSD framework is very general and can be combined with different supervised descriptive rule discovery techniques. As in HotSpot2, the user can define minimal support and minimal improvement of quality measure that also can be tuned automatically.

3 Experimental Settings

In this section, we demonstrate the application of SSD to mining of subgroup rules from publicly available hospital discharge data. Freely available National Hospital Discharge Survey [15] data for 2008 was used for all experiments in this study. The dataset contains hospital discharge records for 207 hospitals, which responded to the survey. Only short-stay hospitals (hospitals with an average length of

stay for all patients of less than 30 days) with at least six beds or those whose specialty is general (medical or surgical) or children's general were included in the survey. From 165,630 records available in the National Hospital Discharge Survey dataset, we selected all adult patients and reduced the dataset to 140,124 records. Each NHDS record contains the personal characteristics of the patient, including birth date or age, sex, race, and marital status; administrative information, including length of stay and discharge status; and medical information, including diagnoses, surgical and nonsurgical procedures. Up to seven diagnosis codes and an admitting diagnosis code were assigned to each sample. In addition, if the medical information included surgical or nonsurgical procedures, a maximum of four codes for these procedures was assigned (62.5% of all samples). The International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9-CM) [16], was used for coding diagnoses and procedures.

Further processing of the original dataset was needed to transform the dataset, especially the diagnosis and procedure codes that were transformed into a multi-label format. A new, binary attribute, was assigned to each diagnosis or procedure code, denoting the presence or absence of the disease or procedure. Only diagnosis and procedure codes that occur in more than 50 records were used in this step, to reduce the size of the final dataset. Altogether, 1575 most frequent diagnosis and 429 procedure codes were used to construct a dataset. Usual ICD-9-CM code representation consists of the first three digits representing a major group of diagnoses, followed by two optional digits, defining a specialization of the diagnosis. Additional to conventional diagnosis codes, ICD-9-CM also uses the so called e-codes that start with letter E. Those are used for the classification of environmental events, circumstances, and conditions as the cause of injury, poisoning, and other adverse effects. Most medical errors can be characterized by E-codes, although every E-code does not necessary represent a medical error. We used diagnosis codes that are most frequently connected to medical errors from the study by Shreve et al. [17]. Only the most frequent group of errors, that represent an actual medical error in more than 90% of cases was used.

In our experiments, all diagnoses were selected as descriptive attributes with all procedure attributes used as model attributes. This experimental setup allows observation of potentially unexpected relations between target classes and used procedures in a subgroup of adverse outcome records. By measuring statistical significance of acquired rules, it also is possible to compare the differences between target and control group classes. Following the experimental setup by Shreve et al., we set the target versus non-target ratio to 4 in the initial experiments. Initial branching factor and maximal tree depth settings were set to 3 and 5, respectively. Because there were almost 1000 samples available for the target class, another experiment was performed where target versus non-target ratio of 1 was used.

4 Results

In this section, we demonstrate a descriptive tree for postoperative infections target attributes for larger dataset (4:1 ratio) and rules from the smaller dataset (1:1 ratio). According to Shreve et al. [14], two target classes can be used to identify postoperative infections that are usually caused by human error – that is, postoperative seroma (D998.51) and other postoperative infection (D998.59).

Figure 1 represents the descriptive tree from the first experimental setting. It reveals some interesting patterns related to postoperative infections, where very compact representation of descriptive rules

was obtained. The first rule has only one condition – that is, $P99.04 \rightarrow \mathrm{D998.51|D998.59}$, where P99.04 stands for transfusion of packed cells. Rules with only one condition usually represent known facts already confirmed by previous studies (e.g., [18]). The second branch of the tree represents the rule $P54.91$ & $P88.01 \rightarrow \mathrm{D998.51|D998.59}$, where P54.91 (percutaneous abdominal drainage) and absence of P88.01 (computerized axial tomography of abdomen) characterize postoperative infections with a rule confidence of 79.01% and a strong statistical significance ($p < 0.01$).

Fig. 1. Descriptive tree for postoperative infections target attributes (D998.51 and D998.59).

In the second experiment, the same number of samples from both classes was used. Top 5 rules from the descriptive tree are presented in Table 1 including some quantitative measures for each rule. Altogether the descriptive tree consisted of 20 rules that still fit the single screen boundaries set for the automated visual tuning. Table 2 contains all diagnosis codes used in Table 1 for easier interpretation of obtained results. More detailed description of used quantitative measures is available in Witten et al. [19].

Table 1. Quantitative measures for top 5 rules on the equal proportions dataset.

In case of the second experiment, one can observe even higher confidence values for obtained rules. Most of the rules achieved confidence scores higher than 90%, with one of the rules reaching the support of 195 cases. More specifically, rule number 2 is related to postoperative infections in 195 of 212 records it covers. According to the study by Azevedo and Jorge [20], one should observe the conviction measure to estimate the predictive potential of rules. Again, rule number 2 demonstrates the highest conviction of 5.89.

Table 2. Procedure code descriptions for interpretation of rules from Table 1.

Similar to a study of Abu-Hanna et al. [21], we conducted an additional experiment where we compare our proposed method to a classical decision tree algorithm. A Weka based implementation of CART [22] called SimpleCart was used to build a decision tree with the same target attribute as in our subgroup discovery experiments. Using the default settings in Weka environment results in large and incomprehensible decision trees. In case of the first experiment (totally balanced data) we obtained a

decision tree consisting of 131 nodes, whereas in using the second dataset, the resulting decision tree consisted of 121 nodes. To simplify the decision trees, we increased a minimal support in leaves setting. In accordance to $\sqrt{n_t}/n$ equation, we set minimal support to 31 in both datasets. Using higher minimal number of samples in nodes, we obtained much simpler and more comprehensive decision trees. Both trees contained 2 decision nodes and 3 leaves. The root node in both trees contained the same procedure code – that is, P38.93 (venous catheterization, not elsewhere classified). The only difference between the trees was the choice of the second decision node – P86.22 (excisional debridement Of wound, infection, or burn) for the first dataset and P54.59 for the second. The obtained results show fairly similar derived models in case of evenly balanced dataset where the number of obtained rules from a decision tree and a tree derived from our proposed method both contain 3 rules. In case of the second dataset SimpleCart produced a simpler model compared with our proposed method. However, the average confidence of the three decision tree rules (0.81, 0.52, and 0.62) was much lower than that of the descriptive tree rules.

5 Discussion and Conclusions

As demonstrated in the previous sections, there is a big potential in intelligent analysis of hospital discharge data with focus on improving safety and quality in healthcare. Our experiments show the analysis workflow for identification of related procedures, based on selected target groups identified by diagnoses highly related to adverse outcomes in hospitals. In section 4, we demonstrate the ability of the proposed method to identify groups of patients with co-occuring diagnoses that were previously not identified in the literature. By temporal analysis of the available data, it also would be possible to track changes and identify specific trends in patient safety risks. A similar approach using temporal organization of hospital discharge data and classification was presented in our recent study [23]. Based on such results, it is possible to organize global campaigns aimed at warning and educating medical personnel in hospitals (e.g., the example from Section 4 could be used to alert the personnel of especially frequent cases of postoperative infections in specific subgroups of patients). On the other hand, intelligent analysis of hospital discharge data can be used to extract new knowledge in the form of different disease or procedure combinations that were never thought of being associated with high risk for patient. The main contributions of the proposed method lie in the following aspects: (1) solving the problem of the highly unbalanced data using the nearest neighbor-based subsampling to allow statistical testing of subgroup significance and (2) visual optimization of descriptive trees that allow better interpretability.

The primary focus of this paper is introduction of SSD methodology that can be applied in all settings where a large number of samples with relatively small sample of target class are present. There are a lot of possibilities to improve the current versions of SSD and HotSpot2. Currently, a very simple subgroup quality metric focusing only on improvement of rule confidence is used. It would be desirable to support more rule quality metrics, although this also would require more knowledge from users, when deciding which metric to use. This study and SSD implementation uses only discrete target attributes when multiple target attributes are used, although it is possible to use numerical target attributes in the original HotSpot implementation. Therefore, it would be possible to offer numerical target support also

for SSD to extend the possibilities of data analysis. Further improvements of the existing algorithms could be achieved by rule pruning algorithms.

References

1. Li J, Fu AW, Fahey P. Efficient discovery of risk patterns in medical data. Artificial Intelligence in Medicine 2009; 45: 77–89.

2. Li B, Evans D, Faris P, Dean S, Quan H. Risk adjustment performance of Charlson and Elixhauser comorbidities in ICD-9 and ICD-10 administrative databases. BMC Health Serv Res 2008; 8:12.

3. Kloesgen W. Explora: a multipattern and multistrategy discovery assistant. In: Advances in Knowledge discovery and data mining. American Association for Artificial Intelligence 1996; 249-271.

4. Wrobel S. An algorithm for multi-relational discovery of subgroups. In: Proceedings of the 1st European symposium on principles of data mining and knowledge discovery 1997, vol 1263. Springer, LNAI, 78-87.

5. Kralj-Novak P, Lavrac N, Webb GI. Supervised descriptive rule discovery: a unifying survey of constrast set, emerging pattern and subgroup mining. J Mach Learn Res 2009; 10:377-403.

6. Lavrac N, Cestnik B, Gamberger D, Flach PA. Decision support through subgroup discovery: three case studies and the lessons learned. Machine Learning 2004; 57:115-143.

7. Nannings B, Abu-Hanna A, de Jonge E. Applying PRIM (Patient Rule Induction Method) and logistic regression for selecting high-risk subgroups in very elderly ICU patients. International Journal of Medical Informatics 2008; 77(4):272-279.

8. Nannings B, Bosman RJ, Abu-Hanna A. A subgroup discovery approach for scrutinizing blood glucose management guidelines by the identification of hyperglycemia determinants in ICU patients. Methods Inf Med. 2008;47(6):480-488.

9. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten, IH. The WEKA Data Mining Software: An Update. SIGKDD Explorations 2009; 11:1.

10. Webb G, Butler S, Newlands D, On Detecting Differences between Groups, in Proceedings of the Ninth ACM SIGKDD Int'l Conference on Knowledge Discovery and Data Mining 2003. 256-265.

11. Gamberger D, Lavrac N. Expert-guided subgroup discovery: Methodology and application. Journal of Artificial Intelligence Research 2002. 17:501-527.

12. Webb GI. Layered critical values: a powerful direct-adjustment approach to discovering significant patterns. Mach. Learn. 2008; 71:307-323.

13. Leman D, Feelders A, Knobbe A. Exceptional model mining. In: Proceedings of the ECML/PKDD 2008; 2:1–16.

14. Miller R and Siegmund D. Maximally selected chi-square statistics. Biometrics 1982; 38:1011-1016.

15. National Center for Health Statistics, National Hospital Discharge Survey (NHDS) data, U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics, Hyattsville, Maryland, available at: http://www.cdc.gov/nchs/nhds.htm (2008)

16. U.S. Department of Health and Human Services. Centers for Disease Control and Prevention, Centers for Medicare and Medicaid Services. Official version International Classification of Diseases, Ninth Revision, Clinical Modification, Sixth Edition. DHHS Pub No. (PHS) 06-1260 (2006)

17. Shreve J, van Den Bos J, Gray T, Halford M, Rustagi K, Ziemkiewicz E. The Economic Measurement of Medical Errors. Society of Actuaries 2010.

18. Curns AT, Steiner CA, Sejvar JJ, Schonberger LB. Hospital charges attributable to a primary diagnosis of infectious diseases in older adults in the United States, 1998 to 2004. J Am Geriatr Soc 2008; 56:969–75.

19. Witten IH, Frank E. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann; 2011.

20. Azevedo PJ, Jorge AM. Comparing rule measures for predictive association rules. In ECML '07: Proceedings of the 18th European conference on Machine Learning 2007; 510-517.

21. Abu-Hanna A, Nannings B, Dongelmans D, Hasman A. PRIM versus CART in subgroup discovery: When patience is harmful. Journal of Biomedical Informatics 2010; 43:701-708.

22. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees, statistics/probability series. Belmont, California, USA: Wadsworth Publishing Company; 1984.

23. Stiglic G, Kokol P. Interpretability of Sudden Concept Drift in Medical Informatics Domain. In ICDM-W ’11: Workshop proceedings of the 11th International Conference on Data Mining 2011; In press.

Fig. 1. Descriptive tree for postoperative infection target attributes (D998.51 and D998.59).

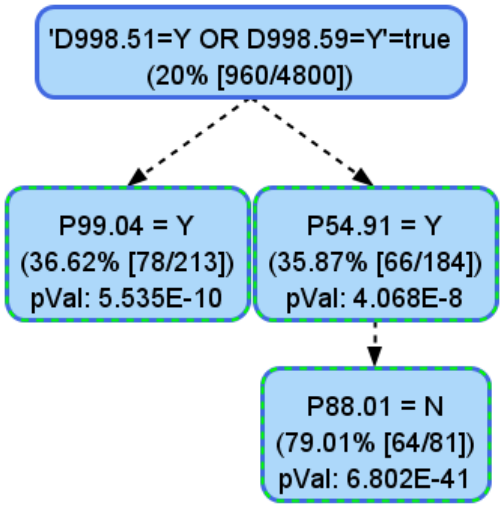

Table 1. Quantitative measures (support, confidence, lift, leverage and conviction) for top 5 rules on the equal proportions dataset.

#	Rule	Supp.	Conf.	Lift	Lev.	Conv.
1	$\overline{37.26}\ AND\ 54.59\ AND\ \overline{40.11}\ AND\ \overline{68.29}\ AND\ \overline{47.19}$	40	0.93	1.85	0.01	5.00
2	$\overline{37.26}\ AND\ 38.93\ AND\ \overline{46.73}\ AND\ \overline{86.69}$	212	0.92	1.84	0.05	5.89
3	$\overline{37.26}\ AND\ 54.59\ AND\ \overline{40.11}\ AND\ \overline{68.29}\ AND\ \overline{45.76}$	37	0.92	1.84	0.01	4.63
4	$\overline{37.26}\ AND\ 54.59\ AND\ \overline{40.11}\ AND\ \overline{47.19}\ AND\ \overline{99.04}$	37	0.92	1.84	0.01	4.63
5	$\overline{37.26}\ AND\ 54.59\ AND\ \overline{40.11}\ AND\ \overline{47.19}\ AND\ \overline{45.76}$	37	0.92	1.84	0.01	4.63

Table 2. Procedure code descriptions for interpretation of rules from Table 1.

Code	ICD-9 CM Description
37.26	Catheter based invasive electrophysiologic testing
38.93	Venous catheterization, not elsewhere classified
45.76	Open and other sigmoidectomy
46.73	Suture of laceration of small intestine, except duodenum
47.19	Other (non-laparoscopic) incidental appendectomy
54.59	Other (non-laparoscopic) lysis of peritoneal adhesions
68.29	Other excision or destruction of lesion of uterus
86.69	Other skin graft to other sites
99.04	Transfusion of packed cells