Technical Report

Department of Computer Science
and Engineering
University of Minnesota
4-192 Keller Hall
200 Union Street SE
Minneapolis, MN 55455-0159 USA

TR 14-028

Clinical Variable Relationship Evaluation using Decision Tree Rule Extraction

Chandrima Sarkar, Jaideep Srivastava

November 10, 2014

Clinical Variable Relationship Evaluation using Decision Tree Rule Extraction

Chandrima Sarkar,[1*] Jaideep Srivastava,[1]

[1] Department of Computer Science and Engineering, University of Minnesota, Twin Cities

[*]To whom correspondence should be addressed; E-mail: sarkar@cs.umn.edu

Evaluating associations and relationships between variable is a very challenging and important problem in the domain of medicine and clinical data analysis. Not many classification methods have been tried in the literature to tackle this problem. Decision Tree is one of the most important data mining methods that can be used, due to their property of implicitly performing variable screening or feature selection and their requirement of relatively little effort from users for data preparation. However, a major drawback associated with the use of decision tree for decision making is their lack of interpret-able capability specially when using tools like Weka. Though decision trees can achieve a high predictive accuracy rate, the reasoning behind how they reach their decisions is not readily available. But this problem can be handled very easily if the decision tree can be utilized by extracting their rules and analyzing these rules. In this paper we present an approach for extracting rules from the decision tree which can be utilized for determining relationship between clinical variables. Furthermore, we also discuss how these rules can be visualized in a compact and intuitive tabular format that facilitates easy analysis.

It is concluded that decision tree rule extraction can be considered as powerful analysis tools that allow us to facilitate analysis of clinical variables and its association.

Introduction

Data mining is the process of finding hidden patterns and trends in databases and using that information to build predictive models (*6*). It can also be considered as data segmentation, selection, exploration and building models using the vast data stores to discover previously unknown patterns in various domains such as healthcare (*8*) (*16*), (*18*), (*15*) (*14*), social media analysis (*13*), (*19*), (*1*), (*10*), (*7*), (*9*), (*11*), finances and various other domains (*17*). From the past few decades, data mining has been used extensively in various areas of decision making and decision analysis by financial institutions, for credit scoring and fraud detection; marketers, for direct marketing and cross-selling or up-selling; retailers, for market segmentation and store layout; and manufacturers, for quality control and maintenance scheduling. Little has been explored in the healthcare domain using data mining. Some of the researches have been done for predictive analysis in certain specifies healthcare domains like diagnosis and treatment (*15*). But little research has been done in order to exploit the immense data resource of clinical domain in healthcare. Due to recent rise in the healthcare costs, many reforms have been made which aims at suing every kinds of technology to eradicate the root cause of the extreme healthcare costs. Even the hospitals in US are trying to look into this matter by reducing their hospital costs. They are aiming at reducing healthcare expenditure as well as costs of patients. Due to the huge data resource that is available nowadays to the researchers, it is extremely important to utilize it using data mining for finding hidden patterns and analyze them for finding the association and relationship between clinical variables. The reason behind aiming for data mining in healthcare is that, it is extremely important from a clinicians perspective that the true

and hidden relationships between feature domain variables be explored. This exploration and explanation regarding a clinical features behavior can help the clinicians and physicians, trace and map the characteristics behind a variables certain behavior. For example in a healthcare database, data such as hospital records where information of a particular patient is stored. This kind of information of a patient contains important details of a patients mental and physical condition. This data also contain every kind of patterns of patient behavior such as number of visits, total number of health problems encountered and recovery patterns. Thus this kind of clinical data is huge and sparse. This is because, all patients have every kind of information listed with their account. For example Blood pressure, Blood count, pain recurrence pattern, uric acid level, age, sex etc. The healthcare data generally has huge numbers of features (The data has over 1000) and should be considered as ultra-high dimensional data. But handling such ultra-high dimensional data is not easy. They are most of the time sparse. There has been research that has shown that, sparseness increases with increase in dimension of data (*20*). Clinical and hospital data are almost always very sparse and have large number of features. As a result these kinds of data require special handling using improved data mining techniques. In this paper, we therefore present an improved approach of using decision tree by decision rule extraction and analysis of these rules for prediction of potentially preventable events defined in the next few sections.

1 Related Works

Little work is known to have been done that has explicitly done an intensive and extensive analysis of the Decision tree rules for determining relationships between feature domain variables (*2*), (*5*). To the best of our knowledge, no work has been done in healthcare utilizing rule extraction technique for generating meaningful and accurate rule sets and hence, using these rule set for finding relationships between clinical variables. A lot of research has been conducted

for improving the performance, scalability and robustness of decision trees (*21*), (*3*). But to the best of our knowledge, extracting decision tree rules and analysis of these rules for obtaining a more meaningful conclusion has hardly been done. Hence, in this paper we aim at providing researchers a tool and an approach which can be utilized for a more meaningful insight into the intricacies of decision tree rules and thus obtaining results that has been previously unknown. Using simple statistical analysis of individual features for analyzing the significance and important relationship between the variables is nearly impossible due to the extensive structure of the data with huge number of variables. As a result data mining is the most important resort for such scenarios where finding out the interaction between variables is the most important concern. One such example is the predicting and analyzing the rules that govern the prediction of potentially preventable events.

2 Preliminaries

A. Decision trees

A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm. Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal. Advantages of decision tree are - Decision tree are easy to understand and interpret. Have value even with little hard data. Important insights can be generated based on experts describing a situation (its alternatives, probabilities, and costs) and their preferences for outcomes. New constraints can be added Worst, best and expected values can be determined for different scenarios Uses a white box model Can be combined with other decision techniques. Disadvantages of decision trees are - In certain cases datasets (including categorical variables) with large number of different levels, information gain in decision trees are biased in favor of those attributes with more

levels.[8] In cases where many values are uncertain and/or if many outcomes are linked, calculations gets highly complex. B. Potentially Preventable Events (PPE) Potentially preventable events (PPEs) are hospital admissions, readmissions, complications or emergency department (ED) visits that could have been avoided if the patient had been given and complied with appropriate interventions in the ambulatory setting. For identifying PPE for a particular patient it is very important to find out the hidden primary conditions which are responsible for identifying this patient to this particular risk group. Finding out the perfect combination of patient conditions and its empirical standing that has caused a patient to fall in the risk group of PPE is an important problem for determining the root behind the prediction. In this paper we use rule extraction for finding this for analyzing Potentially Preventable Events. C. Clinical Variables / Features Clinical variables are particular patient health condition records or other activity records in a hospital data. These patient conditions are an important source of information for finding hidden factors that can lead us to detect and analyze the main reason for a patient who can be detected as susceptible to potentially preventable events. Finding out the relationship between these clinical variables and empirical standing of these relationship are very crucial for success in predicting PPE and hence, reducing some of the important healthcare issues. Using this approach, the root cause behind PPE can be detected in a global scale as well as a patient specific condition analysis can be executed for the better understanding of PPE. Detecting the combination of patient specific conditions can highly assist care providers in warning a patient regarding the impending danger, incurring significant survival benefits to this patient. The main of this entire model is to assist the care providers in detecting PPE and thus saving patient lives while cutting back on wasteful hospital resources.

3 Proposed Approach

We at first used J48 Decision tree model building facility from Weka (*4*). We use J48 Parser to build the initial base parser from the tree. This facility was provided by (*12*). Then we used Java technology with embedded Weka resources (jars) to build the final parsing of the decision tree. The decision tree was spliced into individual decision rules. This decision tree model was built for predicting potentially preventable events. More structuring of the program was conducted to ensure a visually interpretable and user friendly interface of analyzing the rule extraction results. A tabular format was used for interpreting the results. Results for each individual patient was determined and compared with the normal range of clinical conditions. Each rule is a combination of clinical conditions pertaining to a particular patient either being detected PPE or not detected PPE. Many other Model statistics and Model evaluation measures were determined such as rule accuracy, rule coverage (the number of patients covered by a single rule), error rate of the model, Rule confidence, model specificity and model sensitivity. These statistics are extremely important in determining which variables are most often related in having a PPE. For example rule coverage enables us to determine which rule is the most observed rule in patients responsible for PPE. This rules also gives us an idea as to which features are the most significant for prediction purposes.

4 Data Set

Data was extracted from EHR data from the Enterprise Data Warehouse. Due to strict regulation of health data privacy, actual data source has not been disclosed in this manuscript. Data extracted comprised of demographic information, lab results, flow sheets, problem lists and clinic visits for the patients. This data was pre-processed to handle missing values and aggregate other pieces of data to prepare for data mining tasks. Data collected for developing the model, was

EHR data from 2009 to 2011.

5 RESULTS

Results obtained were sent to the clinicians for further analysis. Counts of number of visits and number of chronic diseases present in a patient are among the other factors. Presence of Mood disorder, Anxiety disorders, back pain and Headache were among some of the detected significant conditions that should be considered for a better prediction outcome for PPE. Results of decision rule mining was very patient specific and yielded outcomes such as for a particular patient secondary malignancy with a value higher than a threshold and acute abdominal pain was detected as a PPE. Due to strict regulation of health data privacy, specific results are not presented in this manuscript.

6 Conclusion

We demonstrated that decision rule mining can be used to efficiently select patient conditions or features important to identify the optimal conditions for a probable PPE risk group. This entire approach has the ability to indicate the significant patient conditions as well as to eliminate conditions which are not reliable for prediction problems and the relationships among these conditions. Future work includes building an approach which will include pre-processing of the sparse data and performing co-clustering before extracting rules from their decision tree model. In this way the drawbacks of sparseness and ultra-high dimensionality due to curse of dimensionality (*20*) can be eradicated and used in healthcare for better findings.

Acknowledgment

We would like to thank Dr. Prasanna Desikan for his feedback, support and critical reviews during developing this research. We would like to thank Atanu Roy for his critical reviews and technical feedback during developing this research.

References and Notes

1. Muhammad Aurangzeb Ahmad, Brian Keegan, Atanu Roy, Dmitri Williams, Jaideep Srivastava, and Noshir S. Contractor, *Guilt by association?: network based propagation approaches for gold farmer detection*, Advances in Social Networks Analysis and Mining 2013, ASONAM '13, Niagara, ON, Canada - August 25 - 29, 2013, 2013, pp. 121–126.

2. Bart Baesens, Rudy Setiono, Christophe Mues, and Jan Vanthienen, *Using neural network rule extraction and decision tables for credit-risk evaluation*, Management Science **49** (2003), no. 3, 312–329.

3. Houtao Deng, George Runger, and Eugene Tuv, *Bias of importance measures for multi-valued attributes and solutions*, Artificial Neural Networks and Machine Learning–ICANN 2011, Springer, 2011, pp. 293–300.

4. Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten, *The weka data mining software: an update*, ACM SIGKDD explorations newsletter **11** (2009), no. 1, 10–18.

5. Jong Woo Kim, Byung Hun Lee, Michael J Shaw, Hsin-Lu Chang, and Matthew Nelson, *Application of decision-tree induction techniques to personalized advertisements on internet storefronts*, International Journal of Electronic Commerce **5** (2001), 45–62.

6. Hian Chye Koh, Gerald Tan, et al., *Data mining applications in healthcare*, Journal of Healthcare Information ManagementVol **19** (2011), no. 2, 65.

7. Alice Leung, William Dron, John P. Hancock, Matthew Aguirre, Jon Purnell, Jiawei Han, Chi Wang, Jaideep Srivastava, Amogh Mahapatra, Atanu Roy, and Lisa Scott, *Social patterns: Community detection using behavior-generated network datasets*, Proceedings of the 2nd IEEE Network Science Workshop, NSW 2013, April 29 - May 1, 2013, Thayer Hotel, West Point, NY, USA, 2013, pp. 82–89.

8. Anne Milley, *Healthcare and data mining using data for clinical, customer service and financial results*, Health Management Technology **21** (2000), no. 8, 44–45.

9. Atanu Roy, Muhammad Aurangzeb Ahmad, Chandrima Sarkar, Brian Keegan, and Jaideep Srivastava, *The ones that got away: False negative estimation based approaches for gold farmer detection*, 2012 International Conference on Privacy, Security, Risk and Trust, PASSAT 2012, and 2012 International Confernece on Social Computing, SocialCom 2012, Amsterdam, Netherlands, September 3-5, 2012, 2012, pp. 328–337.

10. Atanu Roy, Zoheb Hassan Borbora, and Jaideep Srivastava, *Socialization and trust formation: a mutual reinforcement? an exploratory analysis in an online virtual setting*, Advances in Social Networks Analysis and Mining 2013, ASONAM '13, Niagara, ON, Canada - August 25 - 29, 2013, 2013, pp. 653–660.

11. Atanu Roy, Chandrima Sarkar, and Rafal A. Angryk, *Using taxonomies to perform aggregated querying over imprecise data*, ICDMW 2010, The 10th IEEE International Conference on Data Mining Workshops, Sydney, Australia, 13 December 2010, 2010, pp. 989–996.

12. Daniel Salber, Anind K Dey, and Gregory D Abowd, *The context toolkit: aiding the development of context-enabled applications*, Proceedings of the SIGCHI conference on Human Factors in Computing Systems, ACM, 1999, pp. 434–441.

13. Chandrima Sarkar, Sumit Bhatia, Arvind Agarwal, and Juan Li, *Feature analysis for computational personality recognition using youtube personality data set*, Proceedings of the 2014 ACM Multi Media on Workshop on Computational Personality Recognition, ACM, 2014, pp. 11–14.

14. Chandrima Sarkar, Sarah Cooley, and Jaideep Srivastava, *Robust feature selection technique using rank aggregation*, Applied Artificial Intelligence **28** (2014), no. 3, 243–257.

15. Chandrima Sarkar, Sarah A. Cooley, and Jaideep Srivastava, *Improved feature selection for hematopoietic cell transplantation outcome prediction using rank aggregation*, Computer Science and Information Systems (FedCSIS), 2012 Federated Conference on, IEEE, 2012, pp. 221–226.

16. Chandrima Sarkar, Prasanna Desikan, and Jaideep Srivastava, *Correlation based feature selection using rank aggregation for an improved prediction of potentially preventable events*, (2013).

17. Chandrima Sarkar and Jaideeep Srivastava, *Predictive overlapping co-clustering*, arXiv preprint arXiv:1403.1942 (2014).

18. Chandrima Sarkar and Jaideep Srivastava, *Impact of density of lab data in ehr for prediction of potentially preventable events*, Healthcare Informatics (ICHI), 2013 IEEE International Conference on, IEEE, 2013, pp. 529–534.

19. Ayush Singhal, Atanu Roy, and Jaideep Srivastava, *Understanding co-evolution in large multi-relational social networks*, CoRR **abs/1407.2883** (2014).

20. Michael Steinbach, Levent Ertöz, and Vipin Kumar, *The challenges of clustering high dimensional data*, New Directions in Statistical Physics, Springer, 2004, pp. 273–309.

21. Yufei Yuan and Michael J Shaw, *Induction of fuzzy decision trees*, Fuzzy Sets and systems **69** (1995), no. 2, 125–139.