Applying info vis in different stage of data mining: A survey

Keqian Li
keqianli@cs.ubc.ca

1 Description of Task

The exploding growth of digital data sets in the information era and its immensurable potential value has led to the rise of several new academic disciplines. Most analytical techniques such as is commonly defined as neural networks, association rule, numerous clustering and classification methods are covered in the field of Data Mining, the extraction of implicit, previously unknown, and potentially useful information from data.[16] On the other hand, the field of Information Visualization focuses on visual reproduction of the data. The rationale is that "visual representations and interaction techniques take advantage of the human eyes broad bandwidth pathway into the mind to allow users to see, explore, and understand large amounts of information at once." [14] Both fields hold the aim of processing real world data to facilitate further application.However there is a sharp difference between the two subject. info vis primarily focus on producing output of graphic format based on original structure of data which usually involves human sense making[10] to further the analysis. Data mining, on the other hand, use automatic algorithmic techniques in such analysis process, to discover derived structure of data or directly provide off the shelf prediction for the objective of interest[16] .

While the two communities advocates different approaches of problem solving and partly represent contrary side of the long term phylosophical debate of Hypothesis testing vs. exploratory data analysis[6], the noticecable overlap of ultimate goal as well as application scenario has lead to possible light of collaboration. The vision of combining the sophiscated algorithmic techniques from data mining as well as the intuitivity and interactivity of information vis is tempting. In this paper, we attempt to survery recent researches and real world systems integrating the wisdom in two fields towards more effective and efficient data analytics. More specifically, we study the intersection from a data mining point of view, explore how information vis can be applied to complement and improve the different stages of data mining through established theories for optimized visual presentation as well as practical toolsets for rapid development.

We organize the survey by firstly identify three stages of typical process of data mining that can greatly benefit from visualization. Data mining research usually come with the stage of **preliminery analysis of data** to show the basic characterstic of data before heavy data

mining algorithm is applied, followed by the stage of **model construction**, then finally the **model output and evaluation**. We devote section 2 to investigate how info vis can be of use in this context. More specifically, we explore in the task usually termed as **exploratory data analysis**, how info vis can help human discover patterns by browse and navigation through the raw data. Then we devote section 3 to survey how info vis can help **interactive model construction** in which visualization is provided for partial results of the current construction of model while user can provide feedback to change the behavior of the process. After that, we investigate **visualization of model output and evaluation** in section 4. The goal in this stage is to use the info vis to convey the result intuitively to facilitate the process of understanding and further adjusting the model. We then discuss some real world systems integrating the two techniques for data exploration and decision support in section 5 and finally conclude in section 6.

2 Previous Work on Visual preliminery analysis of data

The task of data mining typically come with the a light weight stage of **preliminery analysis of data** before main data mining algorithm is applied. The purpose is to show the basic characterstic of data and help form a quick grasp of intuitive understanding of the data to gather information. The insight formed in this stage can greatly help data miners to formulate the strategy of further heavy weighted analysis. This is an area that much need the wisdom of Information Visualization. The focus of Visual preliminery analysis of data is to use visualization to draw specific conclusions to inspire and facilitate further data mining approach. Below we review visualization techniques for different techniques and different application field to illustrate how using the techniques of info vis can draw significant insight about the deeper structure. [11] [13] [9] [2].

3 Previous Work on Interactive model construction

Data mining is usually known as the highly automated model where the algorithm designed to optimize an objective with techniques from computational or statistically literature do off the scene job. In recent years, however, marrying the classical techniques with the idea of info vis has stimulated the research on **Interactive model construction**, where visualization is provided for partial results of the current construction of model while a user can provide feedback to change the behavior of the process. One of the advantages of this approach is that the user can integrate background knowledge into the modelling stage, which could be very valueable. Also, as a sub field of the general model visualization, this approach can greatly increase interpretability of the model. Below we show several examples of such technique. [15] [12] [8]

4 Previous Work on Visualization of model output and evaluation

The output of data mining is usually a set of decision rules of abstract form. Due the real world complex nature of data, the output usually reach a very large scale. The info vis techniques can be used to address this issue by significantly increase interpretability of the model, which is very critical for human decision makders to apply the model output, or reflect on the results to refine the strategy of model training. More specifically, the research direction of **visualization of model output and evaluation** studies how to use info vis to convey the result intuitively.

One approach of applying info vis to this context is to visualize model agnostic parameters and focus on visualizing the evaluate metrics of the performance of alg. [4][1].

Another pile of work focusing of visualization under specific machine learning model to expose its internal structure. By providing an easy-to-interpret visualization the analysts can gain insight and study the effects of predictive factors. One good example is Association Rules, which are one of the most widespread data mining tools because they can be easily mined, even from very huge database. The counterpart is that a massive effort is required (due to the large number of rules usually mined) in order to make actionable the retained knowledge. [3] [7] [5].

5 Personal Expertise

I'm interested area of data mining and did 1 year research in UBC data mining lab. Data mining is my focus in the master program. However I don't have previous exprience in info vis but the the combining the technique info vis to data mining seems promising and very beneficial to the area of data mining.

6 milestones and schedule

- **Nov. 14.** Do a comprehensive literature search and extensive reading to gain a overview of how each subarea of the survey is developed from recent research work. Draw a concrete introduction and conclusion based on the literature search as well as detailed overview of each subarea and list the noticable work together with basic understandy.

- **Nov. 30.** Drill down each subarea of the survey and linearly progress through the subareas. Finish the first draft of survey report.

- **Dec. 10.** Revise the paper and incorporate possible feedbacks. Prepare for the final version of report and the presentation.

References

[1] B. Alsallakh, A. Hanbury, H. Hauser, S. Miksch, and A. Rauber. Visual methods for analyzing probabilistic classification data. 2014.

[2] M. H. Böhlen, L. Bukauskas, A. Mazeika, and P. Mylov. The 3dvdm approach: A case study with clickstream data. In *Visual Data Mining*, pages 13–29. Springer, 2008.

[3] D. Bruzzese and C. Davino. Visual mining of association rules. In *Visual Data Mining*, pages 103–122. Springer, 2008.

[4] A. Bykov and S. Wang. Interaction visualizations for supervised learning. 2009.

[5] D. Caragea, D. Cook, and V. G. Honavar. Gaining insights into support vector machine pattern classifiers using projection-based tour methods. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 251–256. ACM, 2001.

[6] J. F. W. Herschel. *A preliminary discourse on the study of natural philosophy*, volume 1. Longman, Rees, Orme, Brown and Green, 1831.

[7] A. Jakulin, M. Možina, J. Demšar, I. Bratko, and B. Zupan. Nomograms for visualizing support vector machines. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 108–117. ACM, 2005.

[8] O. D. Lampe and H. Hauser. Interactive model prototyping in visualization space.

[9] A. Mazeika, M. H. Böhlen, and D. Trivellato. Analysis and interpretation of visual hierarchical heavy hitters of binary relations. In *Advances in Databases and Information Systems*, pages 168–183. Springer, 2008.

[10] P. Pirolli and S. Card. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of International Conference on Intelligence Analysis*, volume 5, pages 2–4. Mitre McLean, VA, 2005.

[11] Y. Qian and K. Zhang. The role of visualization in effective data cleaning. In *Proceedings of the 2005 ACM symposium on Applied computing*, pages 1239–1243. ACM, 2005.

[12] C. Seifert and E. Lex. A visualization to investigate and give feedback to classifiers. In *Proceedings of the European Conference on Visualization (EuroVis), Berlin*, 2009.

[13] S. J. Simoff and J. Galloway. Visual discovery of network patterns of interaction between attributes. In *Visual Data Mining*, pages 172–195. Springer, 2008.

[14] J. Thomas and K. Cook. Illuminating the path: the r&d agenda for visual analytics. national visualization and analytics center, institute of electrical and electronics engineers, 2005.

[15] M. Ware, E. Frank, G. Holmes, M. Hall, and I. H. Witten. Interactive machine learning: letting users build classifiers. *International Journal of Human-Computer Studies*, 55(3):281–292, 2001.

[16] I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.