Pre-publication draft of an extended abstract, published in *Lecture Notes in Computer Science* 3809*: Proceedings of the 18th Australian Conference on Artificial Intelligence (AI 05)*, pp 1-2. Berlin: Springer.

K-optimal pattern discovery: An efficient and effective approach to exploratory data mining

Geoffrey I. Webb

Faculty of Information Technology, Monash University, Vic, 3800, Australia
`webb@infotech.monash.edu.au`
`http://www.csse.monash.edu.au/~webb`

Abstract

Most data-mining techniques seek a single model that optimizes an objective function with respect to the data. In many real-world applications several models will equally optimize this function. However, they may not all equally satisfy a user's preferences, which will be affected by background knowledge and pragmatic considerations that are infeasible to quantify into an objective function.

Thus, the program may make arbitrary and potentially suboptimal decisions. In contrast, methods for exploratory pattern discovery seek all models that satisfy user-defined criteria. This allows the user select between these models, rather than relinquishing control to the program. Association rule discovery [1] is the best known example of this approach. However, it is based on the minimum-support technique, by which patterns are only discovered that occur in the data more than a user-specified number of times. While this approach has proved very effective in many applications, it is subject to a number of limitations.

- It creates an arbitrary discontinuity in the interestingness function by which one more or less case supporting a pattern can transform its assessment from uninteresting to most interesting.
- Sometimes the most interesting patterns are very rare [3].
- Minimum support may not be relevant to whether a pattern is interesting.
- It is often difficult to find a minimum support level that results in sufficient but not excessive numbers of patterns being discovered.
- It cannot handle dense data [2].
- It limits the ability to efficiently prune the search space on the basis on constraints that are neither monotone nor anti-monotone with respect to support.

K-optimal pattern discovery [4,5,11,14,15,17-20] is an exploratory technique that finds the *k* patterns that optimize a user-selected objective function while respecting other user-specified constraints. This strategy avoids the above problems while empowering the user to select between preference criteria and to directly control the number of patterns that are discovered. It also supports statistically sound exploratory pattern discovery [8]. Its effectiveness is demonstrated by a large range of applications [5-10,12,13].

References

1. Agrawal, R., Imielinski, T., Swami, A.N.: Mining Association Rules between Sets of Items in Large Databases. In Proc. 1993 ACM SIGMOD Int. Conf. Management of Data, Washington, D.C. (1993) 207-216.
2. Bayardo, Jr., R.J., Agrawal, R., Gunopulos, D.: Constraint-Based Rule Mining in Large, Dense Databases. Data Mining and Knowledge Discovery, 4 (2000) 217-240.
3. Cohen, E., Datar, M., Fujiwara, S., Gionis, A., Indyk, P., Motwani, R., Ullman, J.D., Yang, C.: Finding Interesting Associations without Support Pruning. In Proceedings Int. Conf. Data Engineering, (2000) 489-499.
4. Han, J., Wang, J., Lu, Y., Tzvetkov, P.: Mining Top-K Frequent Closed Patterns without Minimum Support. In Int. Conf. Data Mining (2002) 211-218.
5. Hellström,T.: Learning Robotic Behaviors with Association Rules. WSEAS transactions on systems. ISBN 1109-2777 (2003).
6. Eirinaki, M., Vazirgiannis, M., Varlamis, I.: SEWeP: using site semantics and a taxonomy to enhance the Web personalization process. In Proc. KDD-2003: the SIGKDD Conference of Knowledge Discovery and Datamining, ACM Press, New York (2003) 99-108.
7. Jiao, J., Zhang, Y.: Product portfolio identification based on association rule mining. Computer-Aided Desig,n 37 (2005) 149-172
8. McAullay, D., Williams, G.J., Chen, J., Jin, H.: A Delivery Framework for Health Data Mining and Analytics. Australian Computer Science Conference (2005) 381-390.
9. Mennis, J., Liu, J.W.: Mining association rules in spatio-temporal data: an analysis of urban socioeconomic and land cover change. Transactions in GIS, 9 (2005) 13-18.
10. Raz, O.: Helping Everyday Users Find Anomalies in Data Feeds, Ph.D. Thesis - Software Engineering, Carnegie-Mellon University (2004).
11. Scheffer, T., Wrobel, S.: Finding the Most Interesting Patterns in a Database Quickly by Using Sequential Sampling. Journal of Machine Learning Research 3 (2002) 833-862.
12. Siu, K.K.W., Butler, S.M., Beveridge, T., Gillam, J.E., Hall, C.J., Kaye, A.H., Lewis, R.A., Mannan, K., McLoughlin, G., Pearson, S., Round, A.R., Schultke, E., Webb, G.I., Wilkinson, S.J. Identifying markers of pathology in SAXS data of malignant tissues of the brain**.** Nuclear Instruments and Methods in Physics Research A (In Press).
13. Tsironis L., Bilalis N., Moustakis V.: Using inductive Machine Learning to support Quality Management. In Proc. 3rd Int. Conf. Design and Analysis of Manufacturing Systems, Tinos Island, University of Aegean (2001).
14. Webb, G. I.: Discovering associations with numeric variables. In Proc. 7th ACM SIGKDD Int. Conf. Knowledge Discovery and Data mining. ACM Press, (2001) pp 383-388.
15. Webb, G. I.: OPUS: An efficient admissible algorithm for unordered search. Journal of Artificial Intelligence Research. **3** (1995) 431-465.
16. Webb, G. I.: Preliminary investigations into statistically valid exploratory rule discovery. In Proc. Australasian Data Mining Workshop (AusDM03), University of Technology, Sydney (2003) 1-9.
17. Webb, G. I., Butler, S., Newlands, D:. On detecting differences between groups. In Proc. KDD-2003: The SIGKDD Conference of Knowledge Discovery and Datamining, ACM Press, (2003) pp. 256-265.

18. Webb, G. I., Zhang, S. K-Optimal-Rule-Discovery. Data Mining and Knowledge Discovery, 10 (2005) 39-79.
19. Webb, G. I.: Efficient search for association rules. In Proc. KDD-2000: the SIGKDD Conf. Knowledge Discovery and Datamining, ACM Press, New York (2000) 99-107.
20. Wrobel, S.: An Algorithm for Multi-relational Discovery of Subgroups. In Proc. Principles of Data Mining and Knowledge Discovery, Springer, Berlin (1997) 78-87.