ROLE OF ASSOCIATION RULE MINING IN TERMS OF SIGNIFICANT CORRELATIONS BETWEEN DIFFERENT ATTRIBUTES'

Simerjit Kaur[1]and Indu Singh[2]

Assistant Professor, Dept of Applied Sciences

Rayat-Bahra Institute of Engineering & Biotechnology, Mohali Campus (Punjab)

[2]Research Scholar, Dravidian University, Kuppam. Andhra Pradesh

INTRODUCTION

Association rule mining provides valuable information in terms of significant correlations between different attributes' values that might not be evident at the first glance in large datasets. The experimental part of this work has demonstrated benefits of integration of interactivity in Apriori approach for discovering association rules hidden in the target dataset. The interactive algorithm for discovering association rules starts by asking user's requirement with respect to attributes to be included in the search. Since the dataset has one class attribute that determines the patient *class* (*LIVE* or *DIE*), the clinicians are interested in finding rules that determine the value of patient *class* (*LIVE* or *DIE*). In addition to attribute specification, the user supplies the *minimum support* and *confidence threshold*, the two parameters required by Apriori algorithm. In the experimental runs, *minimum support* and *confidence threshold* have been fixed at *15*% and *80*%, respectively

MATERIAL AND METHOD

The proposed algorithm has been implemented in *java* environment (J2SDK1.4.1). The algorithm works in two steps: a) finds all the frequent itemsets with support greater than the *minimum support*, b) uses the frequent itemsets to generate the association rules. The algorithm finds only those rules that have the patient *class* attribute as a consequent in the rule. All other rules are ignored by the algorithm, for the domain user is not interested in such rules.

Placing aforementioned constraints on the algorithm search pattern is quite usual in the field of medicine where clinicians are not interested in finding all the associations in the datasets. For example, in hepatitis dataset, only the rules that show the association of the test findings with the patient *class* are meaningful for diagnosis purpose. The results presented in the next section have depicted that constraining the behaviour of algorithm on-the-fly (i.e. interactively) helps clinicians to find attributes that determine the patient *class*.

INTERPRETATION OF RESULTS

Executing the algorithm with 15% support and 80% confidence has resulted into discovery of 48 association rules hidden in the hepatitis dataset. Figure 6.2 shows all the rules discovered by the algorithm with user specified support and confidence threshold. All these rules have *class* as target attribute. This experimentation gives rule induction method for prediction of patient class based upon the 19 recorded attributes of the hepatitis patient dataset. Upon close observation of the discovered rules, the attribute *Anorexia, Protime* and *Histology* produced the best results. The method will predict which attributes contribute more to a person's chances of being attacked with hepatitis disease. This technique has been applied because of the ready availability of the subjects with some knowledge of the domain that can provide feedback on the explanations. The identification and interpretation of the discovered rules requires ample domain knowledge. For example, the rule ***{9, 15, 16, 18, 48, 55} -> 1*** is translated as:

Male (X, "Yes") ∩ Anorexia (X, "Yes") ∩ Liver Big (X,"Yes") ∩ Spleen Palpable (X, "Yes") ∩ Protime30 (X,"Yes") ∩ Histology (X, "Yes") → Class (X, "DIE")

[Support = 15%, Confidence = 100%] [Rule 1]

Figure 1: The rules generated with support 15% and confidence 80%.

```
(1) {15 48} -> {1} 100.0%

(2) {9 15 48} -> {1} 100.0%

(3) {12 15 48} -> {1} 100.0%

(4) {9 12 15 48} -> {1} 100.0%

(5) {15 16 48} -> {1} 100.0%

(6) {9 15 16 48} -> {1} 100.0%

(7) {12 15 16 48} -> {1} 100.0%

(8) {9 12 15 16 48} -> {1} 100.0%

(9) {15 18 48} -> {1} 100.0%

(10) {9 15 18 48} -> {1} 100.0%

(11) {15 16 18 48} -> {1} 100.0%

(12) {9 15 16 18 48} -> {1} 100.0%

(13) {48 55} -> {1} 100.0%

(14) {9 48 55} -> {1} 100.0%

(15) {12 48 55} -> {1} 100.0%

(16) {9 12 48 55} -> {1} 100.0%
```

Figure 2 provides a pictorial view of the findings in terms of *Patient IDs* who support the discovered association rule *Rule1*. The rule states that the attribute values on the left side of the rule derive the patient *class* with 100% confidence. The rule induction method has the potential to use retrieved cases for prediction. A close look at each attribute in the rule points out the following findings:

– The first part of the antecedent states that hepatitis is more common in males than in females.

– Anorexia (state of loss of appetite) is a persistent problem with many chronic or serious diseases. The next part of the rule states that patients with acute hepatitis will suffer from anorexia.

– The rule suggests that *big liver* and spleen disease virus (BLSV) is closely related to *hepatitis* virus.

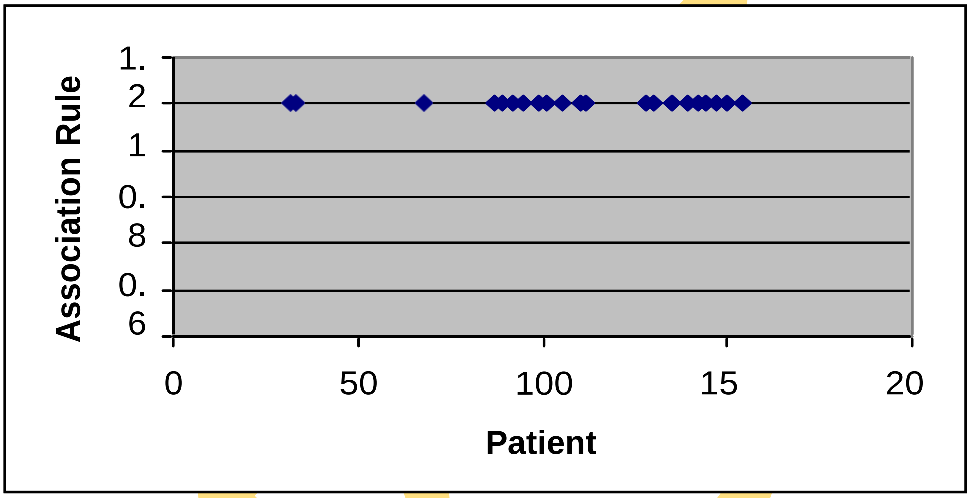

Figure2: Patient IDs satisfying the association rule – *Rule1*.

– The reference range for prothrombin time (a measure of ability of the blood to clot) is usually around 12–15 seconds. A prolonged or increased prothrombin time suggests that the blood under test is taking too long to form a clot which may be caused by conditions such as liver disease.

– *The rule indicates that the patients have gone through the histological*[1] *reconstruction study of liver cell necrosis.*

Once the association rules among various data attributes have been established, the important task that remains is to use these rules in biomedical research and patient treatment. In contrast to traditional data mining, involving domain expert in the analysis and interpretation of patient database can help in the proper diagnostic study of patient data. In the present work, the researchers were interested in finding only those association rules that determine the patient *class* value "*DIE*". Further, it has been found that values of the attributes in antecedent of the rules indicate that the disease is chronic. Such a rule can help clinicians to deal with the disease in a more effective way, in that, if the test results of a patient indicate some similarity with these rules then necessary actions can be taken for the patient so as to cure the infection before reaching it at malignant level.

CONCLUSION

Association rule mining as a data mining technique is very useful in the process of knowledge discovery in medical field, especially in the domain where patients' lab test reports have been electronically stored. In this chapter, association rule method is interactively implemented to predict the hepatitis patient's class. Such an experiment can give medical doctors a tool to quickly get some knowledge from the past patient's database and use them for handling future case. Understanding complex relationships that occur among patient's symptoms, diagnosis and behaviour is one of the most promising areas of data mining. The problem of identifying a patient's class is a major challenge among medical practitioners. Data mining techniques provide a tool to help them quickly make sense out of vast clinical databases.

REFERENCES

Abe H., Yokoi H., Ohsaki M. and Yamaguchi, T. (2007). Developing an Integrated Time-Series Data Mining Environment for Medical Data Mining. Seventh IEEE International Conference on Data Mining, 28-31 Oct. 2007, 127-132.

Agrawal R. and Srikant R. (1994). Fast Algorithms for Mining Association Rule. **Proceedings of the 20th** International Conference on Very Large Databases (VLDB), 487 – 499.

Ankerest M., Ester M. and Kriegel H.P. (2000). Towards an Effective Cooperation of the User and the Computer for Classification. Proceedings of 6th International conference on Knowledge Discovery and Data Mining, Boston, MA.

Bates J.H.T. and Young M.P. (2003). Applying Fuzzy Logic to Medical Decision Making in the Intensive Care Unit. American Journal of Respiratory and Critical Care Medicine, Vol. 167, 948-952.

Berks G., Keyserlingk D.G.V., Jantzen J., Dotoli M. and Axer H. (2000). Fuzzy Clustering - A Versatile Mean to Explore Medical Databases. ESIT, Aachen, Germany, 453-457.

Berson A., Smith S. and Thearling K. (1999). **Building Data Mining Applications for CRM. First Edition,** McGraw-Hill Professional.

Bethel C.L., Hall L.O. and Goldgof D. (2006). Mining for Implications in Medical Data. Proceedings of the 18th International Conference on Pattern Recognition,Vol.1, 1212-1215.

Cheung Y.M. (2003). k-Means: A New Generalised k-Means Clustering Algorithm. N-H Elsevier Pattern Recognition Letters 24, Vol 24(15), 2883–2893.

Chiang I.J., Shieh M.J., Hsu J.Y.J. and Wong J.M. (2005). Building a Medical

Frank H., Klawonn F., Kruse R. and Runkler T. (1999). Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition. New York: John Wiley.

Frawley W.J., Piatetsky-Shapiro G. and Matheus C.(1996). Knowledge Discovery in Databases: An Overview. Knowledge Discovery in Databases, AAAI Press/MIT Press, Cambridge, MA., Menlo Park, C.A, 1-30.

Houtsma M.A.W. and Swami A.N. (1993). **Set-Oriented Mining for Association Rules in Relational Databases. Proceedings of the Eleventh International Conference on Data Engineering, 25-33.**

Leung, K.S., Lee K.H., Wang J.F., Ng E. YT, Chan H. LY, Tsui S. KW, Mok T. SK, Tse P.C.H. and Sung J. J.Y.(2009). Data Mining on DNA Sequences of Hepatitis B Virus. *IEEE/ACM Transactions on Computational Biology and Bioinformatics.* IEEE computer Society Digital Library.

Liu S-H., Chang K-M. and Tyan C-C. (2008). Fuzzy C-Means Clustering for Myocardial Ischemia Identification with Pulse Waveform Analysis. 13th International Conference on Biomedical Engineering, Singapore, Vol. 23, 485-489.

Marx K.A., O'Neil P., Hoffman P. and Ujwal M.L. (2003). Data Mining the NCI Cancer Cell Line Compound GI (50) Values: Identifying Quinine Subtypes Effective against Melanoma and Leukemia Cell Classes. United-States: Journal of Chemical Information and Computer Sciences, Vol. 43, 1652-1667.

Match-Project: http://www.match-project.com/

Mounji, A. (1997). Languages and Tools for Rule-Based Distributed Intrusion Detection. PhD thesis, Faculties Universitaires Notre-Dame dela Paix Namur (Belgium).

Pace R.K. and Zou D. (2000). Closed-Form Maximum Likelihood Estimates of Nearest Neighbor Spatial Dependence. Geoghraphical Anaylsis, Vol. 32(2).

Pechenizkiy M. Tsymbal A. and Puuronen S. (2005). Knowledge Management Challenges in Knowledge Discovery Sytems. 16th IEEE International Workshop on Database and Expert Systems Applications, 433-437.

Pei J., Upadhyaya S.J., Farooq F. and Govindaraju V. (2004). Data Mining for Intrusion Detection: Techniques, Applications and Systems. Proceedings of the 20th International Conference on Data Engineering, p.877.

Rahm E. and Do H. H. (2000). Data Cleaning: Problems and Current Approaches. IEEE Bulletin on Data Engineering, Vol. 23(4).

Saeed M., Lieu C., Raber G. and Mark R.G. (2002). MIMIC: A Massive Temporal ICU Patient Database to Support Research in Intelligent Patient Monitoring. IEEE Computers in Cardiology, Vol. 29, 641-44.

Selfridge P. and SrivastvaD. (1996). A Visual Language for Interactive Data Exploration and Analysis. **Proceedings of the 1996 IEEE Symposium on Visual Languages,** 84.

Soukup T. and Davidson Ian. (2002). Visual Data Mining: Techniques and Tools for Data Visualisation and Mining. Wiley Dreamtech India Pvt. Ltd. First Edition 2002.

Srikant R., Vu Q. and Agrawal R. (1997). Mining Association Rules With Item Constraints. Proceedings of 3rd International Conference on Knowledge Discovery and Data Mining.

The official web site of Central Beauro of Health Intelligence: http://www.cbhidghs.nic.in

Ye N. and Li X. (2003). Application of Decision Tree Classifiers to Computer Intrusion Detection. **Real-Time System Security,** 77 – 93.

Zhang S., Liu S., Wang D., Ou J. and Wang G. (2006). Knowledge Discovery of Improved Apriori-Based High-Rise Structure Intelligent Form Selection. **Proceedings of the 6th International Conference on Intelligent Systems Design and Applications, Vol.1**, 535-539.