STUDY ON ROLE OF DATA MINING IN HEALTH CARE

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

The role of data mining in health care has become the subject matter of wide and varied research activities [Kaur H. et.al. 2006]. Extraction of useful information from health care data offers a lot of challenges in terms of storage, dissemination, privacy and security of patient data. While the privacy issue is more of a legal and ethical issue rather than technological issue, data mining offers broader community-based gains that enable and improve healthcare forecasting, analysis, and visualisation [Payton F.C., 2003]. Guided use of technologies like database systems, data mining and knowledge management can contribute a lot to decision support systems in health care.

The health care environment is generally perceived as being "rich in information" yet poor in knowledge [Lincoln T. et.al., 1999]. Presently, there is a lack of good data analysis tools to uncover hidden relationships in the data. If the data regarding past clinical trials and interviews with the patients is gathered and computerised in a knowledge base, it can be evaluated for effective and safe treatments on human subjects.

TANAGRA – a shareware data mining tool – has been used to implement the ID3 decision tree algorithm to a dataset of diabetes patients. The decision tree algorithm was run under the interactive guidance of a human expert with medical knowledge to select the attributes for the information extraction. An interactive system is an integration of a human user and a machine in which both can communicate and exchange information for achieving a common goal. Interactive data mining can support user"s learning, improve his insight and understanding of the domain, and on the other hand user"s feedback can be used to improve the performance of the machine in terms of efficiency of operations and quality of the output.

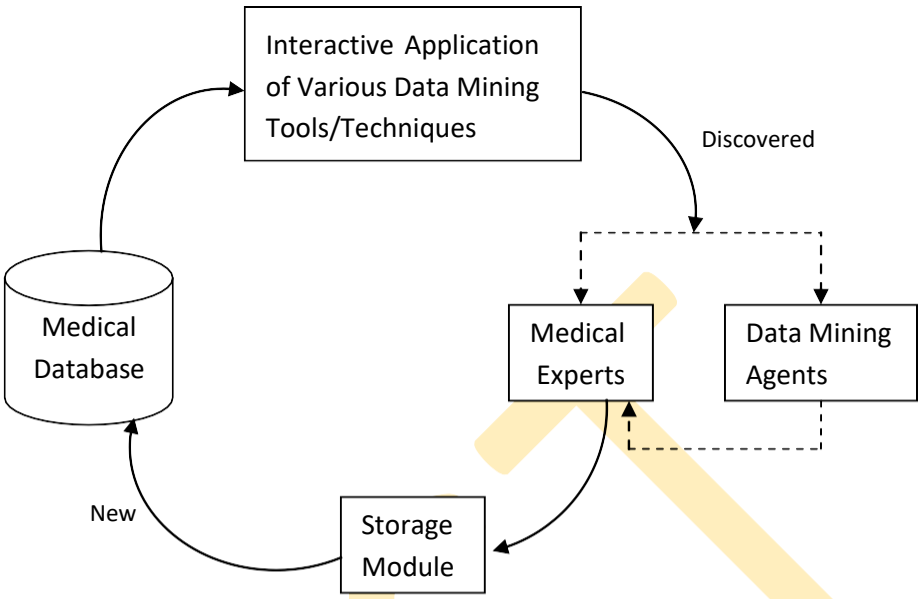

Figure 1.1: Interactive decision support system in medical domain.

Figure 1.1 depicts the flow of information and control in the proposed system. The knowledge discovered from medical databases by the synergy of human and computer interaction is in the form of rules and implications. This knowledge can be stored back in the medical database to further assist the clinicians in diagnosing future cases. Such a system would turn out to be an interactive decision support system in health care management. The system differs from a conventional DSS such as MATCH [Gruzdz A. et.al. 2006] in that it involves domain user in the knowledge discovery process. Medical practitioners can better define the parameters for extraction of useful rules and patterns.

The IDSS proposed and developed as a part of research work reported herein was run on dataset containing details of clinical trial of diabetic patients. The IDSS was supposed to find the major factors responsible for developing the disease. Three experiments were conducted to get some useful knowledge for the clinicians to improve the quality of decisions in the medical diagnosis.

The dataset donated by Vincent Sigillito of John Hopkins University was used in the experimental runs of IDSS. The dataset is available on UCI Machine Learning Repository website http://archive.ics.uci.edu:80/ml/datasets.html. The dataset consists of clinical trials of Indian female patients suffering from Diabetes Mellitus. The dataset comprises of 8 attributes (Table 1) and 768 records. There is an additional attribute named „*Class'* which has two values „YES‟ and „NO‟.

S.No.	Attribute Name/Description (all numeric-valued)
1	Number of times pregnant
2	Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3	Diastolic blood pressure (mm Hg)
4	Triceps skin fold thickness (mm)
5	2-Hour serum insulin (mu U/ml)
6	Body mass index (weight in kg/(height in m)^2)
7	Diabetes pedigree function
8	Age (years)
9	Class variable (0 or 1)

Table 1: Attributes of Diabetes dataset.
(Source: Diabetes Dataset - http://archive.ics.uci.edu:80/ml/datasets.html)

Material and method

TANAGRA - a free, open-source, user-friendly software product developed by Ricco Rakotomalala, has been used to mine the data. TANAGRA supports a host of analytical functions such as binary logistic regression, k-nearest neighbor, neural network trained with back-propagation, Quinlan‟s ID3 (Iterative Dichotomiser 3), linear discriminant analysis, and naive Bayesian classifier. In classification based machine learning approach there are two distinct approaches to learning: supervised and un-supervised learning. The supervised learning deals with problems where a set of data are labeled for training and another set would be used for testing [Li C.H. et.al., 2001].

Table 2 provides a brief description of the various components of TANAGRA used in the experiments. ID3 algorithm of Meta-supervised learning has been run after loading the dataset in Tanagra. TANGARA arranges the Classification results in form of a decision tree which is a predictive model that maps observations about an item to conclusions about the item's target value. In the resulting decision trees, leaves represent the class of the data item and branches represent conjunction of features that has lead to those classifications. After applying the decision tree classification, rules can be designed based upon the results of the algorithm with the help of another component named Rule-based selections and the same can be visualised. The results have been analysed by a domain expert to pick appropriate rules that might be in interest of medical purpose.

Table 2: Components of TANAGRA used in the experiments.

Tab	Operator (Component)	Comment
Feature selection	Define status	Specify the attributes to use
Meta-spv learning	Supervised learning	A container for machine learning

		operators
Spv learning	ID3	A machine learning operator
Instance selection	Rule-based selection	Select a subset of examples based upon a rule.
Data visualisation	View dataset	Visualise the current dataset in a grid.

To conduct the experiment the Diabetes dataset was bifurcated into two equal parts containing 384 records each. The first part was used as *training set* while the second part was used as *test set*. ID3 was run on the training set and the Rule-based selection algorithm was applied to the test set. The goal was to involve the domain expert in the analysis process. The experiment was repeated thrice with different input combinations. The judgment of attributes for different runs of the algorithm has been done both subjectively and interactively.

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.