Journal of Advances in Science and Technology

Vol. IV, No. VII, November-2012, ISSN 2230-9659

REVIEW ARTICLE

BENEFITS OF INVOLVING DOMAIN USER IN THE KNOWLEDGE DISCOVERY PROCESS

www.ignited.in

Journal of Advances in Science and Technology
Vol. IV, No. VII, November-2012, ISSN 2230-9659

Benefits of Involving Domain User in the Knowledge Discovery Process

Rishu Bhardwaj

Research Scholar, CMJ University, Shillong, Meghalaya, India

---------------------------♦-----------------------------

INTRODUCTION

The challenge of performing everything automatically has dominated the awareness of the researchers and developers of the commercial tools up to the present. However, the knowledge discovery is not meant to exclude the human since the discovered knowledge addresses the human. Instead of allowing an automated data mining process to iterate in a trial-and-error manner, a better but largely overlooked way to enhance the knowledge discovery process is to provide a domain expert support through human involvement. With potential rewards of human interaction in the data mining process in mind, the work has been accomplished on design and development of several data mining experiments with human interactivity on medical and other databases. Section 10.1 of this chapter concludes the present work briefly whereas section 10.2 presents an overview of opportunity of further research work in the area of human interactivity in the process of knowledge discovery in various domains.

MATERIAL AND METHOD

During the course of research work reported herein, the potential benefits of involving domain user in the discovery of useful knowledge from large databases, particularly medical databases have been discussed such as: 1) human involvement in the process of knowledge discovery, 2) interactive association rule mining, 3) classification of Hepatitis patient, 4) interactive decision support system, 5) guided clustering technique and 6) interactive fuzzy inference system.

KDD is by no means a push button technology. Human involvement in the knowledge discovery process proves beneficial in terms of temporal efficiency of the data mining tool. The basic task of the knowledge discovery and data mining process is to extract knowledge from data such that the resulting knowledge is useful in a given situation. Furthermore, the quality of results heavily relies on data preparation wherein the domain knowledge can be used to prepare the data and make it suitable for mining. It is therefore, necessary to involve domain user in the knowledge discovery process right from the beginning. Chapter 3 discusses the role of Domain Expert in every phase of knowledge discovery process.

A case study of real-world database has been considered to implement a human-interactive knowledge discovery project (as described in Chapter 3). Dataset of one Blood Transfusion Dataset was chosen for the data mining experiment and was mined by two different data mining techniques. Firstly, outlier mining was applied to the dataset to find those records which are considerably dissimilar, exceptional or inconsistent with respect to the remaining data. The second data mining technique used is the decision tree. Conducting these experiments interactively yielded some useful knowledge in form of rules which can help the patron to manage the functioning of the Blood Transfusion Service Centre better.

INTERACTIVE ASSOCIATION RULE MINING

Association rule mining has the result form: A=>B. Among the various data mining results, rules are most appropriate for integrating human opinions, because human thoughts can be converted into rules relatively easily than into some other forms. Based upon his experiences, the domain user can transfer his specifications into data mining system in form of rules.

Interactive Association Rule (IAR) mining algorithm has been developed (described in Chapter 4) which works on the technique of standard Apriori algorithm, but with a variation, as it prunes the database with respect to the attribute specifications provided by the domain user. As a result, the size of the target dataset gets reduced and IAR algorithm takes less time and space to find the association rules in which the user is interested. The IAR algorithm will provide help in situations where the user is interested in finding relationships between certain attributes instead of the whole dataset.

CLASSIFICATION OF HEPATITIS DATASET

The problem of identifying a patient’s class is a major challenge among medical practitioners. Classification provides great help in understanding complex relationships that occur among patient’s symptoms, diagnosis and behaviour. Interactive association rule has been used as a classification technique to find valuable information from the Hepatitis dataset. The researcher has used the ‘Class’ attribute having two values ‘LIVE’ or “DIE’ as the target attribute from the dataset (described in Chapter 5).

The association rule mining technique was run on the Hepatitis dataset with user specified confidence and support threshold. The algorithm found useful information in form of rules such as 1) Hepatitis is more common in males than in females, 2) patients with acute Hepatitis will suffer from anorexia[1], 3) big liver and spleen disease virus (BLSV) is closely related to Hepatitis virus. This kind of knowledge can help the clinicians to quickly make sense out of vast clinical datasets and provide more imperative input to the decision making process in the medical domain.

INTERACTIVE DECISION SUPPORT SYSTEM FOR MEDICAL DOMAIN

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. Data mining tools, when used in collaboration with domain experts, provide vital knowledge that can help the medical practitioners to deal with the disease in a better way and improve the quality of diagnosis. Experiments based upon Decision tree algorithms, have been done to demonstrate the augmentation of information mined from medical datasets.

The experiment (described in Chapter 6) was conducted on Diabetes dataset in collaboration with a domain expert. The dataset was first bifurcated into two parts, namely: training-set and test-set. Knowledge in the form of rules was driven from the training-set and the same was tested on the test-set. The three experiments conducted on the Diabetes dataset imply that useful knowledge can be derived in such a way and can help the medical practitioners to store the same in knowledge-base to handle the future cases of the same disease in a better way.