INTERNATIONAL JOURNAL OF PURE AND APPLIED RESEARCH IN ENGINEERING AND TECHNOLOGY

A PATH FOR HORIZING YOUR INNOVATIVE WORK

DATA MINING: APPLICATIONS AND USAGE IN HEALTH CARE

S.P. AKARTE, A.A. CHAUDHARI, DR. G. R. BAMNOTE, A.V. DARYAPURKAR

Department of Computer Science & Engineering, Prof. Ram Meghe Institute of Technology & Research, Badnera, Amravati, Maharashtra.

Accepted Date:
27/02/2013
Publish Date:
01/04/2013

Keywords

Breast Cancer
Survivability,
Data Mining,
SEER,
Association Rules,
Breast Cancer,
Classification Clustering,
Diabetes,
Heart Disease.

Corresponding Author

Mr. S.P. Akarte

Abstract

Data large datasets so as to extract "hidden" pieces of information. It is typically used to predict potential future trends or to discover obscured patterns in the "behaviour" of the data. As the name Data Mining implies, the main objective is to search in large databases for valuable, and often hidden, items of information. With the erudition of recent technology, it is possible to process these vast databases in just a few minutes, a procedure that has been in the past both time-consuming and arduous. There are two primary goals of data mining tend to be prediction and description. Mining may befined as that composite of techniques engaged to detect patterns in

ISSN: 2319-507X
IJPRET

1. Introduction

Data mining is a knowledge discovery technique widely used in many domains including finance, commerce, geological surveys, weather pattern prediction and telecommunications. The major reason that data mining has attracted attention in the past few years is due to the wide availability of data in electronic form and the need for turning such data into useful information and knowledge. Sophisticated data mining tools have been developed over the past decade with the progress in database and information technologies. Heterogeneous database systems and the Internet based global information systems such as the World Wide Web now play a major role in information retrieval and processing. Computer hardware technology has also progressed by leaps and bounds in the past three decades, leading to large supplies of powerful and affordable computers and data collection and storage devices.

This has provided a boost to the information industry with huge number of databases and information repositories available for transaction management, information retrieval and data analysis. A database architecture that has recently emerged with the intention of facilitating data mining is the data warehouse, defined by Han and Gambler[2] as “a repository of multiple heterogeneous data sources, organized under a unified schema at a single site in order to facilitate management and decision making”. Data warehouse technology includes data cleaning, data integration and OLAP which enables data to be viewed in an aggregated from in a multidimensional manner. The Disulfide database gathers information about this biological process with structural, evolutionary and neighborhood information on cysteines in proteins. Mining this information with an association rule discovery program permits to extract some strong rules for the prediction of the disulfide-bonding state of cysteines [1].

Data mining is a synonym for another popularly used term “Knowledge Discovery in Databases” or KDD. Knowledge discovery is an iterative process consisting of data cleaning, to remove noisy and inconsistent data, data integration, to combine multiple heterogeneous or homogeneous data sources, data selection, to consider only data relevant to the task and data transformation where data is transformed into forms appropriate for mining functions such as aggregation or summarization. Then data mining algorithms are employed to extract interesting and meaningful patterns from the data and present the knowledge to the domain expert in an informative manner. Based on this, it is intuitive that the typical data mining system has a multi-tiered architecture as shown in Fig 1.

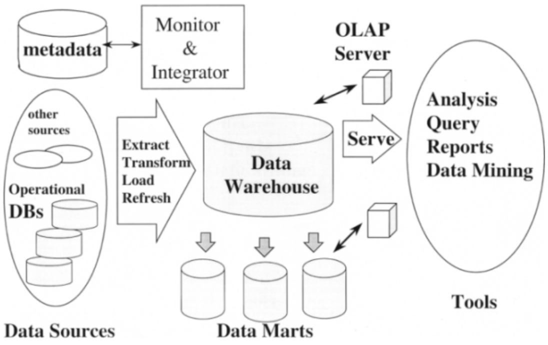

Fig. 1 Architecture of a Data Mining System

The amount of clinical data stored electronically is making it possible to carry out large scale studies that focus on the interaction between genotype, phenotype and diseases at a population level. Matching patient responses (phenotype) with gene expression and known metabolic pathway relationships across large number of individuals may be the best hope for understanding the complex interplay between multiple genes and environment that underlies some of the most debilitating health problems[3]. However, despite the myriad of problems data mining in medicine can solve, the application of data mining to this domain is still relatively new due to obstacles that exist with this domain. There are several special characteristics that make it difficult to analyze medical data in an automated fashion by traditional techniques.

Firstly, the common nature of medical data is such that there is high dimensionality associated with it. There are many data elements, each representing a dimension that varies in value, characterize an itemset of interest such as patient disease or specimen. Usually there could be 50 to 100 different types of data elements. It is important to thus carry out attribute relevance to choose only a subset of the elements since the like hood of patients sharing coincidental data is high and that may lead to incorrect conclusions.

Data mining requires consistent data, meaning that large amounts of data need to converted to compatible representations.

Linking data variables to patient characteristics is not straightforward. Unlike market basket analysis where the buyer patterns can be directly studied, medical observations commonly indicate a probability that a condition exists based on their support and confidence (which are alternatively known as sensitivity and specificity in this domain). A given feature may be consistent for the condition or the condition may exist without the feature. It is not always as easy to predict medical conditions with the excepted confidence and support as it is in other domains. The domain's data records are also in some cases incomplete and with inconsistency. Patients with the same conditions may have substantially different types of timings of observations. Since most of it requires human data entry, the data is prone to inconsistency and noise.

Physician's interpretation generally is the diagnosis in medical conditions. Interpretations of different individuals may differ or even conflict and are often expressed in text that need to be transformed to other forms before data can be mined from them. Even specialists from the same discipline cannot agree on unambiguous terms to be used in describing patient data. There are many synonyms for the same disease and the process of mining data is even more daunting as different grammatical constructs can be used to describe the relationships among medical conditions. The major reason for this ambiguity is because medical data has no canonical form.

If there no equivalent ideas in medicine, then how can indexes and statistical tables constructed and data mining depends on these equivalent concepts.

Usually, algorithms that perform data mining are programmed to work on mathematical form of the data in the form of formula and equations in the area of physical sciences.

Because medical data is collected on human subjects, there is an enormous ethical, legal and social tradition designed to prevent the abuse of patients and misuse on their data. Controversies and confusions that relate to these issues exist that complicate aggregation and data analysis other than for individual patient care. The call for a national framework for the secondary use of health information is largely a recommendation that society, health care providers and government resolve these issues so that the patients can benefit from techniques such as data mining [4].

Despite the difficulties associated with the domain, there have been many efforts to study the potential of data mining in medicine.

2. Heart Disease Prediction

Medical data mining has high potential for exploring the hidden patterns in the data sets of the medical domain. These patterns can be utilized for clinical diagnosis for widely distributed in raw medical data which is heterogeneous in nature and voluminous. These data should be collected in an organized form. This collected data can be integrated to form a hospital information system. Data mining technology provides a user oriented approach to novel and hidden patterns in the data. From the analysis of World Health Organization, they estimated 12 million deaths occur worldwide, every year due to the Heart diseases. Half the deaths occur in United States and other developed countries due to cardio vascular diseases. On the above discussion, it is regarded as the primary reason behind deaths in adults. Heart disease kills one person every 34 seconds in the United States. The following paper reviewed about predicting of heart disease using data mining technique.

Jyoti Soni et. al [3] have produced a large number of rules when association rules are applied to medical dataset and also proposed three different supervised machine learning algorithms . They are **Naïve Bayes, K-NN, and Decision List** algorithm. These algorithms have been used for analyzing the heart disease dataset [5].

Decision tree is one of the popular and important classifier which is easy and simple to implement. It doesn‟t have domain knowledge or parameter setting. It handle huge amount of dimensional data.

Naïve Bayes is a statistical classifier which assigns no dependency between attributes. To determine the class the posterior probability should be maximized.

[2**] K-nearest neighbor's algorithm (k-NN)** is the one of the important method for classifying objects based on closest training data in the feature space. It is simplest among all machines learning algorithm but,

the accuracy of k-NN algorithm can be degraded by presence of noisy features

3. Cancer Information System

Vast collections of raw data are not in themselves useful. To be meaningful, data must be analyzed and converted into information, or even better, into knowledge. The amount of information available is staggering. Traditional methods of data analysis utilizing human beings as pattern detectors and data analysts cannot possibly cope with such a large volume of information.

- **Classification** – This problem involves the need to find rules that can partition the data into disjoint groups. [14]
- **Clustering** -consider clustering a separate class. This is because clustering methods allow data mining algorithms to determine groups automatically Clustering techniques are frequently used to discover structure or similarities in data.[15]
- **Association** – The association data mining problem involves finding all of the rules for which a particular data attribute is either a consequence or an antecedent.
- **Sequences** – This type of data mining problem involves ordered data, most commonly time sequence or temporal data.

4. Breast Cancer Survivability

Breast cancer has become a common cancer in women. For instance, it affects one in every seven women in the United State [8].

There are three predictive focus of cancer prognosis:

1) prediction of cancer susceptibility (risk assessment),
2) prediction of cancer recurrence and
3) prediction of cancer survivability.

The prediction of breast cancer survivability has been a challenging research problem for many researchers. The discovery of the survival rate or survivability of a certain disease is possible by extracting the knowledge from the data related to that disease. One of these data sources is SEER [9,13], The SEER database combines patient-level information on cancer site, tumor pathology, stage, and cause of death [6, 7].

The data mining techniques to predict the survivability rate of breast cancer patients take into account three variables: Survival

Time Recode (STR), Vital Status Recode (VSR), and Cause of Death (COD). Delen et al. preprocessed the SEER data (period of 1973-2000 with 433,272 records named as breast.txt) for breast cancer to remove redundancies and missing information. The resulting data set had 202,932 records, which then pre-classified into two groups of "survived" (93,273) and "not survived" (109,659) depending on the Survival Time Recode (STR) field [16].

5. Dermatological (Skin) Diseases Prediction

Recently, skin diseases have been common to everyone. Many factors influence the onsets of these diseases and each age group usually has its different symptoms. Although skin diseases are detected easier than other diseases and diagnosing symptoms and deciding treatment plans are not as complex as other internal diseases, many people often ignore the importance of them[13].

6. Diabetes

Insulin is one of the most important hormones in the body. It aids the body in converting sugar, starches and other food items into the energy needed for daily life. However, if the body does not produce or properly use insulin, the redundant amount of sugar will be driven out by urination. This disease is referred to diabetes. The cause of diabetes is a mystery, although obesity and lack of exercise appear to possibly play significant roles. Based on the American Diabetes Association [4] in November 2007, 20.8million children and adults in the United States (i.e., approximately 7% of the population) were diagnosed with diabetes. In early the ability to diagnose diabetes plays an important role for the patient's treatment process.

In [10] the author predicts whether a new patient would test positive for diabetes. This paper studied a new approach, called the Homogeneity- Based Algorithm (or HBA) to determine optimally control the over fitting and overgeneralization behaviors of classification on this dataset. Some experimental results seem to indicate that the proposed approach significantly outperforms current approaches. From the experiment the author concluded that it is very important both for accurately predicting diabetes and also for the data mining community, in general.

7. Conclusion

In this paper, the various diseases are elaborate with the use of data mining tools. DM tools which are working fine with medical data and its specific. There is no universal tool facing all problems we can have, but hybrids may cover more. The data mining tools are very helpful in the prediction of the various types of cancer as well as heart diseases. we discuss about the heart disease prediction, in that machine learning algorithms namely naïve bayes, K-NN, Decision Lis

References

1. Tessie D, Bordeaux B, Larre C, and Popinjay Y, 2004 Data mining techniques to study the disulfide-bonding state in proteins: signal peptide is a strong descriptor. Bioinformatics 20: 2509-2512.

2. Han, J., Kamber, M.: "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, 2006

3. Cox DR. Analysis of survival data. London: Chapman & Hall; 1984.

4. Benjamin F. Hankey, et. al. The Surveillance,Epidemiology, and End Results Program: A National Resource. Cancer Epidemiology Biomarkers & Prevention 1999; 8:1117-1121.

5. Asha Rajkumar, G.Sophia Reena, "Diagnosis Of Heart Disease Using Datamining Algorithm", Global Journal of Computer Science and Technology 38 Vol. 10 Issue 10 Ver. 1.0 September 2010.

6. Carloz Ordonez, "Association Rule Discovery with Train and Test approach for heart disease prediction", IEEE Transactions on Information Technology in Biomedicine, Volume 10, No. 2, April 2006.pp 334-343.

7. Jyoti Soni, Ujma Ansari, Dipesh Sharma, Sunita Soni "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction" IJCSE Vol. 3 No. 6 June 2011

8. Wingo PA, Tong T, Bolden S, "Cancer statistics", 1995, CA Cancer J Clin three data mining methods. Artificial Intelligence in Medicine. 2005 Jun; 34(2):113-27.

9. Huy Nguyen Anh Pham and Evangelos Triantaphyllou "Prediction of Diabetes by Employing a New Data Mining Approach Which Balances Fitting and Generalization"

Department of Computer Science, 298 Coates Hall, Louisiana State University, Baton Rouge, LA 70803.

10. Jiawei Han and Micheline Kamber, Data Mining Concepts and Techniques. San Francisco, CA: Elsevier Inc, 2006.

11. U. M., Piatetsky-Shapiro, G. & Smyth, P. & Uthurusamy, R. Fayyad, "From Data Mining to Knowledge Discovery: An Overview," in Advances in Knowledge Discovery and Data Mining , 1996a, pp. 1-36.

12. Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) Public-Use Data (1973-2002), National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch, released April 2005, based on the November 2004 submission.

13. Fayyad, U. M., Piatetsky-Shapiro, G. & Smyth, P. (1996a). From Data Mining to Knowledge Discovery: An Overview. In Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P. & Uthurusamy, R. (eds.) Advances in Knowledge Discovery and Data Mining, 1–36. AAAI Press/MIT Press.

14. Agrawal, R., Imielinski, T. & Swami, A. (1993). Database Mining: A Performance Perspective. IEEE Transactions on Knowledge and Data Engineering 5(6): 914–925.

15. Lundin M., Lundin J., BurkeB.H.,Toikkanen S., Pylkkänen L. and Joensuu H. , “Artificial Neural Networks Applied to Survival Prediction in Breast Cancer”, Oncology International Journal for Cancer Resaerch and Treatment, vol. 57, 1999.