An Automated and Efficient Mining of the Healthcare Data for the Prognosis of Heart Attack using the HUI Miner and Naïve Bayes Classifier

S. Sherlin, D.S halini Devi, T. Vetriselvi
Computer Science and Engineering,
K.Ramakrishnan College of Technology, Trichy, Tamilnadu

***Abstract*— The healthcare environment is generally perceived as being 'information rich' yet 'knowledge poor'. There is a wealth of data available within the healthcare systems. However, there is a lack of effective analysis tools to discover hidden relationships and trends in data. Knowledge discovery and data mining have found numerous applications in business and scientific domain. Valuable knowledge can be discovered form application of data mining techniques in healthcare system. The healthcare industry collects huge amounts of healthcare data which, unfortunately, are not "mined" to discover hidden information. For data preprocessing and effective decision making Naïve Bayes classifier is used. It is an extension of Naïve Bayes to imprecise probabilities that aims at delivering robust classifications also when dealing with small or incomplete data sets. The HUI miner is used to find the high utility item sets from a database. Discovery of hidden patterns and relationships often gets unexploited. Using medical profiles such as age, sex, blood pressure and blood sugar it can predict the likelihood of patients getting a heart disease. It enables significant knowledge, e.g patterns, relationships between medical factors related to heart disease, to be established.**

Keywords— High utility item set, mining algorithm

INTRODUCTION

Knowledge discovery in databases is well-defined process consisting of several distinct steps. Data mining is the core step, which results in the discovery of hidden but useful knowledge from massive databases. A formal definition of knowledge discovery in databases is given as follows:"Data mining is the non trivial extraction of implicit previously unknown and potentially useful information about data".

Data mining technology provides a user-oriented approach to novel and hidden patterns in the data. The discovered knowledge can be used by the healthcare administrators to improve the quality of service. The discovered knowledge can also be used by the medical practitioners to reduce the number of adverse drugs effect, to suggest less expensive therapeutically equivalent alternatives.

PROBLEM DEFINITION

Anticipating patient's future behavior on the given history is one of the important applications of data mining techniques that can be used in health care management. A major challenge facing healthcare organizations (hospitals, medical centers) is the provision of quality services at affordable costs.

Quality service implies diagnosing patients correctly and administering treatments that are effective. Poor clinical decisions can lead to disastrous consequences which are therefore unacceptable. Hospitals must also minimize the cost of clinical tests. They can achieve this result by employing appropriate computer-based information and/or decision support systems. Health care data is massive. It includes patient centric data, resource management data and transformed data.

The availability of integrated information via the huge patient repositories, there is a shift in the perception of clinicians, patients and payers from quantitative assessment of information with the supporting of all clinical and imaging data.

KNOWLEDGE DISCOVERY IN MEDICAL DATABASES:

Data mining is an essential step of knowledge discovery. In recent years it has attracted great deal of interest in information industry. Knowledge discovery process consists of an iterative sequence of data cleaning, data integration, data selection, data mining pattern recognition and knowledge presentation. In particulars, data mining may accomplish class description, association, classification, clustering, prediction and time series analysis. Data mining in contrast to traditional data analysis is discovery driven.

EXISTING SYSTEM:

The healthcare industry collects huge amounts of healthcare data which, unfortunately, are not "mined" to discover hidden information.

Clinical decisions are often made based on doctors' intuition and experience rather than on the knowledge rich data hidden in the database.

This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients.

Many healthcare organizations struggle with the utilization of data collected through an organization.

Online transaction processing (OLTP) system is not integrated for decision making and pattern analysis.

PROPOSED SYSTEM:

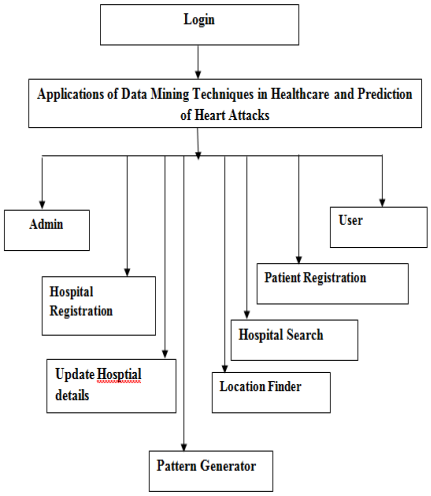
The System enables the search of the common causes and symptoms that lead to a heart attack and also provides the occurrences in form of an analysis chart. Knowledge discovery in databases is well-defined process consisting of several distinct steps.

Data mining is the core step, which results in the discovery of hidden but useful knowledge from massive databases.For successful healthcare organization it is important to empower the management and staff with data warehousing based on critical thinking.

Data warehousing can be supported by decision support tools such as data mart, OLAP and data mining tools.

With stored data in two-dimensional format OLAP makes it possible to analyze potentially large amount of data with very fast response times.

It provides the ability for users to go through the data and drill down or roll up through various dimensions as defined by the data structure.



LITERACY SPECIFICATION :

A. On the optimality of the simple Bayesian classifier under zero-one loss:

The simple Bayesian classifier has traditionally not been a focus of research in machine learning.

B. However, it has sometimes been used as a "straw man" against which to compare more sophisticated algorithms. Clark and Niblett [20] compared it with two rule learners and a decision-tree learner, and found that it did surprisingly well. Cestnik (1990) reached similar conclusions. Kononenko (1990) reported that, in addition, at least one class of users (doctors) finds the Bayesian classifier's representation quite intuitive and easy to understand, something which is often a significant concern in machine learning. John and Langley (1995) showed that the Bayesian classifier's performance can be much improved if the traditional treatment of numeric attributes, which assumes Gaussian distributions, is replaced by kernel density estimation. Although the reasons for the Bayesian classifier's good performance were not clearly understood, these results were evidence that it might constitute a good starting point for further development.

C. Clark and Niblett [20] compared it with two rule learners and a decision-tree learner, and found that it did surprisingly well. Cestnik (1990) reached similar conclusions. Kononenko (1990) reported that, in addition, at least one class of users (doctors) finds the Bayesian classifier's representation quite intuitive and easy to understand, something which is often a significant concern in machine learning. Langley, Iba, and Thompson (1992) compared the Bayesian classifier with a decision tree learner, and found it was more accurate in four of the five data sets used. John and Langley (1995) showed that the Bayesian classifier's performance can be much improved if the traditional treatment of numeric attributes, which assumes Gaussian distributions, is replaced by kernel density estimation. This showed that the Bayesian classifier's limited performance in many domains was not in fact intrinsic to it, but due to the additional use of unwarranted Gaussian assumptions.

D. The previous works in this area are based on a fixed database and did not consider that one or more transactions could be deleted, inserted, or modified in the database. By using incremental and interactive high utility pattern (HUP) mining, we can use the previous data structures and mining results, and avoid unnecessary calculations when the database is updated or the mining threshold is changed.

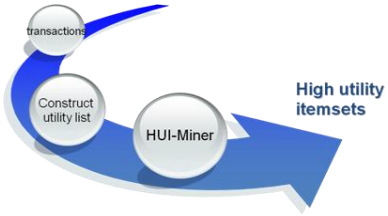
E. The algorithm Frequent Pattern-Growth method (novel algorithm) for mining frequent item sets was proposed by Han et al. It is a bottom-up depth first search algorithm. This uses FP-Tree to store frequency information of the original data base in a compressed form .Compression is achieved by building the tree in such a way that overlapping item sets share prefixes of the corresponding branches. Mining frequent itemsets is to identify the sets of items
that appear frequently in transactions in a database. The frequency of an itemset is measured with the support of the itemset, i.e., the number of transactions containing the itemset.

F. ZP, ZSP, FSH, ShFSH, DCG, Two-Phase, FUM, and DCG+ mine high utility itemsets as the famous Apriori algorithm mines frequent itemsets. Mining of frequent itemsets only takes the presence and absence of items into account. Other information about items is not considered, such as the independent utility of
an item and the context utility of an item in a transaction.

ALGORITHMS USED:

Two algorithms are used to get the efficient results from the input data. One is the HUI miner algorithm and the other is the naïve bayes classifier algorithm.

HUI-Miner uses a novel structure, called utility-list, to store both the utility information about a data and the heuristic information for pruning the search space of HUI-Miner.



An overview of the HUI miner

One of the most common approaches to mining frequent patterns is the Apriori method and when a transactional database represented as a set of sequences of transactions performed by one entity is used, the manipulation of temporal sequences requires that some adaptations be made to the Naive Bayes algorithm.

The algorithm for HUI miner is as follows:

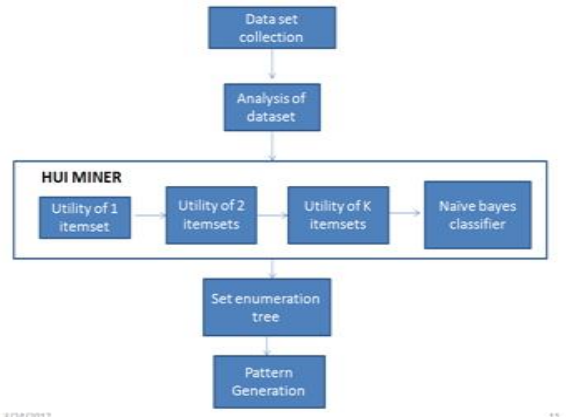
Input: P.UL, the utility-list of item set P;
Px.UL, the utility-list of item set Px;
Py.UL, the utility-list of item set Py.
Output: Pxy.UL, the utility-list of item set Pxy.
Pxy.UL = NULL;
foreach element Ex ∈ Px.UL do
if ∃Ey∈Py.UL and Ex.tid==Ey.tid then
if P.UL is not empty then
search such element E∈P.UL that
E.tid==Ex.tid;
Exy=<Ex.tid, Ex.iutil+Ey.iutil -E.iutil,
Ey.rutil>;
else
Exy=<Ex.tid, Ex.iutil+Ey.iutil, Ey.rutil>;
end
append Exy to Pxy.UL;
end
end
return Pxy.UL;

HUI-Miner Algorithm

Input: P.UL, the utility-list of item set P, initially empty;
ULs, the set of utility-lists of all P's 1-extensions;
minutil, the minimum utility threshold.
Output: all the high utility item sets with P as prefix.
foreach utility-list X in ULs do
if SUM(X.iutils)≥minutil then
output the extension associated with X;
end
if SUM(X.iutils)+SUM(X.rutils)≥minutil then
exULs = NULL;
foreach utility-list Y after X in ULs do
exULs = exULs+Construct(P.UL, X,Y);
end
HUI-Miner(X, exULs, minutil);
end
end

ARCHITECTURE:

In a database, firstly, all 1-itemsets are candidate high utility item sets. After scanning the database, the algorithms eliminate unpromising 1-itemsets and generate 2-itemsets from the remaining 1-itemsets as candidate high item sets. After the second scan over the database, unpromising 2-itemsets
are eliminated and 3-itemsets as candidates are generated from the remaining 2-itemsets.. The procedure is performed repeatedly until there is no generated candidate item set.



IMPLEMENTATION DETAILS

MODULES:

ADMIN MODULE

The hospitals to be registered needs to contact the administrator. The administrator provides the login ID for the doctors working in the hospitals which have registered. The doctors who want to be a member of the website, but working in the non registered hospitals also can create their account directly. This doesn't mean there is no proper authentication, but paves the way to provide a better knowledge discovery.
Those outside doctors can register their details and has to send the patient details to the administrator.
The administrator then checks out the details and finally after the verification is over, admin sends the ID and password to the respected doctor's email.
If any other problem arises then immediately the admin is called, the internal operations are properly handled by the admin only.

USER MODULE

The user can login and view the hospital details and the generated analysis.

HOSPITAL MODULE

The doctors can login with their user name and they are provided to view the patient record. The patient's record are updated and inserted by the doctor and make an entry in the database.
Through this the patient's record is maintained perfectly. The hospitals that are registered are provided with the

registration ID, this provides the users to view the registered hospitals , doctors list and patient details.
This will be useful to generate the pattern using NAIVE BAYESIAN classifier. This pattern will provide the appropriate occurrence diseases and its effects.

REPORT GENERATION MODULE

The patient's record maintained is then made use in generating the pattern. The rate of death due to heart diseases and their risk factors are shown in the pattern and effectively generated.
This pattern creates the user to give right decision through effective mining of data's from a certain number of hospitals and the % of causes.
The patterns are generated with different charts
Year chart
Age chart
Gender chart
Bio chemistry chart

CONCLUSION :

In this paper, we have presented an intelligent and effective heart attack prediction methods using data mining.

Firstly, we have provided an efficient approach for the extraction of significant patterns from the heart disease data warehouses for the efficient prediction of heart attack.

Based on the calculated significant weight age, the frequent patterns having value greater than a predefined threshold were chosen for the valuable prediction of heart attack.

The goals are to be evaluated against the trained models. All these models could answer complex queries in predicting heart attack.

FUTURE ENHANCEMENT:

In our future work, this can be further enhanced and expanded by adding still more algorithms to get accurate and exact results.

For predicting heart attack significantly 15 attributes are listed. Besides the 15 listed in medical literature we can also incorporate other data mining techniques, e.g., Time Series, Clustering and Association Rules. Continuous data can also be used instead of just categorical data.

We can also use Text Mining to mine the vast amount of unstructured data available in healthcare databases.

REFERENCES

[1] Frawley and Piatetsky-Shapiro, 1996. Knowledge Discovery in Databases: An Overview. The AAAI/MIT Press, Menlo Park, C.A.
[2] Miller, A., B. Blott and T. Hames, 1992. Review of neural network Applications in medical Imaging and signal processing . Med. Biol.Engg. Comp., 30: 449-464.
[3] Chen, J., Greiner, R.: Comparing Bayesian Network Classifiers. In Proc. Of UAI-99, pp.101–108 ,1999.
[4] Glymour, C., D. Madigan, D. Pregidon and P.Smyth, 1996. Statistical inference and data mining. Communication of the ACM, pp: 35-41.
[5] Chen, J., Greiner, R.: Comparing Bayesian Network Classifiers. In Proc. Of UAI-99, pp.101–108 , 1999.
[6] "Hospitalization for Heart Attack , Stroke, or Congestive Heart Failure among Persons with Diabetes", Special report: 2001 – 2003, New Mexico.
[7] "Heart disease" from http://wikipedia.org
[8] Rumelhart, D.E., McClelland, J.L., and the PDF Research Group (1986), Parallel Distributed Processing, MA: MIT Press, Cambridge.1994.
[9] Heckerman, D., A Tutorial on Learning With Bayesian Networks.1995, Microsoft Research.
[10] Neapolitan, R., Learning Bayesian Networks. 2004, London: Pearson Printice Hall.
[11] Krishnapuram, B., et al., A Bayesian approach to joint feature Selection and classifier design. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2004. 6(9): p. 1105-1111.
[12] Shantakumar B.Patil, Y.S. Kumaraswamy, Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network, European Journal of Scientific Research ISSN1450- 216X Vol.31 No.4 (2009), pp.642-656 © EuroJournals Publishing,Inc. 2009.
[13] Sellappan Palaniappan, Rafiah Awang, Intelligent Heart Disease Prediction System Using Data Mining Techniques, 978-1-4244-1968-5/08/$25.00 ©2008 IEEE.
[14] Pedro Domingos , Michae l Pazzani , On the Optimality of the Simple Bayesian Classifier under Zero -One Loss, Machine Learning, 29,103–130 (1997) c° 1997 Kluwer Academic Publishers. Manufactured in The Netherlands.
[15] Richard N. Fogoros, M.D, The 9 Factors that Predict Heart Attack 90% of heart attacks are determined by these modifiable risk factors,About.com Guide.
[16] Harleen Kaur and Siri Krishan Wasan, Empirical Study on Application of Data Mining Techniques in Healthcare Journal of Computer Science 2 (2): 194-200, 2006 ISSN 1549-3636 © 2006 Science Publications.
[17] Bressan, M. and J. Vitria, On the selection and classification of independent features. Pattern Analysis and Machine Intelligence,IEEE Transactions on, 2003. 25(10): p. 1312-1317.
[18] Domingos, P. and M. Pazzani, On the optimality of the simple Bayesian classifier under zeroone loss. Machine Learning, 1997.29(2-3): p. 103-30.
[19] Juan Bernabé Moreno, One Dependence Augmented Naive Bayes, University of Granada, Department of Computer Science and Artificial Intelligence.
[20] The CN2 Induction Algorithm, Peter Clark &Tim Niblett Article DOI: 10.1023/A:1022641700528