Performance Analysis of Data Mining Techniques over Sleeping Problem Dataset

Devinder Singh¹, Iqbal Singh²
M.Tech Student¹, Senior Engineer²
Department of Information Technology
CDAC, Mohali, India

Abstract:
Based upon the various classification algorithms such as C4.5, CHC Adaptive, steady state genetic algorithm and Generational Genetic Algorithm, the efficiency of various data mining techniques on sleep habits dataset was analyzed and compared using Euclidian Distance, HVDM, K-fold validation and 5-fold validation. The classification Techniques were evaluated with the help of various data mining tools such as WEKA,KEEL etc, which consist of machine learning language that provides the mechanism for classifying the data through the use of different algorithms. Further, using different classification algorithm, the time taken metrics were observed, analyzed and summarized to the relationships which had been identified. During the research Euclidian distance in CHC preprocessing Technique was found to be most efficient with time taken 27.89 seconds and K-fold validation with Genetic Genration algorithm was least efficient with 107.64 seconds.

Keywords: HVDM, IHDPS, HDPS, CHC, GGA, SGGA

1. INTRODUCTION

Huge amount of healthcare data is collected in the field of healthcare which, unfortunately is not “mined” to discover the hidden information for effective decision making. Discovery of hidden pattern and relationships often goes unexploited. Advanced data mining techniques can be used to resolve this problem. Data mining refers to different techniques of knowledge discovery in the database. Discovering relations that connects variables in a database is a subject of data mining. Data mining is the non-trivial extraction of implicit, previously unknown and potentially useful information from the data. The discovered knowledge can be used to improve quality of service. The role of IT is well established in today’s world. Data warehousing and knowledge management technologies can contribute in decision support system. Knowledge discovery in database is well defined process consisting of several distinct steps. Data mining is the core step, which results in discovery of hidden but useful knowledge from massive databases. Data mining technology provides a user-oriented approach to novel and hidden patterns in data. The discovered information can be used to improve quality of service. Sleep problems are now affecting peoples all over the world. The current research will apply different data mining techniques on sleep behavior related data of Population. This will generate the knowledge about risk of sleep disorder among the population. Following are some important areas of interest where data mining techniques can be of good use in the health care sector

<table>
<thead>
<tr>
<th>Data Mining for healthcare application</th>
<th>Health Insurance Sector</th>
</tr>
</thead>
<tbody>
<tr>
<td>E-Governance in health sector</td>
<td>Public Health Information System</td>
</tr>
<tr>
<td>Forecasting future behavior of patient in accordance to his/her history</td>
<td>Estimating Treatment Expenses</td>
</tr>
</tbody>
</table>

Data mining techniques are now being used in the field of medical sciences to improve healthcare services and to predict the likelihood of persons getting affected with various diseases. There have been few attempts in this regard but the scope of exploration is immense. The gaps and the lacunae in this area need to be addressed with different data mining techniques. Huge information in regard to healthcare is collected on daily bases in electronic form such as Electronic registers, excel sheets. The data is collected for many purposes. It can be for analyzing health and social condition among the population subset. Data may be city-wide, village-wide, state-wide or country-wide. Some data is collected after decades of survey and research. These types of data is used for formulating healthcare policies, for researches, future planning of healthcare, for evaluating the success of current healthcare policies in place and for other wide variety of purposes. Healthcare sector is nowadays dealing with a very peculiar problem of keeping check on patient’s need, improving the healthcare services and parallel reducing the cost of treatment. Data storage, record keeping and automating has its own challenges. Data in increasing day by day. Using this data in an efficient way will increase the efficiency of hospitals, clinics and will help in treatment cost reduction.

http://ijesc.org/
2. PERFORMANCE ANALYSIS RELATED WORK

The concept of classification method has been applied in the study of diabetes. It is a disease that can produce terrible complication such as blindness, kidney failure and premature cardiovascular deaths. In this we diagnosed about diabetes metiltus.Type-1(insulin-dependent) occurs before age of 30, although it can strike at any age. The person affected with this type is usually thin and needs insulin injections to live and dietary modification to control his/her blood sugar levels.Type-2(non-insulin dependent) occurs in obese adults over the age of 40.It is treated with diet and exercise, the blood sugar levels is lowered with drugs.Type-1 is a chronic disease of body metabolism characterized by inability to produce enough insulin to process carbohydrates, fat and proteins efficiently [1].

Associated classifications are a recent and rewarding technique which integrates association rule mining and classification to a model for predicting and achieving maximum accuracy. Associative classifiers are especially fit to applications where maximum accuracy is desired to a model for prediction. In this paper they propose as affiant associative classification algorithm using genetic approach for heart disease prediction. The main motivation for using genetic algorithm in discovery of high level prediction rules is that the discovered rules are highly comprehensible, having the high prediction accuracy. A decision support system is proposed to identify the risk score for predicting the heart disease [2].

Intelligent Heart Disease Prediction System (IHDPS) prototype has been developed using data mining techniques, namely Decision Tree, Naïve Bayes and Neural Network. Results shows each Technique has its unique strengthen realizing the objectives of the defined mining goals. IHDPS can answer complex what if queried which Traditional systems cannot [3]. IHDPS is developed using Neural Network. The HDPS system predicts the likelihood of patient getting a heart disease. For predictions the system uses sex, blood pressure, cholesterol like 13 parameters. Here two more parameters are added i.e. obesity and smoking for better accuracy. From the results it has been seen that neural network predicts the heart disease with nearly 100% accuracy. They have presented HDPS using data mining and AN techniques. From the ANN a multilayer perceptron neural network along with back propagation algorithm is used to develop the system. Because MLPNN model proves the better results and helps the domain experts and even person related with the field to plan for better diagnose and provide the patient with early diagnosis results as it performs realistically well even without retraining. The experimental result shows that using neural networks the system predicts Heart disease with nearly 100% accuracy[4].

The study of different data mining techniques can be employed in automatic heart disease prediction system. The observation reveals that Neural Network with 15 attributes has outperformed over all others data mining Techniques. The analysis shows that Neural Network with 15 attributes has shown the highest accuracy i.e. 100% so far. On the other hand decision support system has also performed well with 99.62% accuracy by using 15 attributes. Moreover in combination with Genetic Algorithm and 6 attributes, Decision Tree has shown 99.2 % efficiency[5].

2. PREDICTION TECHNIQUES OF DATA MINING

- **Decision Tree**

A decision tree is used to visually represent Decision and decision making. It describes the data but not the decisions the resultant classification tree can be used as an input for decisions. It is most common in data mining. A tree can be formed by breaking a dataset which is a source into a number of small data subsets. Decision Tree is composed of nodes and children nodes. Each node represents value of target variable

- **Lazy Learning**

Lazy Learning is a learning method in which generalization is delayed unless a query is formed whereas in other methods system tires to gneralize training set before a query has been reached.

The advantage of this technique is that it will approximate the target function locally. It can solve the problems simultaneously and can deal successfully deal with the continuous changes in problem domain.

It also has its own disadvantages notably it requires large space as huge amount of data is collected. Noisy and irrelevant dataset increases the database without any purpose. It is a slower method

3. TOOLS USED

The evaluation has been conducted on Intel i7 platform consisting of 500GB HDD and 8 gb RAM. The software used is KEEL and the operating system is Window 7.KEEL is data mining software written in Java Language.

4. RESULTS AND OBSERVATIONS

The comparison was made in two parts on sleeping habits dataset. The classification algorithm has been applied by using two distance function namely Euclidian and HVDM. Three preprocessing Techniques namely CHC Adaptive, generational Genetic Algorithm and Steady state genetic algorithm using k-fold cross validation .5-fold validation and without fold validation to find out the most efficient among them. In K-fold Validation mode the time taken by C4.5 Decision Tree algorithm is 45.45 it was less as compared to Genetic Generational Algorithm and steady state genetic algorithm which was 107.64 and 92.89 respectively. In 5-fold Validation mode The C4.5 Decision Tree algorithm whom implemented using 5-fold validation mode. The time taken by Euclidian distance in CHC preprocessing Technique is 27.89 by using Genetetic Genaration algorithm and steady state genetic algorithm is 42.31 and 38.5 respectively. In without validation mode the time taken by C45 using Euclidian Distance in CHC preprocessing technique is 28.61 and Genetic Genaration algorithm and steady state genetic algorithm is 48.31 and 35.5 respectively
Figure 1. TIME Difference between different pre-processing Technique by using C45 algorithm in k fold validation mode

Figure 2. TIME Difference between different pre-processing Techniques by using C45 algorithm in without fold less mode

5. REFERENCES


[6]. YosawinKangwanariyakulet(2010).data Mining of Magneto cardiograms for prediction of ischemic heart disease. EXCLI Journal,4:82-95
[7]. Subbalakshmi G Ramesh K and Rao MC(2011)Decision Support in Heart disease prediction using Naïve Bayes


