Analysis On Data Mining Techniques For Heart Disease Dataset

Subhashri.K [1], Arockia Panimalar.S [2], Ashwin.S[3], Vignesh.P[4]

[1,2] *Assistant Professor, Department of BCA & M.Sc SS, Sri Krishna Arts and Science College, Tamilnadu, India*
[3,4] *III BCA, Department of BCA & M.Sc SS, Sri Krishna Arts and Science College, Tamilnadu, India*

---***---

Abstract - *Data Mining is an analytic process designed to explore data (usually large amounts of data - typically business or market related - also known as "big data") in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data. The ultimate goal of data mining is prediction - and predictive data mining is the most common type of data mining and one that has the most direct business applications. Classification trees are used to predict membership of cases or objects in the classes of a categorical dependent variable from their measurements on one or more predictor variables. Classification tree analysis is one of the main techniques used in Data Mining. During my research, I had analyzed the various classification algorithms and compared the performance of classification algorithms on aspects for time taken to build the model, by using different distance function. The result is being tested on data set which is taken from UCI repositories. The aim is to judge the efficiency of different data mining algorithms on Heart Disease dataset and determine the optimum algorithm. The performance analysis depends on many factors encompassing validation mode, distance function, different nature of dataset.*

Key Words: **Data Mining, Classification, Classification Techniques, Distance function, KEEL Tool, Performance Analysis.**

1. INTRODUCTION

The healthcare industry collects huge amounts of healthcare data which, unfortunately are not **"mined"** to discover hidden information for effective decision making. Discovery of hidden patterns and relationships often goes exploited. Data mining refers to using a variety of techniques to identify suggest of information or decision making knowledge in the database and extracting these in a way that they can put to use in areas such as decision support, prediction ,forecasting and estimation. Discovering relations that connect variables in a database is the subject of data mining. Data mining is the non-trivial extraction of implicit, previously unknown and potentially useful information from 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 and also used by the medical practitioners to reduce the number of adverse drug effect. In information technology, knowledge is one of the most significant assets of any organization. The role of IT in healthcare is well established. Knowledge Management in Health care offers many challenges in creation, dissemination and preservation of health care knowledge using advanced technologies. Pragmatic use of database system, Data Warehousing and Knowledge Management technologies can contribute a lot to decision support systems in health care.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. Following are some of the important areas of interests where data mining techniques can be of tremendous use in health care management. **(Gnanadesikan, et al...(1977).**

1. Data modelling for health care applications.
2. Executives Information System for health care.
3. Forecasting treatment costs and demand of resources.
4. Anticipating patient's future behaviour given their history.
5. Public health Informatics.
6. E-governance structures in health care.
7. Health Insurance.

2. CLASSIFICATION

Classification is the task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam". An algorithm that implements classification, especially, in a concrete implementation, is known as a classifier.The term "classifier" sometimes also refers to the mathematical function, implemented by a classification algorithm that maps input data to a category.Classification and clustering are examples of the more general problem of pattern recognition, which is the assignment of some sort of output value to a given input value.

Classification Algorithm

A. Decision Tree

In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. In data mining, a decision tree describes data but not decisions; rather the resulting classification tree can be an input for decision making. Decision tree learning is a method commonly used in data mining. The goal is to create a model that predicts the value of a target variable based on several input variables. Each interior node corresponds to one of input variables . An example is shown on the right. Each leaf represents a value of the target variable given the values of the input variables represented by the path from the root to the leaf.

A tree can be "learned" by splitting the source set into subsets based on an attribute value test. In data mining, decision trees can be described also as the combination of mathematical and computational techniques to aid the description, categorisation and generalisation of a given set of data. (Kaur H and Wasan KS et al...2006) .

B. Lazy Learning

In artificial intelligence, lazy learning is a learning method in which generalization beyond the training data is delayed until a query is made to the system, as opposed to in eager learning, where the system tries to generalize the training data before receiving queries.

The main advantage gained in employing a lazy learning method, such as Case based reasoning, is that target function will be approximated locally, such as in the k-nearest neighbour algorithm. Because the target function is approximated locally for each query to the system, lazy learning systems can simultaneously solve multiple problems and deal successfully with changes in the problem domain. The disadvantages with lazy learning include the large space requirement to store the entire training dataset. Particularly noisy training data increases the case base unnecessarily, because no abstraction is made during the training phase.

Selected Distance Function

- **Euclidian Distance Function**
- **HVDM Distance Function**

i. Euclidean Distance

In mathematics, the Euclidean distance or Euclidean metric is the "ordinary" distance between two points that one would measure with a ruler, and is given by the Pythagorean formula. By using this formula as distance, Euclidean space (or even any inner product space) becomes a metric space. The associated norm is called the Euclidean norm. Older literature refers to the metric as Pythagorean metric.

Definition

The **Euclidean distance** between points **p** and **q** is the length of the line segment connecting them: **p.q**
In Cartesian coordinates,
if **p** = (*p* , *p* ,..., *p*) and **q** = (*q* , *q* ,..., *q*) are two points in Euclidean *n* space, then the distance from **p** to **q**, or from **q** to **p** is given by the following hetrogenous value difference metric.

ii. Hetrogenous Value Difference Metric

Instance-based learning technique typically handle continuous and linear input values well,but often do not handle nominal input attributes appropriately. The Value Difference Metric (VDM) was designed to find reasonabledistance values between nominal attribute values, but it discretization to map continuous values into nominal values.

This paper proposes three new heterogeneous distance functions, called the Heterogeneous Value Difference Metric (HVDM), the Interpolated Value Difference Metric (IVDM), and the Windowed Value . These new distance functions are designed to handle applications with nominal attributes, continuous attributes, or both. In experiments on 48 applications the new distance metrics achieve higher classification accuracy on average than three previous distance functions on those datasets that have both nominal and continuous attributes.

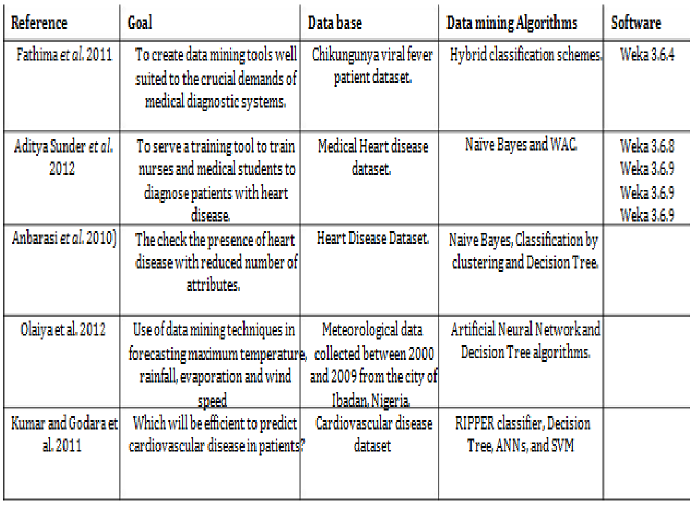
So HVDM is used as shown below:

$$HVDM(x, y) = \sqrt{\sum_{a=1}^{m} d_a{}^2(x_a, y_a)}$$

3. RELATED WORK

The clinical and physical diagnosis of Chikungunya viral fever patients and its comparison with dengue viral fever has been proposed.

Table 1: Summary of selected reference with goal



Reference	Goal	Data base	Data mining Algorithms	Software
Fathima *et al.* 2011	To create data mining tools well suited to the crucial demands of medical diagnostic systems.	Chikungunya viral fever patient dataset.	Hybrid classification schemes.	Weka 3.6.4
Aditya Sunder *et al.* 2012	To serve a training tool to train nurses and medical students to diagnose patients with heart disease.	Medical Heart disease dataset.	Naïve Bayes and WAC.	Weka 3.6.8 Weka 3.6.9 Weka 3.6.9 Weka 3.6.9
Anbarasi *et al.* 2010)	The check the presence of heart disease with reduced number of attributes.	Heart Disease Dataset.	Naive Bayes, Classification by clustering and Decision Tree.	
Olaiya et al. 2012	Use of data mining techniques in forecasting maximum temperature, rainfall, evaporation and wind speed	Meteorological data collected between 2000 and 2009 from the city of Ibadan, Nigeria.	Artificial Neural Network and Decision Tree algorithms.	
Kumar and Godara et al. 2011	Which will be efficient to predict cardiovascular disease in patients?	Cardiovascular disease dataset	RIPPER classifier, Decision Tree, ANNs, and SVM	

Our project aims to integrate different sourcesof information and to discover patterns of diagnosis, for predicting the viral infected patients and their results. The aim is to apply hybrid classification schemes and create data mining tools well suited to the crucial demands of medical diagnostic systems. The approaches in review are diverse in data mining methods. (Fathima *et al.* 2011).

The prototype has been described using data mining techniques, namely Naïve Bayes and WAC (weighted associative classifier). It enables significant knowledge .Eg. patterns, relationships between medical factors related to heart disease, to be established. It can serve a training tool to train nurses and medical students to diagnose patients with heart disease. It is a web based user friendly system and can be used in hospitals if they have a data ware house for their hospital. The models were validated using Classification Matrix **(Aditya Sunder *et al.* 2012).**

The proposed work is to predict more accurately the presence of heart disease with reduced number of attributes. This was carried out using Artificial Neural Network and Decision Tree algorithms and meteorological data collected between 2000 and 2009 from the city of Ibadan, Nigeria.It has been described that the data mining classification techniques RIPPER classifier, Decision Tree, ANNs, and SVM are analyzed on cardiovascular disease dataset.There analysis shows that out of these four classification models SVM predicts cardiovascular disease with least error rate and highest accuracy. (Kumar and Godara et al. 2011).

4. DATASETS AND TOOLS

A. Hardware

We conduct our evaluation on Intel Pentium P6200 platform which consist of 1 GB memory and 320 GB hard disk.

B. Software

In this experiment, we used KEEL tool and window 7 to evaluate the performance of classification algorithms using time taken to build the model according to respective no of clusters. KEEL is machine learning/data mining software written in Java language (distributed under the GNU Public License).

KEEL is a collection of machine learning algorithmsfor data mining tasks. KEEL contains tools for developing new machine learning schemes. It can be used for Pre-processing, Classification, Clustering, Association and Visualization.

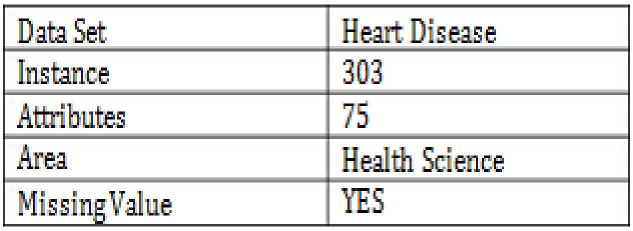
C. Data Set

The input data set is an integral part of data mining application. The data used in my experiment is either real world data obtained from UCI machine learning repository and widely accepted data set available in KEEL toolkit. Heart Disease data set comprises 303 instances and 75 attributes in the area of Health Science and some of them contain missing value.

5. EXPERIMENTS RESULT AND DISCUSSION

To evaluate the selected tool using Heart Disease dataset and comparisons are performed in two parts. In first Comparison, I have applied these Classification algorithms by using two distance function namely Euclidean Distance and HVDM Distance in three different Pre-processing techniques namely CHC Adaptive search for advanced selection, GGA-TSS Generational Genetic Algorithm for Instance selection, SGGA-TSS Steady-state genetic algorithm for Instance selection, using different validation modes namely K-Fold cross validation, 5-Fold Validation and without validation to found the most efficient algorithm among two algorithms.

Table 2: The UCI datasets used for the experiments and their properties



Data Set	Heart Disease
Instance	303
Attributes	75
Area	Health Science
Missing Value	YES

In K-fold Cross Validation Mode: By using Euclidean Distance function, in K-fold Cross validation mode the minimum time taken by C4.5 Decision tree algorithm was 44.443 in CHC pre-processing technique is least as compares to GGA and SGGA pre-processing techniques. The time taken by GGA and SGGA are 106.613 and 89.485 respectively.

When the C4.5 Decision Tree algorithms were applied using HVDM Distance function, the minimum time taken to build the model by CHC pre-processing technique is 157.604 in k-fold cross validation mode as compared to both GGA and SGGA pre-processing technique i.e. 899.583 and 1586.068. While when test was applied using KNN technique in k-fold validation mode by using Euclidean distance in CHC pre-processing technique the time taken is 46.78 but in GGA and SGGA pre-processing technique the time taken is 107.458 and 90.752 resp.

In 5-fold Validation Mode: Then C45 Decision Tree algorithm is implemented by using 5-fold validation mode. The time taken by C45 using Euclidian distance in CHC technique is 72.306,239.223 and 198.667 resp.

Without Fold Mode: In last without validation mode both two techniques were implemented by using without validation mode.By using HVDM distance the time taken in CHC pre-processing technique is 72.306.

In GGA pre-processing technique is 239.223 pre-processing technique is 72.306, in GGA pre-processing technique is 239.223 and in SGGA pre-processing technique it 198.667.

The time taken by KNN technique using Euclidian distance in CHC pre-processing technique is 5.523 and using GGA and SGGA pre-processing technique the time and by is 13.79 and 11.419 resp. Then running time of each algorithm and distance function is evaluated at each validation mode.

6. CONCLUSION

By using Euclidian distance the time taken to build themodel in K-fold validation mode is 80.180 obtained by C45.In 5-fold validation mode and 0 validation mode, C45 is attaining 35.109 and 35.109 respectively. When the algorithm is applied using HVDM distance function the minimum time taken to build the model by C45 in K-fold validation mode,5-fold validation mode and 0 validation is 881.085,170.065,170.065 respectively. By using KNN algorithm the time taken to build the model by using Euclidian distance in K-fold validation mode is 81.66.In 5-fold validation mode and 0 validation mode, KNN-lazy learning is attaining 19.572 and 10.244 respectively.

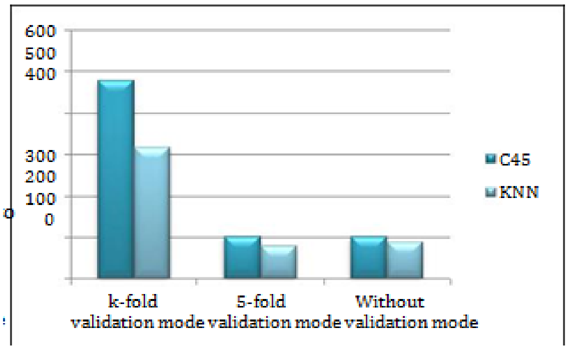
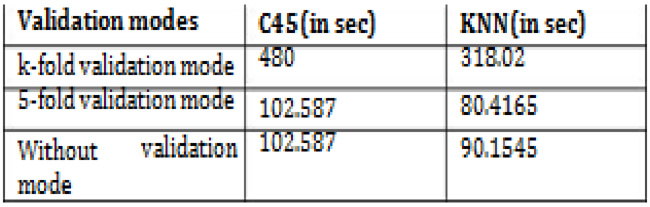


Fig.1: A comparative difference between the two Algorithms by comparing different validation mode.

When the algorithm is applied using HVDM distance function the minimum time taken to build the model by KNN algorithm in K-fold validation mode, 5-fold validation mode and 0 validation is 554.399, 141.261, 170.065 respectively.

Table 2: Time difference between different Validation modes by using C45 and KNN algorithm



Validation modes	C45(in sec)	KNN(in sec)
k-fold validation mode	480	318.02
5-fold validation mode	102.587	80.4165
Without validation mode	102.587	90.1545

We have analyzed the heart disease dataset by using KEEL tool. In KEEL tool different validation mode is selected to perform the operation on a dataset. Then different pre-processing technique is used to remove the noise in a dataset. Finally ClassificationAlgorithm has been selected to perform the analysis of the algorithm by comparing the time taken by different algorithms on a dataset.

More and more Classification algorithm is made available to find the best performance of the heart disease dataset that which algorithm performs fast. Many algorithms have been studied by the researchers to find theoptimum algorithm. Our focus here through Classification algorithm is to determine that which algorithm is optimum to give the best result in a less time by using different validation mode available in a tool. This study confirms that Lazy Learning – KNN is the efficient algorithm in predicting the performance of the heart disease dataset using without validation mode. We aim to carry out this study on other machine learning Classification Algorithm and our focus is on to make a predictive system to find the efficient performance of heart disease dataset in heart disease prediction system.

7. REFERENCES

[1].Gnanadesikan, R. (1977) Methods for Statistical Data Analysis of Multivariate Observations, Wiley. ISBN 0 471-30845-5 (p.83–86).

[2].FathimaSA,Manimegala D and Hundewale N (2011) A Review of Data Mining Classifications Applied for Diagnosis and Prognosis.

[3].International Journal of Computer Science, 6:322-328 Sundar AN, Latha PP, Chandra RM(2012) Performance Analysis of classification Data Mining Techniques Over Heart Disease.

[4].M.Anbarasiet. al. (2010) Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm.

[5].Olaiya F and Adeyem BA (2012) Application of Data Mining Techniques in Weather Prediction Climate Change Studies. International Journal of Information

Engineering and Electronic Business, 3:51-59.

[6].Milan Kumari M and Godara S (2011)Comparative Study ofData Mining Classification Methods in Cardiovascular DiseasePrediction. International Journal of Computer Science andTechnology, 6:304-308.

[7].Meena K, Subramaniam RK, Gomathy M (2012). Performance Analysis of Gender Clustering and Cla ssification Algorithms.International Journal of Computer Science and Engineering,5:442-457.

[8].Kaur H and Wasan KS (2006) Empirical Study on Applications of Data Mining Techniques in Healthcare. International Journal of Computer Science, 4: 194-200.

[9].K.Srinivas et al. (2010) Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks. International Journal on Computer Science and Engineering, 5: 250-255.

[10]. RaniM, SinghV and Bhushan B (2013) Performance Evaluation of Classification Techniques Based on Mean Absolute Error. International Journal of Computing and Business Research, 4: 1-5.