Research Paper
Impact Factor: 4.164
Refereed Journal

IJMDRR
E- ISSN –2395-1885
ISSN -2395-1877

A COMPARATIVE STUDY OF DATA MINING TECHNIQUES IN CLINICAL DATA SETS

P.Meena* **M.Sowmiya****
**Department of Computer Science, Adhiyaman Arts and Science College (w), Uthangarai.*
***Department of Computer Science, Arignar Anna Arts and Science College, Krishnagiri.*

Abstract
Medical science industry has huge amount of data, but unfortunately most of this data is not mined to find out hidden information in data. Advanced data mining techniques can be used to discover hidden pattern in data. Models developed from these techniques will be useful for medical practitioners to take effective decision. Nowadays, the healthcare sector is one of the areas where huge data are daily generated. However, most of the generated data are not properly exploited. Important encapsulated data are currently in the data sets. Therefore, the encapsulated data can be analyzed and put in to useful data. Data mining is very challenging task for the researchers to make diseases prediction from the huge medical databases. To succeed in dealing with this issue, researchers apply data mining techniques such as classification, clustering, association rules and so on. The main objective of this research is to predict heart diseases by the use of classification algorithms namely Naïve Bayes and Support Vector Machine in order to compare them on the basis of the performance factors i.e. probabilities and classification accuracy. In this paper, we also developed a computer-based clinical Decision Support system that can assist medical professionals to predict heart disease status based on the clinical data of the patients using Naive Bayes Algorithm.

Introduction

Data mining is a well Established area of research that has become increasingly popular in health domain in recent years. It plays a vital role the health care towards uncovering latest trends specifically in early disease predictions. Data mining is now becoming helpful for researchers and scientist towards gaining novel and deep insights of any large biomedical datasets. Uncovering new biomedical and healthcare related knowledge in order to support clinical decision making, is another dimension of data mining. Early disease prediction has now become the most demanding area of research in health care sector. As health care domain in bit wider domain and having different disease characteristics, different techniques have their own prediction efficiencies, which can be increased and turned in order to get in to most optimize way. In this research work, authors have comprehensively compared different data mining techniques and their prediction capability on different set of heart disease datasets.

In recent days, the World Health Organization (WHO) has estimated that 12 million deaths occur worldwide, every year due to heart diseases and Heart Diseases remains one of the major causes of death in the world. In 2008, 17.3 people died due to Heart Disease. The World Health Organization Statistics 2012 reports have shown enlightens the fact that one in three adults worldwide has raised blood pressure a condition that cause around half of all deaths from stroke and health disease. In May 2014, WHO also estimated the rate of 93.49 of heart disease in the republic of chad and by 2030, almost 23.6 million worldwide will die due to Heart disease.

Heart disease can be also known as (CVD) Cardiovascular disease, contains a number of conditions that affect the heart including the heart attacks. In addition, Heart Disease possess some functional problem of the heart such as infections of heart muscles like myocarditis (inflammatory heart diseases), heart-valve abnormalities or irregular heart rhythms etc. are the reasons that can be led to heart failure.

The heart is the organ that pumps blood, with its life giving oxygen and nutrients, to all issues of the body. If the pumping action of the heart becomes inefficient, vital organs like the brain and kidneys suffer and if the heart stops working altogether, death occurs within minutes. Life itself is completely dependent on the efficient operation of the heart. Cardiovascular disease is not contagious; you can't catch it like you can the flu or a cold. Instead, there are certain things that increase a person's chances of getting cardiovascular disease. Cardiovascular disease (CVD) refers to any condition that affects the heart. May CVD patients have symptoms such as chest pain (angina) and Fatigue, which occur when the heart isn't receiving adequate oxygen. As per a survey nearly 50 percent of patients, however, have no symptoms until a heart attack occurs. A number of factors have been shown to increase the risk of developing CVD.

Factors

1. Family history of cardiovascular disease.
2. High levels of LDL (bad) cholesterol.
3. Low levels of HDL (good) cholesterol.
4. Hypertension.

5. High fat diet.
6. Lack of regular exercise.
7. Obesity.

With so many factors to analyze for a diagnosis of cardiovascular disease, physicians generally make a diagnosis by evaluating a patient's current test results. Previous diagnoses made on other patients with the same results are also examined by physicians. These complex procedures are not easy. Therefore, a physician must be experienced and highly skilled to diagnose cardiovascular disease in a patient
Heart disease

Heart disease is the type of disease that deals with the operation of the heart by narrowing or blockage of the arteries and vessels that supply oxygen and nutrient-rich. It is caused by the following factors:

1. **The family history of heart disease (Heredity):** people should know that the heart disease can be inherited from the family when one of them is heart disease contractor.
2. **Smoking:** Approximately 40% of person dies from tobacco due to heart attack and blood vessel disease and a smoker's risk of heart attack can rapidly reduce within twelve (12) months of tobacco abstinence.
3. **Cholesterol:** the increment of facts in the blood is a risk factor for heart diseases. Cholesterol is substances consist of lipids in the bloodstream and all over the body's cells. High level of the fat in the body with high level of LDL (low-density lipoprotein) cholesterol can rise up atherosclerosis which will lead to the risk increment of heart disease.
4. **High blood pressure:** high blood pressure also known as HBP or hypertension is a widely misunderstood medical condition. High blood pressure increases the risk of the wall of our blood vessels walls becoming overstretched and injured. Also, increase the risk of having a heart attack or stroke and of developing heart failure, kidney failure, and peripheral vascular disease.
5. **Obesity:** the term obesity is used to describe the health condition of anyone significantly above his or her ideal healthy weight. Being obese puts anybody at a higher risk for health problems such as heart disease, stroke, high blood pressure, diabetes and more.
6. **Lack of physical exercise:** lack of exercise is a risk factor for developing coronary artery disease (CAD). Lack of physical exercise increases the risk of CAD because it also increases the risk for diabetes and high blood pressure.

Heart Diseases are generally provoked by the abovementioned factors and the world health organization survey shows in 2012, an estimated 56 million people died worldwide, every year because of Heart Disease [1].

The summarized version related to the application of different data mining techniques in different disease identification or used as a data sets, is as under:

Technique	Mining Problem Domain
Association Rules	Patterns Identification
J48, C4.5, C5, and CART	Decision Support for Heart Disease, Hypothyroid, Dengue
K-means SVM and Naïve Bayes	Classification for Dengue disease dataset
K-NN	Classification for Diabetes, Cancer datasets
Appriori	Association rule mining
Bayesian Ying Yang	Classification on Liver Disease dataset
Neural Network	Patterns and Trends Extraction
Outlier Prediction Technique	Classification
Fuzzy Cluster Analysis	Medical images
Classification Algorithm	Disease Classification for Cardio Vascular Disease datasets
Bayesian Network Algorithm	Analysis of medical data for Coronary Heart Disease
Naïve bayesian	Classification for Coronary Heart Disease datasets
sGenetic Algorithm	Classification for Diabetes datasets

Time Series Technique	Disease diagnosis
Clustering and Classification	Clustering and classification of biomedical datasets
SVM	Classification for Diabetes datasets
Fuzzy	Drugs and Health effects classification
SVM, ANN and ID3	Classification
Naive Bayes, SVM	Kidney Disease
BPN, RBF, RF	Chronic Kidney Disease
Naive Bayes, MPL, SVM, J48 Conjunction Rule, Decision Table	Chronic Kidney Disease
SVM, KNN	Chronic Kidney Disease
AD Trees, J48 KStar, Naïve Bays, Random Forest	Kidney Disease

Comparing Naive Bayes, Support Vector Machine and Decision Tree Experimental Results

In the final experiment also Rapid miner has been used as a tool for evaluating and comparing three classification techniques using three classes high, medium and low with diabetic patient dataset to determine the possible ways to predict the risk of heart disease for diabetic patients.

In general, The Bayes theorem formula is P(h/D)= P(D/h) P(h) / P(D) where

P (h) - Prior probability of hypothesis h
P (D) - Prior probability of training data D
P (h/D) - Probability of h given D and
P (D/h) - Probability of D given h

Naive Bayes algorithm uses the Bayes formula, which calculates the probability of a patient record Y having the class label Cj.

The label could be “High”, “Medium” and “Low”.
P(label= Cj | Y) = P(Y|label= Cj) * P(Cj) /P(Y)z

Naive Bayes, Support Vector Machine, Decision Tree performances

In order to validate the final results obtained in the research presented, experiments were carried out by combining the three techniques and the performance of Bayes theorem, SVM and Decision tree.

Accuracy of Bayes Theorem Performance

	True low	**True medium**	**True high**	**Class Precision**
pred. low	631	62	21	88.38%
pred. Medium	39	98	26	60.12%
pred. high	11	25	86	70.49%
class recall	92.66%	52.97%	64.66%	

Accuracy of Support vector machine performance

	True low	**True Medium**	**True high**	**Class precision**
pred. low	398	25	15	90.87%

pred. medium	283	155	59	31.19%
pred. high	0	5	59	92.19%
class recall	58.44%	83.78%	44.36%	

The decision tree using various split methods such as Gain ratio, Information gain and Gini index has been which gives different levels of accuracy.

Accuracy by split methods using decision tree

Split method **Criteria**	**Accuracy in** **percentage**	**Classification** **error in** **Percentage**
Gain ratio	88.19	11.81
Information gain	90.79	9.21
Gini Index	87.69	12.31

For classification problems, it is natural to measure a classifier's performance in terms of the error rate. The classifier predicts the class of each instance. The correct class is counted as success and, if not, it is taken as an error. The error rate is just the proportion of errors made over a whole set of instances, and it measures the overall performance of the classifier. With the use of the information gain as split parameter in decision trees, the results are exhibited by average precision, recall and accuracy of this technique was found to be 90.79 %.

Accuracy of Decision tree Performance

	True low	**True Medium**	**True high**	**Class Precision**
pred. low	657	25	11	94.81%
pred. Medium	18	148	20	79.57%
pred. high	6	12	102	85.00%
class recall	96.48%	80.00%	76.69%	

Niyati Gupta et al. (2013) have defined the accuracy as the proportion of instances that are correctly classified. It is calculated by the total number of correctly predicted "high risk" (true positive) and correctly predicted "low risk" (true negative) over the total number of classifications.
It can be calculated as

Accuracy = (TP + TN) / (TP + TN + FP + FN)

For a multiclass classification problem, TP, FP, TN and FN for each class i are indicated as definitions. TPi, FPi, TNi, and FNi for the class i are also defined. Then, certain parameters can be calculated to evaluate the multiclass classification results accordingly. For e.g., True Positive Rate TPR, Precision and f-Measure value for each class and the overall accuracy can be calculated.

Accuracy of various classification techniques (High, Medium, Low)

Technique	**Accuracy In percentage**
Decision tree	**90.79**
Naïve Bayes	**81.58**
Support Vector Machine	**61.26**

.

Decision tree appears to be most effective as it has the highest percentage of correct predictions (90.79%) for patients with heart diseases, followed by followed by naïve Bayes and support vector machine.

When more than two classes are dealt with, the accuracy alone might not be sufficient. So evaluation of precision, sensitivity, specificity and F-score along with accuracy to determine the right classifier has been done.

According to Sheik Abdullah et al. (2012) precision is the fraction of retrieved instances that are relevant and recall is the fraction of relevant instances that are retrieved.

The precision can be calculated as

Precision = TP / (TP + FP)

However, TP rate alone is not sufficient to fully measure performance of the classifier in a single class. Therefore we compute Precision for class i as,

Precision i = TPi / (TPi + FPi)

The sensitivity is the proportion of positive instances that are correctly classified as positive (e.g. the proportion of sick people that are classified as sick). It is also called as Recall.

It can be calculated as

Sensitivity = TP / (TP + FN)

The specificity is the proportion of negative instances that are correctly classified as negative (e.g. the proportion of healthy people that are classified as healthy). It can be calculated as

Specificity = TN / (TN + FP)

F-score or F-measure is a measure of a test's accuracy and it is the harmonic mean of precision and recall which can be calculated as

F-score = 2 * (Precision * Recall) / (Precision + Recall)

We can also compute F-Measure for class i as,

F=2 *(Precision i + TP_rate i) / (Precision i + TP_rate i)

The risks calculated are arrived at by using various classification techniques

Performance of sensitivity, specificity and F-Score

Classification	Accuracy	Precision	Sensitivity	Specificity	F-Score
Decision tree	90.79	0.86	0.84	0.93	0.84
Naïve Bayes	81.58	0.72	0.69	0.85	0.70
Support Vector Machine	61.26	0.71	0.61	0.80	0.65

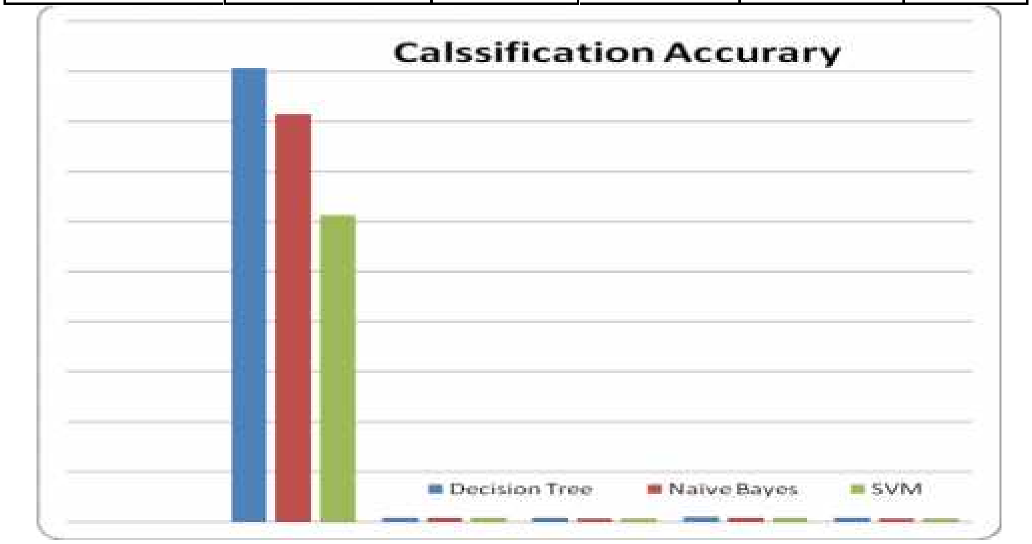

Figure :The Classification accuracy Performance of sensitivity, specificity and precision by Class

Classification Models	Overall Accuracy In Percent	Precision			Recall (Sensitivity)			per class Specificity		
		Med	High	Low	Med	High	Low	Med	High	Low
Decision tree	90.79	0.79	0.85	0.94	0.80	0.76	0.96	0.95	0.97	0.87
Naïve Bayes	81.58	0.60	0.70	0.88	0.52	0.64	0.92	0.92	0.95	0.68
Support Vector Machine	61.26	0.31	0.92	0.90	0.83	0.44	0.58	0.57	0.99	0.84

Conclusion

In information mining, smart strategies are connected keeping in mind the end goal to separate information designs. There are gigantic chances to help doctors manage this vast measure of information. The therapeutic information digging has extraordinary potential for investigating the concealed examples in the informational collections. These examples can be used for clinical analysis. Acknowledgment and characterization of examples in multivariate patient traits empower forecast of future results in view of past encounters. Our examination predicts and orders the information with a sensible precision. It helps in quality medicinal services administrations in view of the patient's needs, manifestations and inclinations. It limits the sitting tight time for restorative treatment.

The gullible Bayes show could characterize 74% of the info occasions effectively. It showed an accuracy of 71% on a normal, review of 74% on a normal, and F-measure of 71.2% on a normal.

SVM classifier has been utilized adequately and the outcomes demonstrate a high characterization exactness i.e 94.60% generally speaking, and a high accuracy for the positive class (97.52%) additionally the review of the positive class is very great (83.10%). On account of negative classes, the classifier shows high accuracy (93.67%) and also high review (99.10%).

The utilization of the choice tree utilizing different split strategies, for example, pick up proportion, data pick up and gini list has been endeavored in the present investigation. The data pick up is utilized as part parameter in choice trees. The outcomes are displayed by normal exactness, and review and we found that precision of this procedure is 90.79 % taken after by gullible Bayes (81.58%) and bolster vector machine (61.26%) for anticipating the coronary illness for diabetic patients utilizing analytic highlights. The exhibitions are looked at through precision, affectability, specificity and F-score.

Choice tree display was steady in its execution and beat innocent Bayes and SVM show. So we at last adjusted the choice tree demonstrate for ideal execution for anticipating the odds of coronary illness for diabetic patients.

References

1. Adriaans. P and Zantinge.D, “Data mining,” Addison-Wesley, 1999.
2. Andonie. R and Kovalerchuk. B, Neural networks for data mining: constrains and open problems.
3. Berson. A and Smith. S. J,” Datawarehousing, Datamining and OLAP, “ Tata Magraw Hill, 2004.
4. Gala.R, Gala.D and Gala.S, “Diabetes, high blood pressure without any fear,” Navneet publications, 2005.
5. “Introduction to Data mining and knowledge discovery,” two crows corporations, 2005.
6. Kononenko. I, “Machine learning for medical diagnosis:” History, state of the art and perspective.
7. Lavrac.N, Keravnou .E and Zupan.B,” Intelligent Data Analysis in medicine”.
8. Liu N.K. and Sin K.Y, “An Integrated data mining approach for maintenance scheduling,” Engineering Intelligent Systems. Vol 8 no 2 june 2000.
9. Magoulas G.D and Prentza.A, “Machine learning in Medical Applications”.
10. Main.J, Dillon .T and Shiv.S, “A tutorial on case based reasoning”, Soft Computing in case based reasoning, springer- verlag ltd, 2001 pp 1-28.
11. Mehrotra. K, Mohan. C. K and Ranka. S, “Elements of artificial neural networks, “ Penram International, 1997.
12. Pujari A.K, “Data mining techniques,” University press, 2005.
13. Tsatsoulis. C and Williams A.B, “Case based reasoning,” knowledge based systems, vol-3, pp. 807-835.
14. WebSite URL http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes, 08/08/2008.
15. Michie.D, Spiegelhalter. D.J and Taylor .C.C. “Machine Learning, Neural and Statistical Classification” Chapter 9 page No 157,158.
16. Misra .B.B, Dehuri. S, 2007, Functional Link Artificial Neural Network for Classification Task in Data Mining, Journal of Computer Science 3 (12):948- 955, ISSN 1549-3636 Science Publications.
17. Jeatrakul .P and Wong .W.K, “Comparing the Performance of Different Neural Networks for Binary Classification Problems”, 2009 Eighth International Symbosium on Natural Language Processing, Page 111-115.
18. Lena Kallin Westin, “Missing data and the preprocessing perceptron”, page 3, ISSN-0348-0542.
19. Bylander .T (2002). How your instructor created a Bayes network from the diabeticdataset.
20. Estebanez .C, Aler .R and Valls. M, “Genetic Programming Base Data Projections for Classification Tasks, World Academy of Science, Engineering and technology” 2005, Pages 56-61.
21. Frawley.W, and Piatetsky-Shapiro .G and Matheus .C, ”Knowledge Discovery in Databases”: An Overview. AI Magazine, Fall 1992, pgs 213-228.

22. Kononenko. I, “Machine learning for medical diagnosis:” History, state of the art and perspective.

23. Tom Mitchell, “Machine Learning,” McGraw Hill, 1997.
24. Abdel-Badeeh M. Salem, Kenneth Revett, El-Sayed A. El-Dahshan, “Machine Learning in Electrocardiogram Diagnosis,” Proceedings of the International Multiconference on Computer Science and Information Technology, pp. 429 – 433 ISSN 1896-7094 ISBN 978-83-60810-22-4.
25. Widrow B., Rumelhard D.E., and Lehr M.A. “Neural networks: Applications in industry, business and science, Commerce”. ACM, vol. 37, p. 93–105, 1994.
26. Altman E.I., Marco G., and Varetto F., “Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks” (the Italian experience), J. Bank. Finance, vol. 18, pp. 505–529, 1994.
27. Guyon I., “Applications of neural networks to character recognition,” Int. J. Pattern Recognit. Artif.Intell.vol. 5, pp. 353–382, 1991.
28. Bourlard H. and MorganN. “Continuous speech recognition by connectionist statistical methods”, IEEE Trans. Neural Networks, vol. 4, pp.893–909, 1993.
29. Lampinen J., Smolander S. and orhonen M.K. “Wood surface inspection system based on generic visual features,” Industrial Applications of Neural Networks, F. F. Soulie and P. Gallinari, Eds, Singapore: World Scientific, 1998, pp. 35–42.
30. Barlett E.B., and UhrigR.E., “Nuclear power plant status diagnostics using artificial neural networks,” Nucl. Technol., vol. 97, pp. 272–281, 1992.