ijecs open access www.ijecs.in

International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 4 Issue 9 Sep 2015, Page No. 14328-14335

Impact Of Data Mining Techniques In Medical System

[1]*Reema Arora*, [2]*Sandeep jaglan*
reema.arora0414@gmail.com

Abstract

The objective of this master's thesis is to recognize and evaluate data mining algorithms which are commonly implemented in modern Medical Decision Support Systems (MDSS). They are helpful in a variety of healthcare units over the entire world. These institutions store enormous amount of medical data. This data may contain suitable medical information buried in different patterns concealed among the records. Inside the research some admired MDSS's are analysed in order to make a decision for the most common data mining algorithms employed for a particular purpose by them. Three algorithms have been identified: Multilayer Perceptron, Naïve bayes and C4.5. A number of testing configurations are examined in order to make a decision of the best setting for the algorithms. After that, an eventual comparison of the algorithms orders them with respect to their manner of functioning. The assessment is rely on a set of performance metrics. In WEKA software analysis has been done and data has been taken from UCI repository medical datasets The data which is taken into consideration are breast cancer, heart disease, hepatitis . The analyses have shown that it is not very simple to name a single data mining algorithm to be the most suitable for the medical data. The consequences of data sets for the algorithms were very linked. However, the concluding evaluation of the outcomes allowed singling out the C4.5 to be the best classifier for the given domain. It was followed by the Naïve Bayes and the Multilayer Perceptron.

Keywords: ***Naïve Bayes, Multilayer Perceptron, C4.5, medical data mining, medical decision support.***

I. *INTRODUCTION*

At present health centres include not only doctors, patients and medical staff but also a wide range of processes has been included together with the patient's treatment. In current years modern systems and techniques have been brought in health-care institutions to make the progress of their operations smooth. A gigantic quantity of medical records are keep aside for future use stored in databases and data warehouses. The data which is being kept in such a system may contain important knowledge concealed in medical records. Correct processing of this information has inherent capacity of enriching every medical unit by providing it with skill of many specialists who contributed their knowledge for the construction of the system. This research projected at recognizing and evaluating the most common data mining algorithms implemented in modern Medical Decision Support Systems (MDSS's). The valuation of many different types data mining methods has been existing in many research papers [13], [15], [14],[23]. Though, they keep observation only on a small number of medical datasets [14], [16], the algorithms use dare not calibrated (tested only on one parameters' settings) [21] or the algorithms compared are not common in the MDSS's [14]. Also, even though a large number of methods have been taken into consideration they were assessed with the use of different metrics on different datasets [16], [19], [14], [22]. This would make the combined evaluation of the algorithms unfeasible. This theory compares and contrasts the three data mining algorithms (determined after an in-depthliterature study) which are enforced in modern

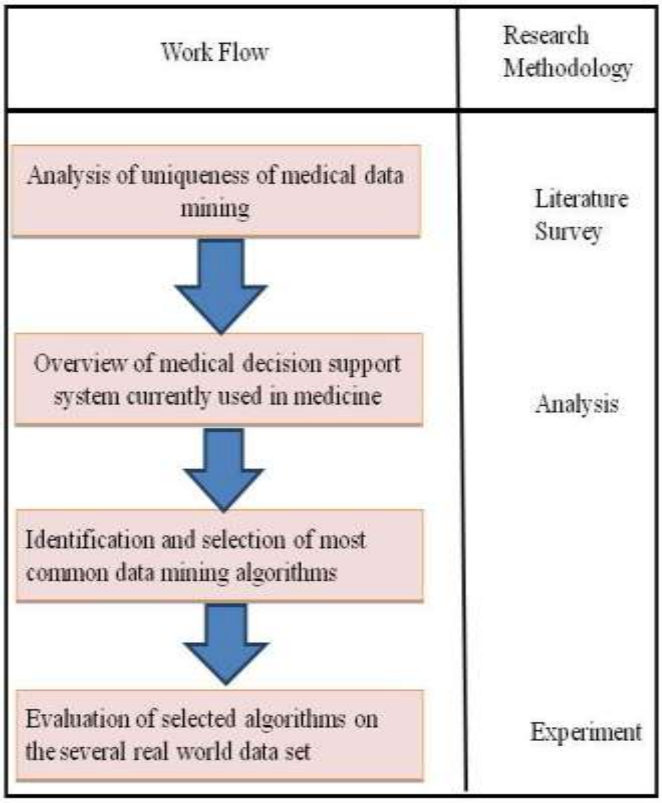
MDSS's. The analyses of specific algorithms are conducted under the same conditions.

Three choosed data mining algorithms are evaluated, which are commonly used in the modern MDSS's, with respect to their manner of functioning. The evaluation performed on five medical data sets obtained from the UCI Repository [11].In order to reach the main goal of the research the following objectives are to be fulfilled:

- Analysis of the inimitability of medical data.
- General idea of Medical Decision Support Systems currently used in medicine.
- Evaluation and comparison of data mining techniques implemented in the modern MDSS

Data Mining Process

The data mining is defined as identifying "valid, novel, potentially useful, and ultimately understandable patterns in data". In order to uncover these regularities several techniques can be taken into consideration. For illustration analysis, machine learning, statistical database technology or human computer interaction. These data mining methods initiated in the AI(artificial intelligence) and the machine learning.


Work Flow
Research Methodology
Analysis of uniqueness of medical data mining
Literature Survey
Overview of medical decision support system currently used in medicine
Analysis
Identification and selection of most common data mining algorithms
Evaluation of selected algorithms on the several real world data set
Experiment

The practise of using concrete data and evidence to support medical decisions (also known as evidence-based medicine or EBM) has existed for centuries. Author in [8] considered being the father of modern epidemiology, In 1854 it took help of maps with early forms of bar graphs to determine the source of cholera and confirm that it was transmitted through the water supply. He counted the number of deaths and plotted the victim's addresses on the map as black bars. He discovered that most of the deaths clustered towards a particular water pump in London.

Today, the size of the population, the sum of electronic data collected including globalization and the speed of disease outbreaks make it almost impractical to achieve what the pioneers did. This is where data mining becomes helpful to healthcare. It has been slowly but steadily more applied to handle many problems of knowledge discovery in the health sector. Data mining and its methods applied to medicines and public health is a relatively young field of study. In 2005, cases were scrutinized where KDD and data mining techniques were practiced in health data. In Pasig City in October 2007 in the Philippines at the Rizal Medical Centre same factors were happened. They are not able to implement strict sanitation and sterilization measures the hospital contributed to the death of several new-born babies due to neonatal sepsis disease which is caused by the infection from bacteria. After investigating hospital records, the Department of Health (DOH) came to know that 12 out of 28 babies born on October4,died of sepsis. With an integrated database and the application of data mining databases, Cheng mentioned the impact of classification algorithms assist in the early revealing of abnormal conditioning of heart , a chief public fitness fear all over the world. One more study used which is the K clustering algorithm which states that analyses of cervical cancer patients and come to the point that clustering found better predictive results than existing medical opinion.

II. LITERATURE REVIEW

The extensive amounts of knowledge and data stored in medical databases consider obligatory to the development of specialized tools for storing and right to use of data, data analysis and effective utilization of stored knowledge of data. The purpose in [1] is to show how methods and tools for intelligent data analysis are useful in narrowing the increasing gap between data assembling and data grasping. This objective is attained by applying Association Rules Technique to facilitate analysing and retrieving buried patterns for a large volume of data collected in a medical database for a great hospital. Hypothetical and realistic quality for the Incremental Enhanced Association Rule Algorithm are presented. These features include Association Rules, Classification, Clarity, , Automation, Accuracy and Raw Data. Coronary heart disease (CHD) is one of the major causes of disability in adults as well as one of the main causes of death in the developed countries. [2] Defines a data mining system based on decision trees for the assessment of Coronary heart disease (CHD) related risk factors targeting in the reduction of CHD events. Diagnosis can be achieved by building a model of a certain organ under surveillance and comparing it with the real time physiological measurements taken from the patient. Data Mining techniques in the computer-aided diagnosis (CAD), pay attention on the cancer detection, in order to lend a hand to doctors to make optimal decisions quickly and accurately.The healthcare industry collects huge amounts of healthcare data which, unfortunately, are not “mined” to discover hidden information for effective decision making. Detection of buried patterns and relationships frequently goes unexploited. Advanced data mining techniques can assist to come across the solution of this situation. [4] Defines a prototype Intelligent Heart Disease Prediction System (IHDPS) using data mining techniques, namely, Decision Trees, Neural Network and Naïve Bayes. Each technique has its unique strength in realizing the objectives of the defined mining goals. IHDPS can answer complex “what if” queries which traditional decision support systems. By using medical profiles such as age, sex, blood group,blood sugar and pressure. It can be forecast the likelihood of patients getting a heart disease. It permit the significant knowledge, e.g. patterns, relationships exists in between medical factors related to heart disease to be established. A novel approach for autonomous decision-making is developed in [5] based on the rough set theory of data mining. The process has been applied on a medical data set for patients with lung abnormalities referred to as solitary pulmonary nodules (SPNs). The two independent algorithms developed in this paper either generate an accurate diagnosis or make no decision.

III EXPERIMENTAL EVALUATION

The outcomes of several experiments carried out with the use of the three data mining algorithms: Naïve Bayes, Multilayer Perceptron and C4.5. The algorithms employed to the medical datasets. The experiments are implemented in WEKA (Waikato Environment for Knowledge Analysis) version 3.7, the environment

This section is devoted to the analyses of calibration of individual algorithms. The intention of these analyses is to decide what parameters settings yield the best models. The experiments are implemented with the help of all of the databases. Records with missing values were removed because, in case of medical accounts, it is tricky to rebuild their values.

The calibration aims at finding optimal settings which maximize the performance of each of the algorithms. The tests are done with a range of dataset split and *n*-fold cross validations (for several *n*’s) – testing configuration. For relative purposes also the complete training set is taken out for testing. These results are exclude while mentioning the results of the analyses, though.

Breast cancer database

The breast cancer database consists of nine conditional attributes. The verdict attribute takes the values 0 or 1. As presented in the Figure 9.1 the distributions of attributes contain even values almost. In case the number of occurrences in which the attributes seize the lowest values is the supreme of all . Each conditional attribute is multi-valued.

Heart diseases database

The heart diseases database consists of thirteen conditional attributes. The decisional attribute takes the values 0, 1, 2, 3, 4 or 5. As presented in the Figure 9.5 the attribute distribution of values like *old peak*, lessen with the raise of the values. The value distribution of the attribute *chest pain type* enhance with the increase of the values. The value distribution of the attributes such as *age*, *trestbps*, *cholesterol* and *maxheart rate*s. The distribution of the decisional attribute *diagnosis* also get lessen with the raise of the values. The attributes *sex*, *fasting blood sugar less than 120*, *exercised inducted angina* are binominal. *Resting ecq*, *slope* and *that* have three values.

Hepatitis database

The hepatitis database contain seventeen conditional attributes in which four of them consists multi-valued and others contain dual value. The determining attribute *die* takes values 0 or 1. As presented in the Figure 9.9 the distributions of values of the attributes. The distribution of values of the attributes such as *age* and *albumin* belongs to bell-shape.

Evaluation and comparison of the data mining algorithms

This part is devoted to the analysis of the results acquired during the calibration of the algorithms (Section 9.1). Here the comparison of the algorithms in terms of performance is also done. This section of the examinations was accomplished by testing in the Experimenter graphical interface of the WEKA environment.

The results attained for the single unit testing configuration (see Section 9.1) make it impossible to choose the finest one. The upshot for single unit metrics were very alike for all of the configurations for most of the cases. Despite the fact that, the outcome depicts that the 10-fold crossvalidation and the 66% split contain a slight plus over the rest of the testing configurations. However, there are cases where they acquiesce one of the worst results. On the other hand, very repeatedly 50% split turned out to give bad results yet, it is unfeasible to say which of the testing configurations provides the most excellent results. Thus for the final evaluation of the algorithms the 10-fold cross-validation has been chosen. The reason for this is high popularity of this configuration.

The effect of the comparison of the algorithms are presented in the Table 9.10. The table illustrate the manner of functioning of the algorithm in case of each of the performance measures and databases. After seeing the results we come to conclusion that C4.5 is undoubtedly doing great job in many of the cases, but worst results this algorithm gained for the breast cancer databases. However, overall performance was always better in comparison to other algorithms. Now,lets see the performance of naïve bayes it attained second position in the table of functioning. The results obtained by this algorithm for most of the databases and metrics were slightly not as good as for the Multilayer Perceptron in generally of the cases. For the hepatitis data this algorithm gained the worst results.. At last, the Multilayer Perceptron decision tree delievered the worst results. The results of this algorithm shows that this performed worst in aspect of errors and AUC when compared with all of the algorithms. Its heterogeneity and complexity for the attribute values can obstruct data withdrawal. In this part the naïve bayes and Multilayer Perceptron may be over trained.

Table 9.10 Performance of the algorithms with respect to the measures and databases with the use of 10-fold cross-validation

		Algorithm		
Measure	**Database**	**NaiveBayes**	**C4.5**	**Multilayer Perceptron**
Percent Correct	**Breast Cancer**	60.2%	62.3%	62.2%
	Heart Disease	56.7%	66.6%	63%
	Hepatitis	75.3%	76.7%	73.5%
Mean absolute error	**Breast Cancer**	45.6%	37.6%	37.6%
	Heart Disease	46.3%	38.1%	39.2%
	Hepatitis	32.2%	30%	27.9%
Root mean squared error	**Breast Cancer**	51.08%	45.4%	48.5%
	Heart Disease	48.9%	49.4%	52.2%
	Hepatitis	41.9%	44.9%	48.8%
Relative absolute error	**Breast Cancer**	93.3%	81.3%	75.2%
	Heart Disease	93.3%	76.5%	79%
	Hepatitis	66.1%	61.5%	57.2%
Root relative squared error	**Breast Cancer**	101.16%	92.5%	94.7%
	Heart Disease	98.1%	99%	104.7%
	Hepatitis	84.9%	91.1%	98.9%
True positive rate	**Breast Cancer**	60%	62.5%	64.2%
	Heart	56.8%	66.7%	63%

	Disease			
	Hepatitis	75.3%	76.8%	73.5%
False positive rate	**Breast Cancer**	40.4%	35.9%	35.2%
	Heart Disease	45.3%	33.7%	37.7%
	Hepatitis	26.2%	23.6%	28.5%
Precision	**Breast Cancer**	59.9%	65.4%	65%
	Heart Disease	56.3%	66.6%	63%
	Hepatitis	75.3%	77%	73.4%
Recall	**Breast Cancer**	60%	65.5%	64.2%
	Heart Disease	56.8%	66.7%	63%
	Hepatitis	75.3 %	76.8%	73.5%
F-measure	**Breast Cancer**	59.9%	65.04%	64.3%
	Heart Disease	56.1%	66.6%	63%
	Hepatitis	75.3%	76.9%	73.5%
AUC	**Breast Cancer**	61%	72.6%	72.3%
	Heart Disease	63.1%	69.3%	67.7%
	Hepatitis	70.9%	70.1%	78.3%

The results from the Table 9.10 have been also presented (for better visualization) in the figures in the Table 9.12. These graphs confirm high performance of the C4.5. Thus, overall best algorithm is the C4.5, with the Naïve Bayes being the second.

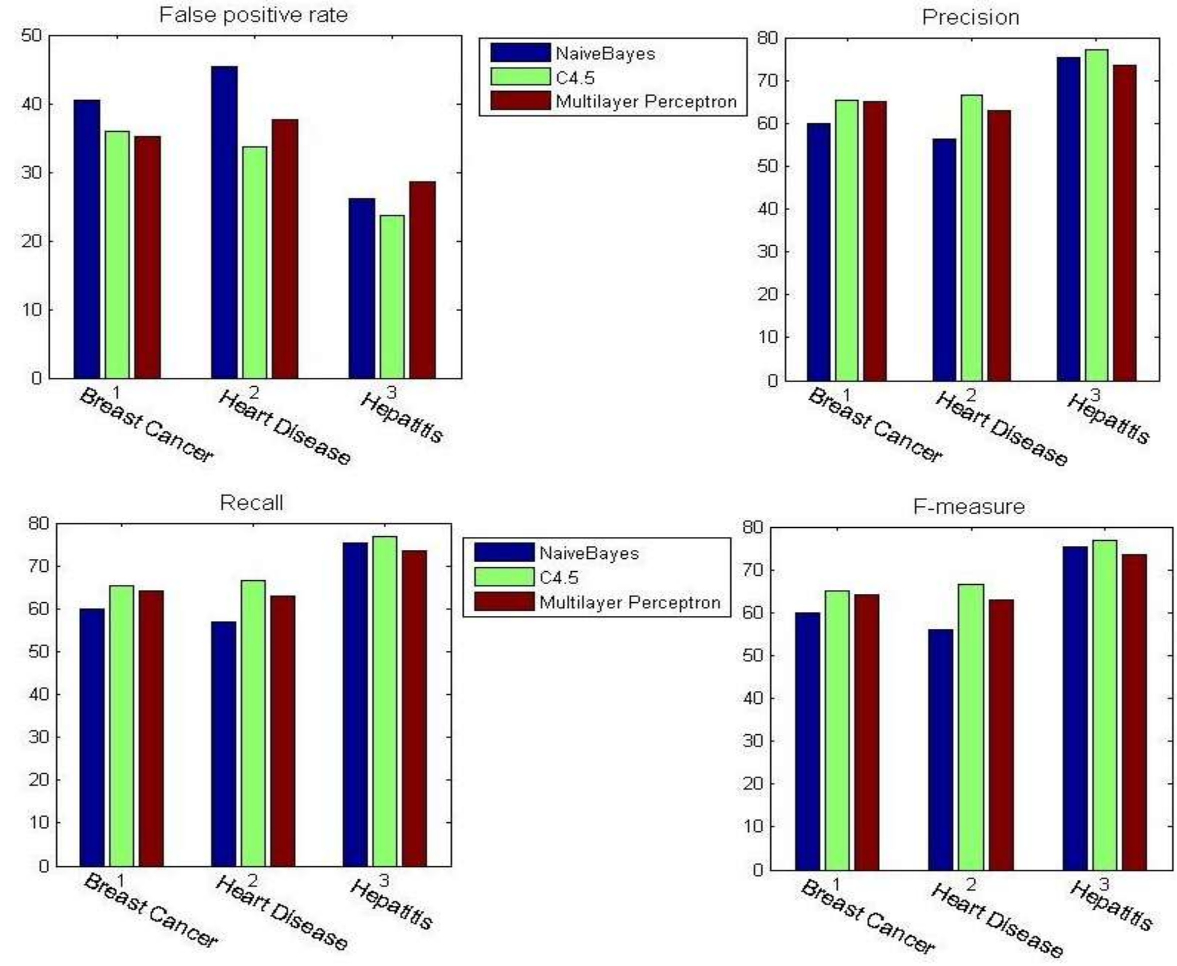


Figure 9.13(a) Evaluation of performance of the data mining algorithms for the medical databases

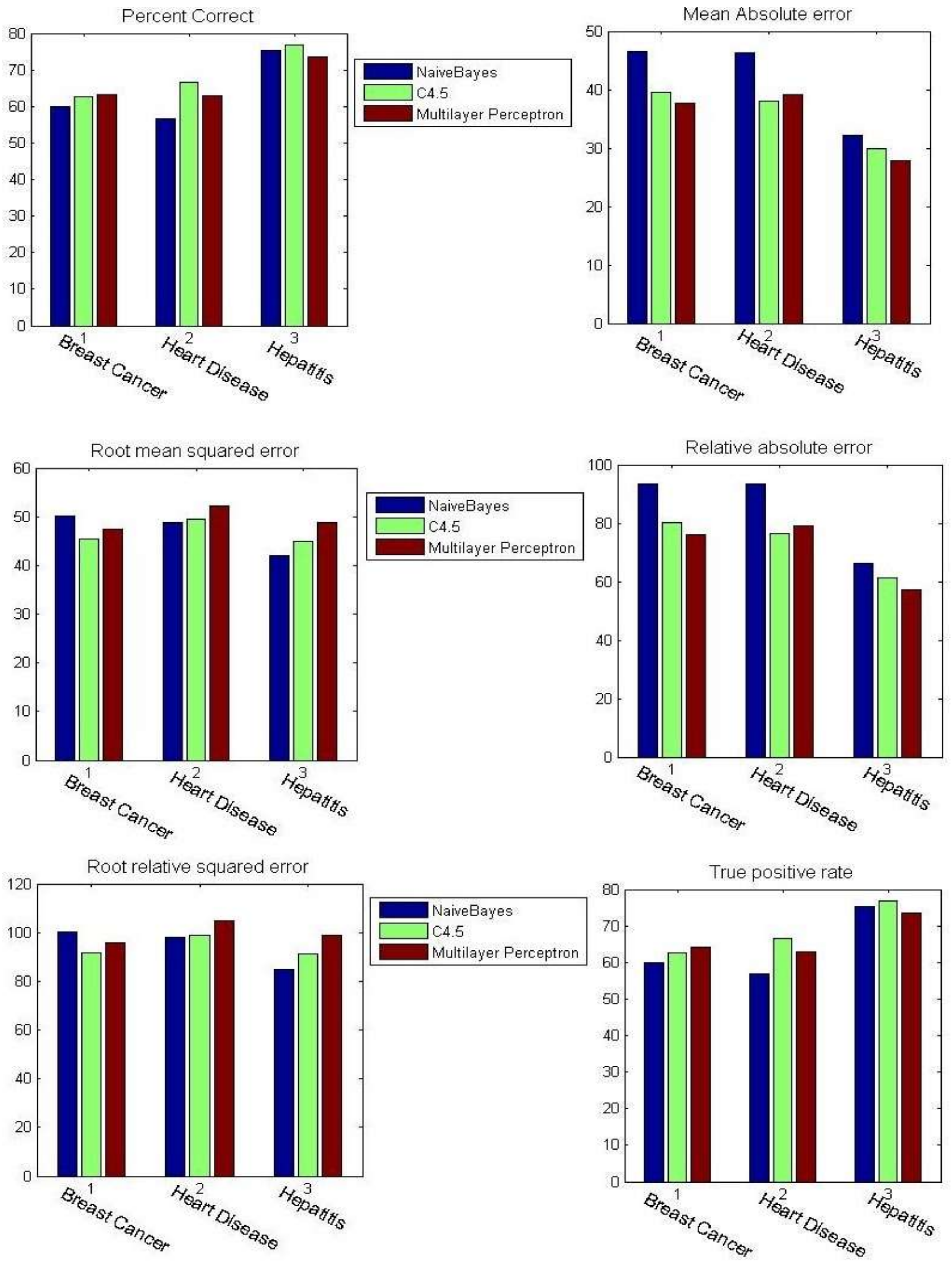


Figure 9.13(b) Evaluation of performance of the data mining algorithms for the medical databases

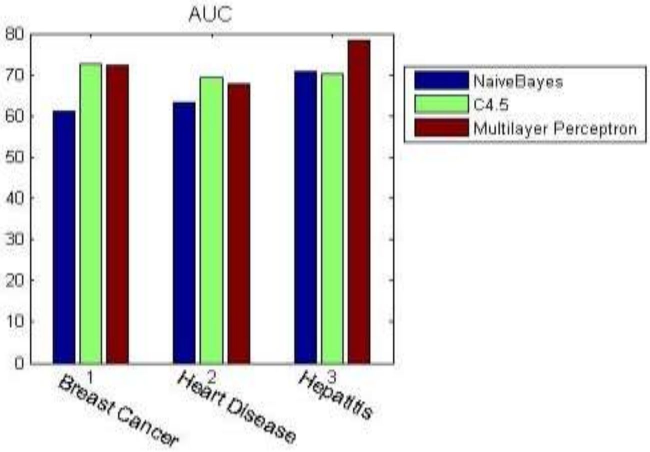


Figure 9.13(c) Evaluation of performance of the data mining algorithms for the medical databases

As it is clearly seen that the C4.5 classifier gained the highest score of the performance for the medical database and data mining algorithm. The Naïve Bayes algorithm obtained the second classification. At the end the performance of the Multilayer Perceptron is valuated. This can be clearly seen from these algorithms C4.5 classifier is the most precise method in Medical Decision SupportSystems, after the C4.5 algorithm accurate result is given by the Naïve Bayes algorithm and in the end Multilayer Perceptron algorithm.

CONCLUSIONS

The main goal of the research was to recognize the most familiar data mining algorithms, which is put into practice in modern Medical Decision Support Systems, and assess their performance on numerous medical datasets. The algorithms which were selected: C4.5, Multilayer Perceptron and Naïve Bayes. For the assessment three UCI databases were used: heart disease, hepatitis, breast cancer. A number of performance metrics were operated: percent of correct classifications, *True/False Positive* rates, AUC, *Precision*, *Recall*, *F-measure* and a set of errors. The reason behind doing such research was the fact that no work was found which would analyse these three algorithms under identical conditions.

The variety of Medical Decision Support Systems makes it difficult to choose the most common data mining algorithms. Sometimes a system may be in a test phase and some part of its functionality may not be working yet. Nevertheless, the classification of data mining algorithms applied in MDSS's was done on worldwide known systems like: Help, DXplain and ERA.

The review of the popular Medical Decision Support Systems brought a list of data mining algorithms. The research utilized three the most popular data mining algorithms commonly implements in modern MDSS's. These included C4.5 decision tree algorithm, Multilayer Perceptron and Naïve Bayes. These three algorithms prove to be highly efficient in the MDSS's.

REFERENCES

[1] Aftarczuk K., Kozierkiewicz A., The method of supporting medical diagnosis based on consensus theory. Report of Institute of Information Science & Engineering, University of Technology. Wroclaw, 2006 Series PRE No. 1.

[2] Aftarczuk K., Kozierkiewicz A, Nguyen N. T., Using Representation Choice Methods for a Medical Diagnosis Problem. Knowledge-Based Intelligent Information & Engineering Systems 2006, 805-812.

[3] Alter S.L. Decision Support System: Current Practice and Continuing Challenge. Addisson-Wesley, 1980.

[4] Autio L., Juhola M., Laurikkala J., on the neural network classification of medical data and an endeavor to balance non-uniform data sets with artificial data extension. Computers in Biology and Medicine, 2007, vol. 37, no. 3, 388-397.

[5] Banfield R.E., Hall L.O., Bowyer K. W., Kegelmeyer W.P., A Comparison of Decision Tree Ensemble Creation Techniques. IEEE Computer Society, vol. 29, 2007.

[6] Berrar D., Bradbury I. and Dubitzky W., Avoiding model selection bias in small-sample genomic datasets. Oxford University Press, 2006.

[7] Berry J., Linoff G., Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management. 2004, 2nd edition, Indianapolis, Wiley.

[8] Brin S., Motwani R., Ullman J. D., Tsur S. Dynamic itemset counting and implication rules for market basket

data. ACM SIGMOD Record, 1997, vol. 26, no. 2, 255–264.

[9] Cancer Facts & Figures 2007. Retrieved on 26 04 2007.

[10] Chae Y. M., Kim H. S., Tark K. C., Park H. J., Ho S. H., Analysis of healthcare quality indicator using data mining and decision support system. Expert Systems with Applications, 2003, 167–172.

[11] Child Ch., Stathis K., The Apriori Stochastic Dependency Detection (ASDD) Algorithm for Learning

[12] Stochastic Logic Rules. J. Dix, J. Leite, and P. Torroni (eds), 2004, Proceedings of the 4th International Workshop on Computation, 201-216.

[13] Cios K., Moore G., Uniqueness of Medical Data Mining. Artificial Intelligence in Medicine, 2002, vol. 26, 1-24.

[14] Comak E., Arslan A., Turkoglu I., A decision support system based on support vector machines for diagnosis of the heart valve diseases. Elsevier, 2007, vol. 37, 21-27.

[15] Cosic D., Loncaric S., Rule-based labeling of CT head image. Lecture Notes in Artificial Intelligence, Berlin, Germany, Springer-Verlag, 1999, vol. 1211, 453–456.

[16] Cunningham P., Carney J., Jacob S., Stability problems with artificial neural networks and the ensemble solution. Artificial Intelligence in Medicine, 2000, vol. 20, no. 3, 217–225.

[17] Duch W., Adamczak R., Grabczewski K., Zal G., Hayashi Y., Fuzzy and crisp logical rule extraction methods in application to medical data. Computational Intelligence and Applications, Berlin, Germany, Springer-Verlag, 2000, vol. 23, 593–616.

[18] Fayyad U. M., Data mining and knowledge discovery: Making sense out of data. IEEE Expert, 1996, vol. 11, no. 5, 20-25.

[19] Haug P. J., Rocha B. H.S.C. and Evans R. S., Decision support in medicine: lessons from the HELP system. International Journal of Medical Informatics, 2003, vol. 69, 273-284.

[20] Hayashi Y., Setiono R., Yoshida K., A comparison between two neural network rule extraction techniques for the diagnosis of hepatobiliary disorders. Artificial Intelligence in Medicine, 2000, vol. 20, no. 3, 205–216.

[21] Herron P., Machine Learning for Medical Decision Support: Evaluating Diagnostic Performance of Machine Learning Classification Algorithms, INLS 110, Data Mining, 2004.

[22] Holsapple C.W., Whinston A. B., Decision Support Systems: A Knowledge-Based Approach. St. Paul, West Publishing. 1996

[23] Kutlu, Y., Isler, Y., Kuntalp, D., Kuntalp, M., Detection of Spikes with Multiple Layer Perceptron Network Structures. Signal Processing and Communications Applications, 2006, 1-4.

[24] Newman D.J., Hettich S., Blake C.L., Merz C.J., UCI Repository of machine learning databases. 1998 [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science (retrieved on 2.05.2007).

[25] Zhou Z.-H., Jiang Y., Yang Y.-B., Chen S.-F., Lung cancer cell identification based on artificial neural network ensembles. Artificial Intelligence in Medicine, 2002, vol. 24, no. 1, 25–36.

[26] Yousef W. A., Wagner R. F., Loew M. H., Estimating the uncertainty in the estimated mean area under the ROC curve of a classifier. Pattern Recognition Letters, 2005, vol. 26, no. 16, 2600-2610.