ISSN:0975-9646

Performance Analysis and Optimization of Supervised Learning Techniques for Medical Diagnosis Using Open Source Tools

Vidyullatha Pellakuri [*1], Deepthi Gurram[*2], Dr.D. Rajeswara Rao[#3], Dr.M.R.Narasinga Rao[#4]

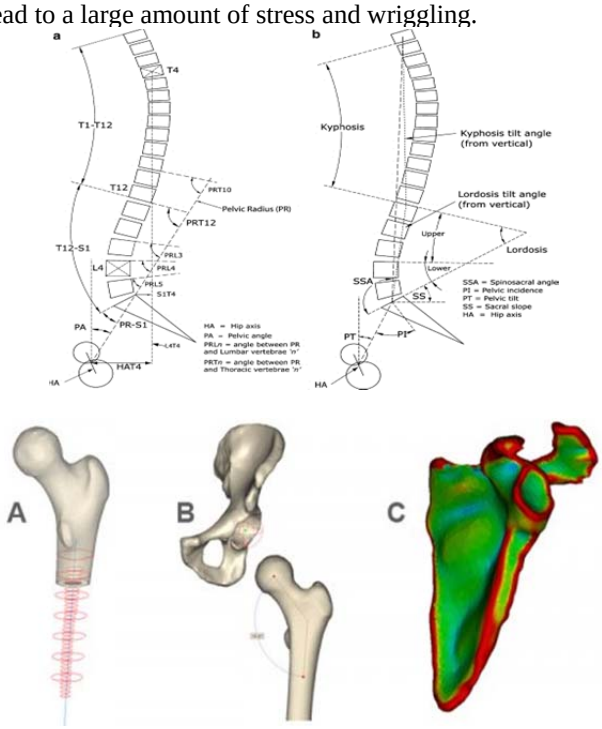
[*1, *2] *Research Scholar, Department of Computer Science & Engineering, KL University, Andhra Pradesh.*
[#3,#4] *Professor, Department of Computer Science & Engineering, KL University, Andhra Pradesh.*

Abstract:Data mining initiates the recent advances and applications in the promising areas of medicine and biology around the world. In the medical diagnosis, it is difficult for the experts to observe disease with assurance. By using data mining techniques, this problem can be solved. This paper intends to provide the current techniques in data mining which are in use in today's medical research particularly in Orthopaedic diagnosis. This paper presents a comparative study of different classification techniques using two open source data mining tools named WEKA and TANAGRA. The aim of this paper is to analyze the performance of different classification techniques for a set of Orthopaedic (muscular-skeletal) data.

Keywords: Classification, Data Mining, Orthopaedic, Tanagra, Weka.

1. INTRODUCTION:

The science of extracting useful information from large data sets is termed as data mining. However data mining concepts have an huge history, the term "Data Mining", is introduced almost new, in mid 90's. Data mining is an interdisciplinary field which cover-up the areas of statistics, machine learning, data management and databases, pattern recognition, artificial intelligence, etc. All of these are involved with certain aspects of data analysis, so they have much in familiar but each also has its own distinct problems and various solutions. The major motivation behind data mining is autonomously extracting useful information or knowledge from large data stores or sets. This paper focuses on Orthopaedic injuries (muscular-skeletal system) which have a huge impact on a person's life and it leads very cost effective for diagnosis and treatments. On diagnosis decisions, the machine learning techniques such as classification methods are very useful. The Orthopaedics Biomedical data set contains details of normal patients and disk hernia or spondylolisthesis termed as abnormal. The dataset contains 100 normal patients and 210 abnormal patients are taken in to two different data mining tools such as Weka and Tanagra. In the dataset, each patient is represented by six attributes which depends upon the shape and orientation of the pelvis and lumbar bone having attributes like pelvic incidence, pelvic tilt, lumbar lordosis angle, sacral slope, pelvic radius and grade of spondylolisthesis and two class labels such as normal and abnormal. Spondylolisthesis [11] is the situation in which one of the bones of the spine (vertebrae) blunder on position, if it blundered so much, the bone might press on a nerve, generate pain. Commonly, the bones of the lower back are affected so much. Spondylolisthesis [12] is the most common cause of back pain in youngerster's. Lumbar lordosis is the inward (ventral) curvature of the lumbar spine found by the wedging of lumbar vertebral bodies and the intervertebral disks. Dorsal wedging of the vertebral bodies and disks raises the lordosis angle, whereas more ventral wedging of these structures reduces the lordosis angle. Pelvic incidence (PI), or pelvisacral angle, is described as the angle between a line perpendicular to the sacral plate at its midpoint and a line connecting the same point to the centre of the bicoxofemoral axis which is shown in the figure. The Sacrum is positioned farther back of the pelvis. Five bones integrated into a trilateral shape, form the sacrum. The sacrum is put on the middle of two hipbones linking the spine to the pelvis positioned just beneath the lumbar vertebrae. Back pain or leg pain can mostly proceed due to injury where the lumbar spine and sacral region connect because this region of the spine is lead to a large amount of stress and wriggling.



Literature survey:

Jahanvi Joshi etal. [1] discussed on Diagnosis and prognosis breast cancer using classification rules on 36

algorithms using Weka tool. Jyoti Soni etal. [2] provide a survey of current techniques of knowledge discovery in databases using data mining techniques that are in use in today's medical research particularly in Heart Disease Prediction. Rashedur etal. [3] examine the performance of different classification methods that could generate accuracy and some error to diagnosis the data set using three data mining tools named WEKA, TANAGRA and MATLAB. Bendi Venkata Ramana et al. [4] compared popular Classification Algorithms for evaluating their classification performance in terms of Accuracy, Precision, Sensitivity and Specificity in classifying liver patient's dataset. Shymaa Mohammed Jameel [5] solves a several cases taken from datasets such as Breast Cancer, Pima Indian Diabetes, Hagerman Surgery Survival, Liver disorders, Wisconsin Breast Cancer, Statlog Heart, Australian Credit Approval, Parkinsons SPECTF, German Credit Data and Appendicitis and take the correct decision within a good enough computational time. Gopala Krishna Murthy Nookala [6] made a comprehensive comparative analysis of 14 different classification algorithms and their performance has been evaluated by using 3 different cancer data sets to predict cancer based on the gene expression data.Abdullah H. Wahbeh etal. [7] conducted a comparative study on the performance of knowledge discovery tools and proved that WEKA toolkit has achieved the highest improvement in classification performance followed by Orange, KNIME and finally Tanagra respectively. Ritu Ganda and Vijay Chahar [8] make use of Cardiology Dataset and compared the results of simple clustering technique and K-means using WEKA and TANAGRA data mining tools. Y. Ramamohan [9] presents an overview of the data mining tools like Weka,Tanagra, Rapid Miner, Orange to make proactive and knowledge-driven decisions. Nikhil N. Salvithal, Dr. R. B. Kulkarni [10] judge the performance analysis which depends on many factors test mode, different nature of data sets, type of class and size of data set by using different data mining classification algorithms on various datasets.

2. METHODOLOGY

The data mining concepts are categorized into predictive and descriptive methods. Predictive methods are analyzed using the previous data and make predictions of future data. This method includes classification, regression, time series analysis, and prediction. For classification, the data is classified into groups or classes which require algorithms based on data attribute values and their rankings. Classification maps data into predefined groups or classes. It is often referred to as supervised learning. Classification algorithms require that the classes be defined based on data attribute values. In this paper different classification algorithms are considered.

2.1 Classification Algorithms

A classification algorithm assigns a class to a group of data records having specific attributes and attribute-values. The classification techniques in healthcare can be applied for diagnostic purposes. A classification model receives a set of relevant attribute-values, such as clinical observations or measurements and gives a class of data records as output. As an example, the classes can identify "whether a patient has been diagnosed with a particular disk problems or not", and the classifier model assigns each patient's case to one of these classes. Some classification techniques in Weka that are applied on healthcare includes Decision stump, Naïve Bayes Classifier, J48 Decision Trees, LMT Tree, Random Forest, Random Tree, REPTree, JRiP, ZeroR, OneR**.** The classification algorithms in Tanagra are Multi layer perceptron, Linear discriminate analysis, multiple linear regression, Naive Bayes, RndTree, ID3, C4.5, CRT, CS-CRT,CS-MC4.

In machine learning, the RndTree (random tree) classifier takes the input vector, classifies with each tree in the forest and gives the output of the class label which receives the majority of "votes"**. C4.5** is a classification algorithm as well as a statistical classifier that generates a decision tree proposed by Ross Quinlan, which is an extension of Quinlan's earlier ID3 algorithm. The C4.5 is similar to ID3 that builds decision trees from a set of training data using the concept of entropy. Using this entropy calculation, the splitting of samples in to subsets is efficiently implemented in C4.5. The attribute with the highest normalized information gain (entropy) is chosen to make the decision of the root node. **J48** is an another open source Java implementation of the C4.5 algorithm in the data mining tool Weka. **The ID3** is an iterative dichotomiser algorithm used to generate a decision tree which is a precursor to the c4.5 algorithm. The **C-RT** and **CS-RT** are the cart method under Tanagra is a very popular classification tree learning algorithm. Cart builds a decision tree by splitting the records at each node. For best splitting of records it uses Gini Index. The CS-RT is similar to cart but with cost sensitive classification. **The CS-MC4** is a cost sensitive decision tree algorithm uses m-estimate smoothed probability estimation which is a generalization of Laplace estimation. It minimizes the expected loss using misclassification cost matrix for the decision of the best prediction with in leaves. The pre-condition required for this algorithm is that at least one discrete target value and one or more continuous / discrete values for input must be available.

3. RESULTS AND DISCUSSIONS

Classification is one of the data mining techniques, which gives the decision for diagnosing process. There are different algorithms which are practised in two different data mining tools such as Weka and Tanagra. The performance parameters such as accuracy and through put are calculated using both the data mining tools. Among them, the classification algorithms are well executed in Tanagra compare to weka with the accuracy of 100%. In Weka, J48 and PART algorithms reach more than 80% accuracy, and the time taken to build the algorithms is almost low. In Tanagra, the each and every algorithm shows more than 90% accuracy and takes less time to execute an algorithm. Almost twenty algorithms are implemented on the Orthopaedic data set using the Open Source tools. Performance of the algorithms are compared and shown in the following figures.

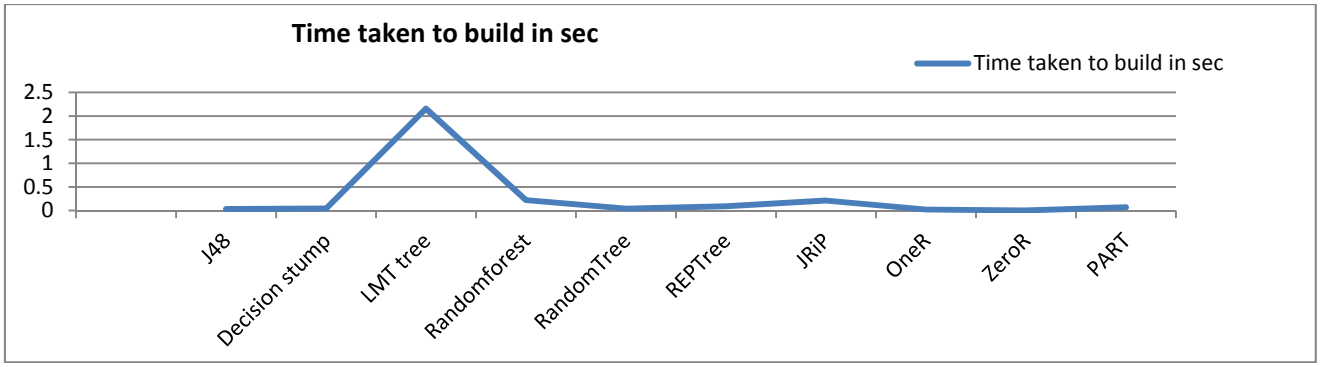


Fig.1 Time taken to build model in WEKA

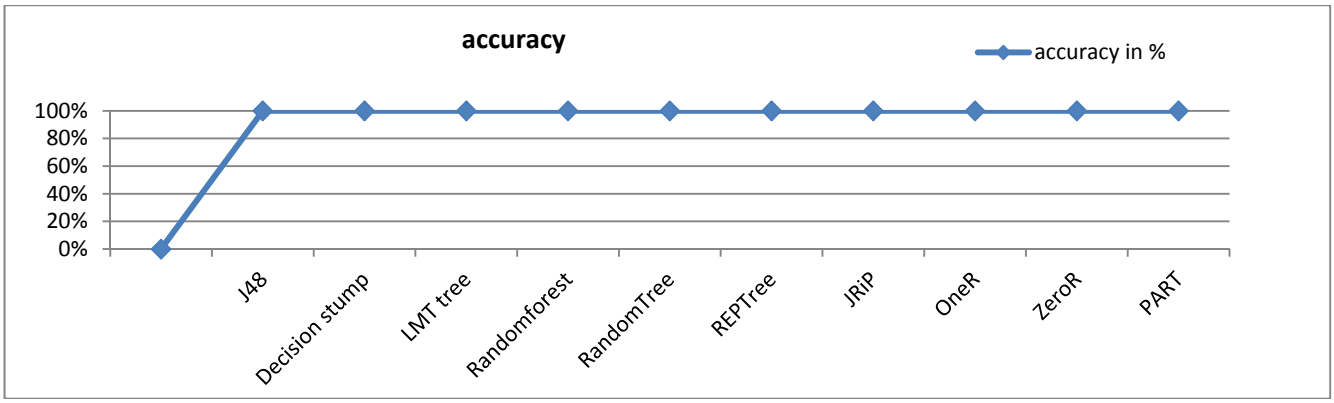


Fig.2 Accuracy of the algorithms in WEKA

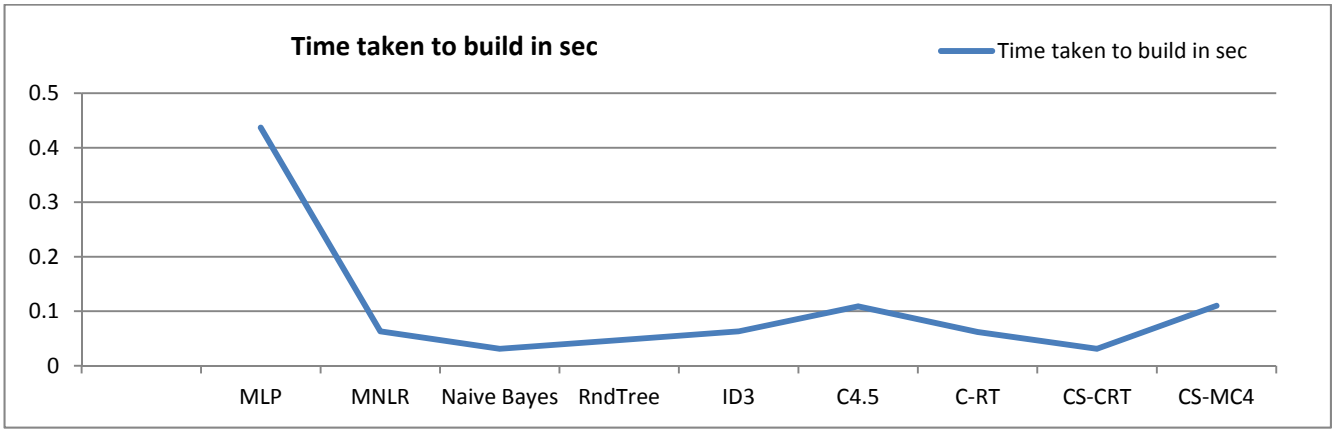


Fig.3 Time taken to build model in TANAGRA

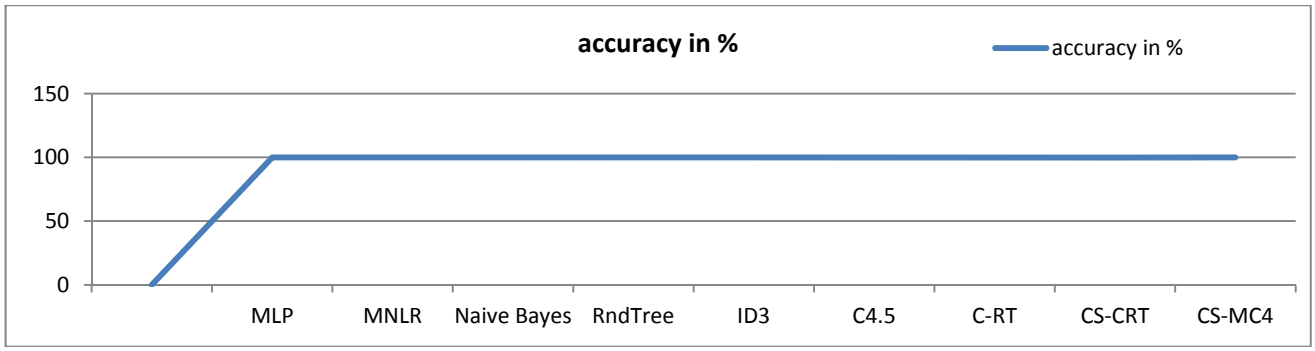


Fig.4 Accuracy of the algorithms in TANAGRA

Attrib No.	Attributes	Gain Ratio	Ranking
6	Degree spondylolisthesis	0.3405	1
1	Pelvic incidence	0.1363	2
3	Lumbar lordosis angle	0.1138	3
2	Pelvic tilt	0.1015	4
4	Sacral slope	0.0936	5
5	Pelvic radius	0.0935	6

Table 1: Attributes Ranking and Gain Ratio

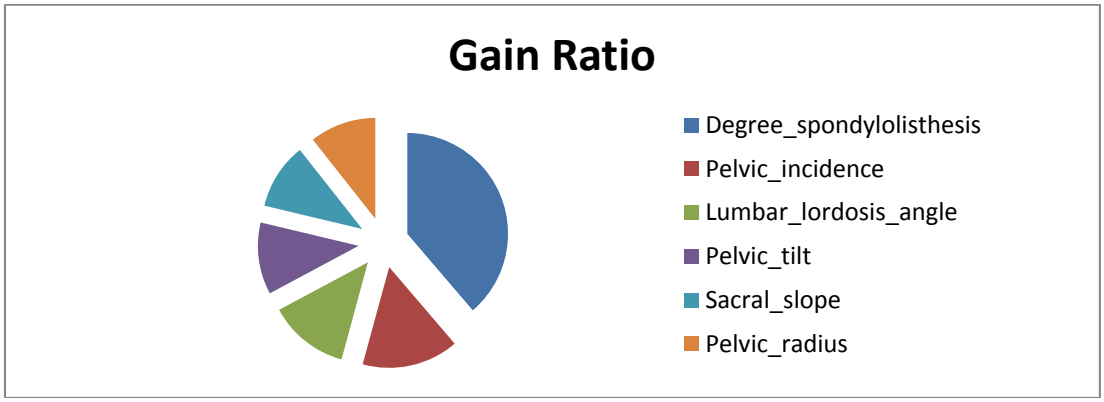


Fig .5 Attribute ranking using Gain Ratio

In Fig.5 the attributes in the data set are ranked using the Gain Ratio. Depending on the rank of the attributes the medical experts diagnosis the orthopaedic conditions of the patient and predict the status of the patients whether they are in normal or abnormal conditions. From the below Gain Ratio table, the degree_spondylolisthesis is the sixth attribute in the Orthopaedic data set and its gain ratio is 0.3405 which is higher than other attributes. The decision tree is built using the attribute with highest gain ratio as root node.

CONCLUSION

Data mining is becoming progressively more widespread in banking, insurance, medicine, and retailing industries. In this paper the problem of orthopaedic (muscular skeletal system) is predicted by different classification algorithms using open source data mining tools. The outcome of this paper is prediction of the orthopaedic problems by implementing almost twenty algorithms on two different open source tools such as Weka and Tanagra to estimate the accuracy among all the algorithms and also the attribute ranking is developed to make a decision on the orthopaedic problems. Among all the classification algorithms, the results are more accurate in Tanagra tool compared to Weka.

REFERENCES

1. Miss Jahanvi Joshi Mr. RinalDoshiDr. Jigar Patel, "Diagnosis and prognosis breast cancer using classification rules", International Journal of Engineering Research and General Science, ISSN 2091-2730, Volume 2, Issue 6, October-November, 2014.
2. Jyoti Soni Ujma Ansari Dipesh Sharma, "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", *International Journal of Computer Applications (0975 – 8887) Volume 17– No.8, March 2013.*
3. Rashedur M. Rahman, Farhana Afroz, "Comparison of Various Classification Techniques Using Different Data Mining Tools for Diabetes Diagnosis", in "*Journal of Software Engineering and Applications"*, 2013, 6, 85-97, March 2013
4. Bendi Venkata Ramana1, Prof. M.Surendra Prasad Babu2, Prof. N. B. Venkateswarlu, "A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis", International Journal of Database Management Systems (IJDMS), Vol.3, No.2, May 2011.
5. Shymaa Mohammed Jameel, "Investigation of Differential Evaluation Optimization Algorithm for Medical Data Classifications", IJCSNS International Journal of Computer Science and Network Security, VOL.14 No.1, January 2014.
6. "Performance Analysis and Evaluation of Different Data Mining Algorithms used for Cancer Classification", (IJARAI) International Journal of Advanced Research in Artificial Intelligence, Vol. 2, No.5, 2013, Gopala Krishna Murthy Nookala, Bharath Kumar Pottumuthu, Nagaraju Orsu, Suresh B. Mudunuri
7. "A Comparison Study between Data Mining Tools over some Classification Methods", (IJACSA) International Journal of Advanced Computer Science and Applications, Special Issue on Artificial Intelligence, pg no: 18-24,Abdullah H. Wahbeh, Qasem A. Al-Radaideh, Mohammed N. Al-Kabi, and Emad M. Al-Shawakfa.
8. "A Comparative Study on Feature Selection Using Data Mining Tools", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 9, pg no:26-33,September 2013, Ritu Ganda, Vijay Chahar.
9. "A Study of Data Mining Tools in Knowledge Discovery Process", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-3, July 2012, Y. Ramamohan, K. Vasantharao, C. Kalyana Chakravarti, A.S.K.Ratnam
10. "Evaluating Performance of Data Mining Classification Algorithm in Weka", International Journal of Application or Innovation in Engineering & Management (IJAIEM), Volume 2, Issue 10, October 2013 ISSN 2319 – 4847, Nikhil N. Salvithal, Dr. R. B. Kulkarni
11. http://en.wikipedia.org/wiki/Spondylolisthesis
12. http://my.clevelandclinic.org/health/diseases_conditions/hic_your_back_and_neck/hic_Spondylolisthesis