International Journal of Engineering & Technology, 7 (4) (2018) 2806-2809

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: 10.14419/ijet.v7i4.18096
Research paper



Systemic lupus erythematosus prediction tool using optimal cluster based classification (OCBC) algorithm

Gomathi. S [1] *, Narayani. V [2]

[1] *Research Scholar, Department of Computer Science, Research and Development Center, Bharathiar University, Coimbatore*
[2] *Assistant professor, Department of Computer Science, St.Xavier's College, Palayamkottai, India*
**Corresponding author E-mail: mailtogomathisrinivasan@gmai.com*

Abstract

The key objective of the paper focuses on developing the Systemic Lupus Erythematosus prediction tool for the prediction of disease in the early stage using Optimal Cluster based Classification Algorithm (OCBC). Systemic Lupus Erythematosus is an autoimmune, chronic, multi-organ disorder which affects more or less all parts of the body with changing symptoms. There is no cure for the disease; the lifetime of the patients can be extended if diagnosed in the early stage. The death occurs due to various reasons like unawareness, late diagnosis, meeting the right specialist in the severe stage etc. The SLE dataset is tested and applied to various classification algorithms such as ID3, C4.5, J48 and OCBC using SLE prediction tool. As a result, statistics are generated based on all classification algorithms and comparison of all four classifiers is also done in order to predict the accuracy, specificity, sensitivity, precision, recall, F-measure, kappa statistics and to find the best performing classification algorithm among all. In this paper, the screenshots of the tool are also shown. This paper outlines the importance of classification and prediction based data mining algorithms in the healthcare field.

Keywords: *Classification; Autoimmune; Chronic; Lupus; ID3; C4.5; J48; OCBC; Data Mining.*

1. Introduction

Data mining has fascinated a lot of consideration in the research industry and in society as a whole in recent years, due to massive availability of huge amount of data and the necessity for converting such data into valuable information and knowledge. Data mining, which is also called Knowledge Discovery in Databases, is the field of discovering new and possibly useful information from huge databases [1]. The main objective of applying data mining in the healthcare industry is to analyze such data and to resolve medical research issues. Health care data mining deals with developing new methods to explore the medical data, and by means of Data Mining methods to better understand patients dataset. The data mining technique converts raw data generated from the healthcare industry into useful information that could possibly have a great influence on medical research and practice. Data mining in healthcare is used in the variety of areas, including early prediction of disease, insurance fraudulent etc. There are growing research interests in using data mining in healthcare domain. Data mining uses many techniques such as Naïve Bayes, Decision Trees, K-nearest neighbor, Neural Networks, Apriori, and many others. Prediction and analysis of disease is an important milestone in the healthcare field. The most challenging area is to predict the chronic diseases like Systemic Lupus Erythematosus. This paper predicts the disease in advance to extend the lifetime of the patients. Through this paper, the accuracy of some classification techniques for predicting the performance of the disease is also examined. The main objectives of this work are: to generate data source of predictive variables, Data mining methodologies to study lupus patients dataset, identification of the patients who are affected with mild symptoms, identification of the patients with moderate symptoms and severely affected patients and to find the best algorithm.

2. Background and prior work

Data mining is the arena of determining novel and potentially useful information from large amounts of data. Data mining in healthcare is an ever-growing technique. Data mining is emerging in the field of the healthcare sector. Vikas Chaurasia, Saurabh Pal [1] used three popular data mining classification algorithms CART, ID3, and Decision Tree. Authors show the accuracy of the three algorithms. The data set for heart disease were collected from the Cleveland Clinic Foundation and implemented in WEKA tool. The tool was written in Java and contains the collection of machine learning algorithms. The evaluation criteria to measure the three algorithms include time taken to build the model, correctly classified instances, and accuracy. Training and simulator error evaluation like Kappa statistics, mean absolute error, root mean squared error, relative absolute error and root relative squared error also measured [1]. The appropriate management of lupus is critically dependent upon the proper assessment of disease activity, quality of life and organ damage which was briefed by Lam and Petri [2]. Assessment of lupus is not based on single test and which involves accurate physical and laboratory diagnosis, recording of accurate morbidity, monitoring of disease activity and combination of these with the patient's own discernments of health status and quality of life. The authors [3] used prediction technique on 152 students' dataset for student performance analysis. The authors researched and compared five mining technique to find the best from multi-layer perceptron, naïve bayes, Sequential minimal optimization, J48 and Reduced Error Pruning Decision Tree.

Copyright © 2018 Gomathi. S, Narayani. V. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

3. Proposed methodology

A survey, Questionnaire and experimental methodology is used. Through extensive search of the literature and discussion with experts on predicting the disease, a number of factors that are considered to have influence on developing the prediction tool and proposing OCBC algorithm are identified. These influencing factors are categorized as input variables. For this work, recent real world data is collected from public. This data is then pre-processed using the proposed tool. Then data is transformed into a standard format which is to be processed by the tool. After that, the four classification techniques are used to find the accuracy and various statistical data to find the best algorithm. After implementation results are produced and analyzed. Stepwise description of algorithm used in the tool is represented with the help of flowchart as shown in Fig 1. The clinical profile of 112 patients and related variables are defined in the Table 1.

Table 1: Clinical Profile of SLE Patients

Clinical Profile	N=112
Gender (male : female)	14:98
Age in years (avg)	28.5
Duration of disease in years (avg)	8.36
Photosensitivity rash	72
Malar rash	73
Discoid rash	52
Oral ulcer	40
Arthritis	65
Myocarditis	7
Serositis	9
Nephritis	49
Vasculitis	18
Positive ANA	90
Positive dsDNA	75
CRP	56
CBC	45
Measurement of Glomerular Filtration Rate and Proteinuria	59
Protein/Creatinine Ratio	59
Urinalysis	60

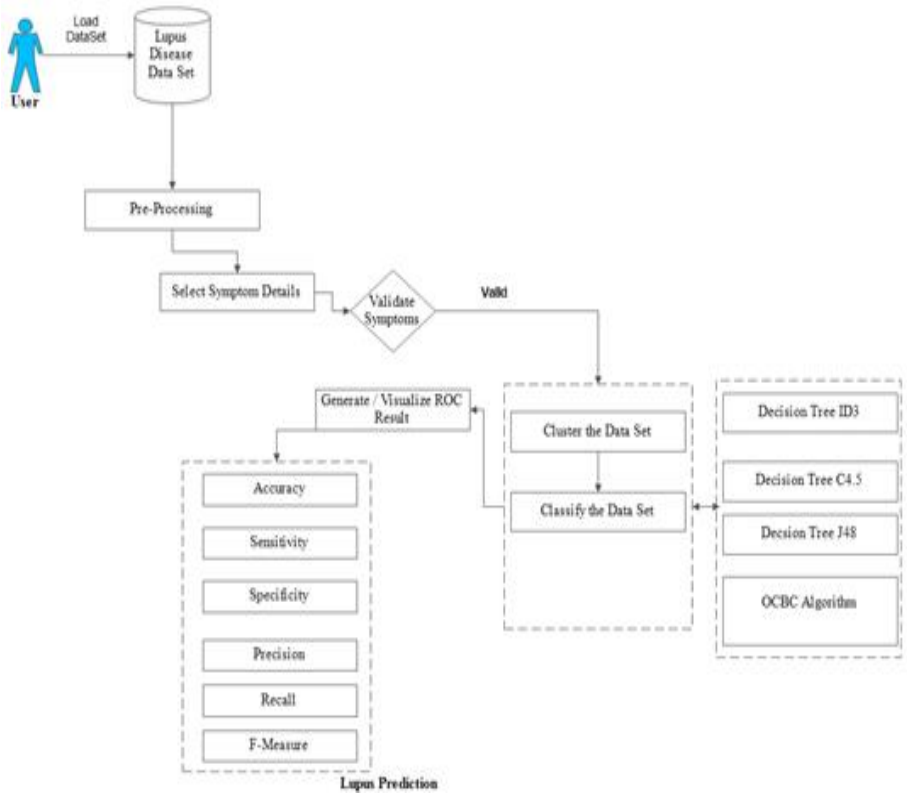


Fig. 1: Flow Chart of the Proposed Model.

3.1. Tools and techniques used

In this paper three existing Data Mining techniques are used to compare with the proposed algorithm for prediction of the disease. The techniques are Classification and Clustering. The output dataset is tested and analysed with ID3, C4.5, J48 and OCBC. For implementation of all these classification a new tool, SLE prediction tool is designed and developed. The output with chart and PDF are generated and stored using the tool.

3.2. ID3

ID3 is an Iterative Dichotomiser3 greedy algorithm which forms decision trees based on top down approach. ID3 ac-cepts input categorical data and yields output as categorical da-ta. All types of attributes can be applied to generate ID3 deci-sion trees, thus creates wide and shallow trees. ID3 computes the entropy of each attribute with the data set, split the data set into subsets, make a

decision tree node containing that attribute and recurse on subsets via left over attributes.

3.3. C4.5

C4.5 is an extension of ID3 algorithm. C4.5 is a statistical classifier which is used for classification to generate decision tree. This algorithm uses information entropy to build the deci-sion tree. C4.5 handles both continuous and discrete attributes, missing attributes and prune trees after creation.

3.4. J48

J48 is an open source Java implementation of C4.5 algo-rithm in WEKA tool. The features managing missing values, pruning decision trees, value ranges, continuous attribute, rules derivation, etc.

a) Optimal Cluster Based Classification Algorithm (OCBC)

The proposed algorithm used the classifier model based on the K-Means and Decision Tree algorithm. The classification accuracy of the dataset is the important one in data mining research. There is numerous classification techniques proposed under the area of data mining. But still, the classification accuracy for SLE datasets is not efficient since the disease predicted to be like many other diseases and the disease have various data formats like structured, unstructured, and semi-structured. Classifying all types of Lupus data is the difficult process for prevailing data mining algorithms.

To lessen such kinds of issues, the proposed algorithm initially cluster the dataset based on the concept of Euclidian distance, and then it classifies the results obtained from the clustering to increase the classification accuracy and efficiency. OCBC can classify all kinds of data formats with high classification accuracy. The procedures of the algorithm are: At the start, it picks the mean values randomly, then it clusters each data values based on the acquired mean.

When the new mean is found, it will update the mean and will cluster the value based on the updated mean. The process will be in the loop until the duplicate mean occurs.

The attained result from this clustering technique is further classified to make the decision. The classification is done by calculating the normalized information gain ratio of each data. The highest normalized information gain is obtained from the calculated value. Based on the final value, the disease will be predicted.

b) Algorithm

Algorithm: K-Means Clustering Build OCBC Decision Tree
Input: K- the number of clusters and R the records of the dataset, the training data T, the attributes_available for computing the next branch
Output: A OCBC decision tree
Method:
Step 1: Randomly choose K objects and make them the K cluster centroids
Step 2: Do
Step 3: For each record in R
Step 4: Calculate distance between each cluster centroid and the record.
Step 5: Assign the record to the cluster that has the minimum distance.
Step 6: Recalculate the cluster means (the values of attributes in the cluster / number of records in the cluster).
Step 7: End for loop
Step 8: While records assignment to clusters do not change
Step 9: End function
Step 10: create a node N.
Step 11: if all records in T have same target class
Step 12: return N as a leaf node with target class.
Step 13: if attributes_available is empty
Step 14: return N as leaf node with maximum target class for the records.
Step 15: Get best_attribute (T, attributes_available).
Step 16: attributes_available = attributes_available – best_attribute.
Step 17: Split the records based on best_attribute(best_attribute, T) //for each split, grown a subtree by calling the //Build OCBC Decision Tree function
Step 18: for each split Ti of T on best_attribute
Step 19: attach a new node returned by build HKMDT Decision-Tree(split records Ti , attributes_available)
Step 20: end for
Step 21: end function

4. Results and discussion

The dataset during this work is tested and analyzed with ID3, C4.5, J48 and OCBC Classification algorithms using cross validation. Also a comparison of accuracy of all classifiers is done and finally it has been investigated that OCBC algorithm performs best with 98% accuracy. The accuracy level of all the algorithms are given below in Table 2.

Table 2: Comparison of Classifiers based on Correctly Classified Instances with Cross Validation

Algorithm	Accuracy
ID3	93.4%
C4.5	95.6%
J48	97.9%
OCBC	98%

Comparison result of all classifiers with the help of SLE prediction tool is shown in fig 2. In this case also OCBC performs best among all classifiers with F-Measure 96.3%. The chart is created which is shown in fig 3.

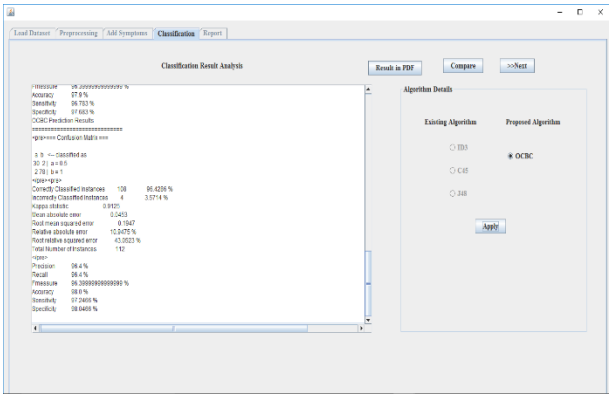
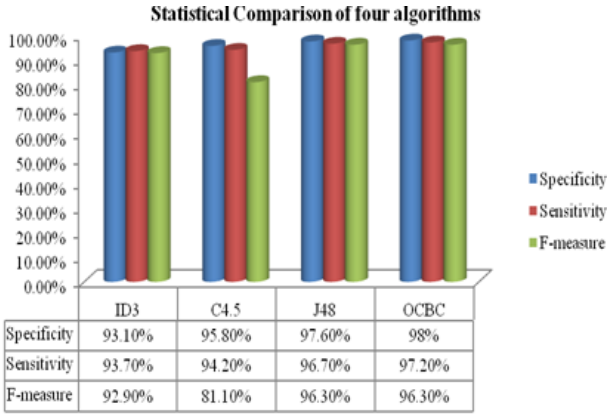


Fig. 2: The Classification Algorithms, Which Are Implemented in the Tool.



	ID3	C4.5	J48	OCBC
Specificity	93.10%	95.80%	97.60%	98%
Sensitivity	93.70%	94.20%	96.70%	97.20%
F-measure	92.90%	81.10%	96.30%	96.30%

Fig. 3: Statistical Comparison of the Classifiers.

Fig.4 shows the report generated after execution and Fig 5 shows the pre-processing module in the tool.

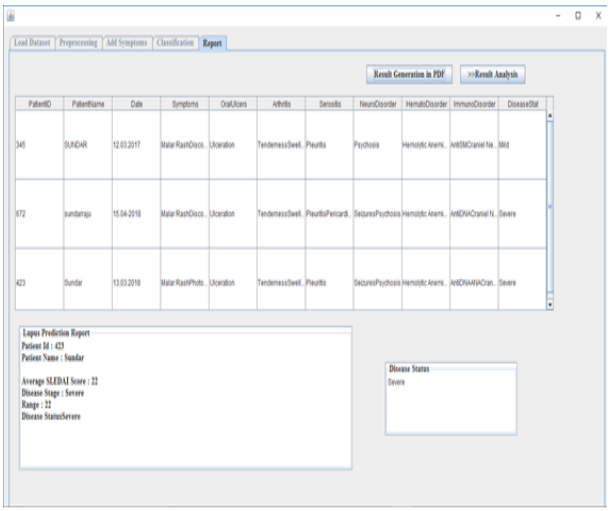


Fig. 4: Report Generated By the Tool.

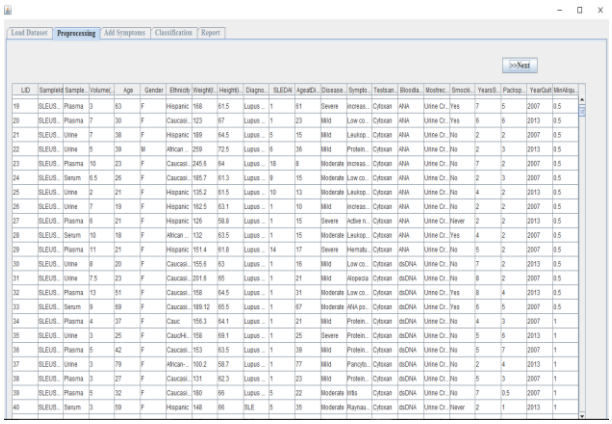


Fig. 5: Pre-Processing the Dataset.

5. Conclusion

In this paper, classification techniques are used for prediction on the dataset of 112 patients, to predict and analyze lupus disease. In this study, a tool was developed based on some selected patient related input variables collected from real world (lab). Among all data mining classifiers OCBC outperforms best with 98% accuracy and therefore that proves to be potentially effective and efficient classifier algorithm. Also comparison of all four classifiers with the help of SLE prediction tool is also done, in this case also OCBC proves to be best with F-measure of 96.3%. Therefore, performance of OCBC is relatively higher than other classifiers. The performance chart is also plotted. This research helps to identify patients' disease severity. Integrating of data mining techniques with healthcare has shown the outstanding results in predicting the disease. The future work is to integrate more algorithms in the tool.

References

[1] Chauraisa, V, and Saurabh Pal, "Early Prediction of Heart Diseases Using Data Mining Techniques", Carib.j. SciTech Vol 1 2013 p 208-217.

[2] Lam, GKW, and M Petri, "Assessment of systemic lupus erythematosus", Clinical and experimental rheumatology 23.5 (2005): S120.

[3] Kaur, Parneet, Manpreet Singh, and Gurpreet Singh Josan, "Classification and prediction based data mining algorithms to predict slow learners in education sector." Procedia Computer Science 57 (2015): 500-508. https://doi.org/10.1016/j.procs.2015.07.372.

[4] NeeshaJothi, Nur'Aini Abdul Rashid, Wahidah Husain, "Data Mining in Healthcare – A Review", Procedia Computer Science,Volume 72,2015,P 306-313. https://doi.org/10.1016/j.procs.2015.12.145.