Analysis, Implementation and Comparison of Machine Learning Algorithms on Breast Cancer Dataset using WEKA Tool

K. Srikanth, S. Zahoor Ul Huq, A.P. Siva Kumar

Abstract*— In modern world among the fatal diseases, cancer is at the peak level. The reasons for cancer disease are modern lifestyle, environment factors or genetic factors. Cancer has become the prime reason of death in developed countries. Among the leading cancers, Breast cancer is at the top in the women. In women, breast cancer is one of the major causes of death. Cancer examination is generally experimental and/or natural in nature. To mine significant data patterns, Data mining acts prominent role in information detection. In data mining applications, forecasting the output of a disease is a major challenge. In this paper, a performance comparison between different machine learning algorithms: Sequential Minimal Optimization (SMO), Naive Bayes (NB), J48 (C4.5 decision tree), K-Means, K-Nearest Neighbours (k-NN) has been performed. To measure the performance of these algorithms, Wisconsin Breast Cancer (Original) dataset has been taken. The main intention is to measure the competence of each algorithm based on precision, specificity, sensitivity and accuracy. Investigated outcomes prove that SMO tops among the list in terms of correctness and low fault rate. All experiments are performed and implemented using WEKA tool.

***Index Terms*—Breast Cancer, Data Mining, Machine Learning, WEKA.**

1. INTRODUCTION

In developed countries, many deaths of people are due to cancer. To decrease cancer deaths, one way is to detect the cancer in premature phase. It is very difficult process to discover cancer in initial stage but it is revealed it can be curable. Among all types of cancers in women, lung cancer top the list and Breast cancer is at second place [1]. According to US reports in the year 2016 [2], nearly 2,46,600 breast cancer cases of women has been registered for diagnosis and among them 40,450 are the expected cases of women's death. Nearly 12% of latest cancer cases in women are of Breast cancer [3]. In India, 1 out of 28 women face breast cancer in their lifetime and it is superior in metropolitan cities. In the western countries, the attack of Breast cancer is high in women at the age 53 to 57 years where as in India it is 43 to 46 years which is a point to be concerned [4].

In reality, Big data has advanced not only the size of data but also creating value from it. Now the word big data is considered to be identical for BA (Business Analytics), DM (Data Mining), BI (Business Intelligence) which plays a major role in estimating the outcomes [5]. From marketing to science, Machine learning and Data mining has high impact in analyzing the data. To save the lives of people, Data mining methods are introduced into health care field to detect the diseases in the early stage and decreasing the cost of expenses. Medical research groups make use of huge amount of health facts available in medical repositories to predict the disease in initial stages. Using the historical data that is offered in repositories, medical researchers use Knowledge Discovery in Databases (KDD) research tool to forecast the result of infection in an efficient manner.

There are many algorithms for classification and prediction of breast cancer outcomes. The present paper gives a comparison of accuracy of the classifiers: SMO [6], NB [7], J48 [8], k-Means [9], k-NN [10] which are among the most significant machine learning algorithms [11, 12]. The paper compares five different machine learning algorithms listed above based on accuracy and correctness using simulation tool, WEKA.

The outstanding paper is prepared as follows. Part 2, provides the reader with the background information on breast cancer research, and literature survey, Part 3 presents a summary of classification techniques namely SMO, NB, J48, k-Means, k-NN, Section 4 gives an outline of WEKA simulation tool and Wisconsin Breast Cancer dataset, Part 5 discusses experiment results obtained and their comparison. Part 6 ends with a conclusion.

2. BACKGROUND

Among a diversity of Data Mining techniques, classification plays a crucial role in Data Mining research. To classify Breast Cancer, more research has been conducted in the medical field by applying machine learning and data mining methods and most of those algorithms have given exceptional precision.

B.Padmapriya and T.Velmurugan [13], done survey on what type of research currently is taken to detect Breast Cancer in the early stages using different data mining techniques. They both work on two popular algorithms namely ID3 and C4.5. In their outcomes, it is clear that C4.5 works better than ID3 on Breast Cancer datasets. Hind Elouedi, Walid Meliani, Zied Elouedi, Nahla Ben Amor [14], utilized the Breast Cancer dataset, to show that the grouping of dangerous examples can be enhanced by

Revised Manuscript Received on March 10, 2019.

K. Srikanth, Research Scholar, Department of Computer Science & Engineering, JNTUA, Anantapuramu, India.(E-mail: kapse.srikanth@gmail.com)

S. Zahoor Ul Huq, Professor, Department of Computer Science & Engineering, G.Pulla Reddy Engineering College, Kurnool, Andhra Pradesh, India.(E-mail: szahoor@gmail.com)

A.P. Siva Kumar, Assistant Professor, Department of Computer Science & Engineering, JNTUA, Anantapuramu, India.(E-mail: sivakumar.ap@gmail.com)

methods for the K-implies bunching calculation and the C4.5 one for arrangement. They showed signs of improvement results by part up the harmful occurrences into two groups, and submitting them to the ID3 calculation for characterization. They considered it as a primer work to understand the significance of refining harmful bosom malignant growth cases that will be affirmed later on a genuine refined dataset. Souad Demigha [15], introduced the ideas and procedures used to build up an information mining framework especially in restorative field and imaging. He cited the utilization of "Choice tree" method, which is exceptionally helpful in the data identification process. He likewise expressed that the capability of Data Mining techniques in therapeutic field permits enhancing the quality and diminishing the expense. Vikas Chaurasia and A.Priyanga and S.Prakasam [16], actualized DMBCPS (Data Mining Based Cancer Prediction System) which appraises the danger of the skin, bosom and lung tumors. The fundamental point of their procedure is to caution the patients about Breast Cancer indications in beginning periods which lead to value productive to the client. The trial results uncovered that the execution of ID3 is better when contrasted with J48 and Naive Bayes grouping calculations. R.Delshi Howsalya Devi and P.Deepika [17], looked at the execution of different bunching procedures to recognize whether the malignancy assault is available or not. As per the aftereffects of their exploratory work, Farthest First Clustering has higher forecast exactness i.e.., 72% than DBSCAN, Canopy, LVQ and Hierarchical grouping strategies.

With respect to all related work mentioned above, here to analyze and estimate the performance of varied top data mining algorithms namely SMO, Naive Bayes, J48, k-Nearest Neighbor, k-Means on Breast Cancer dataset. The objective of the work is to accomplish the best machine learning algorithm in predicting Breast Cancer. For this, first preprocess the Breast Cancer dataset according to WEKA input format, analyze the respective algorithm, implement the algorithm in WEKA platform and assess the effectiveness of each data mining algorithm and compare it. The experimental outcomes prove that SMO attain the top precision (96.92%) with the less fault rate (0.03%) unlike Naive Bayes, J48, k-Means and k-Nearest Neighbors got an precision of 96.33%, 96.04%, 96.04 and 95.75% and fault rate of 0.03%, 0.04%, 0.05% and 0.05% respectively.

III. INTRODUCTION TO CLASSIFICATION TECHNIQUES

A. Sequential Minimal Optimization (SMO)

To train Support Vector Machine (SVM), Sequential Minimal Optimization (SMO) is a novel calculation. In 1998, Sequential Minimal Optimization (SMO) calculation was proposed by John Platt [6], is a straightforward and quick technique for preparing a SVM. In this technique, at each progression, tackle the double quadratic improvement issue. The advantage of SMO is that it very well may be executed essentially and logically. Need the aftereffect of a major quadratic preparing streamlining issue to prepare the SVM. The SMO calculation split this huge quadratic programming issue into a progression of little quadratic programming issues. By tackling these little quadratic programming issues, time-making numerical quadratic programming enhancement stride can be kept away from. SMO can deal with exceptionally enormous preparing sets in light of the fact that the amount of memory required is direct to the preparation set size. In the preparation set size, SMO runs somewhere close to straight and cubic on the grounds that network count is sidestepped. SMO's figuring time is vanquished by SVM estimation; thus SMO is snappier for straight SVMs and meager informational indexes.

B. Naïve Bayes Classifier

The Naïve Bayes calculation is a classifier which depends on likelihood which assesses an arrangement of probabilities by computing the event and stage of qualities in a given preparing informational collection. This classifier utilizes Bayes hypothesis, additionally expect that every one of the properties are autonomous and the estimation of class variable is given. Continuously applications, this suspicion is infrequently holds. In different managed characterization issues, still the calculation needs to execute well and be prepared rapidly since the order is Naïve [18].

Decision tree algorithm J48:

J48 classifier is otherwise called C4.5 choice tree which is utilized for grouping. In 1993, Quinlan Ross proposed C4.5 calculation which is a progression of the ID3 calculation. The calculation is identified with Hunt's calculation and like the ID3, it is successively executed. In contrast to the ID3, the C4.5 works for both consistent and clear cut credits to manufacture choice tree. In arrangement issue, the choice tree acts a key errand for order process. When the choice tree is developed, each tuple in the database is given to choice tree to order [19][20]. For missing qualities in the informational index, J48 allots an esteem dependent on qualities for the rest of the records previously building a choice tree. The key idea is to isolate the information into gathering dependent on the component esteems for that point that is started in the preparation set. J48 arranges the example dependent on choice trees or through standards delivered from them [21][22].

C. K-Means Clustering:

The k-means algorithm [9] defines the centroid of a cluster as the mean value of the points within the cluster. First, it chooses k patterns randomly in the training set D as cluster center or mean. For the remaining objects within the training data set, every pattern is allocated to a cluster based on Euclidean distance between the pattern and cluster center to that it's nearly related. After allotting each pattern in the training data set to a cluster, find the new mean using the objects that are allotted to the cluster as in the above process. Repeat the process and update the object assignment using the new cluster center. The process repeats till the allotment of patterns doesn't change even after the cluster center updated or up to specified repetitions.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

D.IBK (k-Nearest Neighbors Classifier)

Based on similarity, k-NN algorithm [23] classifies the given pattern. The given pattern is classified based on nearest neighbors. When an unknown pattern is given, the k-NN classifier classifies the unknown pattern based on k nearest neighbors in the given training set. The k input to be taken and it can vary. Based on the k value determined, the neighbors will be decided. Most common class label is assigned to unknown pattern surrounded by its k adjacent neighbors. Time taken to classify a test pattern with k-NN classifier raises linearly with the amount of training objects present in the training data set. High storage required for k-NN classifier [24]. The performance of k-NN decreases with raising noise objects within the data set. The performance of k-NN also affects with the value of k i.e., the amount of adjacent neighbors to be used. In many cases, the k value will be treated as one by default.

4. OVERVIEW TO WEKA TOOL AND BREAST CANCER DATA SET

An experiment is conducted to compare the performances of SMO, Naive Bayes, J48, K-Means, K-NN, in terms of effectiveness and the efficiency of the algorithms. In this paper all experiments are implemented using the libraries of WEKA simulation tool [25]. WEKA tool can be applied for classification, data pre-processing, clustering and association rules. The simulation environment, WEKA software is used to apply for various real world applications. The software is a good environment for researchers to implement and evaluate their results on variety of data sets. In this evaluation process, Wisconsin Breast Cancer (original) dataset [26] collected from UCI Machine Learning Library has been used. An overview of the data set is discussed here. Wisconsin Breast Cancer dataset consists of 699 instances, in which 11 attributes are present; one of them is class attribute. Out of 699 instances, malignant class instances are 241 (34.5%) and benign class instances are 458 (65.5%). We tried to remove 16 patterns from the original dataset for which some attribute values are unknown to create a latest dataset consisting of only 683 patterns.

5. IMPLEMENTATION AND RESULTS

Before results discussion, we first try to visualize the attribute values distribution based on class attribute as depicted in Fig. 1.

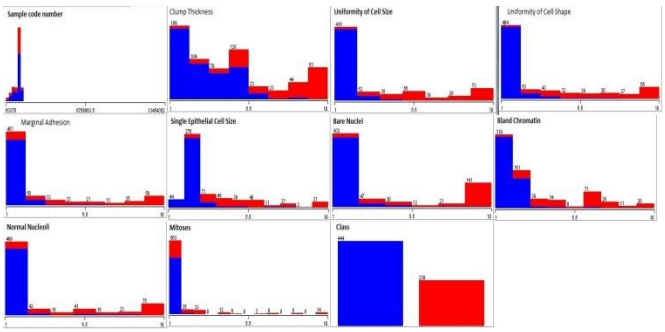

Fig. 1. Illustration outline of breast cancer endurance with all features

Consider 10-fold cross validation test in which the original data set is divided into training and test data set in 10 different formats. To make the classifier to learn, training dataset is utilized and test data set is used to predict the output using the trained classifier. Before analyzing the data visually, first pre-process the data according to WEKA software. After pre-processing the data set, using WEKA software estimate the efficiency of different machine learning algorithms that are discussed above based on precision and efficiency and visualize the results as shown in Fig.2.

Evaluation Criteria	Classifiers				
	NB	k-NN	J48	SMO	k-Means
Time to build a model	0.03	0.01	0.05	0.11	0.06
Correctly classified instances	658	654	656	662	656
Incorrectly classified instances	25	29	27	21	27
Accuracy	96.3	95.8	96	96.9	96.04

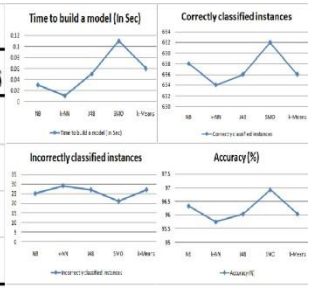

Fig. 2. Performance and comparative graph of the classifiers

Table 1 shows the comparison of machine learning algorithms namely SMO, Naïve Bayes, J48, k-Means, and k-NN in terms of TP, FP, Precision, Recall and F-measure for both classes Benign and Malignant.

Table 1 Comparison of accuracy measures of SMO, Naïve Bayes, J48, k-Means, and k-NN

Fig. 3 represents ROC curve for the classifiers discussed above and confusion matrix. ROC curves are used to evaluate the effectiveness of the classifier and find the optimal classifier and discard others. Confusion matrix gives the outline of forecast outcomes for a classifier

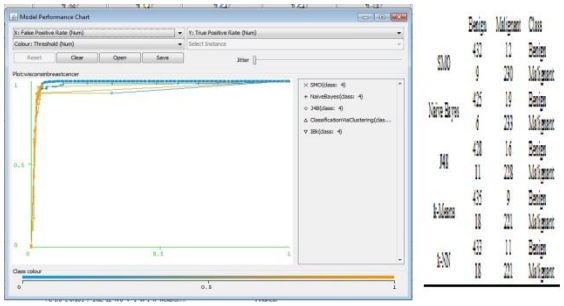

Fig. 3 ROC Curve of the classifiers and Confusion matrix

6. CONCLUSION

To examine medicinal information, different information mining and machine learning strategies are accessible. A basic test in data mining and machine learning districts is to create correct and computationally capable classifiers for Medical applications. In this examination, we used five rule computations: SMO, Naïve Bayes, J48, k-Means and k-NN on the Wisconsin Breast Cancer (one of a kind) datasets. We

endeavored to look at productivity and viability of those calculations as far as exactness, accuracy, affectability and explicitness to locate the best arrangement precision. Taking everything into account, SMO has demonstrated its productivity in Breast Cancer forecast and finding and accomplishes the best execution as far as exactness and low blunder rate.

REFERENCES

1. Nagesh Shukla, Markus Hagenbuchner, Khin Than Win, Jack Yang, "Breast Cancer data analysis for survivability studies and prediction", Computer Methods and Programs in Biomedicine, Volume 155, March 2018, Pages 199-208
2. Madhu Kumari, Vijendra Singh, "Breast Cancer Prediction System", International Conference on Computational Intelligence and Data Science (ICCIDS 2018), Volume 132, 2018, Pages 371-376
3. S.Siva Kumar, Soumya Ranjan Nayak, S.Vidyanandini, J. Ashok Kumar, G. Palai, "An empirical study of supervised learning methods for breast cancer disease, Optik, Volume 175, December 2018, Pages 105-114
4. Siegel RL, Miller KD, Jemal A. Cancer Statistics, 2016. 2016;00(00): 1-24. doi:10.3322/caac.21332
5. Asri H, Mousannif H, Al Moatassime H, Noel T. Big data in healthcare: Challenges and opportunities. 2015 Int Conf Cloud Technol Appl. 2015: 1-7. Doi:10.1109/CloudTech.2015.7337020.
6. R.Delshi Howsalya Devi and P.Deepika, "Performance comparison of various clustering techniques for diagnosis of breast cancer", IEEE International conference on computational intelligence and computing research, 2015.
7. "UCI Machine Learning Repository: Breast Cancer Wisconsin (Original) Data Set." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29. [Accessed: 29-Dec-2015].
8. U.S. Cancer Statistics Working Group. United States Cancer Statistics: 1999-2008 Incidence and Mortality Web-based Report. Atlanta (GA): Department of Health and Human Services, Centers for Disease Control and Prevention, and National Cancer Institute; 2012.
9. "Globocan 2012 – Home." [Online]. Available: http://globocan.iarc.fr/Default.aspx. [Accessed: 28-Dec-2015].
10. Vikas Chaurasia, Saurabh Pal, "A novel approach for Breast Cancer detection using Data Mining techniques", IJIRCCE, vol. 2, Issue 1, January 2014, pp. 2320-9798.
11. Platt, J.C.:Sequential Minimal Optimization: A fast algorithm for training Support Vector Machines. Technical Report MSR-TR-98-14, Microsoft Research, 1998.
12. Rish I. An empirical study of the naive Bayes classifier. IJCAI Work Empir methods Artif Intell. 2001;3(November):41-46.
13. Dr.Neeraj Bhargava, Girja Sharma, Dr.Ritu Bhargava, Manish Mathuria, "Decision Tree Analysis on J48 algorithm for data mining", International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), Vol. 3, Issue 6, 2013, pp. 1114-1120.
14. J. Han and M. Kamber, Data Mining – Concepts and Technique (The Morgan Kaufmann Series in Data Management Systems), 3rd ed. San Mateo, CA: Morgan Kaufmann, 2012, pp. 451-454.
15. Larose DT. Discovering Knowledge in Data. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2004.
16. X. Wu, V. Kumar, J.R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G.J. Mclachlan, A. Ng, B. Liu, P.S. Yu, Z.Z. Michael, S. David, and J.H. Dan, Top 10 algorithms in data mining. 2008, pp. 1-37
17. Datafloq – Top 10 Data Mining Algorithms, Demystified. https://datafloq.com/read/top-10-data-mining-algorithms-demystified/1144. Accessed December 29, 2015.
18. B.Padmapriya and T.Velmurugan, "A Survey on Breast Cancer Analysis Using Data Mining Techniques", IEEE International conference on computational intelligence and computing research, 2014.
19. Hind Elouedi, Walid Meliani, Zied Elouedi, Nahla Ben Amor, "A hybrid approach based on decision trees and clusteing for breast cancer classification", IEEE International conference of soft computing and pattern recognition, 2014, pp. 226-231.
20. Souad Demigha, "Data mining for breast cancer screening", IEEE International conference on computer science & education, 2015, pp. 65-69.
21. A.Priyanga and S.Prakasam, "Effectiveness of data mining based cancer prediction system", Internation journal of computer applications, Vol 83-No. 10, 2013, pp. 11-17.
22. George Dimitoglou, James A. Adams, and Carol M. Jim, "Comparison of the C4.5 and a Naïve Bayes classifier for the prediction of lung cancer survivability.
23. Margaret H. Danham, S. Sridhar, "Data mining, introductory and advanced topics", Pearson education, 1st ed., 2006.
24. Aman Kumar Sharma, Suruchi Sahni, "A comparative study of classification algorithms for spam email data analysis", IJCSE, Vol. 3, No. 5, 2011, pp. 1890-1895.
25. http://www.jstor.org/discover/10.2307/40398417?uid=3738256&uid=2134&uid=368470121&uid=2&uid=70&uid=3&uid=368470111&uid=60&sid=21101751936641.
26. http://stackoverflow.com/questions/10317885/decision-tree-vs-naive-bayes-classifier.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication