ABSTRACT
Breast cancer is cancer that forms in the cells of the breasts which is one of the major reason for death rate in both men and women in the world. It represented that about 12 percent of all new cancer cases and 25 percent of all cancers in women. Classification and data mining methods are effective way to classify and analyze data. Especially in medical field, the machine learning algorithms are used for better prediction. The main objective of WEKA tool is to fetch the highest rate of accuracy from the data collected. This deals with analyzing using single base algorithms to gain highest accuracy rate. This paper presents the K-Nearest Neighbors (IBK) gives the highest accuracy with the lowest error rate. This experiment is done under machine learning data mining tool WEKA.

KEYWORDS: Breast cancer, Classifiers, K-Nearest Neighbor (IBK), WEKA-3.9.

1. INTRODUCTION
Breast cancer is a malignant tumor that forms in the cells of the breast. Although mostly found in women, men can get breast cancer too. Breast cancer has been described as an alarmingly health problem in India. According to the reports, breast cancers have badly attacked women population in India. A survey carried out by Indian Council of Medical Research (ICMR) in the metropolitan cities from 1982 to 2005 has shown that the incidences of breast cancer have doubled. According to the Union Health Ministry, breast cancer ranks as the number one cancer among Indian females with rate as high as 25.8 per 100,000 women and mortality of 12.7 per 100,000 women. According to estimates, at least 17.97,900 women in India may have breast cancer by 2020. In this paper, we include machine learning techniques to predict the recurrence and non-recurrence rate of breast cancer in patients. In our study, we have used the data from University of California, Irvine repository. We analyzed this dataset with the machine learning data mining tool WEKA. WEKA stands for Waikato Environment for Knowledge Analysis is a suite for machine learning software written in Java, developed at the University of Waikato, New Zealand licensed under the GNU General Public License. WEKA contains a collection of visualization tools and algorithms for data analysis and predictive modeling together with graphical user interface for easy access to these functions. Using this tool the paper renders the comparison between four base algorithms K-Nearest Neighbor, Decision Table, Naïve Bayes, Simple Logistics and Multi-Layer Perception.

The rest of this paper is organized as follows. Section 2 is about related work. Section 3 presents the context of the experiment. Section 4 deals with their experimental comparison. Section 5 discusses experiments results obtained. Finally, section 6 concludes the paper.
2. RELATED WORKS

In machine learning and data mining, classification is one of the most important and essential task. After a lot of research has been conducted to apply data mining and machine learning on different medical datasets to classify Breast Cancer. So many of them show good classification accuracy. Vikas Chaurasia and Saurabh Pal11 compare the performance criterion of supervised learning classifiers; such as Naïve Bayes, SVM-RBF kernel, RBF neural networks, Decision trees (J48) and simple CART; to find the best classifier in breast cancer datasets. The experimental result shows that SVM-RBF kernel is more accurate than other classifiers; it scores accuracy of 96.84% in Wisconsin Breast Cancer (original) datasets. Djebbari et al.12 consider the effect of ensemble of machine learning techniques to predict the survival time in breast cancer. Their technique shows better accuracy on their breast cancer data set comparing to previous results. S. Aruna and L.V Nandakishore13, compare the performance of C4.5, Naïve Bayes, Support Vector Machine (SVM) and K- Nearest Neighbor (K-NN) to find the best classifier in WBC. SVM proves to be the most accurate classifier with accuracy of 96.99%. Angeline Christobel. Y and Dr. Sivaprakasam14, achieve accuracy of 69.23% using decision tree classifier (CART) in breast cancer datasets. The accuracy of data mining algorithms SVM, IBK, BF Tree is compared by A. Prades15. The performance of SMO shows a higher value compared with other classifiers. T. Joachims16 reaches accuracy of 95.06% with neuron-fuzzy techniques when using Wisconsin Breast Cancer (original) datasets. In this study, a hybrid method is proposed to enhance the classification accuracy of Wisconsin Breast Cancer (original) datasets (95.96) with 10 fold cross validation. Liu Ya-Qin’s, W. Cheng, and Z. Lu17 experimented on breast cancer data using C5 algorithm with bagging; by preferring additional data for training from the original set using combinations with repetitions to produce multisets of the same size as you’re the original data; to predict breast cancer survivability. Delen et al. Lu18 take 202,932 breast cancer patients records , which then pre-classified into two groups of “survived” (93,273) and “not survived” (109,659). The results of predicting the survivability were in the range of 93% accuracy. The above mentioned all related work with respective, our work compare the behavior of data mining algorithm SVM, NB, k-NN and C4.5 using Wisconsin Breast Cancer (original) datasets in both diagnosis and analysis to make decisions. The goal is to achieve the best accuracy with the lowest error rate in analyzing data. To do so, we compare efficiency and effectiveness of those approaches in terms of many criteria, including: accuracy, precision, sensitivity and specificity, correctly and incorrectly classified instances and time to build model, among others. Our experimental results show that KNN achieves the highest accuracy (97.90%) with the lowest error rate (0.02%) unlike MLP and LR that have an accuracy that varies between 96.50% and 76.22 % and an error rate that varies between 0.03 and 0.06.

3. EXPERIMENT

In order to compare the behaviors of MLP, NB, LR, DT and k-NN we conducted an experiment that focuses on assessing both the effectiveness, and the efficiency of the algorithms. More precisely, the research questions posed for the experiment are: Which algorithm exploits better effectiveness? Which algorithm is more efficient? Which algorithm provides a higher accuracy? And the solution is KNN.

<table>
<thead>
<tr>
<th>S.no</th>
<th>Algorithm</th>
<th>Accuracy</th>
<th>TP Rate</th>
<th>FP Rate</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
<th>ROC area</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>IBK</td>
<td>97.9021</td>
<td>0.979</td>
<td>0.009</td>
<td>0.980</td>
<td>1.000</td>
<td>0.966</td>
<td>0.999</td>
</tr>
<tr>
<td>2</td>
<td>Multi Layer Perception</td>
<td>96.5035</td>
<td>0.965</td>
<td>0.069</td>
<td>0.965</td>
<td>0.965</td>
<td>0.965</td>
<td>0.964</td>
</tr>
<tr>
<td>3</td>
<td>Decision Table</td>
<td>76.2238</td>
<td>0.762</td>
<td>0.474</td>
<td>0.751</td>
<td>0.762</td>
<td>0.734</td>
<td>0.718</td>
</tr>
</tbody>
</table>
4. EXPERIMENT ENVIRONMENT

All experiments on the classifiers described in this paper were conducted using libraries from WEKA machine learning environment\(^1\). WEKA contains a collection of machine learning algorithms for data preprocessing, classification, regression, clustering and association rules. Machine learning techniques implemented in WEKA are applied to a variety of real world problems. The program offers a well-defined framework for experimenters and developers to build and evaluate their models.

5. PROPOSED FRAMWORK

![Diagram: Dataset -> Preprocessing -> Classification Algorithm -> Final Prediction]

6. DESCRIPTION

This includes two phases. In the first phase, the dataset is preprocessed from strings to numerical values. In phase two, the classification algorithm like NB, LR, DT, MLP, KNN are used to find the highest accuracy and finally select the suitable algorithm which render us the accuracy with lowest error rate and time of calculation for the dataset we had taken.

7. DATASET

The Wisconsin Breast Cancer (original) datasets from the UCI Machine Learning Repository is used in this study. Dr. William H. Wolberg (physician) of University of Wisconsin Hospitals Madison, Wisconsin, USA created this dataset. This database therefore reflects this chronological grouping of the data. This dataset consist 286 instance with 10 attributes such as inv-nodes, deg-malign, node caps etc.

8. DISCUSSIONS

By evaluating on training data, that KNN takes about 0.1 s. It can be explained by the fact that MLP is a lazy learner and does not do much during training process unlike others classifiers that build the models. In other hand, the accuracy obtained by KNN (97.90\%) is better than the accuracy obtained by MLP (96.50\%). Naïve Bayes and Simple logistic have an accuracy that varies between 72.72\% and 73.07\%. It can also be easily seen that KNN has the highest value of correctly classified instances of 280 and the lower value of incorrectly classified instances of 6. We can also notice that KNN has the best compatibility between the reliability of the data collected and their validity. MLP and k-NN has the lowest value of error rate; which explains the large number of incorrectly classified instances for each algorithm (6 incorrect instances for KNN and 10 incorrect instances for MLP). After creating the predicted model, we can now analyze results obtained in evaluating efficiency of our algorithms. In fact, that KNN got the highest value (97.9021 \%), of TP for benign class but MLP correctly predicts 276 of instance that belong to malignant class. The FP rate is lower when using KNN classifiers, and then other algorithms follow: Decision table, simple logistic and naïve Bayes. From these results, we can understand why KNN has outperformed other classifiers.

<table>
<thead>
<tr>
<th></th>
<th>Simple Logistic</th>
<th>73.0769</th>
<th>0.731</th>
<th>0.467</th>
<th>0.712</th>
<th>0.731</th>
<th>0.714</th>
<th>0.740</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Naïve Bayes</td>
<td>72.7273</td>
<td>0.727</td>
<td>0.428</td>
<td>0.716</td>
<td>0.727</td>
<td>0.720</td>
<td>0.759</td>
</tr>
</tbody>
</table>

**TABLE 1.1**
3D cylinder graph helps to better understand the power of a machine learning algorithm. We can easily observe in Fig1.1. Then other algorithms follow: DT, LR and NB. KNN correctly predicts 280 instances out of 286 instances, and 6 instances incorrectly predicted. That is why the accuracy of KNN is better than other classification techniques used with lower error rate value. Compared to a good amount of research on Breast-cancer-Wisconsin found in literature that compare Benign Malignant class C4.5 438 20 Benign 14 227 Malignant SVM 446 12 Benign 9 232 Malignant NB 436 22 Benign 6 235 Malignant k-NN 445 13 Benign 20 221 Malignant SVM FP (1-Specificity) TP (Sensitivity) Fig. 3. ROC curve. Hiba Asri et al. / Procedia Computer Science 83 (2016) 1064 – 1069 classification accuracies of data mining algorithms, our experimental results make the highest value of accuracy (97.9021 %) in classifying breast cancer dataset. It can be noticed that KNN outperforms other classifiers with respect to accuracy, sensitivity, specificity and precision; in classifying breast cancer dataset.

9. CONCLUSION

To predict the results in medical field, data mining methods and machine learning tools are effectively used. The main concepts is to get highest accuracy using computational methods. From this study, we analyzed the both single algorithms like NB, LR, DT, MLP and KNN which gave high accuracy of 97.9021% in comparison with all other algorithms. In conclusion, KNN resulted the best and efficient breast cancer algorithm for the analyze and diagnosis of this disease with lowest error rate.

REFERENCES


