Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 83 (2016) 1064 – 1069

Procedia Computer Science

The 6th International Symposium on Frontiers in Ambient and Mobile Systems (FAMS 2016)

Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis

Hiba Asri[a]*,Hajar Mousannif[b],Hassan Al Moatassime[c],Thomas Noel[d]

[a]*OSER Research Team,FSTG Cadi Ayyad University,Marrakech 40000,Morocco*
[b]*LISI Laboratory,FSSM Cadi Ayyad University,Marrakech 40000,Morocco*
[c]*OSER Research Team,FSTG,Cadi Ayyad University,Marrakech 40000,Morocco*
[d]*ICube Laboratory,University of Strasbourg,Strasbourg 67400,France*

Abstract

Breast cancer represents one of the diseases that make a high number of deaths every year. It is the most common type of all cancers and the main cause of women's deaths worldwide. Classification and data mining methods are an effective way to classify data. Especially in medical field, where those methods are widely used in diagnosis and analysis to make decisions. In this paper, a performance comparison between different machine learning algorithms: Support Vector Machine (SVM), Decision Tree (C4.5), Naive Bayes (NB) and k Nearest Neighbors (k-NN) on the Wisconsin Breast Cancer (original) datasets is conducted. The main objective is to assess the correctness in classifying data with respect to efficiency and effectiveness of each algorithm in terms of accuracy, precision, sensitivity and specificity. Experimental results show that SVM gives the highest accuracy (97.13%) with lowest error rate. All experiments are executed within a simulation environment and conducted in WEKA data mining tool.

Peer-review under responsibility of the Conference Program Chairs

Keywords: Breast cancer; SVM;NB;C4.5;k-NN;Classification;Efficiency;Effectiveness.

* Corresponding author. Tel.:+212-666-41-79-11; fax: +212-524-43-31-70.
E-mail address: hiba.asri@ced.uca.ma

Peer-review under responsibility of the Conference Program Chairs
doi:10.1016/j.procs.2016.04.224

1. Introduction

The second major cause of women's death is breast cancer (after lung cancer) [1]. 246,660 of women's new cases of invasive breast cancer are expected to be diagnosed in the US during 2016 and 40,450 of women's death is estimated[2]. Breast cancer represents about 12% of all new cancer cases and 25% of all cancers in women[3].

Information and Communication Technologies (ICT) can play potential roles in cancer care. In fact, Big data has advanced not only the size of data but also creating value from it; Big data, that becomes a synonymous of data mining, business analytics and business intelligence, has made a big change in BI from reporting and decision to prediction results[4]. Data mining approaches, for instance, applied to medical science topics rise rapidly due to their high performance in predicting outcomes, reducing costs of medicine, promoting patients' health, improving healthcare value and quality and in making real time decision to save people's lives.

There are many algorithms for classification and prediction of breast cancer outcomes. The present paper gives a comparaison between the performance of four classifiers: SVM[5], NB[6], C4.5[7] and k-NN[8] which are among the most influential data mining algorithms in the research community and among the top 10 data mining algorithms[9,10]. Our aim is to evaluate efficiency and effectiveness of those algorithms in terms of accuracy, sensitivity, specificity and precision.

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 work

Classification is one of the most important and essential task in machine learning and data mining. About a lot of research has been conducted to apply data mining and machine learning on different medical datasets to classify Breast Cancer. Many of them show good classification accuracy.

Vikas Chaurasia and Saurabh Pal[11] 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 Nandakishore[13], 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. Sivaprakasam[14], 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. Pradesh[15]. The performance of SMO shows a higher value compared with other classifiers. T.Joachims[16] 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. Lu[17] experimented on breast cancer data using C5 algorithm with bagging; by generating 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. Lu[18] 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.

With respect to all related work mentioned above, our work compares the behaviour of data mining algorithms 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 analysing 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 SVM achieves the highest accuracy (97.13%) with the lowest error rate (0.02%) unlike C4.5, Naïve Bayes and k-NN that have an accuracy that varies between 95.12 % and 95.28 % and an error rate that varies between 0.03 and 0.06.

3. Experiment

In order to compare the behaviours of SVM, NB, C4.5 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?

3.1. Experiment Environment

All experiments on the classifiers described in this paper were conducted using libraries from Weka machine learning environment[19]. WEKA contains a collection of machine learning algorithms for data pre-processing, 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.

3.2. Breast cancer dataset

The Wisconsin Breast Cancer (original) datasets[20] from the UCI Machine Learning Repository is used in this study. Breast-cancer-Wisconsin has 699 instances (Benign: 458 Malignant: 241), 2 classes (65.5% malignant and 34.5% benign), and 11 integer-valued attributes.

4. Experimental results

In this section, the results of the data analysis are reported. To apply our classifiers and evaluate them, we apply the 10-fold cross validation test which is a technique used in evaluating predictive models that split the original set into a training sample to train the model, and a test set to evaluate it.

After applying the pre-processing and preparation methods, we try to analyse the data visually and figure out the distribution of values in terms of effectiveness and efficiency.

4.1. Effectiveness

In This section, we evaluate the effectiveness of all classifiers in terms of time to build the model, correctly classified instances, incorrectly classified instances and accuracy. The results are shown in Table 1 and Fig. 1.

Table 1. Performance of the classifiers

Evaluation criteria	Classifiers			
	C4.5	SVM	NB	k-NN
Time to build a model (s)	0.06	0.07	0.05	0.01
Correctly classified instances	665	678	671	666
Incorrectly classified instances	34	21	28	33
Accuracy (%)	95.13	97.13	95.99	95.27

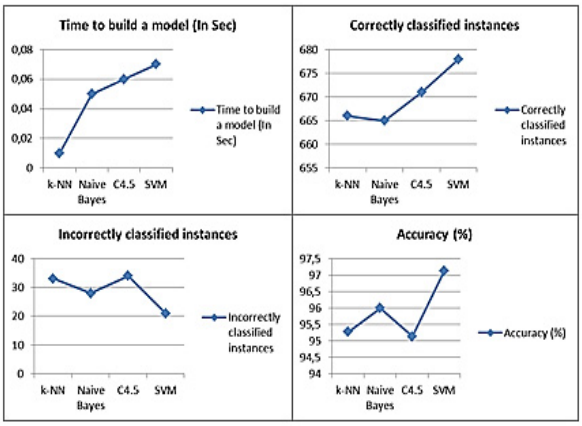

Fig. 1. Comparative graph of different classifiers.

In order to better measure the performance of classifiers, simulation error is also considered in this study. To do so, we evaluate the effectiveness of our classifier in terms of:

- Kappa statistic (KS) as a chance-corrected measure of agreement between the classifications and the true classes,
- Mean Absolute Error (MAE) as how close forecasts or predictions are to the eventual outcomes,
- Root Mean Squared Error (RMSE),
- Relative Absolute Error (RAE),
- Root Relative Squared Error (RRSE).

KS, MAE and RMSE are in numeric values. RAE and RRSE are in percentage. The results are shown in Table 2 and Fig. 2 .

Table 2. Training and simulation error.

Evaluation criteria	Classifiers			
	C4.5	SVM	NB	k-NN
Kappa Statistic (KS)	0.89	0.93	0.91	0.89
Mean Absolute Error (MAE)	0.06	0.02	0.03	0.04
Root Mean Square Error (RMSE)	0.21	0.16	0.19	0.21
Relative Absolute Error (RAE) %	14	6.33	8.59	10.46
Root Relative Squared Error (RRSE) %	45	35.58	40.95	44.77

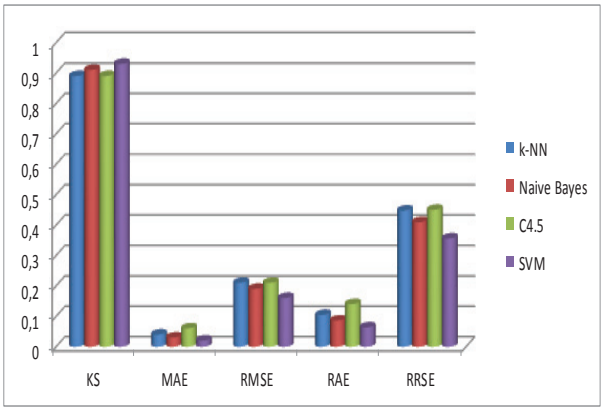

Fig. 2. Comparative diagram of machine learning algorithms with respect to evaluation criteria: KS, MAE, RMSE, RAE and RRSE.

4.2. Efficiency

Once the predictive model is built, we can check how efficient it is. For that, we compare the accuracy measures based on precision, recall, TP rate and FP rate values for C4.5, SVM, NB and k-NN as shown in Table 3.

Table 3. Comparison of accuracy measures for C4.5, SVM, NB and k-NN.

	TP	FP	Precision	Recall	F-Measure	Class
C4.5	0.95	0.05	0.96	0.95	0.96	Benign
	0.94	0.04	0.91	0.94	0.93	Malignant
SVM	0.97	0.03	0.98	0.97	0.97	Benign
	0.96	0.02	0.95	0.96	0.95	Malignant
NB	0.95	0.02	0.98	0.95	0.96	Benign
	0.97	0.04	0.91	0.97	0.94	Malignant
k-NN	0.97	0.08	0.95	0.97	0.96	Benign
	0.91	0.02	0.94	0.91	0.93	Malignant

To better understand efficiency, Fig. 3 presents the ROC curve of our classifiers that better illustrate the precision of each classifier. The ROC curve gives a graphical graph that illustrates the performance of different classifiers. From the plot we can easily select optimal models and discard others to best classification. Since Confusion matrices represent a useful way for evaluating classifier, each row of Table 4 represents rates in an actual class while each column shows predictions.

Table 4. Confusion matrix.

	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

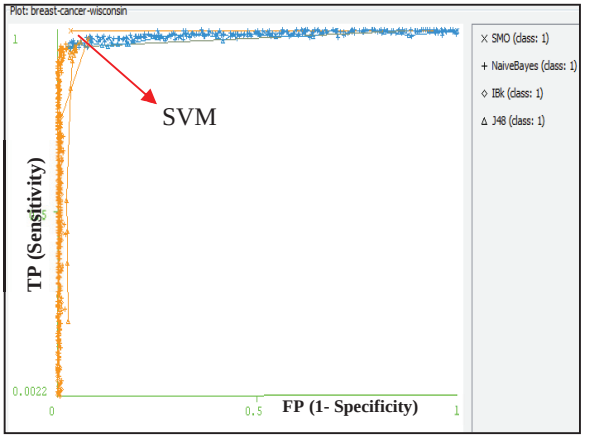

Fig. 3. ROC curve.

5. Discussion:

We can notice from Table 1 that SVM takes about 0.07 s to build its model unlike k-NN that takes just 0.01 s. It can be explained by the fact that k-NN 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 SVM (97.13%) is better than the accuracy obtained by C4.5, Naïve Bayes and k-NN that have an accuracy that varies between 95.12 % and 95.28 %. It can also be easily seen that SVM has the highest value of correctly classified instances and the lower value of incorrectly classified instances than the other classifiers (see Fig. 1).

From Table 2, we can better see that the chance of having a best classification (0.93 %) with the least warning error rate (0.02) is produced by SVM. We can also notice that SVM has the best compatibility between the reliability of the data collected and their validity. C4.5 and k-NN has the highest value of error rate; as shown in Fig. 2, which explains the large number of incorrectly classified instances for each algorithm (34 incorrect instances for C4.5 and 33 incorrect instances for k-NN) (see Fig. 1).

After creating the predicted model, we can now analyse results obtained in evaluating efficiency of our algorithms. In fact, Table 3 shows that SVM and C4.5 got the highest value (97 %) of TP for benign class but k-NN correctly predicts 97% of instance that belong to malignant class. The FP rate is lower when using SVM classifiers (0.03 for benign class and 0.02 for malignant class), and then other algorithms follow: k-NN, C4.5 and NB. From these results, we can understand why SVM has outperformed other classifiers.

ROC curve helps to better understand the power of a machine learning algorithm. We can easily observe in Fig. 3 that SVM is the perfect classifier as it begins from the left corner, to straight up to the upper left corner and then to the upper right corner (99% sensitive and 99% specific). Then other algorithms follow: NB, C4.5 and k-NN.

Let us now compare actual class and predicted results obtained using confusion matrix as shown Table 4. SVM correctly predicts 678 instances out of 699 instances (448 benign instances that are effectively benign and 221 malignant instance that are actually malignant), and 21 instances incorrectly predicted (12 instance of benign class predicted as malignant and 9 instances of malignant class predicted as benign). That is why the accuracy of SVM is better than other classification techniques used with lower error rate value.

In summary, SVM was able to show its power in terms of effectiveness and efficiency based on accuracy and recall. Compared to a good amount of research on Breast-cancer-Wisconsin found in literature that compare

classification accuracies of data mining algorithms, our experimental results make the highest value of accuracy (97.28 %) in classifying breast cancer dataset. It can be noticed that SVM outperforms other classifiers with respect to accuracy, sensitivity, specificity and precision; in classifying breast cancer dataset.

6. Conclusion

To analyze medical data, various data mining and machine learning methods are available. An important challenge in data mining and machine learning areas is to build accurate and computationally efficient classifiers for Medical applications. In this study, we employed four main algorithms: SVM, NB, k-NN and C4.5 on the Wisconsin Breast Cancer (original) datasets. We tried to compare efficiency and effectiveness of those algorithms in terms of accuracy, precision, sensitivity and specificity to find the best classification accuracy. of SVM reaches and accuracy of 97.13% and outperforms, therefore, all other algorithms. In conclusion, SVM has proven its efficiency in Breast Cancer prediction and diagnosis and achieves the best performance in terms of precision and low error rate.

References

1. 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.
2. Siegel RL, Miller KD, Jemal A. Cancer Statistics , 2016. 2016;00(00):1-24. doi:10.3322/caac.21332.
3. “Globocan 2012 - Home.” [Online]. Available: http://globocan.iarc.fr/Default.aspx. [Accessed: 28-Dec-2015].
4. 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.
5. Noble WS. What is a support vector machine? *Nat Biotechnol*. 2006;24(12):1565-1567. doi:10.1038/nbt1206-1565.
6. Rish I. An empirical study of the naive Bayes classifier. *IJCAI Work Empir methods Artif Intell*. 2001;3(November):41-46.
7. Quinlan JR. *C4.5: Programs for Machine Learning*.; 2014:302. https://books.google.com/books?hl=fr&lr=&id=b3ujBQAAQBAJ&pgis=1. Accessed January 5, 2016.
8. Larose DT. *Discovering Knowledge in Data*. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2004.
9. 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.
10. Datafloq - Top 10 Data Mining Algorithms, Demystified. https://datafloq.com/read/top-10-data-mining-algorithms-demystified/1144. Accessed December 29, 2015.
11. V. Chaurasia and S. Pal, “Data Mining Techniques : To Predict and Resolve Breast Cancer Survivability,” vol. 3, no. 1, pp. 10–22, 2014.
12. Djebbari, A., Liu, Z., Phan, S., AND Famili, F. International journal of computational biology and drug design (ijcbdd). 21st Annual Conference on Neural Information Processing Systems (2008).
13. S. Aruna and L. V Nandakishore, “KNOWLEDGE B ASED A NALYSIS OF V ARIOUS S TATISTICAL T OOLS IN D ETECTING B REAST,” pp. 37–45, 2011.
14. A. C. Y, “An Empirical Comparison of Data Mining Classification Methods,” vol. 3, no. 2, pp. 24–28, 2011.
15. A. Pradesh, “Analysis of Feature Selection with Classification : Breast Cancer Datasets,” *Indian J. Comput. Sci. Eng.*, vol. 2, no. 5, pp. 756–763, 2011.
16. Thorsten J. Transductive Inference for Text Classification Using Support Vector Machines. *Icml*. 1999;99:200-209. doi:10.4218/etrij.10.0109.0425.
17. L. Ya-qin, W. Cheng, and Z. Lu, “Decision tree based predictive models for breast cancer survivability on imbalanced data,” pp. 1–4, 2009.
18. D. Delen, G. Walker, and A. Kadam, “Predicting breast cancer survivability: a comparison of three data mining methods,” *Artif. Intell. Med.*, vol. 34, pp. 113–127, 2005.
19. W. Version, “Machine Learning with WEKA,” 2004.
20. “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].
21. “SUGI 31 Statistics and Data Analysis Receiver Operating Characteristic (ROC) Curves Mithat Gönen , Memorial Sloan-Kettering Cancer Center SUGI 31 Statistics and Data Analysis FN + FP,” pp. 1–18, 2001.