International Journal of Engineering & Technology, 7 (2.15) (2018) 136-138

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research Paper

Analyzing performance of classifiers for medical datasets

Rosaida Rosly*, Mokhairi Makhtar*, Mohd Khalid Awang, Mohd Isa Awang, Mohd Nordin Abdul Rahman

Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin, Terengganu, Malaysia
**Corresponding author E-mail: rosaidarosly@gmail.com,mokhairi@unisza.edu.my*

Abstract

This paper analyses the performance of classification models using single classification and combination of ensemble method, which are Breast Cancer Wisconsin and Hepatitis data sets as training datasets. This paper presents a comparison of different classifiers based on a 10-fold cross validation using a data mining tool. In this experiment, various classifiers are implemented including three popular ensemble methods which are boosting, bagging and stacking for the combination. The result shows that for the classification of the Breast Cancer Wisconsin data set, the single classification of Naïve Bayes (NB) and a combination of bagging+NB algorithm displayed the highest accuracy at the same percentage (97.51%) compared to other combinations of ensemble classifiers. For the classification of the Hepatitisdata set, the result showed that the combination of stacking+Multi-Layer Perception (MLP) algorithm achieved a higher accuracy at 86.25%. By using the ensemble classifiers, the result may be improved. In future, a multi-classifier approach will be proposed by introducing a fusion at the classification level between these classifiers to obtain classification with higher accuracies.

Keywords:*Classification model; Data mining; Medical dataset.*

1. Introduction

Medical data is one of the applications that contain a huge amount of data collected and stored in its databases. This issue is prominently famous in data mining literature [8]. Data mining also called knowledge discovery is the process of extracting information from large data sets using various techniques [5]. It can reduce cost or inflate revenue [9]. The main focus of this paper is to classify and analyse the performance of classifiers for medical data sets using the data mining approach. Medical data sets involved in this analysis were the Breast Cancer Wisconsin (Original) and the Hepatitis data sets. Breast cancer ranks highest among women's health concerns. It is the most frequently diagnosed cancer in women [11]. In [7], hepatitis is one of the most common diseases among Egyptians as it represents 22% of hepatitis cases around the world. The diagnosis of some diseases like breast cancer and hepatitis is very difficult task for a doctor, especially in making their decision. Data mining tools such as WEKA are used in this paper for classification techniques. Weka is one of the frameworks used for classification that contains many well-known data mining algorithms [5]. Due to the high dimensionality of the medical data set, the classification methods for data analysis need to be employed. Therefore, various machine learning classification algorithms have been applied to the medical data analysis. These include the use of Naïve Bayes (NB), decision tree (J48), Instance Based for K-Nearest neighbour (IBK), Sequential Minimal Optimization (SMO), Multi-Layer Perception (MLP) and ensemble methods (boosting, bagging and stacking). These classifications were applied here because they are the most popularly used by researchers [11].

In [2], it was argued that NB has a good performance in most medical problems. It has been used widely in medical applications such as in breast cancer diagnosis, especially in classifying the accuracy of performance. NB has been widely used for data classification because it is easy to code and conduct, intuitive and can be easily handled even with missing features [1]. J48 can also handle training data with missing attribute values and continuous and discrete attributes. IBK classification categorises instances based on their similarity. It is widely used in the medical field such as breast cancer diagnosis. Support Vector Machine (SVM) called SMO in WEKA is a very powerful method and famously applied in a wide variety of applications [11]. MLP is an example of an artificial neural network (ANN) applied to solve a number of different problems in applications [3].

The tasks involved in the classification method were preprocessing data, selecting attributes, conducting single classification and lastly combining the ensemble classifiers. The processed data was run through a 10 fold cross-validation.

The rest of this paper is organised as follows. Section 2 discusses the details of the data sets, followed by the classification method explained in section 3 and the experimental results shown in section 4. Lastly, section 5 concludes this paper.

2. Data sets

The data sets used in this paper were publicly available in the UCI Machine Learning Repository [6]. There were two data sets used in this paper, which are the Breast Cancer Wisconsin and Hepatitis data sets.

2.1. Breast cancer Wisconsin data set

The data set was created by Dr. Wiliam H. Wolberg who was a physician at the University of Wisconsin Hospital, Madison, Wisconsin, USA. It was donated by Olvi Mangsarian on 15 July 1992. The Breast Cancer Wisconsin (Original) data set from the UCI machine learning repository was a classification data set which recorded the measurements for breast cancer cases. There were two classes of cases which are benign (non-cancerous) and malignant (cancerous).

Copyright © 2018 Authors. 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.

Table 1 shows the samples taken periodically as Dr. Wolberg reported his clinical cases. The data was shown as chronological groups that describe the period they were created, starting from January 1989 until the last instance which was created in November 1991.

Table 1: Breast Cancer Wisconsin Data set information

Group	Instance	Date
Group 1	367	January 1989
Group 2	70	October 1989
Group 3	31	February 1990
Group 4	17	April 1990
Group 5	48	August 1990
Group 6	49	Updated January 1991
Group 7	31	June 1991
Group 8	86	November 1991
Total: 699 points		

Before being publicly available, the data set had 701 points but on January of 1989, after being revised, two instances from group 1(originally contained 369 instances) were considered inconsistent and removed from the data set.
There were 10 features that differed significantly between benign and malignant samples which were clump thickness, uniformity of cell size and shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, mitoses and class (2 for benign and 4 for malignant). Each feature is evaluated on a scale of 1 to 10 with 1 being the closest to benign and 10 closest to malignant.

2.2. Hepatitis data set

The data set was donated by G. Gong from Carnegie-Mellon University via Bojan Cestnik from the Jozef Stefan Institute, Yugoslavia on 1 November 1988 [4]. The Hepatitis data set from the UCI machine learning repository is a classification data set which records the measurements for Hepatitis cases. There are two classes for classification which are either die or survive.
The data set has 155 instances with some missing values. This data set contains a mixture of integer and real valued attributes, with information about patients affected by the Hepatitis disease. There are19 features in this data set which are age, sex, steroid, antivirals, fatigue, malaise, anorexia, liverbig, liverfirm, spleenPalpable, spiders, ascites, varices, bilirubin, alkPhosphate, sgot, albumin, protime, histology and class.

3. Methodology

The classification of these medical data sets started with the pre-processing method. It focused on the handling of missing values, discretization of numeric attributes and selection of attribute subsets. In handling the missing values, the removal of instances with missing values was applied. All instances with missing values were removed in the first step of pre-processing.
These data sets contain integer attributes, so Weka tools cannot support this format. Thus, the changing of integer attributes to numeric attributes were applied in the discretization task. After discretization, the task was continued with selecting the attribute subsets. It showed the number of attributes used to train the classifiers. A good hypothesis cannot be achieved if there are too many parameters [10].
To classify the data sets, the single classification was used first using some of the classifiers. They are NB, J48, IBK, SMO and MLP. Then, during the task of combining the ensemble classifiers, different ensemble methods such as boosting (AdaBoost), stacking and bagging were used to test the single algorithms to see whether this approach can give better accuracy of the data sets. The ensemble method was combined with the single classifiers to get the best accuracy. The single classifiers with the highest accuracy were used as base classifiers to be combined with the ensemble method. Lastly, the accuracy performance of the combination was determined.
In Weka, multiple classifiers were selected to be used in the Weka.classifiers.meta.vote. The average of probabilities were selected as the combination rule which can work with any types of classes, thus returning the mean of the probability distributions for each of the base classifier (learnt within Vote or built outside and loaded by Vote).

4. Results and discussion

4.1. Breast cancer Wisconsin data set

4.1.1. Single classification task for Breast Cancer Wisconsin data set

This research used the WEKA data mining tool to run this experiment. Figure 1 shows the accuracy performance for the five classifiers with selected feature obtained using a 10 fold cross-validation: NB, J48, MLP, SMO and IBk. The x-axis shows types of classifiers and y-axis shows accuracy performance of classifiers. NB was found to achieve the highest accuracy with 97.51%, followed by SMO and IBk with the same percentage (96.93%), indicating a result better than those produced by MLP and J48.

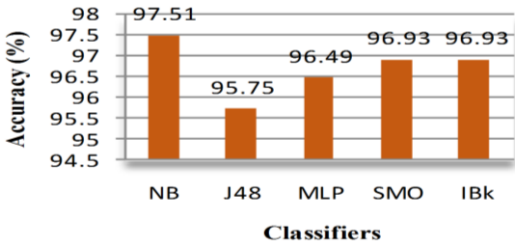


Fig. 1: Single classifier in Breast Cancer Wisconsin (original) data set

4.1.2. Combination of ensemble classifiers for Breast Cancer Wisconsin data set

In this experiment, improvement of results from the NB algorithms was investigated using the combined ensemble method. Figure 2 shows the comparison of ensemble classifiers (boosting, bagging and stacking) combined with NB based on a 10-fold cross validation with feature selection. The x-axis shows accuracy performance of combination ensemble classifiers and y-axis shows types of ensemble classifiers. NB was selected as the base classifier to be combined with the ensemble classifiers because it presented the highest accuracy in the single classification. The classifier used for the combination was voted from Weka. From the figure, bagging+NB achieved the highest accuracy (97.51%) compared to stacking+NB and boosting+NB.

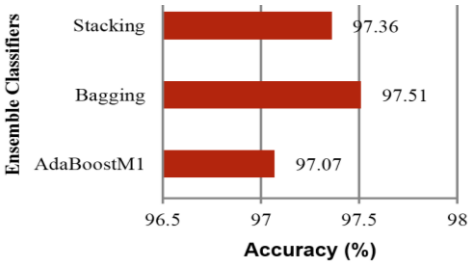


Fig. 2: Ensemble classifiers that were combined with the NB algorithm for the Breast Cancer Wisconsin data set

Overall, the result shows that the combination of bagging+NB has the same percentage with the single classification of NB with 97.51%.

4.2. Hepatitis data set

4.2.1. Single classification task for Hepatitis data set

Figure 3 displays the result for the five classifiers with feature selection obtained using a 10 fold cross-validation: NB, J48, MLP, SMO and IBk. The x-axis shows types of classifiers and y-axis shows accuracy performance of classifiers. It was noticed from the test that MLP had the highest accuracy of 82.50%, followed by NB, J48 and IBk with the same percentage (81.25%) indicating a result better than those produced by SMO.

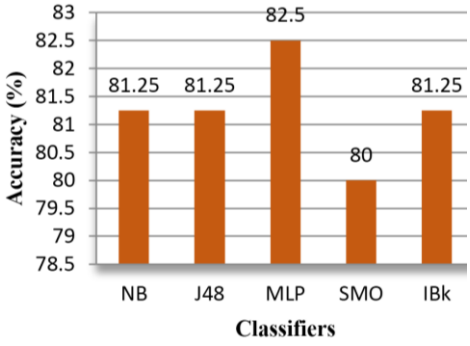


Fig. 3: Single Classifier in Hepatitis data set

4.2.2. Combination of ensemble classifiers for Hepatitis data set

In this experiment, the improvement of results from the MLP algorithms was investigated with the combined ensemble method. Figure 4 displays the comparison of ensemble classifiers (boosting, bagging and stacking) that were combined with MLP based on a 10-fold cross validation with feature selection. The x-axis shows accuracy performance of combination ensemble classifiers and y-axis shows types of ensemble classifiers. MLP was selected as the base classifier to be combined with the ensemble classifiers because it showed the highest accuracy in the single classification of the Hepatitis data set. Stacking+MLP achieved the highest accuracy (86.25%), followed by bagging+MLP (85%), indicating a result better than that produced by boosting+MLP.

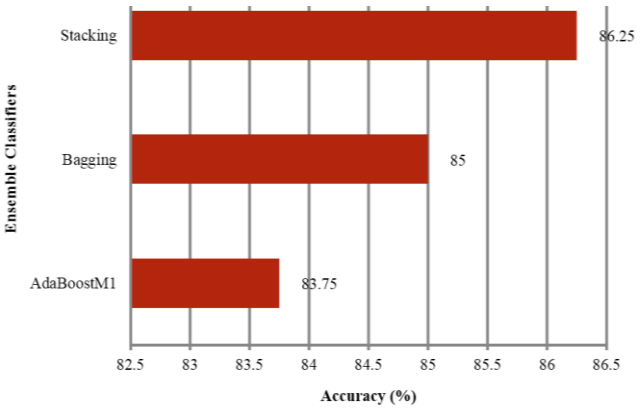


Fig. 4: Ensemble classifiers combined with MLP algorithm for Hepatitis data set

From the experiment on the Hepatitis data set, the result shows that the combination of stacking+MLP has the highest accuracy at 86.25% compared to the single classification and other combinations for this data set.

5. Conclusion

The experimental results show that for the classification of the Breast Cancer Wisconsin data set, the single classification of NB and combination of bagging+NB algorithm achieved the highest accuracy with the same percentage (97.51%) compared to other combinations of ensemble classifiers. While for the classification of the Hepatitis data set, the combination of stacking+MLP algorithm achieved the highest accuracy at 86.25%. Using ensemble classifiers, the results may be improved. In the future, a multi-classifier approach should be proposed by introducing fusion between these classifiers at the classification level in order to obtain a higher accuracy of classification.

Acknowledgement

This work is partially supported by UniSZA and KPM (Grant No. FRGS/1/2016/ICT02/UNISZA/02/2).

References

[1] Abraham R, Simha JB & Iyengar SS (2007), Medical datamining with a new algorithm for Feature Selection and Naïve Bayesian classifier. *Proceedings of the IEEE 10th International Conference on Information Technology*, pp. 44–49.

[2] Al-Aidaroos KM, Bakar AA & Othman Z (2012), Medical Data Classification with Naive Bayes Approach. *Information Technology Journal* 11, 1166–1174.

[3] Areerachakul S & Sanguansintukul S (2010), Classification and regression trees and MLP neural network to classify water quality of canals in Bangkok, Thailand. *International Journal of Intelligent Computing Research* 1, 43–50.

[4] Bache K & Lichman M (2013), *UCI machine learning repository*, University of California.

[5] Bhuvaneswari E & Dhulipala VS (2013), The study and analysis of classification algorithm for animal kingdom dataset. *Information Engineering* 2, 412–414.

[6] Blake CL & Merz CJ (1998), *UCI repository of machine learning databases*, University of California.

[7] EL-Bohy AM, Hashad AI & Taha HS (2015), Performance evaluation of hepatitis diagnosis using single and multi-classifiers fusion. *International Journal of Engineering Research and Technology* 4, 293–298.

[8] Hickey SJ (2013), Naive Bayes classification of public health data with greedy feature selection. *Communications of the IIMA* 13, 87–98.

[9] Nandhini M & Scholar PD (2016), Boosting and meta-learning techniques for distributed data mining on electronic medical datasets. *International Journal of Computer Technology and Applications* 7, 403–410.

[10] Rosly R, Makhtar M, Awang MK, Rahman MN & Deris MM (2006), Multi-classifier models to improve accuracy of water quality application. *ARPN Journal of Engineering and Applied Sciences* 11, 3208–3211.

[11] Salama GI, Abdelhalim M & Zeid MA (2012), Breast cancer diagnosis on three different datasets using multi-classifiers. *International Journal of Computer and Information Technology* 1, 36–43.