

Comparative Analysis of Algorithms in Supervised Classification: A Case study of Bank Notes Dataset

Anahita Ghazvini^{#1}, Jamilu Awwalu^{#2}, and Azuraliza Abu Bakar^{*3}

^{#1}Postgraduate Student at Centre for Artificial Intelligence and Technology (CAIT)

^{#2}Faculty of Computing and Information Science, Baze University, Abuja, Nigeria.

^{*3}Professor at Centre for Artificial Intelligence and Technology (CAIT)

Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia (UKM) 43600, Bangi Selangor, MALAYSIA.

Abstract

There are different techniques in conducting data mining that range from clustering, association rule mining, prediction and classification. These techniques are applied using learning algorithms such as Support Vector Machines (SVM), Naïve Bayes, and Artificial Neural Network (ANN). When conducting data mining, the choice of algorithm to use is an important decision because it depends on factors such as the nature or type of data under examination, and the target outcome of the data mining activity. In this study, we compare Naïve Bayes and Multilayer Perceptron using the classification technique as a case study on the Bank Notes dataset from the University of California Irvine (UCI) from two standpoints, which are; holdout and cross validation. Result from experiments show Multilayer Perceptron outperforms Naïve Bayes in terms of accuracy from both standpoints of holdout and cross validation.

Keywords

Holdout, Cross validation, Naïve Bayes, Multilayer Perceptron

I. INTRODUCTION

Data mining as one of the fields on study in Artificial Intelligence is applicable to different domains that range from industrial, education, or medical fields due to its great potential in aiding data miners or data scientist focus on the important information in their data warehouses to conduct classification, prediction of trends, associative mining, pattern analysis and behaviours allowing them to make a proactive, knowledge-driven decisions. Algorithms used in Machine Learning are commonly used in data mining for classification, prediction, association rule mining, and detection.

These algorithms can be applied in data mining from two standpoints; Holdout and Cross Validation. However, the decision of which of the available algorithms to use in carrying out a mining task such as classification or prediction affects the

reliability of the outcome of the data mining task e.g. classification or prediction. Reliability is a key factor in result of data mining tasks such as classification or prediction, because the more accurate the classification is, the more reliable it is, and vice versa. Data mining as described by [1] is the process of using sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data set. These tools can include statistical models, mathematical algorithm and machine learning methods. Consequently, data mining consists of more than collection and managing data, it also includes analysis and prediction[2].

The detection and classification of fake or counterfeit Banknotes from real ones is an important task in every economy or society usually carried out using different techniques. Counterfeit Banknotes is produced with different motivations as stated by [3], such as the difficulty for the visually impaired people to identify the validity of the Banknote and its value.

The successful implementation of data mining is composed of two step as stated by Bulent (2006); first, is coming up with a precise formulation of the problem are trying to solve. Which includes a focused statement usually results in the best payoff. The second key is using the right data by choosing from the data available, or perhaps buying data from an external source. In this study, classification as a Data mining task is used on the Banknote Authentication Dataset to identify and classify counterfeit Banknotes from real ones using features or attributes collected to form the dataset by training and testing two algorithms, namely; Naïve

Bayes and Multilayer Perceptron to compare their performance on accuracy and speed of the classification.

II. LITERATURE REVIEW

Banknote verification or validation has been studied by researchers from different perspectives ranging from algorithms used in validating Banknotes or detecting counterfeit Banknotes, framework building, to pattern recognition in identifying values validity and values of Banknotes.

A study by [4] for counterfeit banknote recognition used Multiple Kernel Support Vector Machines (SVM). In the study, each banknote was divided into sections and the image histograms for each section is taken as input to the SVM. The SVM architecture permits false positive pattern to have a bigger penalty than a false negative was developed in order to minimize the approximate balanced error rate. The application of multiple kernels, optimal weights with kernel metrics combination were obtained through semi-definite programming (SDP).

In a related study by [5], Neural Network was used in Banknote recognition by optimizing the masks exploited by the Network to perform validity and value recognition. Result showed that the Neural Network was able to several pieces of banknotes. However, there were worries by the authors about the fluctuating masks sets and threshold on the reliability of the system.

Comparison of algorithms in data mining is important in order to identify which algorithm outperforms which, in a given scenario or data. In a study by [6], several algorithms like decision tree, Naïve Bayes, Neural Networks, Nearest Neighbour, and Support Vector Machines were compared with the aim of identifying how each algorithm works, advantages, disadvantages, and research issues on each algorithm. Also, [7] compared Bayesian and Lazy classifiers. The algorithms used were Bayes Net and Naïve Bayes for the Bayesian Classification, while the Lazy algorithms are Instance Based Learning (IBL), IBK (K-Nearest Neighbour), and K-Star. Result showed that IBK from the Lazy classifier achieved better result than

the other classifiers from both the Bayesian and Lazy algorithms.

III. MATERIALS AND METHODOLOGY

This section begins by first describing the bank notes dataset. Then, the algorithms employed in classifying the dataset.

A. Dataset

The dataset used in this study is obtained from the University of California Irvine (UCI) publicly available dataset repository, donated by Volker Lohweg in August 2012. A tabular description of the dataset in shown in table 1

TABLE 1 TABULAR DATASET DESCRIPTION

Dataset Characteristics	Multivariate	Number of Instances	1372
Attributes Characteristics	Real	Number of Attributes	5
Date Donated	2013/04/16	Missing Values	None

The dataset contains five attributes as shown in table 2, the attributes are described as:

TABLE 2 TABULAR DESCRIPTION OF DATASET ATTRIBUTES

Attribute	Type
Variance of Wavelet Transformed image	Continuous
Skewness of Wavelet Transformed image	Continuous
Curtosis of Wavelet Transformed image	Continuous
Entropy of image	Continuous
Class	Integer

The dataset was formed from captured images of genuine and forged bank notes specimen, and wavelet transform tool was used to extract features from the captured images.

IV. CLASSIFICATION

A. Naïve Bayes

The Naïve Bayesian algorithm is a statistical method that uses probability to predict the membership of a given value to a certain class. Developed in 1912 by Thomas Bayes, it is called Naïve because it assumes all variables contribute

towards classification and are mutually correlated, also known as class independence.

The Naïve Bayesian has the following advantages as stated by [8] and [9]:

It requires minimal training time. Easy to interpret in knowledge representation. Robust and a good classifier

Also, the Naïve Bayes has the following disadvantages:

The conditional independence of class assumption by the Naïve Bayes is not always true, thus leading to low accuracy in some cases.

B. Artificial Neural Network (ANN) Back Propagation

Artificial Neural Network is a connected set of input/output units each having an assigned weight. Back Propagation is a type of Artificial Neural Network.

The advantages of Back Prop as stated by [9] are:

Able to tolerate noisy data, and classify from untrained data. Good for continuous valued inputs or output. Can be used when little is known about attributes and classes. Its parallelization technique can be used to speed up computation time.

Some disadvantages of Back Prop are:

It takes long learning time, hence more suitable for an application where that is feasible. It is black box, therefore very hard to interpret. Requires number of parameters that are to be determined empirically

Data Transformation: The bank note dataset attributes contains real numbers that range from negative to positive decimal values. The dataset was normalized resulting in all values converted to the range of 0 and 1, instead of negative to positive values as it was originally in the dataset. Then the attributes of the dataset were discretized, and binned into four bins.

Dataset Split: this is the experimentation on different dataset splits for training and testing, the compared algorithms are used on each dataset split, and the result compared in the experiment. The different dataset splits is shown in table 3.

TABLE 3 HOLDOUT PERCENTAGE SPLIT

S/No.	Percentage Split
1	90:10
2	60:40
3	30:60

V. EXPERIMENT AND RESULT

The classification was conducted on using the previously discussed algorithms, i.e. Back Propagation, and Naïve Bayes algorithms. The experiment conducted are in two phases, name; the holdout phase where a section of the dataset is used to train the classifier and the other section is used to test the classifier. Then, the Cross validation. The result from the classification of using each algorithm is presented and discussed in this section.

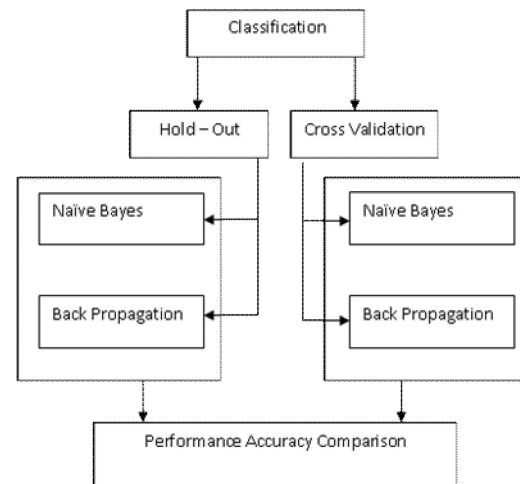


Figure 1 Experiment Flow

A. Accuracy Measure

Result of holdout and cross validation of Naïve Bayes experiment is explained in this section. The results are explained in terms of correctly classified and incorrectly classified results. The result obtained from applying Naïve Bayes algorithm on the pre-processed dataset is shown in table 4.

TABLE 4 NAÏVE BAYES HOLDOUT

Hold out Percentage Split		Build Duration in Seconds	Result	
Training	Testing		Correct	Incorrect
90	10	0.02	90.51	9.48
60	40	<0	87.43	12.56
30	60	<0	89.37	10.62

Naïve Bayes hold out result as shown in table 4 shows that the variation of accuracy and speed based on the percentage split used in training and testing. From the three hold out percentages used in table 4, it is clear that the first hold out which is 90% training and 10% testing achieved the highest accuracy, followed by 30% training 60% testing, and lastly 60% training and 40% testing. The following observations are peculiar to table 4:

The gap between accuracy measure of the three holdout percentages in not wide, especially between the accuracies of hold out one (90.51) and hold out three (89.37) where the gap is 1.14. And the widest gap is between holdout one and hold out two where the gap is 3.08.

The build duration as shown in table 4 is same for hold out two and three where they are both less than zero seconds, and holdout one duration is 0.02 seconds. As such the highest accuracy which is holdout one comes with the cost of taking the longest build duration.

However, the accuracy of Naïve Bayes as shown in table 5 using cross validation is comparatively lower than the accuracy obtained from the holdout classification. But the difference between the two techniques i.e. holdout and cross validation based on tables 4 and 5 is not much, considering that the difference is approximately 2% only.

TABLE 5 NAÏVE BAYES CROSS VALIDATION

No. of Fold	Duration in Seconds	Result	
		Correct	Incorrect
10	< 0	88.04	11.95
7	< 0	87.60	12.39
5	0.02	88.33	11.66

B. Multilayer Perceptron

Result from applying the Multilayer Perceptron in classifying the preprocessed data using Cross validation experiment and result is shown in table 7, and Hold out result obtained from shown in table 6.

Table 6 Multilayer Perceptron Holdout

Percentage Split		Build Duration in Seconds	Result	
Training	Testing		Correct	Incorrect
90	10	22.55	97.08	2.91
60	40	21.77	95.81	4.18
30	60	22.72	95.83	4.16

TABLE 7 MULTILAYER PERCEPTRON CROSS VALIDATION

No. of Folds	Duration in Seconds	Result	
		Correct	Incorrect
10	12.25	95.99	4.00
7	12.22	95.91	4.08
5	12.06	95.99	4.00

The Multilayer Perceptron hold out result as shown in table 6 shows differences of accuracy and build duration speed based on the percentage split used in training and testing. From the three hold out percentages used in table 6, it is clear that the first hold out which is 90% training and 10% testing achieved the highest accuracy i.e. 97.08, followed by 30% training 60% testing i.e. 95.83, and lastly 60% training and 40% testing i.e. 95.81. The following observations are peculiar to table 5:

The gap between accuracy measure of the three holdout percentages in not big, especially between the accuracies of hold out two (95.81) and hold out three (95.83) where the gap is 0.02. And the widest gap is between holdout one and hold out two where the gap is 1.25.

The build duration as shown in table 5 is different for each of the three hold out, but share a similarity which is all are greater than twenty seconds, and the differences between each holdout build duration is minimal ranging from 0.17 to 0.95.

VI. DISCUSSION

The result obtained of preprocessing and classifying the Bank Note dataset using the Naïve Bayes and the Multilayer Perceptron as shown in section 4 varies. However, the variation is a result of the different experiment settings used. Holdout results from both algorithms i.e. Naïve Bayes and Multilayer Perceptron shows the 90% training split to have the best result and closely followed by the 30% training split which are; 1.14% and 0.17% respectively. As such, in order to avoid over-fitting where the algorithm can only perform well in on only data of same type of training set and fail in data where it has not, the highest holdout percentage used is training i.e. 90% is not recommendable because using an excessive amount of training data and very less in testing causes over fitting. Therefore, in order to recommend a result

that is less prone to over fitting or under fitting, and with good accuracy level from the holdout classification, the 30% training percentage and 60% testing is recommended, this is because in both algorithms i.e. Naïve Bayes and Multilayer Perceptron, the second best result with high accuracy is the 30% training and 60% testing with very little difference with the split that is prone to over fitting i.e. 90% training and 10% testing, also the third percentage split (30:60) has a competitive advantage of build duration over the first percentage split (90:10) in Naïve Bayes and 0.17 seconds more in Multilayer Perceptron than the first hold out percentage, thus the delay of 0.17 seconds for the Multilayer Perceptron is not much delay that cannot be tolerated in order to avoid over-fitting.

However, in terms of cross validation the best result was obtained from the 5 folds cross validation from the set of 10, 7, and 5 cross validations. The result from Naïve Bayes shows the 5 folds which is 88.33% exceeds the 7 by 0.73% and 10 by 0.29% folds in terms of accuracy and less time. Also, the Multilayer Perceptron shows the 5 folds to achieve same accuracy level which is 95.99% with the 10 folds in less time it takes the 10 folds by 0.19 seconds and 0.08% more accurate than the 7 folds.

VII. CONCLUSION

In this study, the Naïve Bayes and Multilayer Perceptron algorithms were compared on the Bank notes dataset from two standpoint of classification which are; holdout and cross validation. The result shows that Multilayer Perceptron yields better result than the Naïve Bayes in both standpoints of the experiment in terms of accuracy, however; the Naïve Bayes proves to be faster than the Multilayer Perceptron, but with a short coming in terms of accuracy because the Multilayer Perceptron takes the lead in accuracy.

REFERENCES

- [1] P. P. Tanna and Y. Ghodasara, "Foundation for Frequent Pattern Mining Algorithms ' Implementation," *Int. J. Comput. Trends Technol.*, vol. 4, no. 7, pp. 2159–2163, 2013.
- [2] K. Arts, "A Study On Classification Of Imbalanced Data Set 1," *Int. J. Innov. Sci. Eng. Technol.*, vol. 1, no. 7, pp. 247–250, 2014.
- [3] A. Bruna, G. M. Farinella, G. C. Guarnera, and S. Battiato, "Forgery detection and value identification of Euro banknotes.," *Sensors (Basel)*, vol. 13, no. 2, pp. 2515–29, Jan. 2013.
- [4] C.-Y. Yeh, W.-P. Su, and S.-J. Lee, "Employing multiple-kernel support vector machines for counterfeit banknote recognition," *Appl. Soft Comput.*, vol. 11, no. 1, pp. 1439–1447, Jan. 2011.
- [5] L. Sakoobunthu, "Thai Banknote Recognition Using Neural Network 1055119," *Knowledge-Based Intell. Inf. Eng. Syst.*, vol. 2773, 2003.
- [6] H. Bhavsar and A. Ganatra, "A Comparative Study of Training Algorithms for Supervised Machine Learning," *Int. J. Soft Comput. Eng.*, vol. 2, no. 4, 2012.
- [7] S. Vijayarani and M. Muthulakshmi, "Comparative Analysis of Bayes and Lazy Classification Algorithms," *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 2, no. 8, 2013.
- [8] H. Jiawei and K. Micheline, *Data mining: concepts and techniques*, vol. 49, no. 06. Morgan Kaufmann Publishers, 2006, pp. 49–3305–49–3305.
- [9] S. B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques," *Informatica*, vol. 31, pp. 249–268, 2007.