International Journal of Engineering Science Invention
ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726
www.ijesi.org ||Volume 5 Issue 2|| February 2016 || PP.62-71

Performance Analysis of Various Data Mining Techniques on Banknote Authentication

Nadia Ibrahim Nife
University of Kirkuk, Iraq
nadia.ibra@uokirkuk.edu.iq

ABSTRACT*: In this paper, we describe the functionality features for authenticating in Euro banknotes. We applied different data mining algorithms such as KMeans, Naive Bayes, Multilayer Perceptron, Decision trees (J48), and Expectation-Maximization(EM) to classifying banknote authentication dataset. The experiments are conducted in WEKA. The goal of this project is to obtain the higher authentication rate in banknote classification.*

KEYWORDS -*Banknote authentication dataset, applying data mining algorithms, classification, clustering in Weka.*

I. INTRODUCTION

Banknote authentication stays an important challenge for the central banks in order to keep the strength of the financial system around the world, and to keeping assurance in confidence documents, mostly banknotes. The researchers is described a manner for examination the authenticity of documents, in banknote which involve security of authentic documents, beneficial on the security characteristics of documents Which include image characteristics that used for making the security documents. The method comprises procedure of digitally processing image to be authenticated the surface of applicant document, which state of attention includes at least part of the security features, the digital processing including performing a decomposition of the sample image through means of wavelet transform of sample image.

Decomposition of sample image is based on a wavelet packet transform of the pattern image. We had banknote authentication dataset, these Data extracted from images. These dataset reserved for the estimation of an authentication steps for banknote. Wavelet Transform implement were applied to mine features from images. Authentication obtained through a flow of segmentation and classification measures. The images of banknotes are first fragmented in various parts, and then the results of classification are collective to achieve the final banknote authentication. Inherent algorithm has been used to distinguish valid and counterfeit banknote. The approach considers currency, the applicability is not easy in the environment of Euro banknotes as this currency instructs various approaches to avoid copies hence many theories on features and their location should be done.

II. MOTIVATIONS

One of the most substantial tasks is finding of counterfeit banknotes. Also, there is the trouble for blind and partially sighted people to know both the value and authenticity of banknotes, where there is no method for them to check for the authenticity and for forgeries the banknotes.The validation of banknotes is a difficult task also for people without visualization difficulties; under visible light the Banknotes copying are typically equal to authorized ones.Consumer authentication can be very beneficial in exceeding this issue.This fact makes scientists to develop several forgery discovery algorithms, taking into account various currencies.

III. DATA MINING

It is the analysis stage of the knowledge discovery in databases process [1], and the science of discovery new exciting patterns and relationship in large amount of data.The data mining used to mine information for a dataset and convert it to comprehensible structure for further use.The main task in the Data Mining is the extraction of significant information, samples from hug datasets, mostly in the area of bioinformatics studies.Knowledge indicates data classification, clustering or prediction.

DM has become a well-known in the field of Knowledge Engineering and Artificial Intelligence. Exactly; data mining is the operation of discover connection or samples through lots of attributes in big relational databases and extraction beneficial information from data. The knowledge is to build computer programs that examine over databases automatically, looking for predictabilities or patterns.Robust patterns will

make accurate predictions on future data.The technical of data mining provides through machine learning.It is used to extract information from the databases that is expressed in an understandable form and may be used for a diversity of aims.All attribute in dataset applied through algorithms of machine learning is characterized by the identical collection of features.This study is interested with regression issues in which the output of attributes declares actual values as an alternative of discrete values in classification matters. It is developing field of computational intelligence [2]. The first step of predictive data mining is collecting the data set. Characteristic choice is the operation of recognizing and removing as various unsuitable and redundant characteristics. Several features based on the precision of supervised machine learning models.This problem can be studied by creating new features from simple feature.

DATA SETS

Data sets (banknote authentication) used in our projects are taken from center for machine learning and intelligent systems, this data were mined from images that were taken for the estimation of verification process for banknotes, as shown in Figure (1).

Attribute description:[3]

1. Variance of Wavelet Transformed image (continuous)
2. Skewness of Wavelet Transformed image (continuous)
3. Curtosis of Wavelet Transformed image (continuous)
4. Entropy of image (continuous).
5. Class(integer)

Attribute Characteristics	Real
Instances Number	1372
Attributes Number	5
Date Donated	16/4/2013

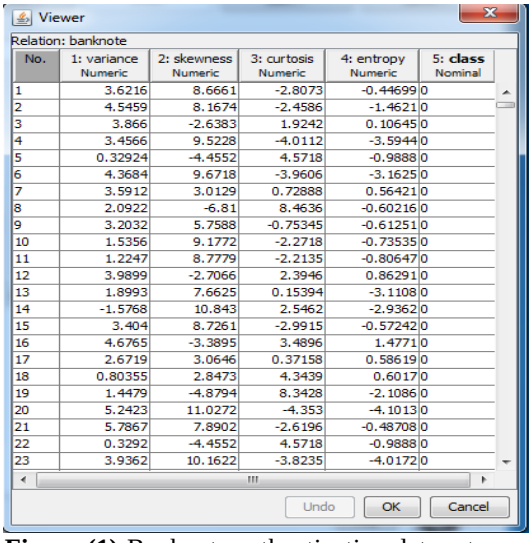

Figure (1): Banknote authentication data sets

IV. DATA MINING ALGORITHMS

In this paper we will give the details of algorithms, in our project we used five Data Mining algorithms that we will apply for our data sets then we obtained the results and evaluate them in both clustering and classifications algorithms. In the subsequent, there are some descriptions about Algorithms that applied in our research:

- ***Decision Trees:***

The C4.5 algorithm is a data mining algorithm, and a statistical classifier that produces a decision tree which can be used to classify test instances. It plays a significant role in the operation of data analysis and data mining [4]. It does so by recursively dividing the data on a single attribute, according to the calculated information gain of each split in the tree represents a spot where a decision must be prepared depend on the input, and you go to the following node and the next till you reach at a leaf that expresses you the predicted output.

- ***Naive Bayes Classifier:***

It is a simple probabilistic [5]. This classifier Naive Bayes is the generality simple text classification methods with different uses in language discovery, arrangement the private email, spam detection into email, and document classification. Although the naive scheme and generalized rules that this method uses, Naive Bayes accomplishes well in several difficult actual world troubles. Naive Bayes classifier is precise proficient as it needs a lesser quantities of training data. Also, the time of training through Naive Bayes is much smaller In comparison with alternate ways. The classification of Bayesian offers prior knowledge, algorithms of process learning, experimental data can be joined, and a beneficial perception for estimating various learning algorithms. It computes obvious eventualities for theory and it is strong in input data.

- ***Multilayer Perception classifier:***

It is the best commonly used of neural network. It is both easy and depended on hard arithmetic field. Input numbers are managed via sequential layers of neurons. The number of variables of the problem equivalent to an input layer with a number of neurons, and an output layer wherever the perceptron answer is made available with a mount of neurons equivalent to the favorite number of amounts calculated from the inputs. The layers amid input layer and output layer are known as hidden layers. Perceptron can simply carry out linear functions without hidden layer. All difficulties which may be resolve, a perceptron may be solved with only one hidden layer but it is sometimes more capable to use two hidden layers. The perceptron calculates an only output as of many real inputs [6]. All neuron of layer other than the input layer calculates initial a bias plus a linear set of the outputs of the neurons for the previous layer. Bias with coefficients of linear groups named the weights.

- ***K-means:***

It is the best common partition clustering technique [7]. It is an algorithm to categorize or to collection your objects depended on characteristics into K number of set. K is a number positive integer. The combination is done by decreasing the sum of squares of distances among the corresponding cluster centroid and data. Hence, the purpose of K-mean clustering is to categorize the data.

- ***Expectation-maximization (EM)***:

It is a technique for obtaining maximum probability or maximum a posteriori evaluations of factors in arithmetical models, where the model influenced by ignored hidden variables. EM offers proficient form of clustering algorithm and more robust [8]. Expectation-maximization usually used to calculate maximum probability evaluations specified uncompleted samples.

V. TESTING AND RESULTS

The sample data set used for this project is "banknote". In this term paper supposes that appropriate data preprocessing has performed and practical five algorithms in WEKA for our dataset. The following testing and results for thesealgorithms as mention bellow:

- ***Classification algorithms :***

- *Decision tree algorithm*:Decision trees are strong and widespread algorithm for classification and prediction. In order to start analyze the dataset "banknote authentication.arff" using DT. You will analyze the data with C4.5 algorithm using J48. Assess classifier depended on what way well it predicts of group of attributes while completed training set. The Classifier Decision tree process output range depicting training and testing results, we got to the results that show in (Table1), (Table2) and (Figure 2).

TABLE 1: Result with Decision Trees

Correctly Instances	Correctly Instances(%)	Incorrectly Instances	Incorrectly Instances(%)
1366	99.5627	6	0.4373
Kappa statistic	Mean absolute error	RMS error	Relative absolute error%
0.9911	0.0086	0.0656	1.7443
Root relative squared error%	Coverage (0.95level)%	Mean rel. region size (0.95level)%	Leaves number
13.2075	99.5627	50	15
Total Instances	Relation	Tree size	Time model created
1372	Banknote	29	0.01 seconds

TABLE 2: Detailed Accuracy through Class

TP Rate	FP Rate	Precision	Recall	Class
0.995	0.003	0.997	0.995	0
0.997	0.005	0.993	0.997	1
0.996	0.004	0.996	0.996	Class
MCC	ROC Area	ROC Area	F-Measure	Class
0.991	0.998	0.998	0.996	0
0.991	0.998	0.995	0.995	1
0.991	0.998	0.997	0.996	

The set of measurements is derived from the training data. In this case only 99.5% of 1372 training instances have been classified correctly. This specifies that the results found from training data are not positive matched with what might have acquired from the separate test set from the same source. Thus Decision tree is a classifier in the method of a tree structure, it classify attributes in dataset via initialing on the tree root then moving over it to a leaf node. Initial criterion of choosing a characteristic in Decision tree is a test in each node to choose a useful feature common to classify data.

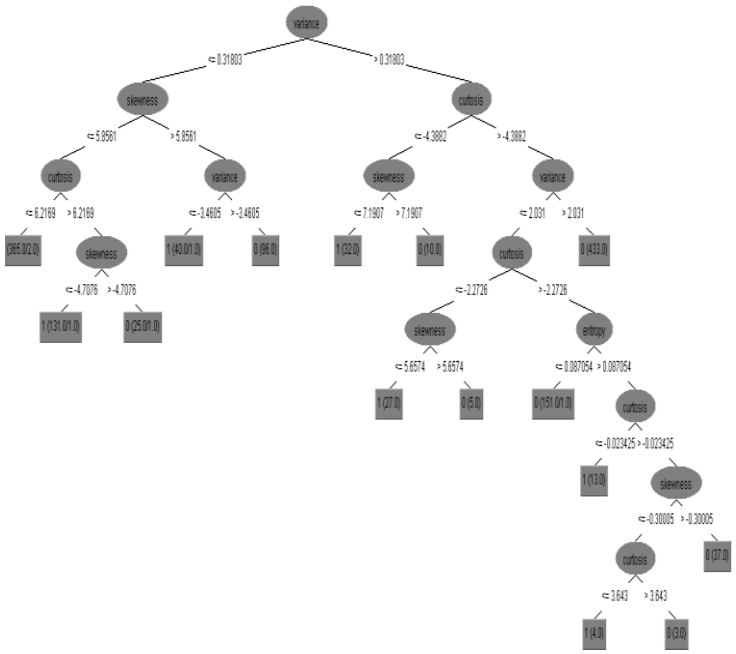

Figure (2): Decision tree chart

- ***Naive Bayes:***It is probabilistic learning method; it is easy classifiers that one may utilize because of the easy mathematics that are interested. The goal of a classifier is to recognize which group fits a sample depended on the given suggestion. We apply Naive Bayes to the dataset to get the results that show in to Table3, Table 4, Table 5, and Figure (3).

TABLE 3: Result with Naive Bayes

Correctly Classified Instances	Correctly Classified Instances(%)	Incorrectly Classified Instances	Incorrectly Classified Instances(%)
1154	84.1108	218	15.8892
Kappa statistic	Mean absolute error	RMS error	Relative absolute error%
0.6764	0.1885	0.3225	38.1726
Root relative squared error%	Coverage (0.95level)%	Mean rel. region size (0.95level)%	Total Instances
64.9043	99.5627	74.3805	1372

TABLE 4: Detailed Accuracy by Class

TP Rate	FP Rate	Precision	Recall	Class
0.881	0.208	0.841	0.995	0
0.792	0.119	0.841	0.997	1
0.841	0.169	0.841	0.996	Class
MCC	ROC Area	ROC Area	F-Measure	Class
0.677	0.940	0.957	0.860	0
0.677	0.940	0.923	0.816	1
0.677	0.940	0.942	0.841	

TABLE 5: Detailed Accuracy by Class

TP Rate	FP Rate	Precision	Recall	Class
1.000	0.000	1.000	0.995	0
1.000	0.000	1.000	0.997	1
1.000	0.000	1.000	0.996	Class
MCC	ROC Area	ROC Area	F-Measure	Class
1.000	1.000	1.000	1.000	0
1.000	1.000	1.000	1.000	1
1.000	1.000	1.000	1.000	

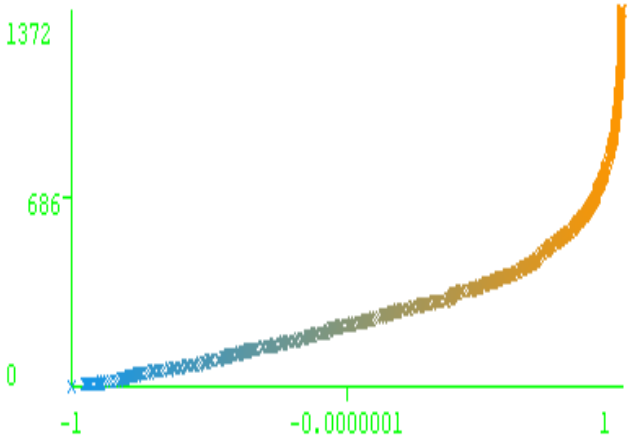

Figure (3): Visualize margin curve

- ***Multilayer Perceptron :***

The multi-layer perceptron (MLP) is the common neural network algorithm. This kind of neural network needs a wanted output so as to learn therefore it is called supervised network. The objective of this form of network is to build a model that properly plots the input to the output by old data so as to the model can then be utilized to produce the output while the wanted output is unidentified. Training dataset with MLP is shown below:

TABLE 6: Result with Multilayer Perceptron

Correctly Instances	Correctly Instances (%)	Incorrectly Instances	Incorrectly Instances (%)
1372	100	0	0
Kappa statistic	Mean absolute error	RMS error	Relative absolute error%
1	0.0026	0.0081	0.5364
Root relative squared error%	Coverage (0.95level)%	Mean rel. region size (0.95level)%	Time model created
1.6382	100	50.4738	1.25

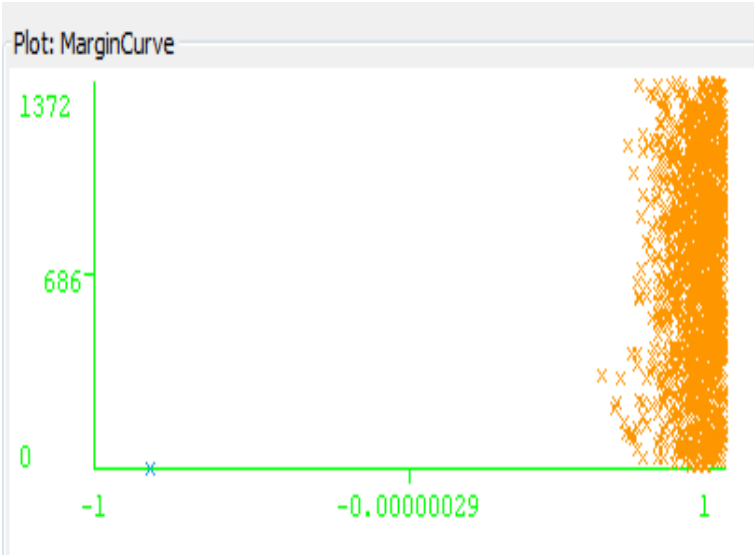

Figure (4): Visualize margin curve

- ***Clustering algorithms:***

- ***KMeans algorithm***

It is an algorithm to association your objects depended on instances into K number of cluster. K is positive integer digit. The combination is complete via decreasing the sum of squares of distances through the corresponding cluster centroid and data. KMean found the most favorable number of clusters. While practical KMean algorithm to the Dataset, we found the results as shown in the following (Figure5), (Figure6) and (Table7):

```
Number of iterations: 3
Within cluster sum of squared errors: 176.5160205561705
Missing values globally replaced with mean/mode

Cluster centroids:
                          Cluster#
Attribute    Full Data          0          1
                (1372)       (610)      (762)
=============================================
variance       0.4337    -1.8684     2.2767
skewness       1.9224    -0.9936     4.2566
curtosis       1.3976     2.1483     0.7967
entropy       -1.1917    -1.2466    -1.1476
class               0          1          0
```

Figure 5:KMean cluster output

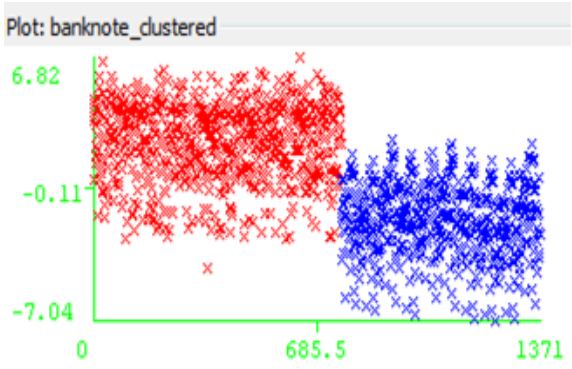

Figure 6: Visualize cluster assignment

TABLE 7: Model and evaluation on training set

Cluster	Instances	Instances%
0	610	44
1	762	56

After creating the clustering then the training attributes into clusters after the cluster illustration and calculates ratio of attributes falling in all clustering. The above clustering produced by k-means shows 44% (610 instances) in cluster 0 and 56% (762 instances) in cluster1, Time taken to build model (full training data): 0.02 seconds.

- ***Expectation maximization (EM) :***

Expectation maximization algorithm discusses calculating the probability that every datum is a member of all categories, maximization raises to changing the factors of every class to make best use of those probabilities. Expectation maximization gives a probability allocation to all attribute which specifies the probability of it to all of the clusters. After us practical EM process, we found the results as shown in the following (Figure 7), and (Table 8):

Table 8: Clustered Instances for EM Algorithm

1	69 (5%)	12	96 (7%)
2	79 (6%)	13	45 (3%)
3	93 (7%)	14	78 (6%)
4	79 (6%)	15	51 (4%)
5	76 (6%)	16	24 (2%)
6	72 (5%)	17	57 (4%)
7	32 (2%)	18	20 (1%)
8	78 (6%)	19	30 (2%)
9	105 (8%)	20	30 (2%)
10	31 (2%)	21	31 (2%)
11	69 (5%)	22	127 (9%)

Figure 7: Visualize cluster assignment

Once we calculating and training data, Expectation–maximization algorithm has taken time 459.54 seconds with LOG probability=-7.95525 Table.1 shows the results in the table 9:

Table 9: Evaluate on training data

Time model created	Clusters Number	Iterations Number	Log likelihood
459.54 seconds	22	82	7.95525-

VI. COMPARISON OF RESULTS

1) Classifications algorithms: compare the results of classification the following Comparison for classifications algorithms in performance sensibility and precision for Banknote authentication, and information evaluation of data which include Coverage of cases, time taken to create model, incorrectly classified attributes, and correctly classified attributes. We observed that Decision trees-J48 classification has the highest error than the others; we may see the variance among algorithms from Table 10, Table 11 and Table12 as follow:

Table 10: Performance (Sensitivity) / Banknote

Algorithms	Sensitivity (%)	
	0	1
Decision trees-J48	99.5%	Decision trees-J48
Naive Bayes	88.1%	Naive Bayes
Multilayer Perceptron	100%	Multilayer Perceptron

Table11: Performance Banknote authentication

Algorithms	Precision (%)	
	0	1
Decision trees-J48	99.7%	99.3%
Naive Bayes	84.1%	84.1%
Multilayer Perceptron	100%	100%

Table 12: Classification evaluation of Banknote

Algorithm	Correctly Attributes	Incorrectly Attributes	Coverage (0.95 level)%	Time model created
Decision trees -J48	99.5%	0.43%	99.5%	0.01 second
Naive Bayes	100%	0%	100%	0.01 second
Multilayer Perceptron	100%	0%	100%	0.001 second

2) Clustering algorithms*:* We can understand the change between numbers of iterations achieved and number of clusters selected through cross authentication, time taken to create model from Table 13 as follow:

Table 13: Times and No. of attributes

Algorithms	clusters number	iterations number performed	Time model created (full training)
KMeans algorithm	2	3	0.02 seconds
EM algorithm	22	82	462.79 seconds

VII. CONCLUSION

In this paper we assessed the performance of classification, and clustering algorithms. The goal of our project is to obtain the optimum algorithm, basically a sample of banknotes was implemented in Weka, and the precision of these various algorithms was recorded. The mostly precise algorithms for this dataset are Decision trees-J48, Multi-Layer Perceptron, EM algorithm, KMeans algorithm, and Naive Bayes, from these calculations we found that Multilayer Perceptron algorithm is superior than other in performance correctly classified attribute and incorrectly classified attribute. In the future we propose examining data by using Multilayer Perceptron algorithm.

REFERENCES

[1] https://en.wikipedia.org/wiki/Data_mining.

[2] Andrew K., Jeffrey A., Kemp H. Kernstine, and Bill T. L.,2000, *"Autonomous Decision-Making: A Data Mining Approach",*IEEE transactions on information technology in biomedicine, vol.4, no.4, , pp. 274-284

[3] http://archive.ics.uci.edu/ml/datasets/banknote+authentication

[4] Dharm S., Naveen C., and Jully S., 2013,*"Analysis of Data Mining Classification with Decision Tree Technique*", Global Journal of Computer Science and Technology Software & Data Engineering, vol.13, issue13.

[5] Naveen K., Sagar P., Deekshitulu, (2012), *"Implementation of Naive Bayesian Classifier and Ada-Boost Algorithm Using Maize Expert System",* (IJIST), vol.2, no.3.

[6] Gaurang P., Amit G., Kosta and Devyani, (2011), *"Behavior Analysis of Multilayer Perceptron's with Multiple Hidden Neurons and Hidden Layers",*International Journal of Computer Theory and Engineering, vol.3, no.2.

[7] Rohtak, H., (2013),*"A Review of K-mean Algorithm",* IJETT, vol.4, issue7.

[8] Aakashsoor, and Vikas, (2014),*"An Improved Method for Robust and Efficient Clustering Using EM Algorithm with Gaussian Kernel",* International Journal of Database Theory and Application vol.7, no.3, pp.191-200.