Research Paper

Volume : 4 | Issue : 7 | July 2015 • ISSN No 2277 - 8179

Applying Data Mining Tool (Weka) to Study Consumer Acceptance of E-Banking'

Computer Science

KEYWORDS : Data Mining, WEKA tool, e-banking, banking sector, classification, Clustering (Simple K Means), etc.

Sulochana Patil — Assistant Professor, Krishna Institute Of Computer Application, And Management, Wathar (Karad), Kolhapur Maharashtra, India

Dr. Sudhakar D. Bhoite — Asso.Professor, Dept. Of M.Phil.Com. & Mgt. Chh. Shahu Institute Of Business Education And Research (SIBER), Kolhapur 416004 MS, India

ABSTRACT

Nowadays the databases are of huge size, so it is difficult to analyze those data. In spite of having ever growing data bases the problem with the banks is that, they fail to fully capitalize the true benefits which can be gained from this great wealth of information. The banking sector has started realizing the need of the techniques like data mining which can help them to compete in the market. This paper highlights the perspective applications of data mining tool WEKA to enhance the performance of some of the core business processes in banking sector. This paper will demonstrate the acceptance of e-banking facility by consumers in sangli district by using two data mining techniques in WEKA: •Classification, •Clustering (Simple K Means).

INTRODUCTION :

Banking sector plays an important role in the development of any nation by encouraging the willingness to save among people by offering them attractive saving and deposit schemes. With the entry of new private sector banks and continuous innovations taking place in the information technology, it has become a necessity for the banks in India to make increasing use of electronic mode for doing their operations (Vivek Bhambri, 2011) [1]. Therefore, the concept of E-banking has emerged in Indian banking sector and is gaining going grounds day by day. To succeed in competitive environment in marketplace, banks must offer a wide array of products with the latest technology. At present, many banks and financial institutions are actively developing new electronic banking products for their customers, throughout the world.

The enormous amount of data that banks have been collecting over the years can greatly influence the success of data mining efforts. By using data mining to analyze patterns and trends, bank executives can predict, with increased accuracy, how customers will react to adjustments in interest rates, which customers will be likely to accept new product offers, which customers will be at a higher risk for defaulting on a loan, and how to make customer relationships more profitable.

DATA MINING

Data mining is the process of discovering previously unknown and potentially interesting patterns in large datasets (Piatetsky-Shapiro and Frawley, 1991)[2]. It is a collective term for dozens of techniques to pick up information from data and turn it into meaningful trends and rules to improve your understanding of the data. Data mining functionalities are characterization and discrimination, mining frequent patterns, association, correlation, classification and prediction, cluster analysis, outlier analysis and evolution analysis (Han J. and Kamber M, 2000) [3]. Two of the major data mining techniques are classification and clustering. Classification is a data mining (machine learning) technique used to predict group membership for data instances. CLUSTERING is a data mining technique to group the similar data into a cluster and dissimilar data into different clusters. we are using WEKA data mining tool for classification and clustering to investigate the acceptance of e-banking facility by bank customers in sangli district with a sample size of 300 customers from public, private and co-operative banks.

WEKA :

WEKA is a landmark system in the history of the data mining and machine learning research communities, because it is the only toolkit that has gained such widespread adoption and survived for an extended period of time [4]. It is a collection of state-of-the-art machine learning algorithms and data preprocessing tools written in Java, developed at the University of Waikato, New Zealand. It is free software that runs on almost any platform and is available under the GNU General Public License. It has a wide range of applications in various data mining techniques. It provides extensive support for the entire process of experimental data mining, including preparing the input data, evaluating learning schemes statistically, and visualizing the input data and the result of learning. The WEKA workbench includes methods for the main data mining problems: regression, classification, clustering, association rule mining, and attribute selection. It can be used in either of the following two interfaces

- Command Line Interface (CLI)
- Graphical User Interface (GUI)

The WEKA GUI Chooser appears like this –

Fig. 1- WEKA GUI Chooser

The buttons from this Chooser can be used to start the following applications –

- Explorer – Environment for exploring data with WEKA. It gives access to all the facilities using menu selection and form filling.
- Experimenter – It can be used to get the answer for a question: Which methods and parameter values work best for the given problem?
- KnowledgeFlow – Same function as explorer. Supports incremental learning. It allows designing configurations for streamed data processing. Incremental algorithms can be used to process very large datasets.
- Simple CLI – It provides a simple Command Line Interface for directly executing WEKA commands.

This research paper will demonstrate "The scenario of e-banking facility acceptance by bank consumers in sangli district by fol-

lowing two data mining techniques using WEKA:

Classification 2) Clustering (Simple K Means)

By creating a classification tree (a decision tree), the data can be mined to determine the acceptance of e-banking by a person. Possible nodes on the tree would be age group, sex, education, occupation. The attributes of a person can be used against the decision tree to determine the likelihood of him accepting e-banking facility.

The data used in this paper is Bank Data available in Attribute Relation File Format (ARFF) format with following fields & 300 records.

age {18-19,20-39,40-59,60-99}
education {uneducated,upto-HSC,diploma,graduate,post-graduate}
sex {m,f}
occupation {student,farmer,service,business,professional,housewife}
e-banking {y,n}

The e-banking.arff is loaded into WEKA and will look like –

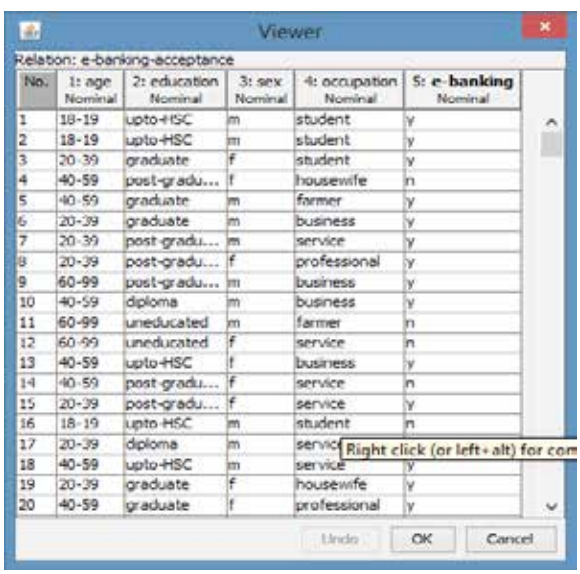

Fig.2 - Bank Data available in ARFF format.

The Instances distribution is visualized with e-banking attribute as class is follows-

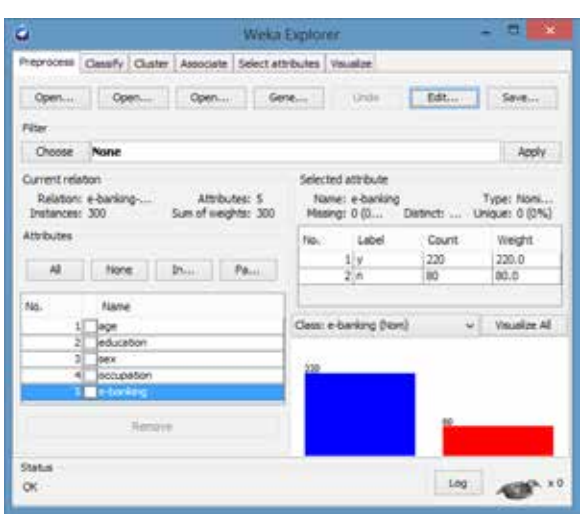

Fig. 3 - WEKA Explorer with e-banking class

CLASSIFICATION OF DATA:

Classification is used to classify each item in a set of data into one of predefined set of classes or groups. In tree classification j48 is an implementation of C4.5 release 8, a standard algorithm that is widely used for practical machine learning (Quinlan 1993)[5]. This implementation produces decision tree models. This algorithm works by forming pruned and unpruned partial decision trees.

The bank data for this article is classified with weka.classifiers. trees.J48 -C 0.25 -M 2 pruned method and got the classifier output as follows.

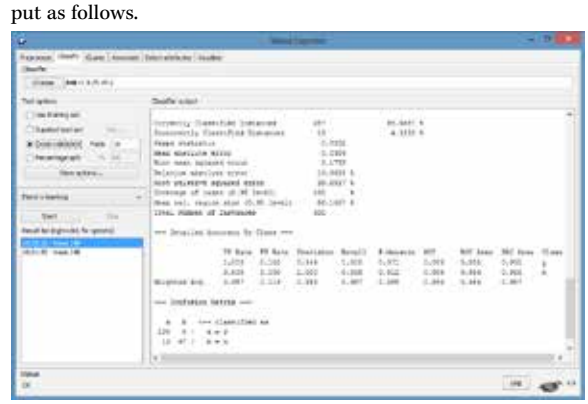

Fig. 4 - weka.classifiers.trees.J48 -C 0.25 -M 2 pruned method (cross-validation) Listing of Output from WEKA's classification model (above window)

J48 pruned tree

```
age = 18-19
|   sex = m: y (19.0/7.0)
|   sex = f: n (6.0)
age = 20-39: y (116.0)
age = 40-59
|   education = uneducated: n (6.0)
|   education = upto-HSC
|   |   occupation = student: y (0.0)
|   |   occupation = farmer
|   |   |   sex = m: y (6.0)
|   |   |   sex = f: n (6.0)
|   |   occupation = service: y (13.0)
|   |   occupation = business: y (7.0)
|   |   occupation = professional: y (0.0)
|   |   occupation = housewife: y (0.0)
|   education = diploma
|   |   occupation = student: n (0.0)
|   |   occupation = farmer: n (6.0)
|   |   occupation = service: n (0.0)
|   |   occupation = business: y (6.0)
|   |   occupation = professional: n (0.0)
|   |   occupation = housewife: n (6.0)
|   education = graduate
|   |   occupation = student: y (0.0)
|   |   occupation = farmer: y (18.0)
|   |   occupation = service: y (18.0/6.0)
|   |   occupation = business: n (6.0)
|   |   occupation = professional: y (6.0)
|   |   occupation = housewife: y (0.0)
|   education = post-graduate: n (13.0)
age = 60-99
|   education = uneducated: n (18.0)
|   education = upto-HSC: y (0.0)
|   education = diploma: y (0.0)
|   education = graduate: y (12.0)
|   education = post-graduate: y (12.0)
```

Number of Leaves : 29
Size of the tree : 37
Time taken to build model: 0.03 seconds
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	287	95.6667 %
Incorrectly Classified Instances	13	4.3333 %
Kappa statistic	0.8832	
Mean absolute error	0.0586	
Root mean squared error	0.1755	
Relative absolute error	14.9615 %	
Root relative squared error	39.6817 %	
Coverage of cases (0.95 level)	100 %	
Mean rel. region size (0.95 level)	56.1667 %	
Total Number of Instances	300	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
1.000	0.163	0.944	1.000	0.971	0.889	0.986	0.995	y
0.838	0.000	1.000	0.838	0.912	0.889	0.986	0.965	n
Weighted Avg.	0.957	0.119	0.959	0.957	0.955	0.889	0.986	0.987

=== Confusion Matrix ===

a	b	<-- classified as
220	0	a = y
13	67	b = n

What do these numbers mean-

Time taken to build model: **0.03 seconds**

Correctly Classified Instances	287	95.6667 %
Incorrectly Classified Instances	13	4.3333 %

Confusion matrix is another important aspect to be considered, from this matrix predictions can be made.

Confusion Matrix

a	b	<-- classified as
220	0	a = y
13	67	b = n

Above 220 instances are correctly classified for the class value Yes. And below 67 instances are correctly classified instances for the class value No. Remaining elements represents the incorrectly classified instances Yes and No respectively.

Based on our accuracy rate of 95.6667 %, we can say that this is a pretty good model to analyze consumer acceptance of e-banking facility.

A **tree Visualization** can be seen on the model just created as follows -

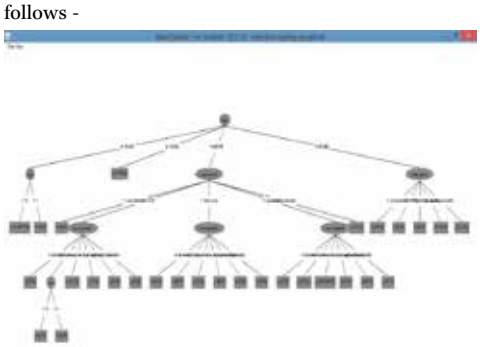

Figure.5- Classification Tree Visualization

CLUSTERING :

Clustering is a data mining technique that makes meaningful or useful cluster of objects that have similar characteristic using automatic technique. Clustering is also called as data segmentation. Since in some applications because clustering partitions large datasets into groups according to similarity. Clustering can also be used as outlier detection, where outliers may be more interesting than common cases. Many clustering algorithms exist in WEKA. The performance of K-Means algorithm produces quality clusters when using huge dataset and is better than Hierarchical Clustering algorithm(Bharat Chaudhari, Manan Parikh,2012)[6].

Considering this, we have used K-Means algorithm to quickly determine patterns in the data.

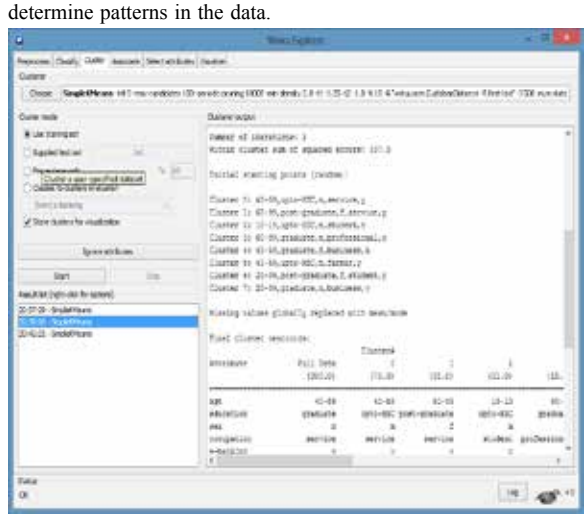

Fig. 6 Clustering by Simple K means algorithm (8 clusters)

Listing of Cluster Output with 8 clusters :

Scheme: weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 8 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10

Relation: e-banking-acceptance
Instances: 300
Attributes: 5
age
education
sex
occupation
e-banking
Test mode: evaluate on training data
Clustering model (full training set) ===kMeans

Number of iterations: 3
Within cluster sum of squared errors: 337.0
Initial starting points (random):
Cluster 0: 40-59,upto-HSC,m,service,y
Cluster 1: 60-99,post-graduate,f,service,y
Cluster 2: 18-19,upto-HSC,m,student,n
Cluster 3: 60-99,graduate,m,professional,y
Cluster 4: 40-59,uneducated,f,housewife,n
Cluster 5: 40-59,graduate,m,farmer,y
Cluster 6: 20-39,post-graduate,f,student,y
Cluster 7: 20-39,graduate,m,business,y
Missing values globally replaced with mean/mode

Final cluster centroids:

Cluster#

Attribute	Full Data (300.0)	0 (75.0)	1 (38.0)	2 (31.0)	3 (18.0)	4 (36.0)	5 (30.0)	6 (54.0)	7 (18.0)
age	40-59	40-59	60-99	18-19	60-99	40-59	40-59	20-39	20-39
education	graduate	upto-HSC	post-graduate	upto-HSC	graduate	uneducated	graduate	graduate	graduate
sex	m	m	f	m	m	f	m	f	m
occupation	service	service	service	student	professional	housewife	farmer	student	business
e-banking	y	y	y	n	y	n	y	y	y

Time taken to build model (full training data) : 0.03 seconds
Model and evaluation on training set Clustered Instances

0	75 (25%)
1	38 (13%)
2	31 (10%)
3	18 (6%)
4	36 (12%)
5	30 (10%)
6	54 (18%)
7	18 (6%)

Description of Clusters-

Cluster 0 - This group represents to males from age group 40-59 with education upto-HSC and are employed somewhere. They are using e-banking service.

Cluster 1 - This group represents to post-graduate, employed females from age group 60-99 and using e-banking service.

Cluster 2 - This group represents to male students having education upto hsc from age group 18-19 and not using e-banking service.

Cluster 3 - This group represents to graduate male professionals from age group 60-99 using e-banking service.

Cluster 4 - This group represents to uneducated housewives from age group 40-59 and not using e-banking service.

Cluster 5 - This group represents to graduate male farmers from age group 40-59 and using e-banking service.

Cluster 6 - This group represents to post-graduate female students from age group 20-39 and using e-banking service.

Cluster 7 - This group represents to graduate businessmen from age group 20-39 and not using e-banking service.

One thing that is clear from the clusters is that behavior of female are clustered in only 3 groups while males behavior are heavily distributed among 5 clusters,

Another interesting way to examine the data in these clusters is to inspect it **visually**. In this example, X axis is set to occupation (Nom), Y axis to education(Nom), and the Color to Cluster (Nom). This showed us in a chart how the clusters are grouped in terms of occupation and education.

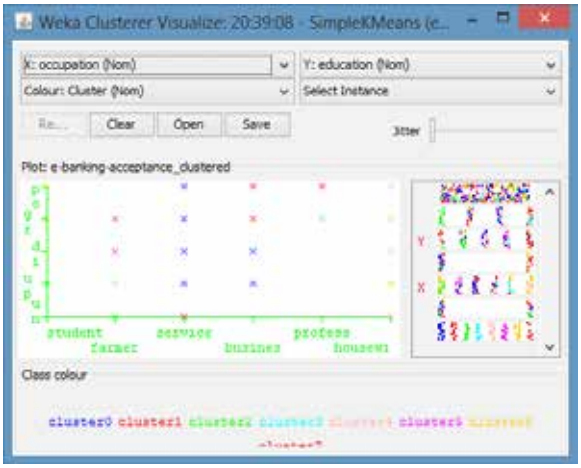

Fig. 6 – WEKA cluster visualization

CONCLUSION :

As illustrated by the study presented here, information 'mined' from data can provide insights into the domain being studied to receive wisdom of a field. Locating these surprising or unusual portions of the model can be the focus for a data mining analysis, so that the results can be applied back in the domain from which the data was drawn. In this case, the results indicate that the subjective attributes age, occupation, sex and education influence the class of the study. Among all the attribute sex and age are the two major attribute that heavily influence the tendency of a person to use e-banking services. Finally, it proves that WEKA is a significant step in the transfer of machine learning technology into the workplace.

REFERENCE

[1] Vivek Bhambri, Dept. of CS, DBIMCS, Mandi Gobindgarh, Punjab, India "Application of Data Mining in Banking Sector" International Journal of Computer Science and Technology ISSN : 2229-4333(Print) | ISSN:0976-8491(Online) www.ijcst.com IJCST Vol. 2, Issue 2, June 2011 | [2] Piatetsky-Shapiro, G., and Frawley, W.J., eds. (1991) Knowledge Discovery in Databases. Menlo Park, CA, AAAI Press. | [3] Han J. and Kamber M.: "Data Mining: Concepts and Techniques," Morgan Kaufmann Publishers, San Francisco, 2000. | [4] Sapna Jain, M Afshar Aalam and M N Doja, "K-means clustering using weka interface", Proceedings of the 4th National Conference; INDIACom-2010. | [5] Quinlan, J.R. (1993) C4.5: Programs for machine learning. Morgan Kaufmann, San Mateo, CA. | [6] Bharat Chaudhari, Manan Parikh, "A Comparative Study of clustering algorithms Using weka tools", Web Site: www.ijaiem.org Email: editor@ijaiem.org, editorijaiem@gmail.com Volume 1, Issue 2, October 2012 ISSN 2319 - 4847 Page 158. | Agrawal, R., Imielinski, T. and Swami, A.N. (1993) "Database mining: a performance perspective." IEEE Transactions on Knowledge and Data Engineering, Vol. 5, 914–925