IOSR Journal of Computer Engineering (IOSRJCE)
ISSN: 2278-0661, ISBN: 2278-8727 Volume 6, Issue 3 (Sep-Oct. 2012), PP 32-41
www.iosrjournals.org

"Performance analysis of Data Mining algorithms in Weka"

[1]Mahendra Tiwari, [2] Manu Bhai Jha, [3] OmPrakash Yadav
[1]Research Scholar(UPRTOU, Allahabad)
[2.3]Assistant Professor(UCER,Allahabad)

Abstract: *The retail industry collects vast amounts of data on sales, customer buying history, goods, and service with ease of use of modern computing technology. This paper elaborates the use of data mining technique to help retailers to identify customer profile for a retail store and behaviors, improve better customer satisfaction and retention. The aim is to judge the accuracy of different data mining algorithms on various data sets. The performance analysis depends on many factors encompassing test mode, different nature of data sets, and size of data set.*

Keywords*-data mining, performance, analysis, retail*

I. Introduction:

The data sizes accumulated from various fields are exponentially increasing, data mining techniques that extract information from huge amount of data have become popular in commercial and scientific domains, including marketing, customer relationship management. During the evaluation, the input datasets and the number of classifiers used are varied to measure the performance of Data Mining algorithm. I present the results based on characteristics such as scalability, accuracy to identify their characteristics in a world famous Data Mining tool-WEKA.

II. Related Work:

We studied various journals and articles regarding performance evaluation of Data Mining algorithms on various different tools, some of them are described here, Ying Liu et all worked on Classification algorithms while Osama abu abbas worked on clustering algorithm, and Abdullah compared various classifiers with different types of data set on WEKA, we presented their result as well as about tool and data set which are used in performing evaluation.

Ying Liu,wei-keng Liao et al [39] in his article "performance evaluation and characterization of scalable data mining algorithms" investigated data mining applications to identify their characteristics in a sequential as well as parallel execution environment .They first establish Mine bench, a benchmarking suite containing data mining applications. The selection principle is to include categories & applications that are commonly used in industry and are likely to be used in the future, thereby achieving a realistic representation of the existing applications. Minebench can be used by both programmers & processor designers for efficient system design. They conduct their evaluation on an Intel IA-32 multiprocessor platform, which consist of an Intel Xeon 8-way shared memory parallel(SMP) machine running Linux OS, a 4 GB shared memory & 1024 KB L2 cache for each processor. Each processor has 16 KB non-blocking integrated L1 instructions and data caches. The number of processors is varied to study the scalability.

In all the experiments, they use VTune performance analyzer for profiling the functions within their applications, & for measuring their breakdown execution times. VTune counter monitor provides a wide assortment of metrics. They look at different characteristics of the applications: execution time, fraction of time spent in the OS space, communication/synchronization complexity , & I/O complexity. The Data comprising 250,000 records. This notion denotes the dataset contains 2,00,000 transactions,the average transaction size is 20, and the average size of the maximal potentially large itemset is 6. The number of items is 1000 and the number of maximal potentially large itemset is 2000.

The algorithms for comparison are ScalParc, Bayesian, K-means, Fuzzy K-means, BIRCH,HOP,Apriori, & ECLAT.

Osama Abu Abbas [38] in his article "comparison between data clustering algorithms" compared four different clustering algorithms (K-means, hierarchical, SOM, EM) according to the size of the dataset, number of the clusters ,type of S/W. The general reasons for selecting these 4 algorithms are:

- Popularity
- Flexibility
- Applicability
- Handling High dimensionality

Osama tested all the algorithms in LNKnet S/W- it is public domain S/W made available from MIT Lincoln lab www.li.mit.edu/ist/lnknet.
For analyzing data from different data set, located at
www.rana.lbl.gov/Eisensoftware.htm
The dataset that is used to test the clustering algorithms and compare among them is obtained from the site www.kdnuggets.com/dataset .This dataset is stored in an ASCII file 600 rows,60 columns with a single chart per line
1-100 normal
101-200 cyclic
201-300 increasing trend
301-400 decreasing trend
401-500 upward shift
501-600 downward shift

Abdullah et al [41] in his article "A comparison study between data mining tools over some classification methods" conducted a comparison study between a number of open source data mining S/W and tools depending on their ability for classifying data correctly and accurately.
The methodology of the study constitute of collecting a set of free data mining & knowledge discovery tools to be tested, specify the datasets to be used, and selecting a set of classification algorithm to test the tool's performance.

For testing, each dataset is described by the data type being used, the types of attributes, whether they are categorical ,real, or integer, the number of instances stored within the data set, the number of attributes that describes each dataset, and the year the data set was created. After selecting the dataset , a number of classification algorithm are chosen that are Naïve Bayes, K-nearest, SVM,C4.5 as well as some classifiers are used that are Zero R, One R, & Decision Tree classifier.
For evaluating purpose two test level modes were used; the K-fold cross validation mode and the percentage split mode.

After running the four tools ,they have obtained some results regarding the ability to run the selected algorithm on the selected tools. All algorithms ran successfully on WEKA, the 6 selected classifiers used the 9 selected data sets.

T. velmurgun [27] in his research paper "performance evaluation of K-means & Fuzzy C-means clustering algorithm for statistical distribution of input data points" studied the performance of K-means & Fuzzy C-means algorithms. These two algorithm are implemented and the performance is analyzed based on their clustering result quality. The behavior of both the algorithms depended on the number of data points as well as on the number of clusters. The input data points are generated by two ways, one by using normal distribution and another by applying uniform distribution (by Box-muller formula). The performance of the algorithm was investigated during different execution of the program on the input data points. The execution time for each algorithm was also analyzed and the results were compared with one another, both unsupervised clustering methods were examined to analyze based on the distance between the various input data points. The clusters were formed according to the distance between data points and clusters centers were formed for each cluster.
The implementation plan would be in two parts, one in normal distribution and other in uniform distribution of input data points. The data points in each cluster were displayed by different colors and the execution time was calculated in milliseconds.

Velmurugan and Santhanam chose 10 (k=10) clusters and 500 data points for experiment. The algorithm was repeated 500 times (for one data point one iteration) to get efficient output. The cluster centers (centroid) were calculated for each clusters by its mean value and clusters were formed depending upon the distance between data points

Jayaprakash et al [37] in their paper "performance characterization of Data Mining applications using Minebench" presented a set of representative data mining applications call Minebench. They evaluated the Minebench application on an 8 way shared memory machine and analyze some important performance characteristics. Minebench encompasses many algorithms commonly formed in data mining. They analyzed the architectural properties of these applications to investigate the performance bottleneck associated with them.
For performance characterization, they chose an Intel IA-32 multiprocessor platform, Intel Xeon 8-way shared memory parallel (SMP) machine running Red Hat advanced server 2.1. The system had 4 GB of shared memory. Each processor had a 16 KB non-blocking integrated L1 cache and a 1024 KB L2 cache. For evaluation they used VTune performance analyzer. Each application was compiled with version 7.1 of the Intel C++ compiler for Linux.

The data used in experiment were either real-world data obtained from various fields or widely accepted synthetic data generated using existing tools that are used in scientific and statistical simulations. During evaluation, multiple data sizes were used to investigate the characteristics of the Minebench applications, For non-bioinformatics applications, the input datasets were classified in to 3 different sizes: small, medium, & large. IBM Quest data generator, ENZO, & real image database by Corel corporation.

Pramod s. and O.P.vyas [24] in their research paper "performance evaluation of some online association rule mining algorithms for sorted & unsorted data sets" evaluated association rule mining algorithm for sorted and unsorted data sets. They worked on Continuous Association Rule Mining Algorithm (CARMA) and Data Stream Combinatorial approximation Algorithm (DSCA) , & estDec method.

The 3 algorithms are implemented in JAVA and the results were plotted, all 3 algorithms were tested with 5 data sets and all of them are available in Frequent Itemset Mining data set (FIM) repository. The transactions of each data set were looked up one by one in sequence to simulate the environment of an online data stream. The DSCA algorithm used sorted transaction items while other 2 algorithms used unsorted transaction items.

P.T. Kavitha and Dr. T. .Sasipraba [30] in their research paper " performance evaluation of algorithms using a distributed data mining frame work based on association rule mining" evaluated the performance of distributed data mining framework on java platform. The aim of framework was to develop an efficient association rule mining tool to support effective decision making. Association rule mining focuses on finding interesting patterns from huge amount of data available in the data warehouses. They used Apriori, AprioriTID, FP growth, & Apriori Hyprid algorithm.

They propose a Java based DDM framework a totally decentralized framework for distributed data mining using association rules as the backbone of the system. This system was completely platform independent including the database support. The use of client-server architecture enabled them to perform distributed data mining ,They define access rights to this framework by classifying users in to groups. They suggested to add or remove algorithms at any client side dynamically . The benchmarking module evaluated performance between algorithms.

Thus the complete platform independency to be achieved using object oriented programming.

The experiment was implemented in java, Pentium 4 processor with the speed of 1.86 GHz and the paradox (.db) format was used for database setting.

Table 1 : Summary of selected references with goals

Reference	Goal	Database/Data description	Data size used	Preprocessing	Data Mining algorithm	Software
Abullah H. wabheh et all. (IJACSA)	Comparative study between a number of free available data mining tools	UCI repository	100 to 20,000 instances	Data integration	NB,OneR, C4.5,SVM ,KNN,ZeroR	Weka ,KNIME ,Orange ,TANAGRA
Ying Liu et all	To investigate data mining applications to identify their characteristic in a sequential as well as parallel execution environment	IBM Quest data generator,ENZO	250,000 records,2, 000,000 transactions		HOP, K-means ,BIRCH, ScalParc Bayesian ,Apriori Eclat	V Tune Performance analyzer
P.T. Kavitha et all (IJCSE)	To develop efficient ARM on DDM framework	Transaction data by Point-of-Sale(PoS) system			Apriori ,AprioriTID ,AprioriHyprid, FP growth	Java
T.velmurugan & T.Santhanam	To analyze K-means &	Normal & uniform	500 to 1000 data		K-means, Fuzzy C-	Applet Viewer

(EJOSR)	Fuzzy C-means clustering result quality by Box-muller formula	distribution of data points	points		means	
Jayaprakash et all	To evaluate MineBench applications on an 8-way shared memory machine	IBM Quest data generator,ENZO , Synthetic data set	Dense database, 1000k to 8000k transcations,73MB real data set	Data cleaning	Scalparc,K-means,HOP, Apriori,Utility,SNP, Genenet, SEMPHY, Research, SVM,PLSA	V tune performance analyzer
Pramod S. & O.P.vyas	To assess the changing behavior of customers through ARM	Frequent Itemset Mining(FIM) data set repository	Sorted & unsorted transaction set	Data cleaning	CARMA, DSCA,estDec	java
Osama abu Abbas	To compare 4 clustering algorithm	www.kdnuggets. com	ASCII file 600 rows 60 columns		K-means, Hierarchical ,SOM,EM	LNKnet

As the number of available tools continues to grow, the choice of one special tool becomes increasingly difficult for each potential user. This decision making process can be supported by performance evaluation of various classifiers and clusterers used in open source data mining tool –Weka.

III. Analysis of Data Mining algorithm:-

- Classification Programs-

A classification algorithm is to use a training data set to build a model such that the model can be used to assign unclassified records in to one of the defined classes. A test set is used to determine the accuracy of the model. Usually ,the given dataset is divided in to training and test sets, with training set used to build the model and test set used to validate it.

There are various classifiers are an efficient and scalable variation of decision tree classification. The decision tree model is built by recursively splitting the training dataset based on an optimal criteria until all records belonging to each of the partitions bear the same class label. Among many trees are particularly suited for data mining , since they are built relatively fast compared to other methods, obtaining similar or often better accuracy.

Bayesian classifiers are statistical based on Bayes' theorem, they predict the probability that a record belongs to a particular class. A simple Bayesian classifier, called Naïve Bayesian classifier is comparable in performance to decision tree and exhibits high accuracy and speed when applied to large databases.

IBK, and KStar of Lazy learners, OneR and ZeroR of Rule, SMO of function are also used in evaluation process.

K-nearest neighbor classifiers are based on learning by analogy. The training samples are described by n dimensional numeric attributes. Each sample represents a point in an n-dimensional space. In this way, all of the training samples are stored in an n-dimensional pattern space. When given an unknown sample, a k-nearest neighbor classifier searches the pattern space for the training samples that are closest to the unknown sample.

- Clustering Program-

Clustering is the process of discovering the groups of similar objects from a database to characterize the underlying data distribution. K-means is a partition based method and arguably the most commonly used clustering technique. K-means clusterer assigns each object to its nearest cluster center based on some similarity function. Once the assignment are completed , new centers are found by the mean of all the objects in each cluster.

BIRCH is a hierarchical clustering method that employs a hierarchical tree to represent the closeness of data objects. BIRCH first scans the database to build a clustering-feature tree to summarize the cluster representation. Density based methods grow clusters according to some other density function. DBscan , originally proposed in astrophysics is a typical density based clustering method.

After assigning an estimation of its density for each particle with its densest neighbors, the assignment process continues until the densest neighbor of a particle is itself. All particles reaching this state are clustered as a group.

3.1 Evaluation Strategy/Methodology:-

- H/W tools:

We conduct our evaluation on Pentium 4 Processor platform which consist of 512 MB memory, Linux enterprise server operating system, a 40GB memory, & 1024kbL1 cache.

- S/W tool:

In all the experiments, We used Weka 3-6-6 we looked at different characteristics of the applications-using classifiers to measure the accuracy in different data sets, using clusterer to generate number of clusters, time taken to build models etc.

Weka toolkit is a widely used toolkit for machine learning and data mining that was originally developed at the university of Waikato in New Zealand . It contains large collection of state-of-the-art machine learning and data mining algorithms written in Java. Weka contains tools for regression, classification, clustering, association rules, visualization, and data processing.

- Input data sets:-

Input data is an integral part of data mining applications. The data used in my experiment is either real-world data obtained from UCI data repository and widely accepted dataset available in Weka toolkit, during evaluation multiple data sizes were used, each dataset is described by the data type being used, the types of attributes, the number of instances stored within the dataset, also the table demonstrates that all the selected data sets are used for the classification and clustering task. These datasets were chosen because they have different characteristics and have addressed different areas.

Zoo dataset in csv format whereas labor ,and Supermarket dataset are in arff format. Zoo, & Labor dataset have 17 number of attributes while Supermarket dataset has 200 attributes. Zoo dataset encompasses 101 instances, Labor comprises 57 instances , & Supermarket has 4627 instances. All datasets are categorical and integer with multivariate characteristics.

- Details of data Set:

We used 3 data set for evaluation with classifier on WEKA ,one of them from UCI Data repository that are Zoo data set, rest labor data set and supermarket data set is inbuilt in WEKA 3-6-6 .Zoo data set in csv file format ,and labor and supermarket data set are in arff file format.

Table 2: Detail of data set:

Name of Data Set	**Type of file**	**Number of attributes**	**Number of instances**	**Attribute characteristics**	**Dataset characteristics**	**Missing value**
Zoo	CSV(comma separated value)	17	101	Categorical,Integer	Multivariate	No
Labor	ARFF(Attribute Relation File Format)	17	57	Categorical,Integer	Multivariate	No
Supermarket	ARFF(Attribute Relation File Format)	217	4627	Categorical,Integer	Multivariate	No

- Experimental result and Discussion:-

To evaluate the selected tool using the given datasets, several experiments are conducted. For evaluation purpose, two test modes are used, the k-fold cross-validation(k-fold cv) mode, & percentage split(holdout method) mode. The k-fold cv refers to a widely used experimental testing procedure where the database is randomly divided in to k disjoint blocks of objects, then the data mining algorithm is trained using k-1 blocks and the remaining block is used to test the performance of the algorithm, this process is repeated k times. At the end, the recorded measures are averaged. It is common to choose k=10 or any other size depending mainly on the size of the original dataset.

In percentage split (holdout method) ,the database is randomly split in to two disjoint datasets. The first set, which the data mining system tries to extract knowledge from called training set. The extracted knowledge may be tested against the second set which is called test set, it is common to randomly split a data set under the mining task in to 2 parts. It is common to have 66% of the objects of the original database as a training set and the rest of objects as a test set. Once the tests is carried out using the selected datasets, then using the available classification and test modes ,results are collected and an overall comparison is conducted.

- Performance Measures:-

For each characteristics, we analyzed how the results vary whenever test mode is changed. Our measure of interest includes the analysis of classifiers and clusterers on different datasets, the results are described in value of correlation coefficient, mean absolute error, root mean squared error, relative absolute error, root relative squared error after applying the cross-validation or holdout method.

For performance issues, after applying the decision stump and REP tree classifiers on Zoo dataset with cross-validation method ,the correlation coefficient values are 0.8231 and 0.3066, the mean absolute error are 0.9518 and 1.8346 respectively. When holdout (percentage split) method is applied the correlation coefficient value is 0 for REP tree. The decision stump tree did not work with percentage split. The mean absolute error is 2.0076 for REP tree.

On same dataset(Zoo) the IBK and KStar of Lazy learner classifier are applied with cross-validation method the correlation coefficient values are 0.9966 and 0.981. With percentage split method the value of correlation coefficient are little changed, the values are 0.9942 for IBK and 0.9612 for KStar. The mean absolute error value are 0.0297 and 0.1036 with cross-validation method, and 0.0606 and 0.1946 with percentage split .
ZeroR classifier of Rules is applied on Zoo dataset the correlation coefficient value is same for both test mode ie 0,while mean absolute error is 1.8165 for cross-validation and 2.0076 for holdout method.
When linear regression of function is applies on Zoo dataset the correlation coefficient of cross-validation is 0.5958 and 0.5116 for holdout method. The mean absolute error is 1.443 for cross-validation and 1.6402 for holdout method.
There are 2 other datasets which I used for measurement they are labor, & Supermarket dataset. The details of applied classifiers on those datasets are as following:

1. Dataset: Labor

Classifier: Lazy-IBK,KStar, Tree-Decision stump, REP, Function- Linear regression, Rule-ZeroR, Bayesian-Naïve Bayes

2. Dataset: Supermarket

Classifier: Lazy-IBK,KStar, Tree-Decision stump, CART, Function- SMO, Rule-ZeroR, OneR, Bayesion-Naïve Bayes.

In all the experiments with classifiers ,results are in the form of statistical analysis along with correctly or incorrectly instances classified by classifiers. The classifier model for all the dataset is ‘Full training set’ and two test mode are used cross-validation and holdout method.

In performance characterization, this research work also deals some most delegated clustering algorithms. The performance of the algorithm is investigated during different test mode on the input data. The test mode in evaluation are Full training data and Percentage split. The results are in the form of Number of generated clusterer, time taken to build the models, and unclustered data.

The details of clusterer with different dataset are as following

1. Dataset: Zoo

Clusterer: DBscan, EM, Hierarchical, K-means

2. Dataset: Labor

Clusterer: DBscan, EM, Hierarchical, K-means

3. Dataset: Supermarket

Clusterer: DBscan, EM,, K-means

IV. Evaluation of Classifiers on different dataset:

I tried to evaluate the performance of various classifiers on two test mode 10 fold cross validation and percentage split with different data sets at WEKA 3-6-6, The results after evaluation is described here:-

4.1 Evaluation on Zoo data set:-

Table 3: Evaluation of classifiers on Zoo data set with cross validation test mode:

Classifier	Classifier model	Test mode	Correlation coefficient	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error
Lazy-IBK	Full training set	10 Fold cross-validation	0.9966	0.0297	0.1723	1.619 %	8.1487 %
Lazy-KStar	Full training set	10 Fold cross-validation	0.981	0.1036	0.4067	5.64 7%	19.231 %
Function-Linear regression	Full training set	10 Fold cross-validation	0.5958	1.4443	1.8315	78.723 %	86.593 %
Rules-ZeroR	Full training set	10 Fold cross-validation	0	1.8165	2.0923	100 %	100 %
Tree-REP	Full training set	10 Fold cross-validation	0.3066	1.8346	2.115	100 %	100 %
Tree-Decisionstump	Full training set	10 Fold cross-validation	0.8231	0.9518	1.1883	51.881 %	56.185 %

Table 4 :Evaluation of classifiers on Zoo data set with percentage split test mode:

Classifier	Classifier Model	Test mode	Correlation Coefficient	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error
Lazy-IBK	Full training set	Percentage split	0.9942	0.0606	0.2462	3.018 %	10.719 %
Lazy-KStar	Full training set	Percentage split	0.9612	0.1946	0.6309	9.691 %	27.471 %
Function-Linear regression	Full training set	Percentage split	0.5116	1.6402	2.0806	81.702 %	90.593 %
Rules-ZeroR	Full training set	Percentage split	0	2.0076	2.2967	100 %	100 %
Tree-REP	Full training set	Percentage split	0	2.0076	2.2967	100 %	100

4.2 Evaluation on Labor data set:-

Table 5: Evaluation of classifiers on labor data set with cross validation test mode:

Classifie	Classifier model	Test mode	Correctly classified instances	Incorrectly classified instances	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error
Lazy-IBK	Full training set	Cross-validation	47/57 (82.45%)	10/57 (17.54%)	0.1876	0.4113	41.014%	86.148%
Lazy-KStar	Full training set	Cross-validation	51/57 (89.47%)	6/57 (10.52%)	0.0948	0.2742	20.718%	57.432%
Naïve Bayes	Full training set	Cross-validation	51/57 (89.47%)	6/57 (10.52%)	0.1042	0.2637	22.776%	55.226%
Rules-OneR	Full training set	Cross-validation	43/57 (75.43%)	14/57 (24.56%)	0.2456	0.4956	53.692%	103.796%

Rules-ZeroR	Full training set	Cross-validation	37/57 (64.91%)	20/57 (35.08%)	0.4574	0.4775	100%	100%
Function-SMO	Full training set	Cross-validation	51/57 (89.47%)	6/57 (10.52%)	0.1053	0.3244	23.011%	67.950%
Tree-CART	Full training set	Cross-validation	45/57 (78.94%)	12/57 (21.05%)	0.2709	0.4292	59.230%	89.896%
Tree-Decision stump	Full training set	Cross-validation	46/57 (80.70%)	11/57 (19.29%)	0.2102	0.3358	45.959%	70.334%

Table 6 : Evaluation of classifiers on labor data set with Percentage split test mode:

Classifier	**Classifier model**	**Test mode**	**Correctly classified instances**	**Incorrectly classified instances**	**Mean absolute error**	**Root mean squared error**	**Relative absolute error**	**Root relative squared error**
Function-SMO	Full training set	Percentage Split	17 (89.47%)	2 (10.52%)	0.1053	0.3244	23.188	69.238
Lazy-KStar	Full training set	Percentage Split	16 (84.21%)	3 (15.78%)	0.154	0.3858	33.931	82.336
Lazy-IBK	Full training set	Percentage Split	15 (78.94%)	4 (21.05%)	0.225	0.4479	49.565	95.588
Rules-OneR	Full training set	Percentage Split	16 (84.21%)	3 (15.78%)	0.1579	0.3974	34.782	84.799
Rules-ZeroR	Full training set	Percentage Split	13 (68.42%)	6 (31.57%)	0.4539	0.4686	100	100
Tree-CART	Full training set	Percentage Split	17 (89.47%)	2 (10.52%)	0.1432	0.272	31.542	58.050
Tree-Decisionstump	Full training set	Percentage Split	16 (84.21%)	3 (15.78%)	0.1997	0.2903	43.991	61.952

4.2 Evaluation on supermarket data set:-

Table 7: Evaluation of classifiers on supermarket data set with cross validation test mode:

Classifier	**Classifier model**	**Test mode**	**Correctly classified instances**	**Incorrectly classified instances**	**Mean absolute error**	**Root mean squared error**	**Relative absolute error**	**Root relative squared error**
Function-SMO	Full training set	Cross-validation	2948 (63.71%)	1679 (36.28%)	0.3629	0.6024	78.473%	125.281%
NaiveBayes	Full training set	Cross-validation	2948 (63.71%)	1679 (36.28%)	0.4624	0.4808	100%	100%
Ruless-ZeroR	Full training set	Cross-validation	2948 (63.71%)	1679 (36.28%)	0.4624	0.4808	100%	100%
Rules-OneR	Full training set	Cross-validation	3110 (67.21%)	1517 (32.78%)	0.3279	0.5726	70.902%	119.084%
Lazy-IBK	Full training set	Cross-validation	1718 (37.12%)	2909 (62.87%)	0.6218	0.7806	134.473%	162.335%
Trees-CART	Full training set	Cross-validation	2948 (63.71%)	1679 (36.28%)	0.4624	0.4808	99.996%	100%
Trees-Decisionstump	Full training set	Cross-validation	2980 (64.40%)	1647 (35.59%)	0.4212	0.4603	91.079%	95.734%

V. Evaluation of clusterer on different data set:-

5.1 Evaluation of clusterer on Zoo data set:

Table 8 : Evaluation of clusterer on Zoo data set with Full training data test mode

Clustering Algorithm	No. of Instances	Test mode	No. of cluster generated	Clustered instances	Time taken to build the model	Unclustered
DBscan	108	Full training data	1	6(100%)	0.04 second	102
EM	108	Full training data	6(8,12,13,22, 20,33)	6(7%,11%,13 %,12%,20%, 19%,31%)	3.54 second	0
Hierarchical	108	Full training data	1	108(100%)	0.03 second	0
k-means	108	Full training data	2(40,68)	2(37%,63%)	0.01 second	0

5.2 Evaluation of clusterer on Labor data set:-

Table 9: Evaluation of clusterer on Labor data set with Percenatge split test mode

Clustering Algorithm	No. of Instances	Test mode	No. of cluster generated	Clustered instances	Time taken to build the model	Unclustered instances
DBscan	57	Percenatge split	0	0	0	20
EM	57	Percenatge split	3(4,12,4)	3(20%,60%,20 %)	0.54 second	0
Hierarchical	57	Percenatge split	2(0,20)	2(100%)	0	0

5.3Evaluation of clusterer on supermarket data set:-

Table 10: Evaluation of clusterer on supermarket data set with Percenatge split test mode

Clustering Algorithm	Instances	No. of cluster generated	Clustered instances	Unclustered instances	Test mode	Time taken to build model
DBscan	4627	2(1007,567)	2(64%,36%)	0	Percentage split	0.23 second
EM	4627	2(0,1574)	2(100%)	0	Percentage split	102.29 second
K-means	4627	2(987,587)	2(63%,37%)	0	Percentage split	0.61 second

VI. Conclusion:

Data Mining has a large family composed of different algorithms, and the scope of research is rapidly increasing to improve the accuracy of existed algorithms. In this paper, we evaluate some Data Mining algorithms contains 8 representative applications: four classification algorithms, and four clustering algorithms,. We analyzed important characteristics of the applications when executed in well known tool WEKA. The work described in this paper comparatively evaluates the performance of algorithms on three test modes that is hold out method , percentage split, and full training.

Our current work is focusing on evaluating the applications on different data sets to allow the retailers to increase customer understanding and make knowledge- driven decisions in order to provide personalized and efficient customer service.

References:

[1] www.boirefillergroup.com/....KDD_CONFERENCE_PAPER_AUG2006.pdf

[2] www.dcc.fc.up.pt/~ricroc/aulas/0708/atdmlp/material/paper_dmbiz06.pdf

[3] www.ecmlpkdd2006.org/ws-pdmaec.pdf

[4] http://www.linkedin.com/in/federicocesconi

[5] www.linkedin.com/in/federicocesconi

[6] C. Ling and C. Li, (1998) “Data mining for direct marketing: Problem and solutions,” in Proc, of the 4th international Conference on Knowledge Discovery & Data Mining, pp. 73-79

[7] John, F., Elder iv and Dean W.(1998) A comparison of leading data mining tools, fourth International conference on Knowledge discovery and data mining pp.1-31

[8] Michel A., et all (1998), Evaluation of fourteen desktop data mining tools , pp 1-6

[9] Kleissner, C.(1998),, data mining for the enterprise, Proceeding of the 31st annual Hawaii International conference on system science

[10] Brijs, T., Swinnen, G.,(1999), using association rules for product assortment decisions: A case study., Data Mining and knowledge discovery 254.

[11] Goebel M., L. Grvenwald(1999), A survey of data mining & knowledge discovery software tools, SIGKDD,vol 1, issue 1

[12] Rabinovitch, L. (1999),America’s first department store mines customer data. Direct marketing (62).

[13] Grossman, R., S. Kasif(1999), Data mining research: opportunities and challenges. A report of three NSF workshops on mining large, massive and distributed data, pp 1-11.

[14] A. Kusiak, (2002) Data Mining and Decision making, in B.V. Dasarathy (Ed.). Proceedings of the SPIE Conference on Data Mining and Knowledge Discovery: Theory, Tools and Technology TV, ol. 4730, SPIE, Orlando, FL, pp. 155-165.

[15] Giraud, C., Povel, O.,(2003), characterizing data mining software, Intell Data anal 7:181-192

[16] Sreejit, Dr. Jagathy Raj V. P. (2007), Organized Retail Market Boom and the Indian Society, *International Marketing Conference on Marketing & Society IIMK , 8-1*

[17] T. Bradlow et all, (2007) Organized Retail Market Boom and the Indian Society, *International Marketing Conference on Marketing & Society IIMK, 8-10*

[18] Michel. C. (2007), Bridging the Gap between Data Mining and Decision Support towards better DM-DS integration, International Joint Conference on Neural Networks, Meta-Learning Workshop

[19] Wang j. et all (2008), a comparison and scenario analysis of leading data mining software, Int. J Knowl Manage

[20] Chaoji V.(2008), An integrated generic approach to pattern mining: Data mining template library, Springer

[21] Hen L., S. Lee(2008), performance analysis of data mining tools cumulating with a proposed data mining middleware, Journal of Computer Science

[22] Bitterer, A., (2009), open –source business intelligence tool production deployment will grow five fold through2010, Gartner RAS research note G00171189.

[23] Phyu T.(2009), Survey of classification techniques in data mining, Proceedings of the International Multiconference of Engineering and Computer Scientist(IMECS), vol 1

[24] Pramod S., O. Vyas(2010), Performance evaluation of some online association rule mining algorithms for sorted & unsorted datasets, International Journal of Computer Applications, vol 2,no. 6

[25] *Mutanen. T et all,* (2010), Data Mining for Business Applications , Customer churn prediction – a case study in retail banking , Frontiers in Artificial Intelligence and Applications, Vol 218

[26] Prof. Das G. (2010), A Comparative study on the consumer behavior in the Indian organized Retail Apparel Market, ITARC

[27] Velmurugan T., T. Santhanam(2010), performance evaluation of k-means & fuzzy c-means clustering algorithm for statistical distribution of input data points., European Journal of Scientific Research, vol 46

[28] Lunenburg. C. (2010), Models of Decision Making FOCUS ON COLLEGES, UNIVERSITIES, AND SCHOOLS VOLUME 4, NUMBER 1.

[29] *Krishna M. (2010),* Data Mining- Statistics Applications: A Key to Managerial Decision Making, *indiastat.com* socio – economic voices

[30] Kavitha P.,T. Sasipraba (2011), Performance evaluation of algorithms using a distributed data mining frame work based on association rule mining, International Journal on Computer Science & Engineering (IJCSE)

[31] Mikut R., M. Reischi(2011), Data Mining tools, Wires. Wiley.com/Widm, vol 00

[32] Allahyari R. et all (2012), Evaluation of data mining methods in order to provide the optimum method for customer churn prediction: case study Insurance Industry , International conference on information & computer applications(ICICA), vol 24

[33] Giering M., SIGKDD exploration Retail Sales prediction & Item Recommendations using customer Demographics at store level, vol 10, Issue 2.

[34] Andersen, M. et all, Mining Your Own Business in Retail Using DB2 Intelligent Miner for Data, ibm.com/redbooks, Prasad P, Latesh, Generating customer profiles for Retail stores using clustering techniques, International Journal on Computer Science & Engineering (IJCSE)

[35] Chen X. et all, A survey of open source data mining systems,National Natural Science Foundation of China (NSFC)

[36] Jayaprakash et all, performance characteristics of data mining applications using minebench, National Science Foundation (NSF).

[37] Osama A. Abbas(2008),Comparison between data clustering algorithm, The International Arab journal of Information Technology, vol 5, N0. 3

[38] www.eecs.northwestern.ed/~yingliu/papers/pdcs.pdf

[39] *www.ics.**uci**.edu/~mlearn/*

[40] *www.thesai.org/.../Paper%204-... - United State*