Performance Evaluation of Data Mining clustering algorithm in WEKA

Mahendra Tiwari
Research Scholar,
Department Of Comp. Science,
UPRTOU Allahabad
tiwarimahendra29@gmail.com

Yashpal Singh
Head, Deptt Of CSE,
BIET Jhansi
yash_biet@yahoo.co.in

ABSTRACT

Data mining is a computerized technology that uses complicated algorithms to find relationships and trends in large data bases, real or perceived, previously unknown to the retailer, to promote decision support.., data mining is touted to be one of the widespread recognition of the potential for analysis of past transaction data to improve the quality of future business decisions. The purpose of this paper is to critique data mining technology in comparison with more familiar data mining algorithm in well known tool Weka for strategic decision making by small to medium size retailers. The context for this study includes current and future industry applications and practices for research performed in data mining applications within the retail sector.

KEYWORDS

WEKA	Algorithm
Cluster	Data Mining

INTRODUCTION

As 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 clusterer 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.

RELATED WORK

I 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, I presented their result as well as about tool and data set which are used in performing evaluation.

Ying Liu,wei-keng Liao et all in his article "performance evaluation and characterization of scalable data mining algorithms by Ying Liu, Jayaprakash, Wei-keng, Alok chaudhary" 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.

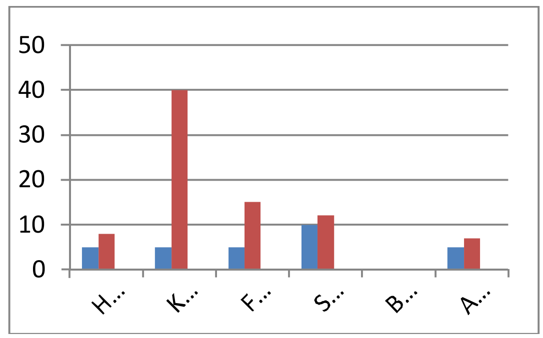

Fig 1: OS overheads of Mine Bench applications as a percentage of the total execution time.

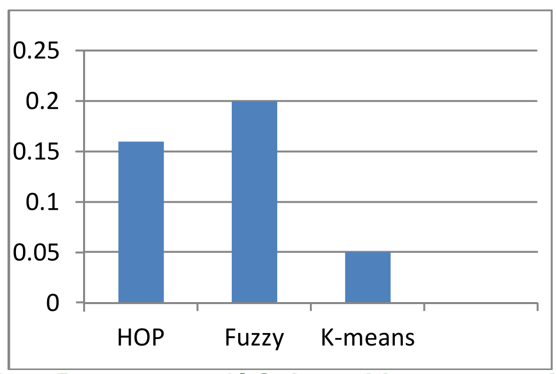

Fig 2: **Percentage of I/O time with respect to the overall execution times.**

Osama Abu Abbas in his article "comparison between data clustering algorithms by Osama Abu Abbas" 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

No. of cluster (K)	Performance			
	SOM	K-means	EM	HCA
18	59	63	62	65
16	67	71	69	74
32	78	84	84	87
64	85	89	89	92

Fig 3 : **Relationship between number of clusters and the performance of algorithm**

K=32				
Data type	SOM	K-means	EM	HCA
Random	830	910	898	850
Ideal	798	810	808	829

Fig 4 : The affect of data type on algorithm

T. velmurgun 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

Cluster		1	2	3	4	5	6	7	8	9	10	Time (ms)
Run 1	N	36	47	74	47	75	26	43	50	65	37	3469
	U	45	44	41	71	37	51	38	65	47	61	3265
Run 2	N	34	34	32	71	43	71	47	81	52	35	3266
	U	60	46	53	48	57	32	63	48	48	45	3250
Run 3	N	61	49	52	38	70	28	32	49	55	56	3156
	U	59	43	43	63	52	57	41	54	45	43	3297
Run 4	N	58	24	46	40	70	41	52	50	71	48	3469
	U	39	50	54	28	63	65	61	46	47	47	3187
Run 5	N	70	29	39	67	65	41	34	53	63	39	3484
	U	59	42	55	44	51	65	52	38	59	35	3282
Run 6	N	41	48	48	34	52	68	35	42	74	58	3281
	U	50	48	46	38	58	53	42	49	51	65	3266
Run 7	N	35	44	58	43	45	43	72	36	70	54	3283
	U	49	53	43	55	58	52	58	45	45	42	3281
Run 8	N	34	55	50	69	45	39	68	57	44	39	3328
	U	51	59	58	48	51	30	41	52	59	51	3282
Run 9	N	26	53	42	41	61	63	79	68	44	23	3328
	U	45	49	56	49	62	45	49	50	48	47	3281
Run 10	N	37	34	54	60	54	58	39	59	31	74	3360
	U	36	44	46	59	41	61	50	52	53	58	3266

Fig 5 : Clusters on 500 data points

Jayaprakash et all 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.

Reference	Goal	Database/Data description	Data size used	Preprocessing	Data Mining algorithm	Software
Abullah H. wabheh et all. (IJACSA)	Comparative study between a number og 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 (EJOSR)	To analyze K-means & Fuzzy C-means clustering result quality by Box-muller formula	Normal & uniform distribution of data points	500 to 1000 data points		K-means, Fuzzy C-means	Applet Viewer
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

Table 1 : Summary of selected references with goals

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 clusterers used in open source data mining tool –Weka.

ANALYSIS OF DATA MINING ALGORITHM

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.

EVALUATION STRATEGY/METHODOLOGY

H/W tools

I conduct my 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, I used Weka 3-6-6, I 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 and Letter image recognition dataset are in csv format whereas labor ,and Supermarket dataset are in arff format. Zoo, Letter, & Labor dataset have 17 number of attributes while Supermarket dataset has 200 attributes. Zoo dataset encompasses 101 instances, Letter image contains 20000 instances but I taken just 174 instances. Labor comprises 57 instances, & Supermarket has 4627 instances. All datasets are categorical and integer with multivariate characteristics.

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 Full training set & percentage split(holdout method) mode. The training set 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 characteristic, I analyzed how the results vary whenever test mode is changed. My measure of interest includes the analysis of clusterers on different datasets, the results are described in value number of cluster generated, clustered instances, time taken to build the model, and unclustered instances. after applying the cross-validation or holdout method.

For performance issues, There are 3 other datasets which I used for measurement they are Letter image recognition, labor, & Supermarket dataset. The details of applied classifiers on those datasets are as following:

Dataset: Letter image recognition
Classifier: Lazy-IBK,KStar, Tree-Decision stump, REP, Function- Linear regression, Rule-ZeroR
Dataset: Labor
Classifier: Lazy-IBK,KStar, Tree-Decision stump, REP, Function- Linear regression, Rule-ZeroR, Bayesian-Naïve Bayes
Dataset: Supermarket
Classifier: Lazy-IBK,KStar, Tree-Decision stump, CART, Function- SMO, Rule-ZeroR, OneR, Bayesion-Naïve Bayes.

The details of clusterer with different dataset are as following

- Dataset: Zoo
- Clusterer: DBscan, EM, Hierarchical, K-means
- Dataset: Letter image recognition
- Clusterer: DBscan, EM, Hierarchical, K-means
    - Dataset: Labor: Clusterer: DBscan, EM, Hierarchical, K-means
    - Dataset: Supermarket: Clusterer: DBscan, EM,, K-means

Clustering in Weka:-

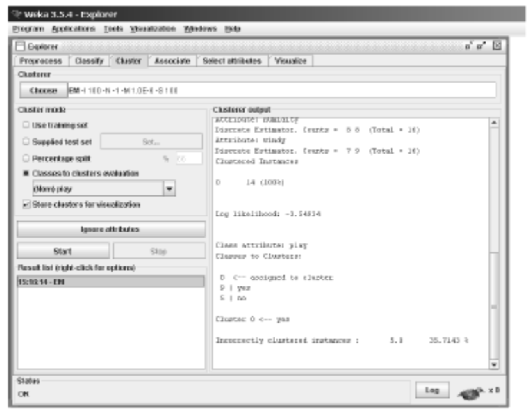

Fig 6 : Clustering window

- ***Selecting a Cluster:*** By now you will be familiar with the process of selecting and configuring objects. Clicking on the clustering scheme listed in the Clusterer box at the top of the window brings up a Generic Object Editor dialog with which to choose a new clustering scheme

- ***Cluster Modes:*** The Cluster mode box is used to choose what to cluster and how to evaluate the results. The first three options are the same as for classification: Use train- ing set, Supplied test set and Percentage split except that now the data is assigned to clusters instead of trying to predict a specific class. The fourth mode, Classes to clusters evaluation, compares how well the chosen clusters match up with a pre-assigned class in the data. The drop-down box below this option selects the class, just as in the Classify pane

- ***Ignoring Attributes:*** Often, some attributes in the data should be ignored when clustering. The Ignore attributes button brings up a small window that allows you to select which attributes are ignored. Clicking on an attribute in the window highlights it, holding down the SHIFT key selects a range of consecutive attributes, and holding down CTRL toggles individual attributes on and off. To cancel the selection, back out with the Cancel button. To activate it, click the Select button.

Working with Filters

The Filtered meta-clusterer offers the user the possibility to apply filters directly before the clusterer is learned. This approach eliminates the manual application of a filter in the Preprocess panel, since the data gets processed on the fly. Useful if one needs to try out different filter setups.

Learning Clusters

The Cluster section, like the Classify section, has Start/Stop buttons, a result text area and a result list. These all behave just like their classification counterparts. Right-clicking an entry in the result list brings up a similar menu, except that it shows only two visualization options: Visualize cluster assignments and Visualize tree.

DETAILS OF DATA SET

I used 4 data set for evaluation with clustering in WEKA ,Two of them from UCI Data repository that are Zoo data set and Letter image recognition, rest two labor data set and supermarket data set is inbuilt in WEKA 3-6-6 .Zoo data set and letter image recognition are in csv file format ,and labor and supermarket data set are in arff file format.
Detail of data set used in evaluation:--

Table 2 : Detail of data set

ZOO DATA SET

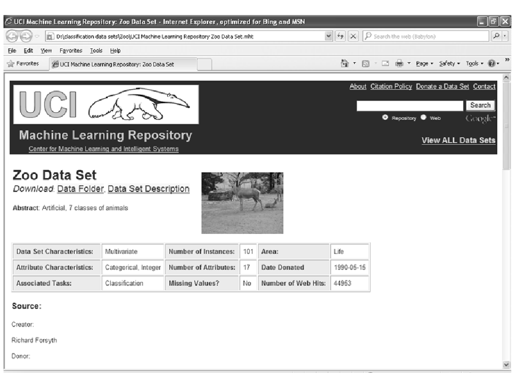

Fig 7 : Zoo data set (UCI repository.)

. Title: Zoo database
. Source Information
-- Creator: Richard Forsyth
-- Donor: Richard S. Forsyth
8 Grosvenor Avenue
Mapperley Park
Nottingham NG3 5DX
0602-621676
-- Date: 5/15/1990

Relevant Information:

-- A simple database containing 17 Boolean-valued attributes. The "type" attribute appears to be the class attribute. Here is a breakdown of which animals are in which type: (I find it unusual that there are 2 instances of "frog" and one of "girl"!) Class# **Set of animals**

1 (41) aardvark, antelope, bear, boar, buffalo, calf, cavy, cheetah, deer, dolphin, elephant, fruitbat, giraffe, girl, goat, gorilla, hamster, hare, leopard, lion, lynx, mink, mole, mongoose, opossum, oryx, platypus, polecat, pony, porpoise, puma, pussycat, raccoon, reindeer, seal, sealion, squirrel, vampire, vole, wallaby, wolf

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
Letter Image Recognition	CSV(comma separated value)	17	174/20000	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

2 (20) chicken, crow, dove, duck, flamingo, gull, hawk, kiwi, lark, ostrich, parakeet, penguin, pheasant, rhea, skimmer, skua, sparrow, swan, vulture, wren

3 (5) pitviper, seasnake, slowworm, tortoise, tuatara

4 (13) bass, carp, catfish, chub, dogfish, haddock, herring, pike, piranha, seahorse, sole, stingray, tuna

5 (4) frog, frog, newt, toad

6 (8) flea, gnat, honeybee, housefly, ladybird, moth, termite, wasp

7 (10) clam, crab, crayfish, lobster, octopus, scorpion, seawasp, slug, starfish, worm

Number of Instances: 101

Number of Attributes: 18 (animal name, 15 Boolean attributes, 2 numerics)

Attribute Information: (name of attribute and type of value domain)

○ **Animal name:**	**Unique for each instance**
○ **hair**	**Boolean**
○ **feathers**	**Boolean**
○ **eggs**	**Boolean**
○ **milk**	**Boolean**
○ **airborne**	**Boolean**
○ **aquatic**	**Boolean**
○ **predator**	**Boolean**
○ **toothed**	**Boolean**
○ **backbone**	**Boolean**
○ **breathes**	**Boolean**
○ **venomous**	**Boolean**
○ **fins**	**Boolean**
○ **legs**	**Numeric (set of values:**
○ **tail**	**Boolean**
○ **domestic**	**Boolean**
○ **catsize**	**Boolean**
○ **type**	

numeric (integer values in range [1,7])

8. Missing Attribute Values: None

9. Class Distribution: Given above

Letter image recognition data set :-

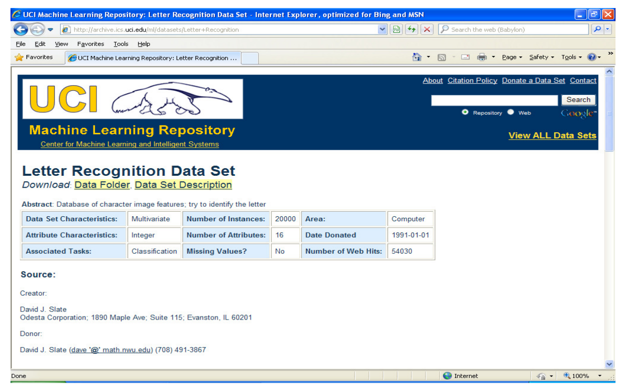

Fig 8: Letter image recognition data set

Title: Letter Image Recognition Data
Source Information
-- Creator: David J. Slate
-- Odesta Corporation; 1890 Maple Ave; Suite 115; Evanston, IL 60201
-- Donor: David J. Slate (dave@math.nwu.edu) (708) 491-3867
-- Date: January, 1991

Past Usage: "Letter Recognition Using Holland-style Adaptive Classifiers".

The research for this article investigated the ability of several variations of Holland-style adaptive classifier systems to learn to correctly guess the letter categories associated with vectors of 16 integer attributes extracted from raster scan images of the letters. The best accuracy obtained was a little over 80%. It would be interesting to see how well other methods do with the same data.

RELEVANT INFORMATION

The objective is to identify each of a large number of black-and-white rectangular pixel displays as one of the 26 capital letters in the English alphabet. The character images were based on 20 different fonts and each letter within these 20 fonts was randomly distorted to produce a file of 20,000 unique stimuli. Each stimulus was converted into 16 primitive numerical attributes (statistical moments and edge counts) which were then scaled to fit into a range of integer values from 0 through 15. We typically train on the first 16000 items and then use the resulting model to predict the letter category for the remaining 4000. See the article cited above for more details.

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

Number of Instances: 20000

Number of Attributes: 17 (Letter category and 16 numeric features)

Attribute Information:

- lettr capital letter (26 values from A to Z)
- x-box horizontal position of box (integer)
- y-box vertical position of box (integer)
- width width of box (integer)
- high height of box (integer)
- onpix total # on pixels (integer)
- x-bar mean x of on pixels in box (integer)
- x2bar mean x variance (integer)
- y2bar mean y variance (integer)
- xybar mean x y correlation (integer)
- x2ybr mean of x * x * y (integer)
- xy2br mean of x * y * y (integer)
- x-ege mean edge count left to right (integer)
- xegvy correlation of x-ege with y (integer)
- y-ege mean edge count bottom to top (integer)
- yegvx correlation of y-ege with x (integer)
- y-bar mean y of on pixels in box (integer)

Missing Attribute Values: None

Class Distribution:

789 A	766 B	736 C	805 D	768 E
775 F	773 G	734 H	755 I	747 J
739 K	761 L	792 M	783 N	753 O
803 P	783 Q	758 R	748 S	796 T
813 U	764 V	752 W	787 X	786 Y
734 Z				

Evaluation of Clusterer on various data set:

Evaluation of Clusterer on Zoo data set:-

Table 3 : 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 instances
DBscan	108	Percenatge split	0	0	0.02 second	37
EM	108	Percenatge split	5(5,5,10,12,5)	5(14%,27%,14%,32%)	1.58 second	0
Hierarchical	108	Percenatge split	2(0,37)	2(100%)	0.01 second	0
k-means	108	Percenatge split	2(8,29)	2(22%,78%)	0 second	0

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

7.2 Evaluation of Clusterer on Letter Image Recognition data set:-

Clustering Algorithm	**No. of Instances**	**Test mode**	**No. of cluster generated**	**Clustered instances**	**Time taken to build the model**	**Unclustered instances**

DBscan	174	Full training data	1	6(100%)	0.09 second	168
EM	174	Full training data	6(56, 25, 6,28,40, 19)	6(32%,14%, 3%,16%,23%,11%)	10.92 second	0
Hierarchical	174	Full training data	1	1(100%)	0.06 second	0
k-means	174	Full training data	2(69, 105)	2(40%,60%)	0.1 second	0

Table 5 : Evaluation of clusterer on Letter image recognition 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 instances
DBscan	174	Percenatge split	0	0	0.04 second	60
EM	174	Percenatge split	4(3,23,15,19)	4(5%,38%, 25%,32%)	3.91 second	0
Hierarchical	174	Percenatge split	1(60)	1(100%)	0.02 second	0
k-means	174	Percenatge split	2(40, 20)	2(67%,33%)	0 .01 second	0

Table 6: Evaluation of clusterer on Letter image recognition with percentage split test mode

7.3 Evaluation of Clusterer on Labor data set:-

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
k-means	57	Percentage split	2(9, 11)	2(45%,55%)	0	0

Table 7: Evaluation of clusterer on Labor data set with percentage 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	Full training	0	0	0.02 second	57
EM	57	Full training	3(14, 7,36)	3(25%,12%,63%)	0.69 second	0
Hierarchical	57	Full training	2(0,57)	1(100%)	0.02 second	0
k-means	57	Full training	2(48, 9)	2(84%,16%)	0 second	0

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

7.4 Evaluation of Clusterer on Supermarket data set:-

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

Table 9 : Evaluation of clusterer on supermarket data set with percentage split test mode

Clustering Algorithm (clusterer)	Instances	No. of cluster generated	Clustered instances	Unclustered instances	Test mode	Time taken to build model
DBscan	4627	2(1679,2948)	2(36%,64%)	0	Full training data	0.37 second
EM	4627	2(0,4627)	2(100%)	0	Full training data	159.54 second
K-means	4627	2(1679,2948)	2(36%,64%)	0	Full training data	1.06 second

Table 10 : Evaluation of clusterer on supermarket data set with Full training test mode

Result of Experiments in Weka

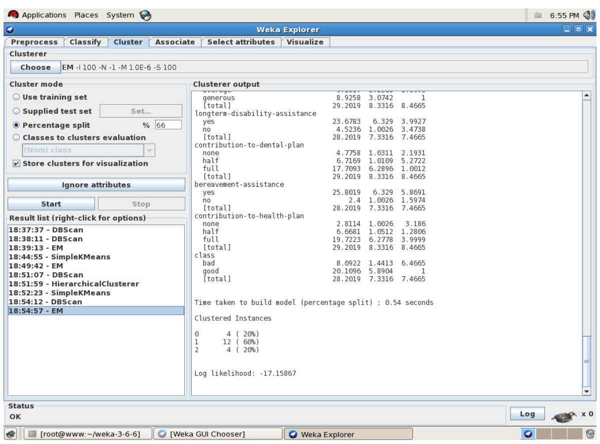

Fig 9: EMclusterer with percentage split test on labor data

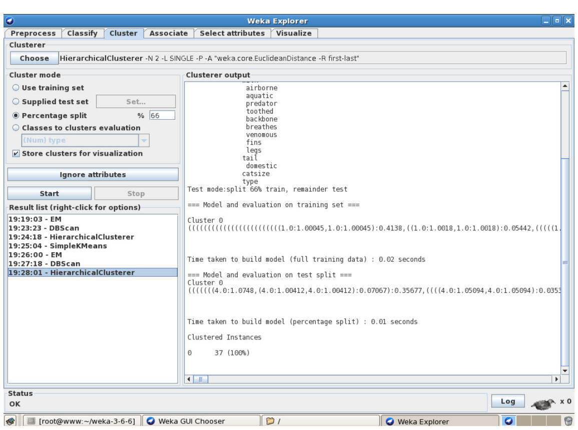

Fig 10: Hierarchical clusterer with percentage split test on zoodata

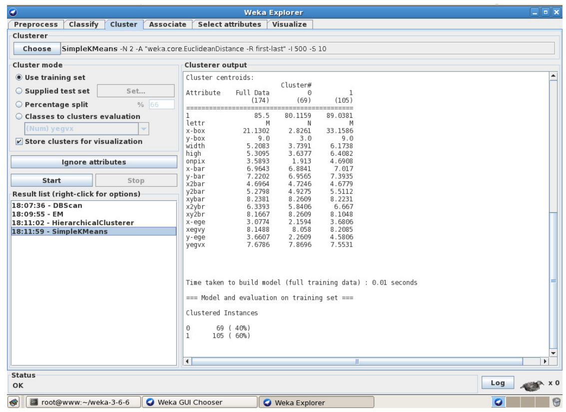

Fig 11: Kmeans clusterer with training set with letter data

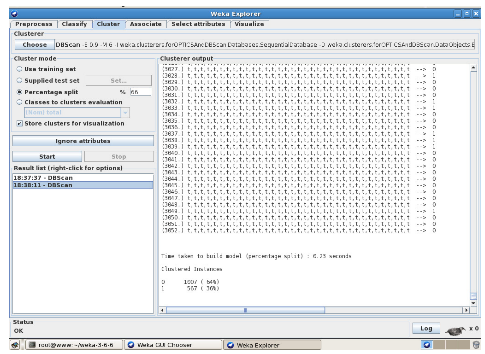

Fig 12:DBscan clusterer on supermarket data with percentage split

REFERENCES

i. www.boirefillergroup.com/....KDD_CONFERENCE_PAPER_AUG2006.pdf
ii. www.dcc.fc.up.pt/~ricroc/aulas/0708/dmlp/material/paper_dmbiz06.pdf
iii. www.ecmlpkdd2006.org/ws-pdmaec.pdf
iv. http://www.linkedin.com/in/federicocesconi
v. www.linkedin.com/in/federicocesconi
vi. www.footwearsinfolinethree.tripod.com/indian_retail_industry_its_growth_
vii. Open source Initiative: The open source definition(2007) www. Opensource.org/docs/definition_plain.html
viii. Retail and Consumer Worlds, Pricewaterhousecoopers, January, 2009
ix. Bose B.S. (2003), "Handbook of Marketing Management", Himalaya Publish in house, New Delhi.
x. Bishop. C.M. (1995) Neural Networks for pattern Recognition. New York: Oxford University Press
xi. Bigus, J.P. (1996) Data Mining with Neural Networks: Soling Business Problem- from Application Development to Decision Support. New York: McGraw-Hill.
xii. Jiawei han, Micheline Kamber, Data mining : concepts & Techniques (2nd edition).
xiii. Decision Trees for Business Intelligence & Data Mining: using SAS Enterprise minor.
xiv. DB2 Intelligent miner library(2002), Using the intelligent miner for data ,IBM, version 8 release 1.
xv. SAS Enterprise miner documentation, what's new in SAS enterprise miner 5.1 SAS Institue Inc.

xvi. SPSS Inc,(2005), maximize your returns with data mining and predictive analysis, Clementine.

xvii. Peter M. chen and David A.(1993), storage performance-metrics and bench marks, Proceeding of the IEEE, 81:1-33

xviii. M.Chen, J. Han, and P. Yu. (1996) Data Mining Techniques for marketing, Sales, and Customer Support. IEEE Transactions on Knowledge and Data Eng., 8(6)

xix. Agrawal R, Mehta M., Shafer J., Srikant R., Aming (1996) A the Quest on Knowledge discovery and Data Mining, pp. 244-249..

xx. Chaudhuri, S.Dayal, U. (1997) An Overview of Data Warehousing and OLAP Technology. SIGMOD Record 26(1) 65-74

xxi. *John F. Elder et all, (1998)* A Comparison of Leading Data Mining Tools, Fourth International Conference on Knowledge Discovery & Data Mining

xxii. 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

xxiii. 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

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

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

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

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

xxviii. Rabinovitch, L. (1999),America's first department store mines customer data. Direct marketing (62).

xxix. 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.

xxx. Brijs T. et all(2000), a data mining framework for optimal product selection in a retail supermarket: The generalized PROFSET model. Data Mining & Knowledge Discovery, 300

xxxi. Dhond A. et all (2000), data mining techniques for optimizing inventories for electronic commerce. Data Mining & Knowledge Discovery 480-486

xxxii. Jain AK, Duin RPW(2000), statistical pattern recognition: a review, IEEE trans pattern anal mach Intell 22:4-36

xxxiii. Zhang, G.(2000), Neural network for classification: a survey, IEEE Transaction on system, man & cybernetics, part c 30(4).

xxxiv. X.Hu, (2002) "Comparison of classification methods for customer attrition analysis" in Proc, of the Third International Conference on Rough Sets and Current Trends in Computing, Springer, pp. 4897-492.

xxxv. 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.

xxxvi. Rygielski. D.,(2002) , data mining techniques for customer relationship management, Technology in society 24.

xxxvii. Anderson, J. (2002), Enhanced Decision Making using Data Mining: Applications for Retailers, Journal of Textile and Apparel, vol 2,issue 3

xxxviii. Madden, M.(2003), The performance of Bayesian Network classifiers constructed using different techniques, Proceeding of European conference on machine learning, workshop on probabilistic graphical models for classification, pp 59-70.

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

xl. Ahmed, S.(2004), applications of data mining in retail business, Proceeding of the International conference on Information Technology : coding & computing.

xli. Bhasin M.L. (2006) Data Mining: A Competitive Tool in the Banking and Retail Industries, The Chartered Accountant

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

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

xliv. 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

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

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

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

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

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

l. 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

li. *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

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

liii. 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 no. 3

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

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

lvi. 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)

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

lviii. 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

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

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

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

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