Enhancing the Performance of Data Mining Algorithm in Letter Image Recognition Data

Mahendra Tiwari[1], Manu Bhai Jha[2]

[1]Dept of Comp. Sci. UPRTOU, Allahabad
[2]UCER, Allahabad

[1]tiwarimahendra29@gmail.com
[2]manujha1979@gmail.com

Abstract—The letter image recognition is a challenging problem for knowledge worker, the basic objective of this data set 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 evaluating the performance of clustering algorithm using letter image recognition data set. The performance of clustering will be calculated using the mode of classes to clusters evaluation.

Keywords—Clustering, numerical attributes, knowledge worker

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 dataset 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.

Analysis of Clustering Algorithm:

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:-

A. 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.

B. 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.

C. Input data sets

Input data is an integral part of data mining applications. The data used in experiment is either real-world data obtained from UCI data repository and widely accepted dataset available in Weka toolkit, during evaluation 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 selected data set is used for the clustering task. This dataset was chosen because ii have different characteristics and have addressed different areas. Letter image recognition dataset is in csv format and contains 20000 instances and having 17 attributes but I taken just 174 instances . The dataset is categorical and integer with multivariate characteristics.

II. EXPERIMENTAL RESULT AND DISCUSSION

To evaluate the selected tool using the given dataset, 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 dataset, then using the available clustering and test modes ,results are collected and an overall comparison is conducted.

Letter Recognition Data Set

Fig.1 Letter image recognition data set

III. 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.

Number of Instances: 20000

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

Attribute Information:

1.	lettr	capital letter	(26
2.	x-box	horizontal position	
3.	y-box	vertical position	
4.	width	width of box	
5.	high	height of box	
6.	onpix	total # on pixels	
7.	x-bar	mean x of on pixels	
8.	y-bar	mean y of on pixels	
9.	x2bar	mean x variance	
10.	y2bar	mean y variance	
11.	xybar	mean x y correlation	
12.	x2ybr	mean of x * x * y	
13.	xy2br	mean of x * y * y	
14.	x-ege	mean edge count	
15.	xegvy	correlation of x-ege	
16.	y-ege	mean edge count	
17.	yegvx	correlation of y-ege	

Missing Attribute Values: None

IV. EVALUATION OF CLUSTERING ALGORITHM

There are four clustering algorithms we have taken For evaluation of performance with letter image recognition data. K-means, EM, Hierarchical, and DBscan are the algorithms which generates the clusters of using data. Prediction accuracy as well as time taken to build the model is varying among them. The letter image data have 20,000 instance, but we chosen just 174 instance,K-means algorithm took minimum time to generate clusters with both test modes i.e. Full training set, and Percentage split whereas EM algorithm took maximum time to build model.

TABLE 1: PERFORMANCE OF CLUSTERING ALGORITHM ON LETTER IMAGE DATA 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	Prediction Accuracy
DBscan	174	Percentage split	0	0	0.04 second	60	---
EM	174	Percentage split	4(3,23,15,19)	4(5%,38%,25 %,32%)	3.91 second	0	100
Hierarchical	174	Percentage split	1(60)	1(100%)	0.02 second	0	100
k-means	174	Percentage split	2(40,20)	2(67%,33%)	0 .01 second	0	100

TABLE 2: PERFORMANCE OF CLUSTERING ALGORITHM ON LETTER IMAGE DATA WITH FULL TRAINING SET TEST MODE

Clustering Algorithm	No. of Instances	Test mode	No. of cluster generated	Clustered instances	Time taken to build the model	Unclustered instances	Prediction Accuracy
DBscan	174	Full training data	1	6(100%)	0.09 second	168	3.4
EM	174	Full training data	6(56,25, 6,28,40,19)	6(32%,14%, 3%,16%,23%,1 1%)	10.92 second	0	100
Hierarchical	174	Full training data	1	1(100%)	0.06 second	0	100
k-means	174	Full training data	2(69,105)	2(40%,60%)	0.1 second	0	100

V. RESULT OF EXPERIMENTS

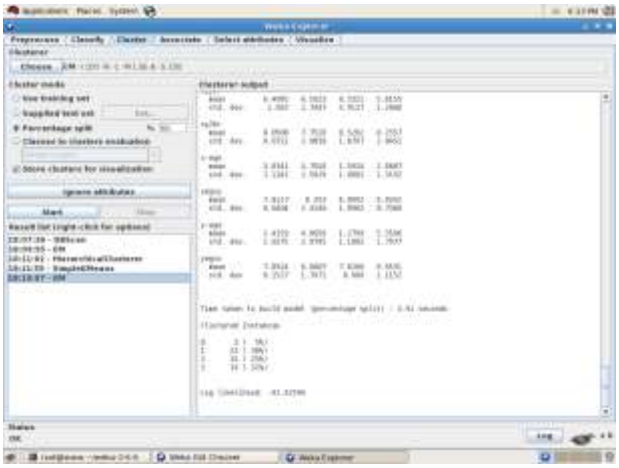

Fig2: EM algorithm on percentage split mode

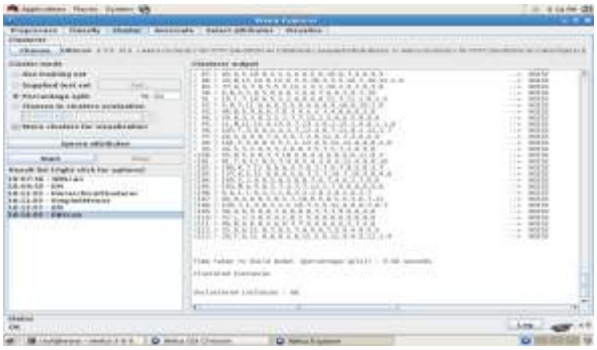

Fig3: DBscan algorithm on percentage split mode

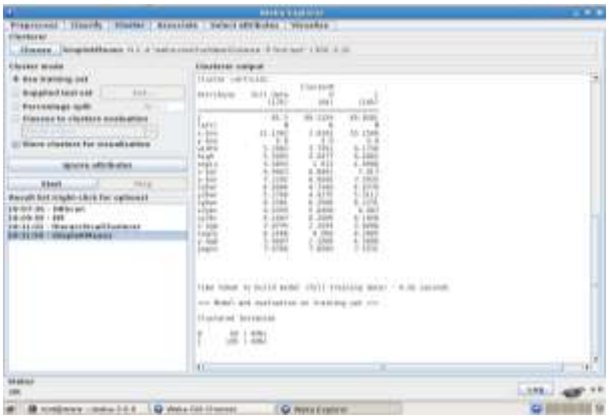

Fig 4: Kmeans clustering with training set mode

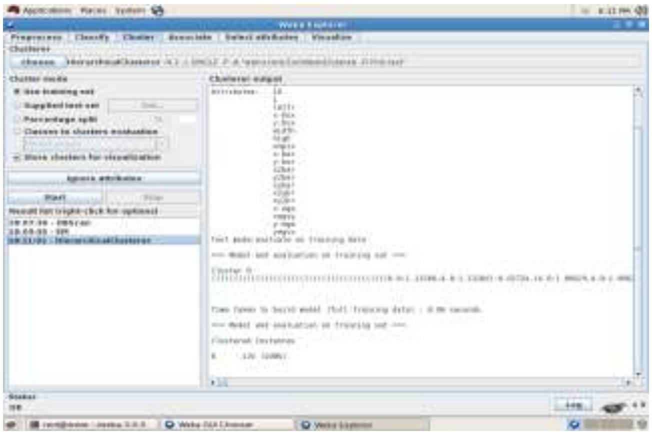

Fig 5: Hierarchical clustering with training set mode

VI. Conclusion

The letter image data is useful in scientific purpose, The data describes the horizontal and vertical position of box, width and height of box, and mean and correlation of box etc. The prediction data is used for find the performance of clustering algorithm. In the result, the prediction accuracy shows that DBscan clustering algorithm is weaker than others in generating cluster instances.

References

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

[2] John F. Elder et al, (1998) A Comparison of Leading Data Mining Tools, Fourth International Conference on Knowledge Discovery & Data Mining

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

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

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

[6] 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

[7] 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

[8] Prasad P, Latesh, Generating customer profiles for Retail stores using clustering techniques, International Journal on Computer Science & Engineering (IJCSE)

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

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

[11] .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)

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

[13] 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

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

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

[16] www.ics.uci.edu/~mlearn/