DATA MINING USING AGGLOMERATIVE MEAN SHIFT CLUSTERING WITH EUCLIDEAN DISTANCE

Suraj s. Damre[1],prof.L.M.R.J. Lobo[2]

[1]Department of Computer Science, Walchand Institute of technology, Solapur, Maharashtra.

[2]Department of Information technology, Walchand Institute of technology, Solapur , Maharashtra.

Email:surajdamre@gmail.com

ABSTRACT:

Agglomerative clustering is a non parametric clustering technique. In the present paper an approach agglomerative mean shift clustering applied to a text document using a query compression technique Clustering is presented. Here a distance based technique is developed. Two types of distances one for document and one for Query are made use of. Euclidean distance is used for finding query distance between two terms.The results achieved are comparable to other distance methods showing better time and distance accuracy than the existing system.

IJSER

1 INTRODUCTION:

A method entitled 'Cluster analysis' comprises of a grouping of allied techniques which are used to classify objects or cases of a domain into relatively like groups i.e a cluster. Each Object in a cluster tends to be similar in characteristics and different to objects in other clusters. Analysis done on Clusters is called classification analysis and is based on numerical taxonomy.[1]

Both the techniques 'cluster analysis' and 'discriminant' analysis are strongly related to classification. However, discriminant analysis requires previously known knowledge of the cluster or group membership for each item or case includedin it, to develop the classification rule. In contrast, cluster analysis deals with situations where there is no *a priori* known information about the group or cluster membership for any of the items. Groups or clusters are suggested by the data, not defined *a priori*. Fig.1 shows the structure formation of such a system.

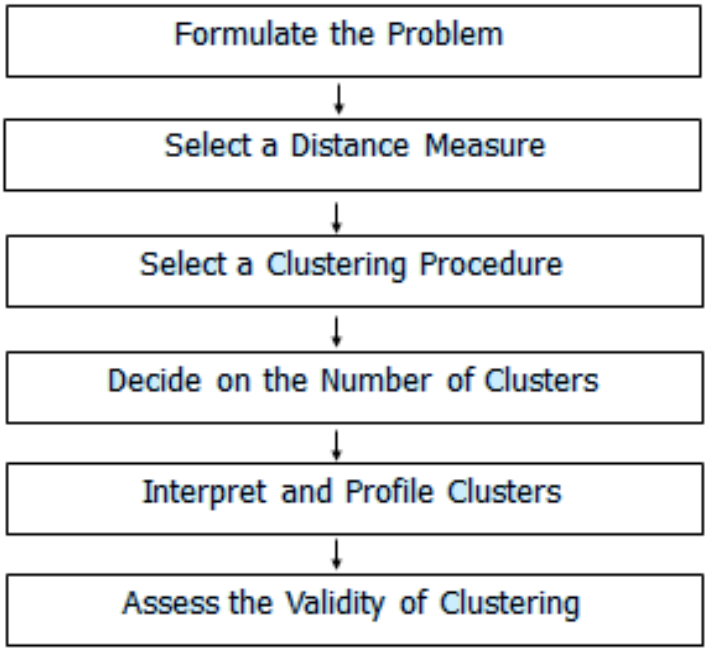

Fig.1 : Cluster analysis structure

IJSER

1.2.1 Formulation of the Problem :

This step is the most important part. It consists of formulating the clustering problem .Here there is a task of selecting the variables on which the clustering is based. Inclusion of even one or two irrelevant variables may distort an otherwise useful clustering solution. The set of variables selected should describe the similarity between objects in terms that are relevant to the marketing research problem. The variables should be chosen based on past research, theory, or a consideration of the hypotheses being tested. In exploratory research, the researcher should exercise judgment and intuition.

1.2.2 Select a Distance :

The most commonly used measure of similarity is the Euclidean distance or its square. The Euclidean distance is the square root of the sum of the squared differences in values for each variable. Other distance measures are also available. The *city-block or Manhattan distance* between two objects is the sum of the absolute differences in values for each variable. The *Chebychev distance* between two objects is the maximum absolute difference in values for any variable.

If the variables are measured in very different units, the clustering solution will be influenced by the units of measurement. In these cases, before clustering respondents, we must standardize the

data by rescaling each variable to have a mean of zero and a standard deviation of unity. It is also desirable to eliminate outliers (cases with typical values).

1.2.3 Selecting a clustering procedure:

Here we have to select the clustering procedure which we are going to use. This may be a agglomerative or divisive so after selecting the procedure apply that procedure on the data for getting the results.

1.2.4 Deciding on the number of clusters:

In this part the operation is decided on the number of results which we get if the number of clusters are huge then use agglomerative concept and if the number of clusters are less then use divisive method because agglomerative is a top down approach and divisive is a bottom up approach.

1.2.5 Interpreting and profile clusters:

Here the clusters are interpreted and that clusters are profiled in a proper manner. Interpretation of clusters is done on the results that how closely they are identical to each other in that basis interpretation is performed.

1.2.6 Accessing the validity of clusters:

Finally the clusters validity is checked using validation technique so we get the final clustered objects. [2]

IJSER

2. RELATED WORK:

Xuemin Lin, Jian Xu, Qing Zhang, Hongjun Lu, Jeffrey Xu Yu, Xiaofang Zhou, and Yidong Yuan[3] presented techniques for online processing of massive continuous quantile queries over data streams. While many research results on continuous queries and data stream computation have been recently reported in the literature, their research involved first attempts to develop scalable techniques to deal with massive query streams and data streams. Their query processing techniques were not only efficient and scalable, but also could guarantee the query precision requirements. Further, these techniques are applicable to both whole streams and sliding windows. Their experiment results demonstrated that the techniques are able to support online processing of massive queries over very rapid data streams.

M. Carreira-Perpinan[6] in 2006 dealt with Gaussian mean-shift(GMS), a clustering algorithm that has been shown to produce good image segmentations.GMS operates by defining a Gaussian kernel density estimate for the data and clustering together points that converge to the same mode under a indexed-point iterative scheme.The acceleration strategies proposed in this could be classified as 3types: discretisation methods(ms1), neighborhood methods(ms2andms3)and hybrid EM-Newton (ms4).All these methods can obtain nearly the same segmentation as GMS with significant speedups.

Y. Cheng, in 1995, provided an appropriate generalization to the mean shift algorithm, so that many interesting and useful properties would be preserved.Compared to gradient descent or ascent

methods, mean shift seems more effective in terms of adapting to the right step size. These methods may not be suitable for problems with prohibitive sizes and dimensionalities. [1][4]

3. METHODOLGY:

Mean-Shift is a powerful nonparametric clustering method. The proposed method, builds a Agglo-MS algorithm upon an iterative query set compression mechanism which is motivated by the quadratic bounding optimization nature of MS algorithm. The given query set is compressed iteratively until it achieved the convergence point. Initially, the query set covering procedure is implemented for grouping a new query set that covers under the given query set. Then the Iterative Query set Compression is applied for each new query set. Finally the convergence analysis is done to analyze whether the query set reached its convergence point Used Mahalanobis Distance for finding both Document Distance and Query Distance. It ensures more computational cost.[8]

Since this is a Iterative Process, Mahalanobis Distance calculation is a complicated one which is a time consuming process.

The **Euclidean distance** or **Euclidean metric** is the "ordinary" distance between two points that one would measure with a ruler, and is given by the Pythagorean formula. By using this formula as distance, Euclidean space (or even any inner product space) becomes a metric space. The associated norm is called the **Euclidean norm.** Older literature refers to the metric as **Pythagorean metric**.The **Euclidean distance** between points **p** and **q** is the length of the line segment connecting them ().

IJSER

In Cartesian coordinates, if **p** = (p_1, p_2,..., p_n) and **q** = (q_1, q_2,..., q_n) are two points in Euclidean n-space, then the distance from **p** to **q**, or from **q** to **p** is given by:

$$d(p, q) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

4. EXPERIMENTAL SETUP:

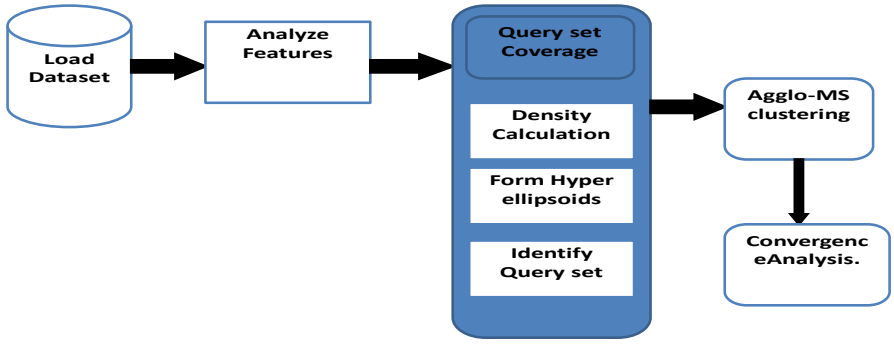

Fig.2 System Arhitecture

Fig.2 shows the system architecture for the implemented system. The basic blocks in this system comprise of the following[1]:

Load dataset:

Selection of the document which we are going to analyze is first done. The document may be text or image. Before performing the clustering operations we have to load the dataset on which we are performing the clustering approach.

IJSER

Analyze features:

The features of the selected dataset are analysed. The data set is preprocessed. The data is converted into the format which is suitable for further operation.

Query set coverage:

The selected document is converged with the query which gives the fast result i.et the given document is converted into a query.

Density calculation:

However, one needs to specify the bandwidth parameters for the Kernel Density Estimation (KDE). These parameters will significantly influence the output of the clustering. As the bandwidth decreases, the basin of attraction for each point will shrink, and when it goes to zero the algorithm will not move the mode-finding vectors from the initial position of the data vectors. Hence the output of a clustering procedure will be a one point cluster for each data vector. In the opposite case, when the bandwidth goes to infinity, the output will be one cluster located at the mean of the data vectors [7][9].

Form of hyper ellipsoids;

The words or text present in that query are used here. The basic idea is to construct a family of d-dimensional hyperellipsoids to cover the current query set Q, in hope that points inside each hyperellipsoid will converge to the same mode via MS iteration. For a given query point, the hyperellipsoid is constructed from a lower QB function of KDE defined at this point. It is observed that the number of the covering hyperellipsoids is much smaller than the size of Q. Then the centers of these hyperellipsoids are taken to form a new query set with size dramatically reduced. Such a query set covering procedure can be iteratively run until convergence is attained [7][9].

Agglomerative algorithm:

The Input for the algorithm is a Sample set, covariance matrix. We then perform initial query set covering operation. Iterative Query set compression phase is now applied on the data. It returns the cluster centers and clustering decisions at each level of iteration [2].

Convergence analysis:

So finally here we get the clustered objects from that dataset.

In this paper, clustering is done upon generating a compression on Query set iteratively. The work in this dissertation motivated by a concept called 'quadratic bounding optimization' of mean shift algorithm. until it achieves the convergence point.

IJSER

Initially, the query set covering procedure is implemented for grouping a new query set that covers under the given query set. Then the Iterative Query set Compression is applied for each new query set. Finally the convergence analysis is done to analyze whether the query set reached its convergence point.

Mahalanobis Distance is used in this implementation. This distance is a well known statistical distance function for finding both Document Distance and Query Distance It ensures more computational cost. Since this is a Iterative Process, Mahalanobis Distance calculation is a complexity intense procedure, one which is a time consuming process.To overcome these mentioned limitations, We are using the Mahalanobis distance only for finding the Document Distance.To find the Query Distance We are using the Euclidian Distance which finds the distance between the two terms more accurately.This is also a cost effective and time efficient process when compared with the Mahalanobis Distance

5. EXPERIMENTAL RESULT:

After experimentation on different systems we have achieved that using Euclidean distance the time accuracy for searching the dataset has increased by 0.098% as compared to Mahalanobis method and we got a distance accuracy of 18% more than Mahalanobis method. These are shown in Fig.3 and Fig.4.

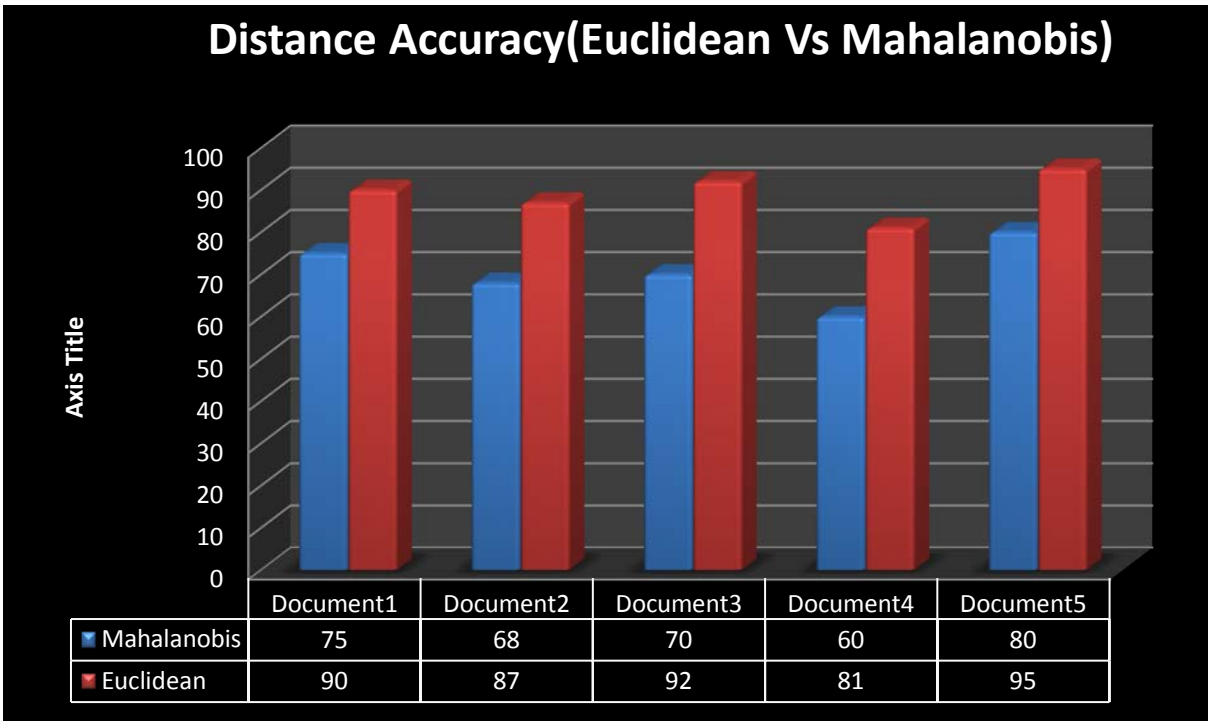

Fig.3:Distance accuracy graph for Euclidean vs Mahalnobis

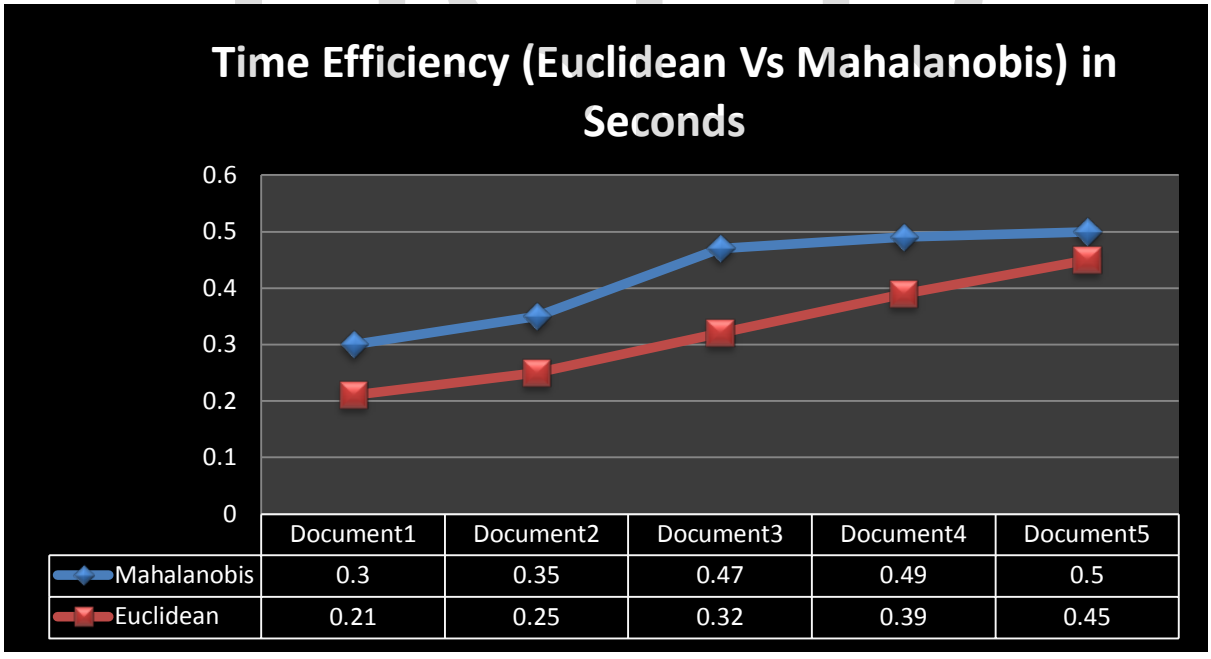

Fig.4:Time efficiency graph for Euclidean vs Mahalnobis

6.REFERENCES:

[1] Suraj s. Damre,prof.L.M.R.J.Lobo "Agglomerative mean shift clustering embedding query set compression" in International Journal of Scientific & Engineering Research, Volume 5, Issue 1, January2014 828ISSN2229-5518 IJSER © 2014

[2] Y. Cheng, "Mean Shift, Mode Seeking, and Clustering," in IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 17, no. 7, pp. 790-799, July 1995.

[3] Y. Zho and G. Karypis, "Hierarchical Clustering Algorithms for Document Datasets," in Data Mining and Knowledge Discovery, vol. 10, no. 2, pp. 141-168, Mar. 2005.

[4] M. Allain, J. Idier, and Y. Goussard, "On Global and Local Convergence of

[5] Half-quadratic Algorithms," in IEEE Trans. Image Processing, vol. 15, no. 5, pp.

[6] 1130-1142, May 2006.

[7] M. Carreira-Perpinan."Fast nonparametric clustering with Gaussian blurring mean-shift". in International Conference on Machine Learning, pages 153–160, 2006.

[8] M. Carreira-Perpinan."Gaussian mean-shift is an em algorithm".in IEEE TPAMI, 29(5):767–776, 2007.

[9] D. Comaniciu and P. Meer, "Mean Shift: A Robust Approach toward Feature Space Analysis," in IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24, no. 5, pp. 603-619, May 2002.

[10] X.-T. Yuan and S.Z. Li, "Stochastic Gradient Kernel Density Mode- Seeking,"in Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, 2009.

IJSER