A new Skeletonization Algorithm for Data Processing in Cloud Computing

Erhan Gökçay[1], Murat Karakaya[2], Atila Bostan[2]
[1]Software Engineering Department, Atılım University, Ankara, Turkey
erhan.gokcay@atilim.edu.tr
[2]Computer Engineering Department, Atılım University, Ankara, Turkey
murat.karakaya@atilim.edu.tr, atila.bostan@atilim.edu.tr

***Abstract*—This work presents a new approach to create abstract information from a huge amount of data, so that the processing and storage of the data stay within acceptable limits in terms of processing power and storage. Skeletonization is an important abstraction in pattern recognition, data storage and compression. In this paper, a novel method is proposed to create a similar skeleton from a cluster in the data, where a cluster is represented by a feature set. Single centroid information or a single statistical distribution is usually not sufficient to represent a nonlinearly distributed dataset, thus, in this work, the data will be represented by multiple centroids. The centroids are found by recursively clustering the data using an information theoretic clustering function. The set of centroids will represent the skeleton of a particular feature set, i.e. a cluster and they can be used to replace the original data for further storage and processing.**

Keywords — information theory; clustering; skeleton; centroid; abstract information; cloud data processing.

I. INTRODUCTION

The data generation and processing requirement are increasing rapidly in every field. The enormous increase in data generation results in new challenges and these challenges have created new fields to study which can be summarized as cloud computing. In very basic terms, cloud computing deals with storing and processing the huge amount of data generated mostly, but not limited to, by IT industry.

A complete review of all clustering algorithms is not possible and certainly is out of scope of the paper but a good review of the clustering algorithms is given in [1][2] and [12]. As a summary, we can put clustering algorithms under several categories like Information Theoretic [3], Hierarchical, Fuzzy, Center-based, Search-based, Graph-based, Density-based, and Model-based clustering algorithms. All algorithms are using a different distance metric to separate the clusters. Many of these algorithms are not designed to handle huge amount of data; hence there is a great deal of need to represent data in a compressed form.

Data representation and compression are another important processing technique in data processing. The Expectation-Maximization (EM) algorithm [4][5] is trying to place the parameters of a model that represents the data, in an iterative optimization. First; an initial guess is made for the parameters, then the maximum likelihood function is maximized and the parameters are guessed again iteratively until convergence. The number of center points is initialized beforehand and they may diverge from the data depending on the random initialization. EM algorithms can represent a nonlinear distribution nicely provided that there are enough centers and they are converged correctly.

Cloud data processing is a fundamental research area and it is discussed in terms of data representation, storage and analysis in [6]-[11]. The storage and processing problems of cloud data need a new approach in data analysis. Many standard algorithms cannot deal with the huge stream of data. Therefore data abstraction plays a very important role in cloud computing where the data is stored not in its pure form but as a reduced data set preserving the same statistical and structural form. In this form clustering algorithms can process the data much faster than the original data.

In the paper the limitation of linear abstraction is given. In the second chapter, the nonlinear clustering function is introduced and the proposed algorithm is given as a pseudo algorithm. The algorithm is applied to several nonlinear data distributions with different number of centers.

The algorithm is calculating centroids from a clustering point of view instead of a pdf calculation. There are several methods like parzen window estimation or EM methods to find out a pdf distribution using more than one kernel combined together. These algorithms focus on finding the underlying distribution and not necessarily to cluster the data. The proposed algorithm is trying to find the centroids by focusing on the clustering the data in the future; hence the centroids are better placed in terms of clustering the data. Another advantage is that the centroids never diverge from the data itself since they are calculated by clustering the data. On the other hand skeletonization algorithms

focus on finding a continuous thin structure to represent the shape and again this calculation is not focusing on clustering. A single line through a shape may miss important nonlinear structures important in a clustering process.

II. CENTROID REPRESENTATION

When the data has a hyper-ellipsoid distribution, then the mean and the covariance matrix are enough to represent the data in an abstract way. For example, the data distribution given in Fig. 1 is a hyper-ellipsoid and the data can be represented by the mean (center point or centroid) and the distribution around the mean in each dimension (covariance). By using these attributes the set of data given in Fig. 1 can be compared, processed and stored in an abstract way.

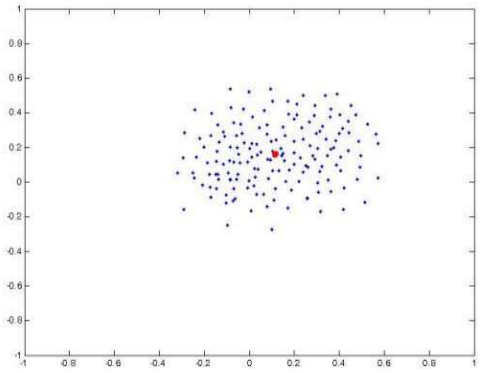

Figure 1: Hyper-Ellipsoid Distribution

The data representation in an abstract way becomes more difficult when the data in question has a nonlinear distribution. An example is given in Fig. 2.

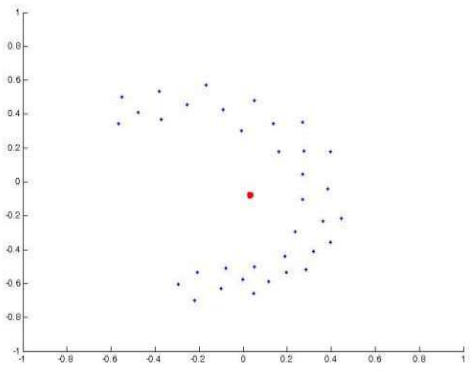

Figure 2: Non Linear Distribution

The data cannot be represented by a mean and a covariance matrix as the distribution is highly nonlinear. There are no data points around the mean which shows that the abstract representation; i.e. mean and covariance matrix cannot represent the given data set correctly. It is obvious that there should be more than one centroid to represent the data properly. Algorithms using multiple centers; like EM, has convergence problems. There is a need for an algorithm to find out the centroid locations without convergence problems.

III. PROPOSED METHOD

In this paper a novel method is proposed to find out the correct locations of multiple centroids where the centroids are basically forming the skeleton of the data set. It is clear that more centroids will represent the data better but also at the same time less number of centroids is better for storage and compression. There is a tradeoff between better representation and better storage and processing. If we have N data points, then the number of centroids will be between a range of 1 and N. The covariance is also part of the information attached to the skeleton, but without the correct location of centroid, the variance is useless.

Assume that we have a huge amount of data about human behavior and from this set of data a cluster is created of happy people where the features are age, height, salary, family and location. If you apply the skeleton algorithm to this data set, the set of happy people represented by the given feature set can be saved and processed by only using the information skeleton which consists of multiple centroids and covariance of each centroid.

A. Information Theoretic Clustering

Information theory has been used as a clustering algorithm successfully in one of the authors' previous paper [3]. Using an information theoretic measure, nonlinear clusters can be separated. The derivation will not be repeated here. Only the final formulation will be used. In (1), p and q are clusters of size N_1 and N_2, respectively where $x_i \in p$ and $x_j \in q$. The Gaussian kernel needs a parameter σ for the kernel size. The proper value of this parameter is important in clustering.

$$CEF(p,q) = \frac{1}{N_1 N_2} \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} G\left(x_i - x_j, 2\sigma^2\right) \qquad (1)$$

Although in the Information Skeleton algorithm, any clustering formulation can be used, CEF is chosen because of its capability to separate nonlinear regions.

B. Algorithm

In order to obtain a good abstraction, each centroid should represent the data distribution around it correctly. The algorithm has some similarities with the Expectation-Maximization (EM) algorithm where multiple Gaussian centers and variances are iteratively calculated and maximum likelihood is maximized to create a better representation of the data set. In the EM algorithm, the initial centers are assigned randomly and the number of centers should be predefined. Depending on the random assignment, the centers may move of the data as seen in

Fig.3 and Fig. 4. Fig. 3 shows the correct convergence of 3 centers using EM algorithm and Fig. 4 shows an unwanted convergence of 3 centers.

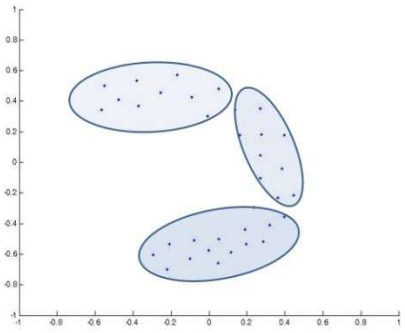

Figure 3: Correct convergence of EM alg.

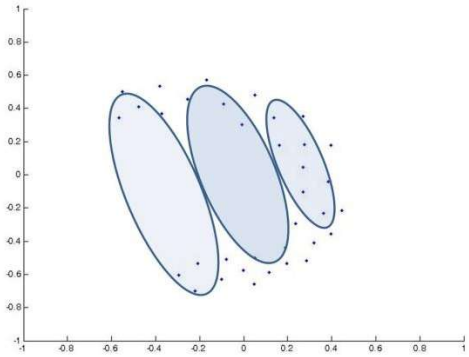

Figure 4: Incorrect convergence of EM alg.

Similarly in this algorithm the data is divided iteratively until a desired number of centroids are obtained. In the Information Skeleton algorithm, the data set is divided into two clusters initially using CEF clustering algorithm and each sub-cluster is divided recursively until the data set is represented correctly by the centroids of each sub-cluster. Since the clustering algorithm is used to create a skeleton of information represented by a feature set, the algorithm is called Information Skeleton (IS). Each time the size of the clusters are getting smaller and algorithm is terminated when the size of each sub-cluster is less than a predefined number which basically defines the number of centroids to represent the given data set.

The algorithm can be formulated as below and a flowchart is given below in Fig. 5. Assume that the data is represented by D, and the optimum sub-clusters are represented by C_1 and C_2. The maximum group size of any sub-cluster is given as a parameter N_{max}. When minimized in terms of its arguments, the *CEF_CLUSTER* function returns the optimum sub-clusters C_1 and C_2 from a dataset D by using CEF distance function. Calling *CEF_CLUSTER* function recursively on the sub-clusters will divide the data into smaller subsets until each sub-cluster size is less than N_{max}.

$$[Centroids] = InfoSkel(D, \sigma)\{$$
$$converged = size(D) < N_{max}$$
if (converged){
Centroids(n++)=mean(D)
Return
}
$$[C_1, C_2] = argmin_D CEF\_CLUSTER(D, \sigma)$$
$$[Centroids] = InfoSkel(C_1, \sigma)$$
$$[Centroids] = InfoSkel(C_2, \sigma)$$
$$\}$$

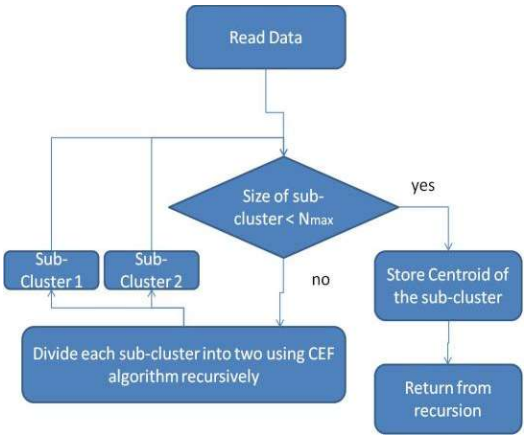

Figure 5: Flow Diagram

Each centroid is calculated from a sub-cluster which is found by clustering the data. At every step, clustering the data into two sub-clusters creates the optimum separation of the data. Therefore, each centroid has the best location to represent a given cluster. A recursive calculation is given in Fig. 6 as an example with a maximum cluster size of 4.

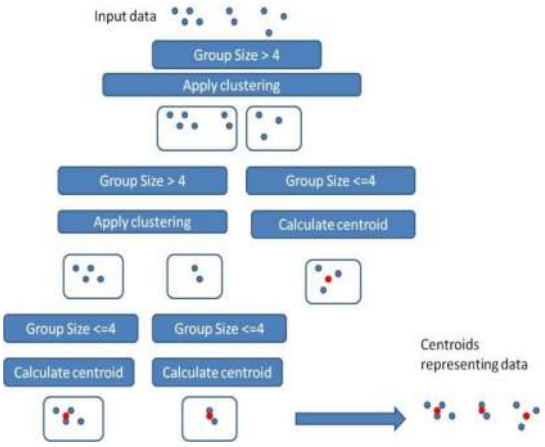

Figure 6: Example Calculation

The algorithm is applied to the data given in Fig. 2 using different number of centroids to show the progress of the representation. The results are given in Fig. 7, Fig. 8, Fig. 9, and Fig. 10.

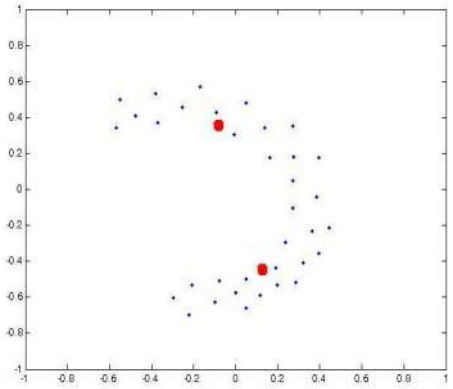

Figure 7: 2 centroids

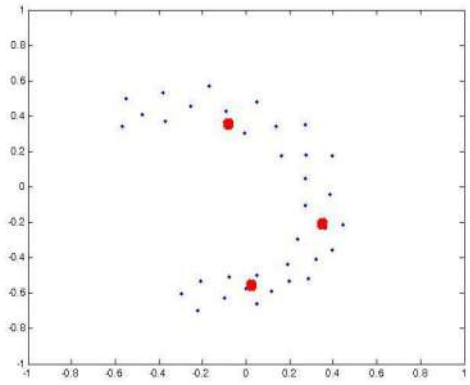

Figure 8: 3 centroids

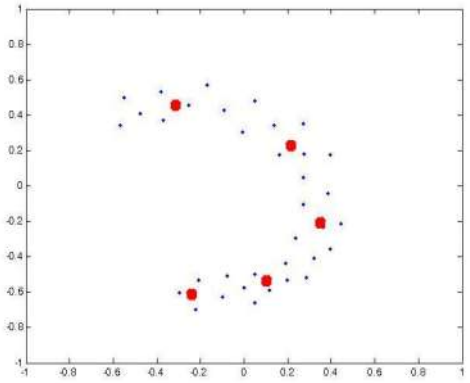

Figure 9: 5 centroids

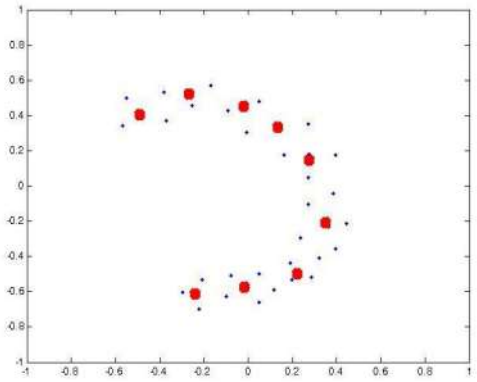

Figure 10: 9 centroids

It is clear that multiple centroids; i.e. Information Skeleton, can represent the nonlinear distribution in a much better way than a single centroid and IS in Fig.8 can be used for processing and storage instead of the original data.

It should be noted that although a nonlinear clustering function is used in the algorithm, this is not a clustering algorithm by itself. The algorithm is trying to find out the best representation with less data. After creating an abstract version of the data, any clustering algorithm could be applied to the data.

Information skeleton is calculated for different datasets given in Fig. 11-12. It can be observed that the centroids are closer to each other when the data is denser.

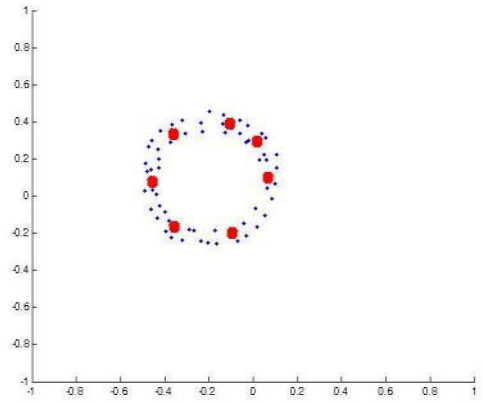

Figure 11: Example Data Set

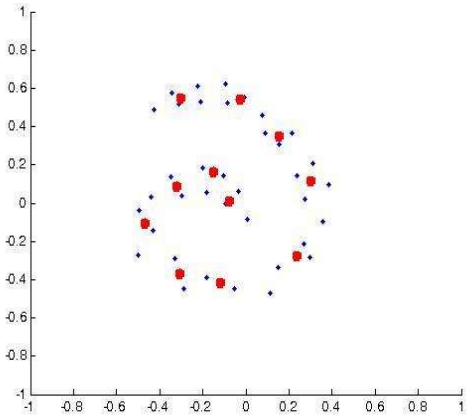

Figure 12: Example Data Set

IV. FUTURE WORK

The maximum number of centroids is equal to the number of data points. In that case there is no abstraction. Therefore; there should be a measure to decide to the optimum number of centroids so that data is represented efficiently and correctly.

V. CONCLUSION

The Information Skeleton algorithm creates an abstract version of a given dataset. Unlike EM algorithm, the centroids are always converging towards the data. The abstract representation can be controlled by the maximum size of each sub-cluster which also determines the number of centroids implicitly. When the data distribution is nonlinear, it is always possible to add more centroids to the representation.

The abstract representation can be used for storage and processing instead of the original data. This is especially important in cloud computing where a huge amount of data is generated.

REFERENCES

[1] N. Ahmed, "Recent review on image clustering," in IET Image Processing, vol. 9, no. 11, pp. 1020-1032, 11 2015. doi: 10.1049/iet-ipr.2014.0885.

[2] Guojun Gan, Chaoqun Ma, and Jianhong Wu, "Data Clustering Theory, Algorithms, and Applications", 2007

[3] Gokcay, E.; Principe, J.C., "Information theoretic clustering," Pattern Analysis and Machine Intelligence, IEEE Transactions on , vol.24, no.2, pp.158,171, Feb 2002

[4] A.P.Dempster, N.M. Laird, and D.B. Rubin. Maximum-likelihood from incomplete data via the EM algorithm. J. Royal Statist. Soc. Ser. B., 39, 1977

[5] M. Jordan and R. Jacobs. Hierarchical mixtures of experts and the EM algorithm. Neural Computation, 6:181–214, 1994

[6] S. Pandey and V. Tokekar, "Prominence of MapReduce in Big Data Processing," Communication Systems and Network Technologies (CSNT), 2014 Fourth International Conference on, Bhopal, 2014, pp. 555-560.

[7] C. K. S. Leung, R. K. MacKinnon and F. Jiang, "Reducing the Search Space for Big Data Mining for Interesting Patterns from Uncertain Data," 2014 IEEE International Congress on Big Data, Anchorage, AK, 2014, pp. 315-322.

[8] K. Park, M. C. Nguyen and H. Won, "Web-based collaborative big data analytics on big data as a service platform," 2015 17th International Conference on Advanced Communication Technology (ICACT), Seoul, 2015, pp. 564-567.

[9] S. R. Qureshi and A. Gupta, "Towards efficient Big Data and data analytics: A review," IT in Business, Industry and Government (CSIBIG), 2014 Conference on, Indore, 2014, pp. 1-6.

[10] H. Xiong, "Structure-Based Learning in Sampling, Representation and Analysis for Multimedia Big Data," Multimedia Big Data (BigMM), 2015 IEEE International Conference on, Beijing, 2015, pp. 24-27.

[11] O. Kurasova, V. Marcinkevicius, V. Medvedev, A. Rapecka and P. Stefanovic, "Strategies for Big Data Clustering," 2014 IEEE 26th International Conference on Tools with Artificial Intelligence, Limassol, 2014, pp. 740-747.

[12] A. Nagpal, A. Jatain and D. Gaur, "Review based on data clustering algorithms," Information & Communication Technologies (ICT), 2013 IEEE Conference on, JeJu Island, 2013, pp. 298-303.