International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 4 Issue 01,January-2015

EM Clustering Approach for Multi-Dimensional Analysis of Big Data Set

Amhmed A. Bhih
School of Electrical and Electronic Engineering
Liverpool John Moores University
Liverpool, UK

Princy Johnson
School of Electrical and Electronic Engineering
Liverpool John Moores University
Liverpool, UK

Martin Randles
School of Computer Science
Liverpool John Moores University
Liverpool, UK

***Abstract*—Data mining is one of the long known research topics, which is making a comeback especially with the advent of Big Data. 'Clustering' technique is an important component in data mining. As we enter the Big Data era where many real-world datasets consist of multi-dimensional features, clustering has been gaining momentum in importance within this topic. The traditional clustering algorithms often fail to detect meaningful clusters in high-dimensional data set. Therefore, they become computationally expensive when dealing with data comprised of multiple dimensions. In this paper, we have proposed a modified technique that will perform well with high dimensional data set. In our proposed method we used Principle Component Analysis for dimension reduction before applying standard EM algorithm. The performance of the proposed set of algorithms is evaluated on the basis of silhouette index and time of execution.**

Keywords—Clustering; dimensionality reduction; Particle Component Analysis; Expectation Maximization

I. INTRODUCTION

In the current digital age, data is regularly being generated and flowing everywhere and every minute. Terabytes (10^{12} bytes) are old news; now we are having to contend with petabytes (10^{15} bytes) and zeta bytes (10^{21} bytes). However, the data is also generated from a variety of sources, and there are two phenomena in particular that are driving this explosion, the "Internet of things" and the social web networks [10].

Big data as a term appeared literally first time towards the end of the 1990's [13] and generally, it is defined as the data whose size required new technologies and methods to make it possible to extract the necessary values from it. The most popular definitions are based on the 3Vs model for describing Big Data, namely volume (amount of data), velocity (update time per new observation) and variety (dimension, or range of sources).

The primary value of big data does not come from the data itself, but from using it in an intelligent way, by processing and analysis it. Data without a model is just noise. Models are used to describe salient features in the data. The science of extracting useful information from data sets is known as data mining. Data mining involves models from statistics, machine learning, artificial intelligence and data base management. The main goal of data mining is to find hidden patterns in datasets and predict models that can explain it [5].

As an important function of data mining, cluster analysis, also known as data clustering, is an important technique especially when dealing with a large number of data analyses. It is a crucial data mining step and performing this task over large databases is essential. Clustering can be described as the process of finding groups of objects whose members are similar or related in some way and different from or unrelated to the objects in other groups. By clustering big data set, people could obtain data distribution, make further study on particular clusters or observe the characteristics of each cluster.

Typically in clustering there is no one perfect solution to the problem, but a good clustering method produces high quality clusters with high intra-cluster similarity and low inter-cluster similarity [3].

Data Clustering has a long and rich history. According to JSTOR, data clustering first appeared in the title of a 1954 article dealing with anthropology. One of the most popular and simple clustering algorithms is K-means. It was first proposed over 50 years ago. Despite the fact that thousands of clustering algorithms have been published since then, K-means is still widely used due to its simplicity and efficiency [7].

However, with the proliferation of data sources, clustering real-world data sets is hampered by the so-called curse of dimensionality; many real world datasets consist of a very high dimensional feature space. Traditional methods for cluster analysis become computationally expensive and do not work well for such high dimension data set. Moreover, the results may not be accurate most of the time due to noise associated with original data. Hence, in the context of Big Data it is of high interest to find a method that is able to obtain clustering with minimized complexity of data by reducing the size of the data but, keeping its main characteristics.

Dimensionality reduction or attribute reduction is an essential pre-processing task for cluster analysis of datasets having a large number of dimensions (features) [3].

Based on the above considerations, in this paper we have used the Principal Component Analysis (PCA) method as a first phase for Expectation Maximization (EM) clustering algorithm which will simplify the analysis of a high dimensional data set.

This paper is organized as follows. In the next section, we will present the concept of Expectation Maximization (EM) clustering algorithm. PCA has been discussed in section 2. Section 3 describes our new proposed algorithm (PCA-EM). Section 4 briefly presents the experiments and performance analysis. Section 5 shows the experimental results which is followed by a set of conclusions in section 6.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

II. REVIEW OF THE LITERATURE AND RELATED STUDIES

A. *Expectation Maximization algorithm*

EM algorithm was introduced by Dempster et al in 1977 [6]. It is an optimization algorithm for using a proper statistical model of data to find the clusters such that the maximum likelihood of each cluster parameters is obtained. EM algorithm uses iterative method to find the maximum likelihood of clusters centroid starting from some initial guess. Generally, each iteration consists of two steps, expectation (E) step and maximization (M) step. In the former step, the current model parameters are used to compute the probability that each data record is a member of each cluster. Whereas, in the latter step, the fraction assessment is given by re-estimating the parameters of each cluster to maximize those probabilities. [4].

In this paper, Gaussian mixture model has been used as a statistical model to find the desired clusters in a given data set. EM algorithm finds clusters by determining a mixture of Gaussians that fit the given data set. Each Gaussian has an associated mean and standard deviation. These parameters could be initialized by randomly selecting means of the Gaussians, or via some heuristic methods. Then the algorithm converges on a locally optimal solution by iteratively updating values for means and covariance matrix.

Convergence is detected by calculating the value of the probability density function (PDF) after each iteration until a halting criterion is reached when the PDF appears not to be changing in a significant manner or the maximum number of iteration is obtained. However, the PDF for cluster h = 1 …k, is parameterized by the d-dimensional mean vector μ_h and d × d covariance matrix Σ_h is defined as follows [2]:

$$f_h\left(x\backslash\mu_h, \sum h\right) = \frac{1}{\sqrt{(2\pi)^d |\sum h|}} \exp\{-\frac{1}{2}(x-\mu_h)^T \left(\sum h\right)^{-1} (x-\mu_h)\}$$

B. *Principal Components Analysis (PCA)*

PCA is a statistical technique which uses sophisticated underlying mathematical principles to find and extract important patterns from high dimension datasets by reducing the number of dimensions without much loss of information. PCA has been considered one of the most important applications of applying linear algebra. However, beside its common use as a reduction method, it has found application in many fields like image compression, de-noising signal, blind source separation and face recognition [11].

The main idea behind PCA is to reduce the dimension of data set by removing unimportant information, such as noise and redundant data sets. This should transform the data to new coordinate system in such a way that the most important features could be extracted easily. The most important features could be described mathematically by using variance. Whereas, the valuable data features could be extracted easily by changing of the basis that represents the data set.

The first task of PCA is to transform the data to a new coordinate axis such that the greatest variance of data set lies on the new coordinate. This process is equivalent to obtaining the least-squares line of best fit through the plotted data. This new axis is called first principal component of the data (PC1). Once the first principal component is obtained, another axis is added orthogonal to the first principal component to represent the next highest variation through the data. This axis is called the second principal component (PC2). The process of adding more principal components (axis) continues. Each component orthogonal to the previous one and each one accounting for less and less of the variance in the data set [12]. The number of principal components should be less than or equal to the number of original dimensions of the data. The more number of components is considered, the more information is taken into account in the analysis [9].

The mathematics behind the principle component analysis is statistics hinges on standard deviation, eigenvalues and the eigenvectors. The eigenvectors define the directions of the new coordinate axes and the eigenvalues indicate the amount of variance experienced by each corresponding eigenvector. Therefore, in order to decide which eigenvectors needed to be dropped to obtain lower-dimensional dataset, we have to take a look at the corresponding eigenvalues of the eigenvectors. In other words, the eigenvectors with the largest eigenvalues are the most dominant principle components of the dataset and the eigenvectors with the lowest eigenvalues could be dropped without losing too much information [8], as they would be of least importance.

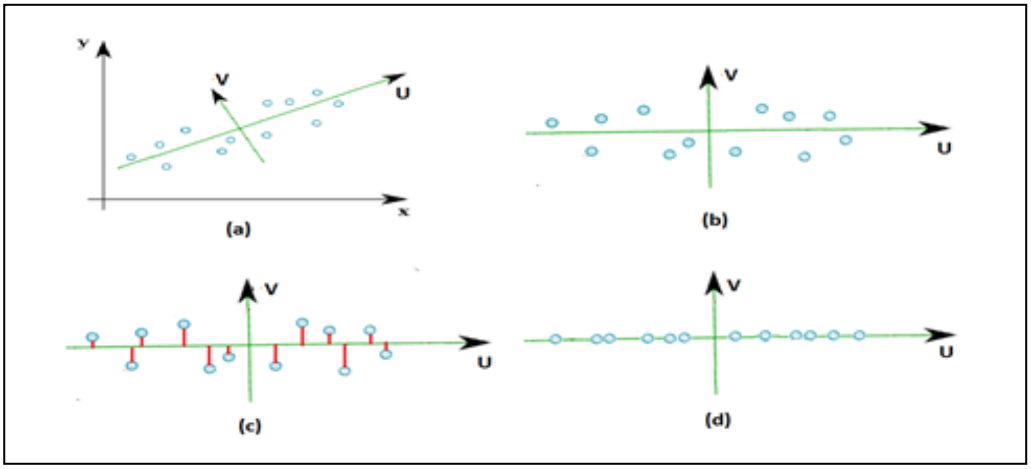

Fig. 1. PCA for two dimensional data reduction: (a) Original data. (b) Data representation. (c) Projection errors using one dimensional PCA. (d) PCA for one dimension dataset

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

Fig. 1 shows a two dimensional data set which has been represented in X and Y coordinates system. The principal direction of data and the second most important direction are illustrated by U and V coordinates respectively in Fig. (1.b).

The data set can be rotated to align the principal axes with x and y axes. This process does not change the data; it just represents it in a fashion that makes it easier to see which factors affect the data.

U and V components contain 100% of the information in the original 2D dataset. However, the first component shows more information than the second component. Thus in the U and V axis system it is possible to represent the data set by one variable U and discard V. Thus we have reduced the dimensionality of the problem by 1.

III. PROPOSED METHOD AND ALGORITHM

As an EM clustering algorithm does not work well for high dimensional data sets, in order to improve its ability and efficiency we applied PCA technology to reduce the dimensions of data set without losing much information. To the Authors' best knowledge this method of dimensions reduction has not been done prior to the EM clustering before. However, the standard EM algorithm initializes with randomly chosen partition. In the proposed algorithm a k-means algorithm, presented in [7], is used as a partition initialization method to improve the performance of clustering. This partition initialization method makes the EM estimation process much faster than starting from a random initialization.

Simply, as seen from Fig. 2, the algorithm can be described as comprising the following inputs, outputs and three phases as discussed below:

Input:

X: a set of n data items.

K: Number of desired clusters.

Output: A set of k clusters.

Phase one: Apply PCA to reduce the dimension of the data set to get PCs. Here, the number of dimension specified to illustrate 95% of the information in the dataset.

Phase two: Find the initial centroids

Phase three: Apply the EM clustering with the initial centroids.

IV. EXPERIMENTATION AND PERFORMANCE ANALYSIS

In this study, and in order to evaluate the efficiency of our proposed clustering algorithm in handling big data the standard EM algorithm has been used as a bench mark. We generate synthetic datasets in matrix forms (m x n) with a fixed dimension (n) value equal to 1500 and gradually increasing the number of observation in data sets (m) from 1,000 to 45,000. Each algorithm has been applied more than once on the data set to obtain 500 clusters (k=500). The experimental results presented are the average of five runs.

In all experiments, MATLAB software has been used as a tool to compute clusters on an Intel core I7 with 3.4GHz CPU and 8 GB main memory with windows 7 operating system. To evaluate the performance of the clustering algorithms, the average silhouette index and execution time have been used.

Silhouette index is a cluster validity index that is usually used to judge the quality of clustering. The silhouette index for each element in dataset is a measure of how similar that element is to other elements in its cluster compared to elements in other clusters [1]. Its values rage from -1 to +1. The higher the values, better the clustering quality. The silhouette index of the element X^i of a cluster S^j is defined as:

$$\text{silhouette}(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

where a(i) is the average distance (similarity) between the X^i and all of the objects in cluster S^j; 'max' is the maximum operator, and b(i) is the minimum average distance between the X^i and the rest of the objects in all the clusters. However, the distance measure used in this paper is the Euclidean distance.

V. RESULTS AND DISCUSSION

In the analysis of our algorithm, Table 1 and table 2 show the number of reduced dimensions for PCA-EM algorithm and the time required for execution for EM and PCA-EM algorithms respectively. As an overview, it can be clearly seen from Table 2 that the PCA-EM algorithm has by far better time consumption. When the data size increases to 40,000 with 1500 dimensions, the standard EM clustering algorithm is unable to extract any useful clusters within reasonable time limits. On an average the EM algorithm takes more than 70 times the PCA-EM algorithm to perform the same amount of work.

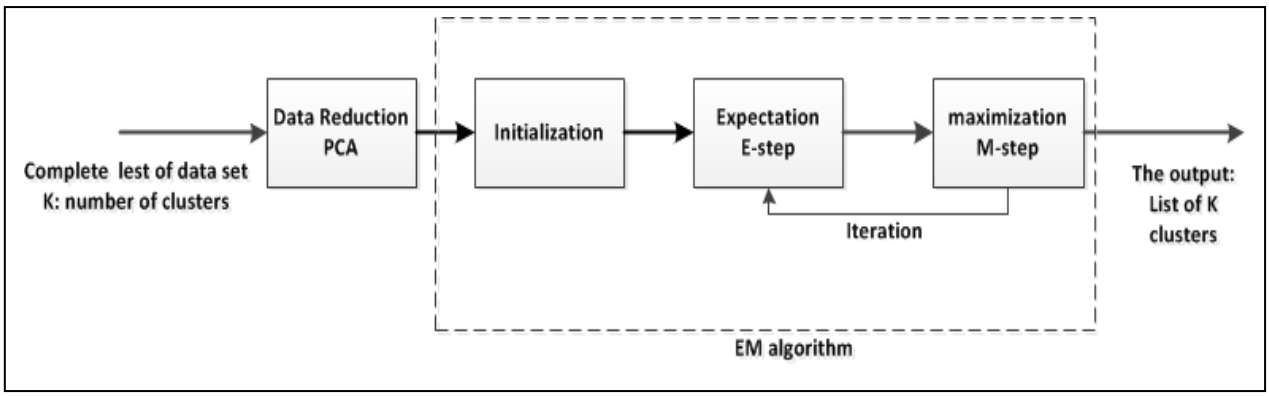

Fig. 2. The model of proposed algorithm

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

TABLE I. NUMBER OF REDUCED DIMENSIONS FOR PCA-EM ALGORITHM

Synthetic data of 1500 Dimension, number of clusters 500										
Data Size	1K	5K	10K	15K	20K	25K	30K	35K	40K	45K
No. of reduced dimensions for PCA-EM algorithm	11	35	53	69	79	83	92	81	91	127

TABLE II. PERFORMANCE OF EM AND PCA- EM ALGORITHM

Synthetic data of 1500 Dimension, number of clusters 500										
Data Size		1K	5K	10K	15K	20K	25K	30K	35K	40K	45K
Clustering time (min)	EM	2.16	134.78	479.19	501.40	425.12	477.36	658.03	896.73	-	-
	PCA-EM algorithm	0.02	0.23	1.07	2.97	4.79	6.20	10.07	11.60	18.09	30.90

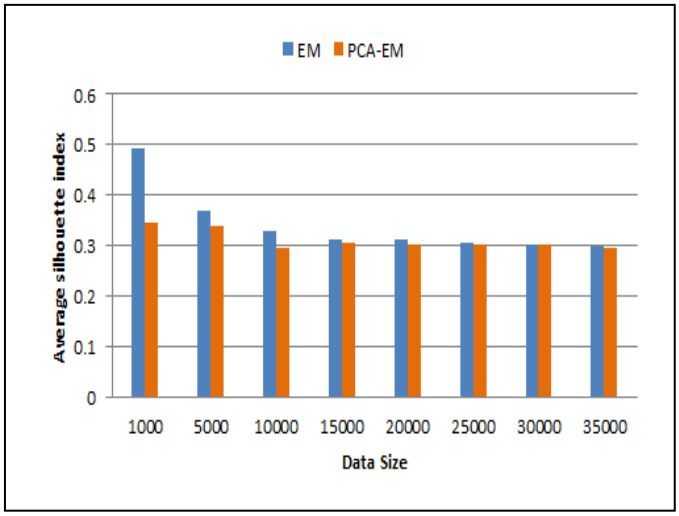

Fig. 3. Average silhouette index

Primarily due to the fact that our algorithm finds the clusters by executing EM algorithm not on the entire dimension on the data set, it does it on the top eigenvectors that represent 95% of features in each observation which means less running time required. Furthermore, to see the effect of the dimension reduction in our proposed algorithm and judge the quality of clustering results an average silhouette index has been calculated for each cluster. However, when comparing EM algorithm with PCA-EM algorithm in terms average silhouette index, (see Fig.3), PCA-EM algorithm performs slightly worse than EM algorithm, when data size is small but the performance differences becomes negligible when the size of data gets bigger.

CONCLUSION

With the prolific growth of Big Data, research on clustering of high dimensional data sets gained a lot of importance. There is a growing need for developing research around algorithms that are able to efficiently filter, structure and analyze Big Data at a very fast rate. So, in this paper, we have developed a new set of algorithms for clustering of high dimensional dataset. The developed set of algorithms combines the PCA and EM algorithms, PCA is used to reduce the data set from high dimensional to a meaningful representation (lower) dimensional. Also, k-means algorithm is used as an initialization method to improve the performance of clustering.

Our experimental results clearly show that as far as time consumption is considered, the proposed algorithm performs significantly better than EM algorithms. In terms of average silhouette index, the proposed algorithm has almost the same clustering quality in experiments with big data size. This indeed indicates that our proposed algorithm is applicable to the application with large data set where response time is critical.

ACKNOWLEDGMENT

We would like to express our thanks to school of electrical and electronic engineering at Liverpool John Moores University for their continuous support and encouragement during this work.

REFERENCES

[1] B.Barileé, “More Work on K -Means Clustering Algorithm: The Dimensionality Problem, ” International Journal of Computer Applications (0975 – 8887),Volume 44– No.2, April 2012.

[2] B.Paul, F.Usama and R.Cory, “ Scaling EM (Expectation-Maximization) Clustering to Large Databases, ” Technical Report MSR-TR-98-35. 1998.

[3] D. Rajashree ,M. Debahuti ,R Amiya ,A Milu, “ A hybridized K-means clustering approach for high dimensional dataset, ” International Journal of Engineering, Science and Technology Vol. 2, No. 2, 2010, pp. 59-66

[4] F.Adil, A.Najlaa and T.Zahir, “ A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis, ” IEEE Transactions on Emerging Topics in Computing publishes papers on emerging aspects of computer science, computing technology, and computing applications not currently covered by other IEEE Computer Society Transactions.2014.

[5] F.Jerome, “Data mining and statistics what is the connection, ” Computing Science and Statistics. Proceedings of the 29th Symposium on the Interface. Interface Foundation of North America. (1998)

[6] H.victoria and A.jim, “Discretisation of data in a binary neural k-nearest neighbour algorithm, ”

[7] J.Anil, “ Data Clustering: 50 Years Beyond K-Means, ” Pattern Recognition Letters Volume Volume 31 Issue 8, June, 2010 Pages 651-666

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

[8] J.Dong, Z.Caroline, R.William and C.Remco, "Understanding Principal Component Analysis Using a Visual Analytics Tool, " Charlotte Visualization Center, UNC Charlotte. 2009.

[9] J.Julie. and H.François, " Selecting the number of components in principal component analysis using cross-validation approximations, " Computational Statistics and Data Analysis 56 (2012) 1869–1879.

[10] M.Ian and W. Mark, " Linked datav Connecting and exploiting big data. White Paper Linked Data: March 2012.

[11] R,Mark, "Principal Component Analysis, "May 2009.

[12] Total Lab [Online] Availablet:http://www.totallab.com/ products/samespots/support/faq/ pca.aspx. [Accessed December 2014].

[13] [1]G, John and R, David "The digital universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East, Study report, IDC, " [Online] available at: www.emc.com/leadership/digital-universe/index.htm [Accessed December 2014].

(This work is licensed under a Creative Commons Attribution 4.0 International License.)