Advanced Science and Technology Letters
Vol.145 (NGCIT 2017), pp.108-111
http://dx.doi.org/10.14257/astl.2017.145.21

A Novel of Comparative Evaluation of Unsupervised Learning Classification Method for Density Data Analysis

Se-Hoon Jung[1*], Kyoung-Jong Kim[2], and Chun-Bo Sim[3]

[1*,2] Dept. of Multimedia Engineering, Sunchon National University, Suncheon, Korea
[3] School of Information Communication and Multimedia Engineering, Sunchon National University, Suncheon, Korea
E-mail : iam1710@hanmail.net, kkj0201@nate.com, cbsim@sunchon.ac.kr

Abstract. This study set out to analyze and compare data of high density within the categories of various Big Data by applying altered K-means and DBSCAN algorithms. The study would propose an algorithm to extract K, the optimal number of clusters, through principal component analysis to supplement the K-means algorithm for its difficulty with the analysis of concentrated data compared with the DBSCAN algorithm and an algorithm to extract an initial center point according to the standard normal distribution to choose an initial cluster center point.

Keywords: K-means, DBSCAN, Big Data, Density, Analysis

1 Introduction

As the production cycle of data grows increasingly shorter in recent years, there is more and more interest in Big Data and the areas of Big Data-based information search and analysis[1]. Data classification and analysis techniques for such Big Data information search are ongoing research topics. Previous studies, however, focused on the algorithms optimized for the classification and analysis of old small-sized data instead of Big Data to develop an algorithm of data classification and analysis. The K-means and DBSCAN algorithms are, in particular, representative ones in the clustering of unsupervised learning. The two algorithms are used to classify data with no labels. While the K-means algorithm proceeds with clustering according to K, the predetermined number of clusters by an analyst, the DBSCAN algorithm uses density without determining the number of clusters in advance based on the distributed data. The latter has an advantage over the former since it raises no need for an analyst to determine the number of clusters in advance, but it does have its share of problems including its lower performance when classifying data of different densities, high-dimensional data of lower density, or Big Data. The present study thus proposed an algorithm capable of classifying dense data by altering the K-means algorithm, an unsupervised learning technique, in a Big Data environment and compared and evaluated the proposed altered K-means algorithm with the DBSCAN algorithm. The

ISSN: 2287-1233 ASTL

altered K-means algorithm was proposed as an algorithm to supplement the problems with the old K-means algorithm and be optimized for Big Data analysis.

2 Related Work

2.1 K-means

K-means algorithm is a clustering technique to classify input data into k clusters based on unsupervised learning similar to supervised learning. Unlike supervised learning, which updates weight vectors every time a vector is entered, the K-means algorithm updates weight vectors simultaneously after all the input vectors are entered. The criteria of clustering classification are the distance between clusters, dissimilarity among clusters, and minimization of the same cost functions. Similarity between data objects increases within the same clusters. Similarity to data objects in other clusters decreases. The algorithm performs clustering by setting the centroid of each cluster and the sum of squares between data objects and distance as cost functions and minimizing the cost function values to repeat cluster classification of each data object.

3 Proposed K-means Algorithm

The present study proposed an algorithm that supplemented the problems with the old K-means algorithm to analyze and classify dense Big Data efficiently. Fig. 1 presents the altered K-means algorithm proposed in the study. It is further divided into two major types: one for the extraction of K, the number of clusters, and the other for the determination of initial central point.

① Principal components are extracted until the point where a certain value to explain the entire data is maintained through the principal component analysis of the entire entry data objects.

② The division of center point technique is applied to C_{kl}, the number of random clusters, and n_k, the number of randomly selected center points, based on the principal components extracted through principal component analysis. m_k, the center point of each initial cluster, is measured with a random cluster index vector.

③ All the vector data are measured for mean , standard deviation , and standard normal distribution $\phi_{u,\sigma^2}(x_k u_k)$.

④ Objects are selected in the space quantile and $P(\overline{X} \geq x_k u_k) = \pm 0.95$ with , the mean of vector data, and σ, the standard deviation of vector data.

④-① When there is an object distributed, the object will be assigned to , the center point of C_1, the first cluster.

④-② When there are two or more objects distributed, the one whose two vectors have the biggest length will be first assigned to m_1, the center point of C_1, the first cluster.

⑤ The distance(A_i) between the remaining object(x_i) and m_1, the center point of C_1, the first cluster, is measured. The object whose distance measurement is the biggest will be assigned to m_2, the center point of C_2, the second cluster.

Fig. 1. The Algorithm presents the altered K-means algorithm proposed in the study.

Big Data that are entered in an initial stage are provided in a multi-dimensional format, and the dimensions of Big Data are reduced through principal component analysis. The covariance of principal component analysis is used to check connections between the clusters chosen from the entire given data and other clusters and designate the scope of K, the maximum number of clusters. The number of the lowest dis-similarity(Euclidean distance) within the scope of K is determined as K. An initial cluster center point is chosen according to the determined K, and the algorithm chooses an initial center point based on the predicted point of an outlier and space division. When some of all the given objects record a lower density and similarity value than other objects, the probability of an outlier increases. Similarity between objects can be measured by comparing numbers with the squared Euclidean distance. When the mean distance from all the objects except for the initial cluster center point is bigger than each cluster mean, the probability of an outlier increases. When the mean distance from surrounding objects is big, the density decreases and the probability of an outlier increases.

4 System Comparative Evaluation

The present study tried to overcome the problems of K-means algorithm, an unsupervised learning

The present study proposed an altered algorithm to supplement the old K-mean algorithm for its problems including the K, the number of clusters, and initial center point choice. The DBSCAN algorithm is universally investigated as an unsupervised learning-based classification technique that does not determine K, the number of clusters, for clustering. In the present study, an experiment was conducted with a comparison and evaluation algorithm to assess the performance of the proposed algorithm. The experiment environment for performance evaluation included Windows 7 64bit, RAM 16GB, and Visual Studio 2012 C#. Total 60 sports articles on baseball, football, and basketball were collected to establish data sets for performance evaluation. The preprocessing stage involved the elimination of stopwords and the extraction of nouns and extracted 303 words, which were categorized into baseball(103), football(113), and basketball(87). The extracted words were substituted with numbers for evaluation. Table 1 Table 2 shows the results of applying the 303 words to the proposed and DBSCAN algorithms three times. The measurements show that the proposed algorithm automatically extracted three cluster analyses in all of

three experiments, recording a classification rate of mean 98%. The DBSCAN algorithm recorded a classification rate of mean 92%. It automatically extracted the number of clusters based on density and made an error in the extraction process in two of three experiments. As for unclassified data noises, the proposed algorithm recorded mean 2.2% or total 20 data noises in the experiments, whereas the DBSCAN algorithm recorded mean 8.1% or total 74 data noises.

Table 1. Altered K-means algorithm(Proposed algorithm) Result .

Cluster Number	Classification	Noise
3	294	9
3	299	4
3	296	7

Table 2. DBSCAN Result.

Cluster Number	Classification	Noise
2	269	34
5	277	26
3	289	14

5 Conclusion

The present study tried to overcome the problems of K-means algorithm, an unsupervised learning technique, with the proposed algorithm in a high-density Big Data environment and compare the proposed algorithm with the DBSCAN algorithm of high density classification performance. Follow-up study will assess the proposed algorithm and DBSCAN algorithm in terms of performance.

Acknowledgments. The research was supported by 'Area Software Convergence Commercialization Program', through the Ministry of Science, ICT and Future Planning(S0417161012).

References

1. S. H. Jung, J. C. Kim, C. B. Sim,: Prediction Data Processing Scheme using an Artificial Neural Network and Data Clustering for Big Data, International Journal of Electrical and Computer Engineering, vol.6, no.1, pp.330--336(2016)

