International Journal of Soft Computing 4 (4): 168-172, 2009
ISSN: 1816-9503


# Impact of Normalization in Distributed K-Means Clustering

[1]N. Karthikeyani Visalakshi and [2]K. Thangavel
[1]Department of Computer Science, Vellalar College for Women, Erode, India
[2]Department of Computer Science, Periyar University, Salem, India

**Abstract:** Distributed clustering is an emerging research area in the broader field of Knowledge discovery in databases. Normalization is an essential preprocessing step in data mining, to standardize values of all attributes or features from different dynamic range into a specified range. In this study, distributed K-Means clustering algorithm is extended by applying global normalization before performing the clustering on distributed datasets, without necessarily downloading all the data into a single site. The performance of proposed normalization based distributed K-Means clustering algorithm is compared against distributed K-Means clustering algorithm and normalization based centralized K-Means clustering algorithm. The quality of clustering is also compared by three normalization procedures, namely Min-max, Z-score and decimal scaling for the proposed distributed clustering algorithm. The comparative analysis shows that the distributed clustering results depend on the type of normalization procedure. The experiments are carried out for various numerical datasets of UCI machine learning data repository.

**Key words:** Distributed clustering, euclidean distance, global centroid, K-Means, local centroid, normalization

## INTRODUCTION

Clustering methods seek to organize a set of objects into clusters such that objects within a given cluster have a high degree of similarity, whereas objects belonging to different clusters have a high degree of dissimilarity. These methods have been widely applied in various areas such as taxonomy, image processing, information retrieval, data mining, etc. (Jain *et al*., 1999). Today's large-scale datasets are usually logically and physically distributed, requiring a distributed approach to clustering. Huge amounts of data are stored in autonomous, geographically distributed sources over networks with limited bandwidth and large number of computational resources (Folino *et al*., 2006).

Traditional clustering methods require all data to be located at the place, where they are analyzed and cannot be applied in the case of multiple distributed datasets, unless all data are transferred to a single location and clustered. Due to technical, economical or security reasons, it is not always possible to transmit all data from different local sites to single location and then perform global clustering. It is obvious that alternate distributed clustering algorithms (Ghosh and Merugu, 2003) reduce the communication overhead, central storage requirements and computation times by exchanging few data and avoiding synchronization as much as possible.

Most of the existing distributed clustering algorithms available in the literature (Sanghamitra *et al*., 2006; Jin *et al*., 2006; Jeong *et al*., 2007) aim to provide hard clusters based on K-Means algorithm. The K-Means (Jain *et al*., 1999) typically uses Euclidean or squared Euclidean distance to measure the distortion between a data object and its cluster centroid. These distances are usually computed from raw data and not from standardized data. While, using Euclidean distances, the distance between any two objects is not affected by the addition of new objects to the analysis. However, the clustering results can be greatly affected by differences in scale among the dimension from, which the distances are computed. Data normalization is the linear transformation of data to a specific range. Therefore, it is worthwhile to enhance clustering quality by normalizing the dynamic range of input data objects into specific range (de Souto *et al*., 2008).

The effects of normalization are evaluated for different conventional clustering methods like K-Means, fuzzy C-Means, Partitioning around Mediods and Hierarchical clustering and showed that the clustering results depend on the normalization method, but only in centralized environment (Doherty *et al*., 2007; Seo Young Kim and Toshimitsu, 2008; de Souto *et al*., 2008). To the best of our knowledge, there is no research on normalization in distributed clustering. Hence, an attempt is made in this study, to propose a novel Normalization based Distributed K-Means (NDKM) clustering algorithm, by extending Genlin's Distributed K-Means (DKM) algorithm (Genlin and Ling, 2007) with

**Corresponding Author:** N. Karthikeyani Visalakshi, Department of Computer Science, Vellalar College for Women, Erode, India

global normalization procedure. A comparative study is made on DKM, NDKM and Normalization based Centralized K-Means (NCKM) clustering, where all the data are merged into a single data source, normalized and clustered using K-Means algorithm. This study also takes an additional effort to compare the performance of three different normalization procedures (Luai *et al*., 2006), while clustering homogeneously distributed numerical datasets.

**K-Means clustering:** Traditionally, clustering algorithms have been classified into two categories: hierarchical and partitional. Commonly used algorithms in the hierarchical category are single linkage and complete linkage algorithms. K-Means is a widely used algorithm in the partitional class (Jain *et al*., 1999).

The K-Means method partitions the data set into K subsets such that all objects in a given subset are closest to the same centroid. In detail, it randomly selects K of the objects to represent the cluster centroids. Based on the selected objects, all remaining objects are assigned to their closer centroid one by one. The Euclidean distance between the object and every centroid is computed and the object is moved to one of the cluster centroid, which yields minimum distance. The value of selected centroid is recalculated by taking the mean of all data objects belonging to the same cluster. The operation is iterated for all the objects. If K cannot be known ahead of time, various values of K can be evaluated until the most suitable one is found.

**Distributed clustering:** Distributed clustering assumes that the objects to be clustered reside on different sites. This process is carried out in two different levels: local level and global level. In local level, all sites carry out clustering process independently from each other. After having completed the clustering, a local model such as cluster centroids is determined, which should reflect an optimum trade-off between complexity and accuracy. Next, the local model is transferred to a central site, where the local models are merged in order to form a global model. The resultant global model is again transmitted to local sites to update the local models (Januzaj *et al*., 2004).

The key idea of distributed clustering is to achieve a global clustering that is as good as the best centralized clustering algorithm with limited communication required to collect the local models or local representatives into a single location, regardless of the crucial choice of any clustering technique in local site. Distributed clustering algorithms (Sanghamitra *et al*., 2006) can be classified along two independent dimensions such as classification based on data distribution and classification based on data communication.

A common classification based on data distribution in the study (Ghosh and Merugu, 2003) is those, which apply to homogeneously distributed or heterogeneously distributed data. Homogeneous datasets contain the same set of attributes across distributed data sites. Examples include local weather, databases at different geographical locations and market-basket data collected at different locations of a grocery chain. Heterogeneous data model supports different data sites with different schemata. For example, a disease emergence detection problem may require collective information from a disease database, a demographic database and biological surveillance databases.

According to the type of data communication, distributed clustering algorithms are classified into two categories: multiple communications round algorithms and centralized ensemble-based algorithms. The first group consists of methods requiring multiple rounds of message passing. These methods require a significant amount of synchronization, whereas the second group works asynchronously. Many of the distributed clustering algorithms work in an asynchronous manner, first generating the local clusters and then combining those at the central site (Park and Kargupta, 2003).

**Normalization:** Preprocessing Luai *et al*. (2006) is often required before using any data mining algorithms to improve the results' performance. Data normalization is one of the preprocessing procedures in data mining, where the attribute data are scaled so as to fall within a small specified range such as -1.0 to 1.0 or 0.0 to 1.0. Normalization before clustering is specially needed for distance metric, such as Euclidian distance, which are sensitive to differences in the magnitude or scales of the attributes. In real applications, because of the differences in range of attributes' value, one attribute might overpower the other one. Normalization prevents outweighing attributes with large range like 'salary' over attributes with smaller range like age. The goal is to equalize the size or magnitude and the variability of these attributes.

There are many methods for data normalization, which include Min-max normalization, Z-score normalization and normalization by decimal scaling. Min-max normalization performs a linear transformation on the original data. Suppose that $min_a$ and $max_a$ are the minimum and the maximum values for attribute A. Min-max normalization maps a value v of A-v in the range (0, 1) by computing:

$$v' = \frac{v - min_a}{(max_a - min_a)} \quad (1)$$

In Z-score normalization, the values for an attribute A are normalized based on the mean and standard deviation of A. A value v of A is normalized to v by computing:

$$v' = \frac{v - \bar{A}}{\sigma_A} \quad (2)$$

where, $\bar{A}$ and $\sigma_A$ are the mean and the standard deviation, respectively of attribute A. This method of normalization is useful when the actual minimum and maximum of attribute A are unknown, or when there are outliers that dominate the min-max normalization.

Normalization by decimal scaling normalizes by moving the decimal point of values of attribute A. The number of decimal points moved depends on the maximum absolute value of A. A value v of A is normalized to v' by computing:

$$v' = \frac{v}{10^j} \quad (3)$$

where, j is the smallest integer such that Max(|v'|)<1.

## MATERIALS AND METHODS

**Normalization based distributed K-Means clustering:** The Distributed K-Means algorithm is centralized ensemble based distributed clustering algorithm introduced in Genlin and Ling (2007) for clustering homogeneously distributed datasets. The DKM first does clustering in local site using K-Means, then sent all centroid values to central site, finally global centroid values of underlying global clustering are obtained by using K-Means again. The NDKM is extended version of DKM, where global normalization is performed to standardize the data objects into specific range. The step by step procedure of proposed algorithm is described in Fig. 1. First minimum and maximum values of each feature vectors are extracted from all local datasets and transmitted to central place, where global minimum and maximum values are identified. These two values are transmitted to local sites to perform global normalization using min-max normalization procedure. Next, normalized objects are clustered using K-Means algorithm to obtain centroids matrix and cluster index for each dataset. All local centroids are merged and clustered using K-Means algorithm, to group similar centroids and obtain global centroids. The global centroids is now transmitted to local site, where the Euclidean distance of each object from the global set of centroids are computed and assigned to the nearest cluster centroid.

The proposed algorithm can also be implemented for other types of normalization procedure, by simply modifying first two steps in the algorithm. In case of Z-score normalization method, global mean and standard deviation values are to be calculated based on local mean and standard deviation of each attribute from each dataset. Similarly, global maximum absolute value of each attribute is to be computed, to perform normalization on individual datasets.

**Algorithm:** NDKM
**Input:** Homogeneous p datasets, each with d dimensions
**Output:** Global partitions of p datasets
**Procedure:**
Step 1: Find maximum and minimum values of each feature from each local dataset and transmit them into central place
Step 2: Compute global maximum and minimum value at central place
Step 3: Normalize real scalar values of local datasets with global maximum and minimum values using Eq. 1
Step 4: Cluster each local dataset by K-Means algorithm and obtain centroids matrix along with cluster index for each dataset
Step 5: Merge cluster centroids of local datasets into a single dataset named as centroids dataset at central place
Step 6: Cluster centroids dataset using K-Means to obtain global centroids
Step 7: Update local cluster indices by assigning each object to nearest cluster centroid, after computing Euclidean distance between the object and global centroids

Fig. 1: Normalization based distributed K-Means algorithm

## RESULTS AND DISCUSSION

The main objective of this research is to explore the impact of normalization in the process of distributed K-Means clustering. The experimental analysis is performed with six benchmark datasets in two aspects. First, the efficiency of NDKM with Min-max normalization is compared against DKM and NCKM with Min-max normalization procedure. Next, the performance of three different normalization procedures is evaluated for NDKM. The information about the datasets available in the UCI machine learning data repository (Merz and Murphy, 1998) is shown in Table 1. The performance of the clustering algorithm is measured in terms of three external validity measures (Halkidi *et al*., 2002; Hui *et al*., 2006) namely rand index, F-measure and entropy. In case of F-measure and rand index, the value 1 indicates that the data clusters are exactly same. But, the value 0 indicates that the data clusters are exactly same for Entropy measure. For the purpose of experimental setup, the dataset is divided into three disjoint subsets and each subset is considered as distributed data source.

**Experiment 1:** This experiment is to show the significance of Min-max normalization in distributed clustering. The results of NDKM, in comparison with the results of DKM and NCKM, in terms of rand index, F-measure and

Table 1: Details of datasets

| Dataset | No. of attributes | No. of classes | No. of instances |
|---|---|---|---|
| Australian | 14 | 2 | 690 |
| Breast cancer | 10 | 2 | 699 |
| Mammography | 5 | 2 | 961 |
| Pen digit | 16 | 10 | 10992 |
| Satellite image | 36 | 7 | 6435 |
| Image segmentation | 19 | 7 | 2310 |

Table 2: Performance analysis of NDKM based on rand index

| Dataset | DKM | NDKM | CKM |
|---|---|---|---|
| Australian | 0.5071 | 0.6301 | 0.6089 |
| Breast cancer | 0.5209 | 0.9178 | 0.9178 |
| Mammography | 0.5606 | 0.6585 | 0.6538 |
| Pen digit | 0.8950 | 0.9039 | 0.9055 |
| Satellite image | 0.7606 | 0.8556 | 0.8572 |
| Image segmentation | 0.8155 | 0.8767 | 0.8605 |

Table 3: Performance analysis of NDKM based on F-Measure

| Dataset | DKM | NDKM | CKM |
|---|---|---|---|
| Australian | 0.6695 | 0.7549 | 0.7327 |
| Breast cancer | 0.6195 | 0.9569 | 0.9569 |
| Mammography | 0.6713 | 0.7761 | 0.7766 |
| Pen digit | 0.6695 | 0.7549 | 0.7327 |
| Satellite image | 0.5598 | 0.7107 | 0.7107 |
| Image segmentation | 0.5128 | 0.7054 | 0.6146 |

Table 4: Performance analysis of NDKM based on entropy

| Dataset | DKM | NDKM | CKM |
|---|---|---|---|
| Australian | 0.6835 | 0.5552 | 0.5787 |
| Breast cancer | 0.6416 | 0.1770 | 0.1770 |
| Mammography | 0.6220 | 0.5126 | 0.5129 |
| Pen digit | 0.8241 | 0.7911 | 0.7600 |
| Satellite image | 1.0182 | 0.6752 | 0.6644 |
| Image segmentation | 1.1099 | 0.7412 | 0.8125 |

entropy are shown in Table 2-4, respectively. From the Table 2-4, it is observed that NDKM algorithm yields better results than DKM for all datasets in terms of rand index, F-measure and entropy. The values of all three measures are highly appreciable for breast cancer datasets, with NDKM algorithm. It is noted that the quality of clusters produced by NDKM is as good as NCKM for all datasets, in terms of all three measures. Moreover, the performance of NDKM is slightly higher than the NCKM for Australian and Image Segmentation datasets.

**Experiment 2:** The second experiment compares the quality of clusters obtained by NDKM with min-max normalization procedure against two other normalization procedures Z-score and decimal scaling. The domino effect of three normalization procedures based on the external validity measures, rand index, F-measure and entropy is shown in Table 5-7, respectively. From these Table 5-7, the following observations are identified. All three normalization procedures produce almost same quality clusters for Breast cancer, Mammography and Pen digit datasets. Both min-max and Z-score methods yield better performance than Decimal Scaling for Satellite image dataset. The Min-max is suitable for Image Segmentation dataset, whereas Z-score is suitable for Australian dataset. Hence, it is identified that there is no unique normalization procedure, which yields better quality clusters for all datasets and so every dataset supports specific normalization method. In general, it is also observed that the methods Min-max and Z-score are suitable for more datasets than Decimal Scaling. So, it may be concluded that the selection of normalization method depends on specific domain.

Table 5: Comparative analysis of normalization methods based on rand index

| Dataset | Min-max | Z-score | Decimal scaling |
|---|---|---|---|
| Australian | 0.6301 | 0.7273 | 0.5003 |
| Breast cancer | 0.9178 | 0.9152 | 0.9178 |
| Mammography | 0.6545 | 0.6618 | 0.6622 |
| Pen digit | 0.9039 | 0.9012 | 0.9034 |
| Satellite image | 0.8556 | 0.8570 | 0.8065 |
| Image segmentation | 0.8767 | 0.8268 | 0.8020 |

Table 6: Comparative analysis of normalization methods based on F-measure

| Dataset | Min-max | Z-score | Decimal scaling |
|---|---|---|---|
| Australian | 0.7549 | 0.8349 | 0.6428 |
| Breast cancer | 0.9569 | 0.9555 | 0.9569 |
| Mammography | 0.7761 | 0.7827 | 0.7828 |
| Pen digit | 0.7549 | 0.6724 | 0.6846 |
| Satellite image | 0.7107 | 0.7138 | 0.6116 |
| Image segmentation | 0.7054 | 0.6291 | 0.5705 |

Table 7: Comparative analysis of normalization methods based on entropy

| Dataset | Min-max | Z-score | Decimal scaling |
|---|---|---|---|
| Australian | 0.5552 | 0.4407 | 0.6685 |
| Breast cancer | 0.1770 | 0.1812 | 0.1770 |
| Mammography | 0.5126 | 0.5052 | 0.5044 |
| Pen digit | 0.7911 | 0.7858 | 0.7713 |
| Satellite image | 0.6752 | 0.6729 | 0.8865 |
| Image segmentation | 0.7412 | 0.8439 | 0.9991 |

## CONCLUSION

A novel method of distributed K-Means clustering using global normalization is proposed to produce optimum quality clusters in distributed environment. Comprehensive experiments on six benchmark numerical datasets have been conducted to study the impact of normalization and to compare the effect of three different normalization procedures in distributed K-Means clustering. It can be concluded that the normalization before distributed clustering leads to obtain better quality clusters. Also, it is important to select specific normalization procedure, according to the nature of datasets. In future, appropriate normalization procedure can be applied in distributed fuzzy clustering and its variants to improve the performance.

## REFERENCES


de Souto, M.C.P., D.S.A. De Araujo, I.G. Costa, R. Soares, T.B. Ludermir and A. Schliep, 2008. Comparative study on normalization procedures for cluster analysis of gene expression datasets. Proc. IEEE Int. Joint Conf. Neural Networks, Hong Kong, pp: 2792-2798.

Doherty, K.A.J., R.G. Adams and N. Davey, 2007. Unsupervised Learning with Normalised Data and Non-Euclidean Norms. Applied Soft Comput., 7 (1): 203-210.

Folino, G., A. Forestiero and G. Spezzano, 2006. Swarm-based distributed clustering, peer-to-peer systems, artificial evolution, Lecture Notes in Computer Science, Springer-Verlag, pp: 37-48.

Ghosh, J. and S. Merugu, 2003. Distributed Clustering with Limited Knowledge Sharing. Proc. 5th Int. Conf. Adv. Pattern Recognition, Calcutta, India, pp: 48-53.

Genlin, J. and Ling Xiaohan, 2007. Ensemble Learning Based Distributed Clustering. In: Washio, T. *et al.* (Eds.). Emerging Technology and Knowledge Discovery and Data Mining, LNCS. Springer-Verlag, pp: 312-321.

Halkidi, M., Y. Batistakis and M. Vazirgiannis, 2002. Cluster Validity Methods: part II, ACM SIGMOD Rec., 31 (3): 19-27.

Hui Xiong, Junjie Wu and Jian Chen, 2006. K-Means Clustering versus Validation Measures: A Data Distribution Perspective. Proc. 12th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining. Philadelphia, PA, USA, pp: 779-878.

Jain, A.K., M.N. Murthy and P.J. Flynn, 1999. Data Clustering. A Rev. ACM Comput. Surveys, 31 (3): 265-323.

Januzaj, E., P. Kriegel Hans and M. Pfeifle, 2004. DBDC: Density based Distributed Clustering, Advances in Databases Technology-EDBT 2004, LNCS, In: Bertino, E., S. Christodoulakis and D. Plexousakis (Eds.). Springer Berlin/ Heidelberg, pp: 529-530.

Jeong, J., B. Ryu, D. Shin and D. Shin, 2007. Integration of Distributed Biological Data using Modified K-Means Algorithm. In: Washio T. *et al.* (Eds.). Emerging Technologies in Knowledge Discovery and Data Mining, LNCS, Springer-Berlin, pp: 469-475.

Jin, R., A. Goswami and G. Agarwal, 2006. Fast and Exact Out-of-Core and Distributed K-Means Clustering. Knowledge and Inform. Syst., 10 (1): 17-40.

Luai A. Shalabi, Zyad Shaaban and Basel Kassabeh, 2006. Data Mining A Preprocessing Engine. J. Computer Sci., 2 (9): 735-739.

Merz, C.J. and P.M. Murphy, 1998. UCI Repository of Machine Learning Databases, Irvine, University of California, <http://www.ics.uci.eedu/~mlearn/>.

Park, B. and H. Kargupta, 2003. Distributed Data Mining, The Hand Book of Data Mining. In: Nong Ye (Ed.). Lawrence Erlabum Associates, Publishers, Mahwah, New Jersey, pp: 341-358.

Sanghamitra, B., C. Giannella, U. Maulik, H. Kargupta, K. Liu and S. Datta, 2006. Clustering distributed data streams in peer-to-peer environments. Inform. Sci., 176 (4): 1952-1985.

Seo Young Kim and Toshimitsu Hamasaki, 2008. Evaluation of Clustering based on Preprocessing in Gene Expression Data. Int. J. Biol. Biomed. Med. Sci., pp: 48-53.