ISSN: 2278-4632
Vol-9 Issue-4 No.01 April 2019

INCREMENTAL CLUSTERING ALGORITHM WITH A NOVEL SIMILARITY MEASURE

Debasish Pradhan1, Ananya PreetiPadma2, Chandrakanta Pradhan[3]

[1]Asst. Prof. Einstein Academy of Technology and Management, Bhubaneswar, India
[2]Asst. Prof. Einstein Academy of Technology and Management, Bhubaneswar, India
[3]Student, Einstein Academy of Technology and Management, Bhubaneswar, India

Abstract

Clustering is one of data mining and text mining techniques which is used to analyze datasets by dividing it into meaningful groups. The objects in the dataset can have certain relationships among them. All clustering algorithms assume this before they are applied to datasets. The existing algorithms for text mining make use of a single viewpoint for measuring similarity between objects. Their drawback is that the clusters can't exhibit the complete set of relationships among objects. To overcome this drawback, we propose a new similarity measure known as multi-viewpoint based similarity measure to ensure the clusters show all relationships among objects. We also proposed two clustering methods. The empirical study revealed that the hypothesis "multi-viewpoint similarity can bring about more informative relationships among objects and thus more meaningful clusters are formed" is proved to be correct and it can be used in the real time applications where text documents are to be searched or processed frequently.

Keywords: Data mining, text mining, similarity measure.

1. Introduction

Data mining is a process of analyzing data in order to bring about trends or patterns from the data. Many techniques are part of data mining. Other mining such as text mining and web mining also exists. Clustering is one of the important data mining or text mining algorithm that is used to group similar objects together. In other words, it is used to organize given objects into some meaningful sub groups that make fur- ther analysis on data easier. Clustered groups make search mechanisms easy and reduce the bulk of operations and computational cost. Many clustering algorithms have been around since the inception of data mining domain. They are used based on the kind of application. One such clustering algorithm being used widely by the IT industry is k- means. It still remains in the top list of widely used clustering algorithms in the world. It has many variants as well.

Basically its functionality is similar. It takes two arguments and forms clusters. The first argument is data set or objects to be clustered while the second argument is the number of clusters to be formed. It has wide range of applications. One such application is credit card fraud detection. In such application, it generates clusters offline and makes a model. And then new transactions are simply added to the model which has clusters indicating high, low and medium range transactions. When a new transaction takes place, it can compare with the general buying patterns of customer and can detect abnormality. Any abnormality is suspected to be a fraudulent transaction. According to also k-means is the most favourite clustering algorithms in the data mining domain. Never- the less, it has its own drawbacks that are well known to the world. They are sensitiveness to cluster size, sensitiveness to initialization; its performance is lesser than many other clustering techniques used in the data mining domain. Provided these drawbacks, it is still considered popular due to its simplicity, scalability and understand ability. As it is less complex with adequate performance, it is widely used in the industry overlooking its known limitations.

Another important quality of k-means algorithm is that it can be easily combined with other algorithms for best results. Generally the problem of clustering can be thought as optimization process. By optimizing similarity measures the optimal clusters can be formed thus performance is improved. Therefore the soundness of clustering algorithms depends on their similarity measure adopted. To meet various requirements k-means has many variants. For instance spherical k-means (uses cosine similarity) is used to cluster text documents while original k-means can be used to clustering using Euclidean distance [3].

According to Leo Wanner, clustering methods are classified into hierarchical clustering, data partitioning, data grouping. The hierarchical clustering is used to establish cluster taxonomy. Data partitioning is used to build a set of flat partitions.

ISSN: 2278-4632
Vol-9 Issue-4 No.01 April 2019

They are also known as non-overlapping clusters. Data group is used to build a set of flat or overlapping clusters. The proposed work in this paper is motivated by the facts ascertained by investigation of the above. Especially similarity measures are considered. From research findings it is understood that the nature of similarity measured used in any clustering technique has profound impact on the results. The aim of the paper is to develop a new method that is used to cluster text documents that have sparse and high dimensional data objects. Afterwards we formulate new clustering criterion functions and corresponding clustering algorithms respectively. Like k-means the proposed algorithms work faster and provide consistent, high quality performance in the process of clustering text documents. The proposed similarity measure is based on multi-viewpoint which is elaborated in the later sections.

2. Multi-View Point Based Similarity

Our approach in finding similarity between documents or objects while performing clustering is multi-view based similarity. It makes use of more than one point of reference as opposed to existing algorithms used for clustering text documents. As per our approach the similarity between two documents is calculated as:

$$Sim(di,dj) = 1/n\text{-}nr \; \Sigma \; Sim\,(di\text{-}dh,\ dj\text{-}dh) \quad (5)$$

$$dt,dj \in Srdh \in S\backslash Sr$$

Here is the description of this approach. Consider two point di and dj in cluster Sr. The similarity between those two points is viewed from a point dh which is outside the cluster. Such similarity is equal to the product of cosine angle between those points with respect to Eucludean distance between the points. An assumption on which this definition is based on is “dh is not the same cluster as di and dj. When distances are smaller the chances are higher that the dh is in the same cluster. Though various viewpoints are useful in increasing the accuracy of similarity measure there is a possibility of having that give negative result. However the possibility of such drawback can be ignored provided plenty of documents to be clustered.

3. Algorithms Proposed

A series of algorithms are proposed to achieve MVS (Multi-
View point Similarity).

1: procedure BUILDMVSMATRIX(A)2: for r ← 1 : c do
3: DSISrI←_di ΣSr di4: nSISr ← |S ISr|
5: end for
6: for i ← 1 : n do7: r ← class of di 8: for j ← 1 : n do9: if dj Sr then
10: aij ← dti dj – dti DSISr nSISr – dt j DS\Sr nSISr + 1
11: else
12: aij←dti dj−dti DSISr –dj nSISr −1–dt j DSISr –dj nSISr −1
13: end if
14: end for
15: end for
16: return A = {aij}n×n17: end procedure

From the condition it is understood that when ***di*** is consideredcloser to ***dl***, the ***dl*** can still be considered being closer to ***di*** asper MVS. For validation purpose listing 2 is used.

Require: 0 < percentage ≤ 1
1: procedure GETVALIDITY(validity,A, percentage)2: for r ← 1 : c do
3: qr ← _percentage × nr
4: if qr = 0 then _ percentage too small5: qr ← 1
6: end if
7: end for
8: for i ← 1 : n do
9: {aiv[1], . . . , aiv[n] } ←Sort {ai1, . . . , ain}
10: s.t. aiv[1] ≥ aiv[2] ≥ . . . ≥ aiv[n] {v[1], . . . , v[n]} ← permute
{1, . . . , n}
11: r ← class of di
12: validity(di) ← |{dv[1], . . . , dv[qr] } ∩ Sr| qr
13: end for
14: validity ← _ni←1 validity(di)n15: return validity
16: end procedure

This process is terminated when iteration finds no document to be moved to new clusters.

ISSN: 2278-4632
Vol-9 Issue-4 No.01 April 2019

4. Performance Evaluation of Mvs

As part of the performance evaluation, the comparison is made between MVSC Ir, MVSC Iv with existing algorithms. The document database, data corpora, has benchmark data- sets for clustering purposes.

Experimental Setup and Evaluation

To demonstrate MVSCs we compared them with 5 other clus- tering algorithms. All the clustering algorithms used in evalu- ation are:

- MVSC Ir : MVSC with criterion function Ir
- MVSC Iv : MVSC with criterion function Iv
- K-means : conventional k-means with Ecludean distance
- Spkmeans: Spherical k-means with CS
- graphCS : CLUTO's graph method with CS
- graphEJ: CLUTO's graph with extended Jaccard
- MMC: Min Max Cut algorithm

5. Results

The experimental results are shown in fig. 2 and fig. 3 for all clustering algorithms using 20 bench mark document data bases. As the results are not fit into one graph they are split into two graphs and eash graph shows results with 10 data-sets. It is evident that with respect to many data sets MVSC is performing better. In some cases only other algorithms like graphEJ performed well. Both MVSC Ir and MVSC Iv outperform many other existing algorithms in most of the cases. As part of experiments we also present the effect of on the performance of MVSC Ir.

The Effect Of On The Performance Of MVSC Ir

Cluster size and balance have impact on the partitional clustering methods that are based on criterion functions. Based on the clustering results in Accuracy, FScore and NMI, this assessment is done. MVSR Ir's performance worst at 0 and 1 while it has significant performance improvement in the middle. MVSR Ir performs within 5% of the best case with respect to any type of evaluation metrics.

6. Conclusion

In this paper we proposed a new similarity measure known as MVS (Multi-Viewpoint based Similarity). When it is compared with cosine similarity, MVS is more useful for finding the similarity of text documents. The empirical results and analysis revealed that the proposed scheme for similarity measure is efficient and it can be used in the real time applications in the text mining domain. IR and IV are the two criterion functions proposed based on MVS. Their respective clustering algorithms are also introduced. The proposed scheme is tested with large datasets with various evolution metrics. The results reveal that the clustering algorithm provides performance that is better than many state – of – the - art clustering algorithms. Similarity measure from multiple viewpoints is the main contrition of this paper. The paper also provides partitioned clustering that can be applied on documents. The future work is that the proposed algorithms can be altered and applied to hierarchical clustering. Our novel approach to measure document similarity is described in the following sections.

References

[1] X. Wu, V. Kumar, J.R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G.J. McLachlan, A. Ng, B. Liu, P.S. Yu, Z.-H. Zhou, M. Steinbach, D.J. Hand, and D. Steinberg, "Top 10 Algorithms in Data Mining," Knowledge Information Systems, vol. 14, no. 1, pp. 1-37, 2007.
[2] I. Guyon, U.V. Luxburg, and R.C. Williamson, "Clustering: Science or Art?," Proc. NIPS Workshop Clustering Theory, 2009.
[3] I. Dhillon and D. Modha, "Concept Decompositions for Large Sparse Text Data Using Clustering," Machine Learning, vol. 42, nos. 1/2, pp. 143-175, Jan. 2001.
[4] S. Zhong, "Efficient Online Spherical K-means Clustering," Proc. IEEE Int'l Joint Conf. Neural Networks (IJCNN), pp. 3180-3185, 2005.