EFFICIENT CLUSTERING OF HIGH DIMENSIONAL DATASETS WITH MULTI VIEWPOINT BASED SIMILARITY MEASURE

M. S. ANBARASI¹, MUGILAN², RAJTHILAK² AND RAMKUMAR²

¹Assistant Professor, Department of Information Technology
²Final Year Information Technology
Pondicherry Engineering College, Puducherry, India
E-mails: anbarasims@gmail.com, mugilan_anamica121@yahoo.com, rajthilak90@gmail.com, rmsg99kumar@gmail.com

Abstract: Many important real time applications involve clustering large datasets. Dataset can be large if there are a large number of elements in the data set, each element can have many features and there can be many clusters to discover. Recent advances in clustering algorithms have been addressed these datasets issues partially. However, there has been much less work on methods of efficiently clustering datasets that are high dimension. Clustering requires more robust Similarity or Dissimilarity Measure. In this paper we introduce Multi-viewpoint Similarity Measure. Using Multiple Viewpoints, more informative assessment of Similarity could be achieved. The advantage of the proposed clustering is predicted by comparing with well known clustering algorithm that uses different similarity measure using the three efficiency metrics Accuracy, NMI and FScore.

Keywords: Similarity Measure, Image clustering, Hierarchical Agglomerative Algorithm, Evaluation Metrics

1. INTRODUCTION

As patients’ disease increases in huge number every year, a time efficient diagnosis is required to make the doctors to diagnose the patient in quicker manner. So, it is required to develop the system to avoid travelling time and waiting time of the patient and a single medical expert to diagnose and treat distributed group of patients of various category.

All clustering methods have to assume some cluster relationship among the data objects that they are applied on. Similarity between a pair of objects can be either explicitly or implicitly. The traditional dissimilarity/similarity measure [13] [10] uses only a single viewpoint, which is the origin, while the recent clustering techniques utilizes many different viewpoints, which are objects, assumed to not be in the same cluster with the two objects being measured. By using multiple viewpoints [3], more informative assessment of similarity could be achieved. The nature of similarity measure plays a very important role in the success and failure of the clustering method.

The objective of our project is to collect medical data from distributed environment. In development phase, the collected medical data undergoes the process of standardization as first module. Processed data is made into medical guidelines and clustered into several groups based on the similarity. These guidelines are provided as real-time diagnostic information for the queried patient.

2. RELATED WORKS

In the existing system (DucThang Nguyen, Lihui Chen, 2011) two real world text documents, Reuters 21578 and 2350 yahoo WebPages are taken as input to the system. The Documents are pre-processed using stop word removal and streaming and are further normalized to unit. A novel method for measuring similarity between data objects in sparse and high-dimensional domain [7][17], particularly text documents. From the proposed similarity measure, we then formulate new clustering criterion functions and introduce their respective clustering algorithms.

Cluster size weighted average of pair wise similarities [8] of documents Ir and Iv known as the criterion function are calculated. From the criterion function the Clustering Algorithm
MVSC-Ir and MVSC-Iv are implemented. Clustering [6] framework by MVSC ie, clustering with Multi-Viewpoint based Similarity. Subsequently, MVSC-Ir and MVSC-Iv, which are MVSC with criterion function Ir and Iv respectively. Clustering by optimization is performed by K-Way algorithm ie, Sequential K-Means algorithm.

During the optimization procedure, in each iteration, the main sources of computational cost are:

- Searching for optimum clusters to move individual documents to: O (nz · k)
- Updating composite vectors as a result of such moves: O (m · k)

The implemented Algorithm is compared with several clustering algorithm that uses different similarity measures [8]. Further Statistical testing is carried on by pairing the MVSC-Ir and MVSC-Iv with other algorithm for performance comparison.

3. LIMITATION ON EXISTING WORK

The detailed surveys on the given base papers were made and main drawback of the existing system is that the method is applied only for small text documents. The Drawback in this system is overcome in our system. When the dataset to be clustered is very large, traditional Clustering Algorithm [5][6] becomes expensive and there are many important problems involved in clustering large datasets. Till now there are very few works on the methods of efficiently clustering large datasets [12]. The motive of this paper is to provide a cheap approximate distance measure [1] for efficient clustering.

4. PROPOSED ECHDMVS SYSTEM

The high level architecture of the ECHDMVS system is shown below which further subdivided into several modules and explained below.

4.1. Input Module

The Patient Database is collected from several databases along with affected body part image (X-Ray or RGB Image) and is stored. The image of the patient is stored using the respective patient name in the database using the BLOB property. All the RGB images are transformed to standard gray scale image format of standard size 600*600. This is followed by the standardization of input images by the standard display function (Grayscale Standard Display Function) that specifies a standardized display function for consistent display of grayscale images. This function provides methods for calibrating a particular display system for the purpose of presenting images consistently on different display media.

4.2. Clustering

In this module, Multi Viewpoint Similarity Measure [3] is first implemented where; the two objects to be measured must be in the same cluster, while the points from where to establish this measurement must be outside of the cluster. The similarity between any two points in a particular cluster view from a reference point outside the cluster is equal to the cosine of angle between one of the point inside cluster and the reference point and Euclidean distance between the another point inside the cluster and the reference point.
The similarity between two documents $d_i$ and $d_j$ is determined w.r.t. the angle between the two points when looking from the origin. To construct a new concept of similarity, it is possible to use more than just one point of reference. We may have a more accurate assessment of how close or distant a pair of points is, if we look at them from many different viewpoints. From a third point $d_h$, the directions and distance to $d_i$ and $d_j$ are indicated respectively by different vectors $(d_i - d_h)$ and $(d_j - d_h)$. By standing at various reference point $d_h$ to view $d_i$ and $d_j$ and working on their different vectors, the similarity between the two documents is defined as

$$\text{Sim}_{d_i,d_j}(d_i,d_j) = \frac{1}{n-n_r}$$

Where,

$$x = \sum_{d_h\in\mathcal{D}/\mathcal{S}_r} \text{Sim}(d_i - d_h, d_j - d_h)$$

The similarity between two documents in the same cluster is defined as average of similarities measured relatively from views of all documents outside that cluster. The two objects to be measured must be in the same cluster while the points from where to establish this measurement must be outside the cluster. This is referred as Multi-View Point Based Similarity (MVS). The final form of MVS in Eq.(2) depends on particular formulation of the individual similarities within the sum. If the relative similarity is defined by dot-product of the difference vectors, then

$$\text{MVS}(d_i,d_j | d_i,d_j \in \mathcal{S}) = z \sum_{d_h \in \mathcal{S}/\mathcal{S}_r} (d_i - d_h)^t (d_j - d_h)$$

$$= z \sum_{d_h \in \mathcal{S}, y} \|d_i - d_h\| \|d_j - d_h\|$$

Where,

$$y = \cos (d_i - d_h, d_j - d_h)$$

$$z = \frac{1}{n-n_r}$$

The similarity between two points $d_i$ and $d_j$ inside the cluster viewed from a point $d_h$ outside the cluster is equal to the product of the cosine of angle between $d_i$ and $d_j$ looking from $d_h$ and the Euclidean distances from $d_h$ to these points. This is based on the assumption that $d_h$ is not in the same cluster with $d_i$ and $d_j$. The smaller the distances $\|d_i - d_h\|$ and $\|d_j - d_h\|$, higher the chance that $d_h$ is in the same cluster with $d_i$ and $d_j$ and the similarity based on $d_h$ should also be small. Therefore through these distances, Eq. (3) provides a measure of intercluster dissimilarity[1] given that points $d_i$ and $d_j$ is determined by taking average for over all the viewpoints not belonging to the cluster. The effect of misleading viewpoints is constrained and reduced by the averaging step. This similarity measure is implemented in hierarchical agglomerative clustering algorithm.

### 4.3. Efficiency Comparison

In this module, our implemented clustering algorithm with multi viewpoint similarity measure is compared with the K-means and Spherical K-Means Algorithm that uses Euclidean [10] and Cosine Similarity Measure respectively. The Efficiency of our proposed algorithm is compared based on three different evaluation metrics: Accuracy, FScore and Normalized Mutual Information. FScore is an equally weighted combination of the Precision (P) and Recall(R) values used in the information retrieval. FScore value is determined using the formula;

$$\text{FScore} = \frac{\sum_{i=1}^{k} n_i}{\max_{i,j} \left(F_{i,j}\right)}$$

Where

$$F_{i,j} = \frac{2 * P_{i,j} * R_{i,j}}{P_{i,j} + R_{i,j}}$$

$$P_{i,j} = \frac{n_{i,j}}{n_i}; R_{i,j} = \frac{n_{i,j}}{n_j}$$

Where $n_i$ denotes the number of documents in class $i$, $n_j$ the number of documents assigned to cluster $j$, and $n_{i,j}$ the number of shared by class $i$ and cluster $j$ in Eq. (4).

Normalized Mutual Information (NMI) measures the information of the true cluster partition and the cluster assignment share. NMI measures based on the following formula;
\[
\text{NMI} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{k} n_{i,j} \log \left( \frac{n - n_{i,j}}{n_i n_j} \right)}{\sqrt{\left( \sum_{i=1}^{k} n_i \log \frac{n_i}{n} \right) \left( \sum_{j=1}^{k} n_j \log \frac{n_j}{n} \right)}}
\] (5)

Accuracy measures the fraction of documents that are correctly labeled, assuming a one-to-one correspondence between the true classes and assigned clusters. Accuracy is calculated for the same datasets and is plotted in graph with dataset along X-axis and value of accuracy in Y-axis. Let \( q \) denote any possible permutation of index set \( \{1, \ldots, k\} \). Accuracy is calculated by Eq (6) as follows:

\[
\text{Accuracy} = \frac{1}{n} \max_{q} \sum_{i=1}^{k} n_{i,q}
\] (6)

5. RESULT ANALYSES

Presented in a different way, Clustering results based on Accuracy, FScore and NMI are reported in Table 1, Table 2 and Table 3 respectively. For each dataset in a row, the value in bold and underline is the best result. It can be observed that MVSC performs consistently well compared with K-Means Algorithm. The assessment was done based on the clustering results in NMI, FScore and Accuracy, each averaged over all the datasets.

![Figure 1.3: Clustering Results in Accuracy](image)

6. CONCLUSIONS

In this paper, we propose a Multi-View Point Based Similarity Measuring Method. Analysis that MVS is potentially more suitable for text and image documents than the popular cosine similarity measure. Compared with other state-of-the-art clustering methods that use different types of similarity measure, on a large number of datasets and under different evaluation metrics, the proposed algorithms show that they could provide significantly improved clustering performance. The key contribution of this paper is the fundamental concept of similarity measure from multiple viewpoints. Further methods could make use of the same principle, but define alternative forms for similarity in Eq (3), or do not use average but other methods to combine the relative similarities according to the different viewpoints. Finally, we have shown the application of MVS and its clustering algorithm for image data. It would be interesting to explore how they work on other types of sparse and High-Dimensional data.

Table 1
Clustering Results in Accuracy

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>k=3</th>
<th>k=4</th>
<th>k=5</th>
<th>k=6</th>
<th>k=7</th>
<th>k=8</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-Means</td>
<td>.708</td>
<td>.689</td>
<td>.668</td>
<td>.620</td>
<td>.605</td>
<td>.578</td>
</tr>
<tr>
<td>MVSC</td>
<td>.884</td>
<td>.867</td>
<td>.875</td>
<td>.840</td>
<td>.832</td>
<td>.780</td>
</tr>
</tbody>
</table>

Table 2
Clustering Results in FScore and NMI

<table>
<thead>
<tr>
<th>DATA</th>
<th>FSCORE</th>
<th>K-MEANS</th>
<th>NMI</th>
<th>KMEANS</th>
</tr>
</thead>
<tbody>
<tr>
<td>hosp1</td>
<td>.645</td>
<td>.578</td>
<td>.606</td>
<td>.584</td>
</tr>
<tr>
<td>tr11</td>
<td>.512</td>
<td>.467</td>
<td>.323</td>
<td>.270</td>
</tr>
<tr>
<td>hosp2</td>
<td>.620</td>
<td>.502</td>
<td>.612</td>
<td>.563</td>
</tr>
<tr>
<td>hosp3</td>
<td>.873</td>
<td>.825</td>
<td>.739</td>
<td>.629</td>
</tr>
<tr>
<td>hosp4</td>
<td>.719</td>
<td>.565</td>
<td>.569</td>
<td>.367</td>
</tr>
<tr>
<td>tr12</td>
<td>.721</td>
<td>.538</td>
<td>.568</td>
<td>.361</td>
</tr>
<tr>
<td>kny1</td>
<td>.460</td>
<td>.421</td>
<td>.399</td>
<td>.388</td>
</tr>
<tr>
<td>akle</td>
<td>.514</td>
<td>.456</td>
<td>.591</td>
<td>.532</td>
</tr>
<tr>
<td>kny1</td>
<td>.728</td>
<td>.585</td>
<td>.613</td>
<td>.488</td>
</tr>
</tbody>
</table>

References


[2] Alex Petland Content based Indexing of Images and Videos, Medical Laboratory, Massachusetts Institute of Technology, 20 Ames Street, Cambridge, MA02139. USA.

[3] Duc Thang Nguyen, Lihui Chen, and Chee Keong Chan, Clustering with Multi-Viewpoint based Similarity
Efficient Clustering of High Dimensional Datasets with Multi Viewpoint based Similarity Measure


[12] Singh Vijendra, Efficient Clustering for High Dimensional Data : SubSpace Based Clustering and Density Based Clustering, Dept. of Computer Science and Engineering, Mody Institute of Technology and Science, Lakshmangapur, Sikar, Rajasthan, India.


