Clustering-based High Trend Identification in Dataset

Dr. Muhammad Zubair Asghar, Gomal University, Dera Ismail Khan
Dr. Auranzeb Khan
Fazal Masud Kundi, Gomal University, Dera Ismail Khan
Nafees ur Rehman
NCICT 2007
National Conference on
Information and Communication Technologies,
Bannu, NWFP - Pakistan

Saturday June 9, 2007

PROCEEDINGS

Organized by

University of Science & Technology Bannu
CLUSTERING-BASED HIGH TREND IDENTIFICATION IN DATA SET

Fazal Masud Kundi¹, Aurangzeb Khan², Muhammad Zubair Asghar1, Nafees Ur Rehman3

¹Institute of Computing and Information Technology,
Gomal University, DI Khan, NWFP, Pakistan.
²Institute of Information Technology, University of Science and Technology,
Bannu, NWFP, Pakistan.
³Cecos University of Emerging Sciences and Information Technology,
Peshawar, NWFP, Pakistan.
paintak@yahoo.com, aurangzebh_khan@yahoo.com, Zubair@gu.edu.pk, mafees_msit@yahoo.com

ABSTRACT

Clustering data into meaningful groups has a vast scope of research in several fields, like: statistics, information theory, machine learning, databases, and bioinformatics. This paper presents the modified form of K-means clustering algorithm called T-means. This algorithm creates, sorted clusters and labels as “high trend” and “low trend”. Then M cases are selected from the high trend cluster (HTC) to construct the final HTC. The algorithm was tested on real univariate data of student’s marks. Experimental results show that T-means can be efficiently used to construct the sorted cluster of significant cases in data set. T-means will help to identify high trend in data set related to business, education, medical, agriculture and environment.

Keywords: Algorithm, Clustering, K-means, T-means, High trend, Low trend, Centroid.

I. INTRODUCTION

Clustering data into meaningful groups is the subject of active research in several fields such as statistics, information theory, machine learning, databases, and bioinformatics. The term cluster analysis first used by Tryon, encompasses a number of different algorithms and methods for grouping objects of similar kind into respective categories [1]. A general question facing researchers in many areas of inquiry is how to organize observed data into meaningful structures, that is, to develop taxonomies. In other words cluster analysis is an exploratory data analysis tool which aims at sorting different objects into groups in a way that the degree of association between two objects is maximal if they belong to the same group and minimal otherwise.

We deal with clustering in almost every aspect of daily life. For example, a group of people sharing the same room in a office may be regarded as a cluster of people. In computer stores items of similar nature, such as different types of storage media are displayed in the same or nearby locations. There are a countless number of examples in which clustering plays an important role. As a branch of statistics, cluster analysis has been studied extensively for many years, focusing mainly on distance-based cluster analysis[2].

T-means algorithm, being presented in this paper presents the modified form of K-means clustering algorithm [3]. K-means is a typical clustering algorithm and its popularity is largely due to its simplicity, low time complexity, and fast convergence [4]. It partitions a set of data into k clusters. The K-means uses Euclidean distance measure. T-means is the extension of typical K-means. This algorithm creates, sorted clusters and labels as “high trend” and “low trend”. By high trend we mean a sorted cluster with significant cases, in which the user is interested, as cluster of high merit in student’s dataset of the problem domain.

The K-means method does not consider the size of the clusters. Some clusters may be large some very small [5]. This problem has been solved in T-means by creating HTC of size M.

Another problem was of sorted clusters. K-means produces unsorted clusters, where T-means sorts clusters before their merging. The algorithm was tested on real univariate data of student’s marks. Experimental results showed that T-means can be efficiently used to construct the sorted cluster of significant cases in dataset.

II. RELATED WORK

A lot of work has been done on various versions of K-means. In the following lines we present a brief introduction of few data mining algorithms based on K-means.

Y-means is clustering algorithm for intrusion detection based on K-means and other related clustering algorithms. It overcomes two shortcomings of K-means: number of clusters dependency and degeneracy [6].
**Constrained K-Means Clustering** is a method for adding constraints to the K-Means to avoid local solutions with empty clusters or clusters having very few points. This situation arises when the number of dimensions are greater than 10, and the number desired clusters is \( k \geq 20 \) [7].

Anjan Goswami, Ruoming Jin, and Gagan Agrawal present a fast K-means based algorithm which typically requires only one or a small number of passes on the entire data set. The algorithm uses sampling to create a initial cluster centers, and then takes one or more passes over the entire dataset to adjust these cluster centers [8].

**Filtering algorithm** is the efficient implementation of Lloyd's K-means clustering algorithm. The algorithm is easy to implement and only requires that kd-tree be built once for the given data points [9].


### III. THE T-Means ALGORITHM

In this section we presented a T-means algorithm, as shown in Figure 1. This algorithm reads a dataset D of N cases, number of clusters (K=2), and number of desired significant cases (M). In the first phase, the algorithm creates two sorted clusters, and labels them as, “High trend” and “Low trend”. In second phase, final HTC is created of size M from cluster with label, “High trend”. If this cluster has less number of cases than M, then M-\(\text{Count1}\) cases from “Low trend” cluster are merged with “High trend” cluster.

### IV. EXPERIMENTS

Experiments have been performed on various real datasets of student’s marks [10]. See Table 1 for detail.

Preprocessing was made for normalization of dataset, where total marks were not the same.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Cases</th>
<th>Preprocessed</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>50</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>50</td>
<td>Yes</td>
</tr>
<tr>
<td>3</td>
<td>147</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>147</td>
<td>Yes</td>
</tr>
<tr>
<td>5</td>
<td>62</td>
<td>No</td>
</tr>
</tbody>
</table>

**Table 1** Datasets of Student’s Marks

**Algorithm T-means**

**Input**

\[ D // data set of N cases \]

\[ K // the number of clusters \]

\[ M // number of desired sorted significant cases \]

**Output**

Cluster of high trend of size M

1. **Begin**
2. Read N cases in D
3. Select centroid
4. Distance of object to centroid
5. Grouping based on minimum distance
6. If object does not move group any more then
7. Sort clusters
8. Label clusters as “High trend” and “Low trend”
9. \(\text{Count1} = \text{getCount(“High trend”)}\)
10. \(\text{Count2} = \text{getCount(“Low trend”)}\)
11. If \(\text{M > Count1}\) then
12. Get M-\(\text{Count1}\) cases from “Low trend” and merge with “High trend” to create HTC
13. Else
14. Get M cases from “High trend” Cluster to create HTC
15. Else
16. Goto step 03
17. **End**

**Figure 1** The T-means algorithm

Table 2 shows the “High trend” cluster for dataset 5, created in first phase. Table 3 shows the final HTC cluster of size M=10.

### VI. CONCLUSION AND FUTURE RESEARCH DIRECTION

The T-means algorithm is K-means based algorithm, which was applied on univariate dataset of student’s marks to get the sorted cluster of M significant cases in the dataset.

Results of the proposed T-means algorithm were highly satisfactory. T-means will be helpful to identify high trend in dataset related to business, education, medical, agriculture and environment.

T-means can be extended to multivariate data and will become more intelligent with the provision of prior knowledge of the problem domain.

<table>
<thead>
<tr>
<th>Sorted</th>
<th>Cluster</th>
<th>Marks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case#</td>
<td>Cluster</td>
<td>Marks</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>723</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>699</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>695</td>
</tr>
<tr>
<td>12</td>
<td>2</td>
<td>693</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>690</td>
</tr>
<tr>
<td>43</td>
<td>2</td>
<td>686</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>682</td>
</tr>
<tr>
<td>22</td>
<td>2</td>
<td>675</td>
</tr>
<tr>
<td>33</td>
<td>2</td>
<td>674</td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td>667</td>
</tr>
<tr>
<td>40</td>
<td>2</td>
<td>663</td>
</tr>
<tr>
<td>47</td>
<td>2</td>
<td>660</td>
</tr>
<tr>
<td>31</td>
<td>2</td>
<td>656</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>654</td>
</tr>
</tbody>
</table>


June 2007
<table>
<thead>
<tr>
<th>Case#</th>
<th>HTC</th>
<th>M=10 Marks</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>2</td>
<td>723</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>699</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>695</td>
</tr>
<tr>
<td>12</td>
<td>2</td>
<td>693</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>690</td>
</tr>
<tr>
<td>43</td>
<td>2</td>
<td>686</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>682</td>
</tr>
<tr>
<td>22</td>
<td>2</td>
<td>675</td>
</tr>
<tr>
<td>33</td>
<td>2</td>
<td>674</td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td>667</td>
</tr>
</tbody>
</table>

Table 3
HTC for M=10

REFERENCES


[10] Institute of Computing and Information Technology, Gomal University, Dera Ismail Khan, Pakistan.
