Review of Error Rate and Computation Time of Clustering Algorithms on Social Networking Sites

Jaskaranjit Kaur
M. Tech,
CSE
DAV University,
Jalandhar

Gurpreet Singh
Research Scholar
(Ph.D, Computer Sc. Engg.)
Pacific Academy of Higher Education and Research University, Udaipur

ABSTRACT
Data mining is a method of finding useful patterns from large volumes of data. It is an extension of traditional data analysis and statistical approaches. Data Clustering is a task of grouping a set of items or objects into subsets (called clusters). It is an algorithm to discover the similarity between objects in the same class (intraclass similarity) and minimizing the similarity between objects of different classes (interclass similarity). This paper discusses the standard KMeans clustering algorithm and Kohonen Self Organizing Map (SOM) clustering algorithm using the Tanagra datamining tool. These algorithms are applied on facebook dataset i.e. which type of information is shared by university students on facebook. And that information is then used for product marketing purposes. And according to our analysis SOM gives best result with high accuracy and less computational time.

Keywords

1. INTRODUCTION
Data mining is a multidisciplinary field which draws work from areas like database technology, statistics, machine learning, pattern recognition, information retrieval, neural networks, knowledge-based systems, artificial intelligence, high-performance computing, and data visualization. Data mining is a way of analyzing datasets in order to find out unsuspected relationships and to summarize the data in better ways that are both understandable and useful to the data owner. The summaries, relationships and groups that we get through a data mining process are referred to as patterns. It is a process of discovering hidden knowledge from large volumes of raw data. The knowledge must be new and appropriate so that one can use it. Data mining finds these patterns, groups and relationships using data analysis tools.

We have two models in data mining:

- Predictive models, in which we use data with known results to build a model that can be used to explicitly predict values.
- Descriptive models, which describe patterns in existing data.

Clustering is an unsupervised learning technique which divides data items into a number of groups, such that items in the same cluster are more similar and items in different clusters are dissimilar, on the basis of some condition or criteria. It is different from supervised learning where the training examples are provided with class labels which describes the membership of every example to the appropriate class. In clustering we have no prior information about the classes.

Based on the information found in the data describing the objects or their relationships cluster analysis groups objects or events. Clustering is a tool to analyze the data. Its objective is to distribute the objects or events into groups, such that the members of the same cluster have strong degree of association than the members of different clusters. So in this way each cluster describes the class to which its members belong [5].

This paper includes two clustering algorithms: K-means and Kohonen SOM clustering algorithm.

2. CLUSTERING TECHNIQUES
2.1 K-Means Clustering
It is an algorithm to group objects based on attributes/features of the object into K number of groups. K is any positive integer. The clusters are formed on the basis of minimizing the sum of squares of distances between data and the corresponding cluster centroid.

K-Means is an unsupervised learning algorithm which clusters only numerical data in an iterative manner. The main basic concept behind the algorithm is to define k centroids, one for each cluster. As different locations of k causes different cluster result so centroids should be positioned in a semantically way. Then to find the k centroids, the K-Means clustering algorithm will iteratively cluster data and assign each object to the nearest centroid. The centroid here is the mean of the coordinates of the objects in the cluster. Then, the k centroids will change their positions step by step until no further changes occur [6]. The k-means algorithm is as follows [7]:

1. Select k points as initial centroids (randomly generated vectors can also be used).
2. Calculate the distance from each cluster centroid to each point.
3. Assign each point to the nearest cluster.
4. Calculate new cluster centroid, where each new centroid is the mean of all points in that cluster.
5. Repeat steps 2-4 until a stopping condition is reached.

2.2 Kohonen SOM Clustering
Self Organizing Map (SOMs) is an unsupervised learning. The basic notion of a SOM is to map the data patterns onto an n-dimensional grid of units. The grid act as an output space and the space where the data patterns are act as an input space. This mapping tries to preserve topological relations, i.e., patterns that are close in the input space will be mapped to units that are close in the output space and vice-versa [12]. It
is a way to find a good mapping from high dimensional input space to the 2-D representative of the nodes. One way to use SOMs for clustering is to consider the objects in the input space represented by the same node grouped into cluster in the output space. At the training process, each object in the input is presented to the map and the best matching node is identified. After the locations of nodes are tuned, they will create meaningful coordinate system for the input features [6].

Algorithm for Kohonen's Self Organizing Map [1]

- Assume output nodes are connected in an array (usually 1 or 2 dimensional)
- Assume that the network is fully connected - all nodes in input layer are connected to all nodes in output layer.
- Use the competitive learning algorithm as follows:
  1. Randomly choose an input vector $x$
  2. Determine the "winning" output node $i$, where $w_i$ is the weight vector connecting the inputs to output node $i$. $w_i \cdot x \geq w_k \cdot x$ only if the weights are normalized.

$$|\omega_i - X| \leq |\omega_k - X| \quad \forall k$$

Given the winning node $i$, the weight update is

$$\omega_k(\text{new}) = \omega_k(\text{old}) + \Delta \omega_k(n)$$

where $\Delta \omega_k(n)$ represents the change in weight.

3. TANAGRA TOOL

TANAGRA is a free DATA MINING tool for data analysis and research purposes. This software is the advancement of SIPINA. SIPINA implements various supervised learning algorithms, mainly the visual construction of decision trees. TANAGRA is more powerful than it as it contains some supervised learning and also other paradigms such as clustering, parametric and nonparametric statistics, association rule, feature selection, factorial analysis and construction algorithms [4]. The main objective of Tanagra is to give researchers and students an easy data mining software, in compliance to the present work of the software development in this domain and allow to analyze data either real or synthetic. The second objective of TANAGRA is to allow researchers to easily add their datamining algorithms or methods and compare their performance with other methods. The third objective is to help developers in diffusing a probable methodology for building this type of software. Developers can take benefit to look how this kind of software is built, the main steps to follow in developing the project, the kind of problems which we have to avoid during development and which kind of tool and code libraries to use. So in this manner, Tanagra can be considered as an educational tool for learning programming techniques [4].

4. DATASET AND RESULTS

4.1 Data Set and Screenshots

The dataset used is a “Facebook” dataset. It has approximately 10 attributes and 1999 instances. The type of information shared among students is taken as a class attribute.

<table>
<thead>
<tr>
<th>Table 1: Attributes of Data Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
</tr>
<tr>
<td>Age</td>
</tr>
<tr>
<td>Area of Education</td>
</tr>
<tr>
<td>Education</td>
</tr>
<tr>
<td>Products</td>
</tr>
<tr>
<td>No. of Facebook Friends</td>
</tr>
<tr>
<td>No. of hours used</td>
</tr>
<tr>
<td>No. of Groups joined</td>
</tr>
<tr>
<td>No. of Social Networking sites joined</td>
</tr>
<tr>
<td>Information Shared</td>
</tr>
</tbody>
</table>

Figure 1. Implemented K-Means Clustering Algorithm with 6 clusters with error rate 0.3580

Figure 2. Implemented K-Means Clustering Algorithm with 6 clusters with computation time 156ms

Figure 3. Implemented Kohonen SOM Clustering Algorithm and generate MAP Topology with 6 clusters, with Error rate 0.3395
4.2 Result

After implementation of these algorithms on Facebook dataset, the following results are obtained:

<table>
<thead>
<tr>
<th>Parameters</th>
<th>K-Means</th>
<th>Kohonen SOM</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of clusters</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Error Rate</td>
<td>0.3580</td>
<td>0.3395</td>
</tr>
<tr>
<td>Computation Time</td>
<td>156ms</td>
<td>16ms</td>
</tr>
</tbody>
</table>

5. CONCLUSION

The data based upon the type of information shared is collected and validated. Efficient clustering algorithms (K-Means and Kohonen SOM) are applied to finalize the number of clusters which resulted in six qualified clusters. Out of these SOM gives more accuracy. If marketers are interested in expanding the market, they should target to promote the products which are of interests to students by measuring their type of information shared. This clustering information will provide direct benefits to marketers, especially when students have large number of friends in the network where outgrowths for commercial can possibly be gained. The marketers get the opportunity to promote their products as these specific groups have high purchasing power while maintaining close contact with their correspondents.
6. REFERENCES


