Neighborhood Clustering of Web Users
With Rough $K$-Means

Author[1]
Dr. RITU SONI
Head Dept. of computer application,
Guru Nanak Girls college Santpura
Yamunanagar Haryana India -135001

Author[2]
RAJEEV NANDA
Msc (I.T,Maths) Yamunanagar Haryana India -135001

Abstract. Data collection and analysis in web mining faces certain unique challenges. Due to a variety of reasons inherent in web browsing and web logging, the likelihood of bad or incomplete data is higher than conventional applications. The analytical techniques in web mining need to accommodate such data. Fuzzy and rough sets provide the ability to deal with incomplete and approximate information. Fuzzy set theory has been shown to be useful in three important aspects of web and data mining, namely clustering, association, and sequential analysis. There is increasing interest in research on clustering based on rough set theory. Clustering is an important part of web mining that involves finding natural groupings of web resources or web users. Researchers have pointed out some important differences between clustering in conventional applications and clustering in web mining. For example, the clusters and associations in web mining do not necessarily have crisp boundaries. As a result, researchers have studied the possibility of using fuzzy sets in web mining clustering applications. Recent attempts have used genetic algorithms based on rough set theory for clustering. However, the genetic algorithms based clustering may not be able to handle the large amount of data typical in a web mining application. This paper proposes a variation of the $K$-means clustering algorithm based on properties of rough sets topology.

Keywords: Clustering, Interval sets, $K$-means algorithm, Rough sets, Unsupervised learning, Web mining, Topology

1. Introduction

Web mining can be viewed as the extraction of structure from an unlabeled, semi-structured data set containing the characteristics of users and information (Joshi and Krishnapuram, 1998). Logs of web access available on most servers are good examples of the data set used in web mining. Three important operations in web mining are clustering, association, and sequential analysis. This paper focuses on clustering, which is a process of identifying natural groupings of objects.

The clustering process is an important step in establishing user profiles. User profiling on the web consists of studying important
characteristics of the web visitors. Due to the ease of movement from one portal to another, web users can be very mobile. If a particular web site doesn't satisfy the needs of the user in a relatively short period of time, the user will quickly move on to another web site. Therefore, it is very important to understand the needs and characteristics of web users.

Clustering faces several additional challenges in web mining, compared to traditional applications (Joshi and Krishnapuram, 1998). The clusters tend to have vague or imprecise boundaries. The membership of an object in a cluster may not be precisely defined. There is a likelihood that an object may be a candidate for more than one cluster. In addition, due to noise in the recording of data and incomplete logs, the possibility of the presence of outliers in the data set is quite high. Joshi and Krishnapuram (1998) argued that the clustering operation in web mining involves modeling an unknown number of overlapping sets. They proposed the use of fuzzy clustering (Hathaway and Bezdek, 1993; Krishnapuram et al., 1995; Krishnapuram and Keller, 1993) for grouping the web users. This paper proposes neighborhood rough set clustering using a modified K-means algorithm. Any classification scheme can be represented as a partition of a given set of objects. Objects in each equivalence class of the partition are assumed to be identical or similar. In web mining, it is not possible to provide an exact representation of each class in the partition (Joshi and Krishnapuram, 1998). Rough sets (Pawlak, 1982, 1984, 1992) enable us to represent such classes using upper and lower bounds. There are increasing number of research efforts on clustering in relation to rough set theory (Peters et al., 2002; doPrado et al., 2002; Hirano and Tsumoto, 2000). Lingras (2001) described how a rough set theoretic classification scheme can be represented using a rough set genome. The resulting genetic algorithms (GAs) were used to evolve groupings of highway sections represented as interval or rough sets. Lingras (2002) applied the unsupervised rough set clustering based on Gas to group web users. The preliminary experimentation by Lingras (2002) illustrated the feasibility of rough set clustering for developing user profiles on the web. However, the clustering process based on GAs seemed computationally expensive for scaling to a larger data set. One of the most popular and efficient clustering algorithms in conventional applications is K-means clustering (Hartigan, 1979; MacQueen, 1967). In the K-means approach, randomly selected objects are used as the centroids of clusters. The objects are then assigned to different clusters based on their distance from the centroid. The newly formed clusters are then used to determine new centroids. The process continues until the clusters Stabilize Lingras and Huang (2002) provided a theoretical and experimental analysis of various clustering techniques for two datasets of different sizes. They clearly illustrated the computational advantages of the K-means approach for large datasets. However, it is necessary to adapt the K-means algorithm for creating intervals of clusters based on rough set theory. In studies based on marketing data, K-means clustering has been considered in the context of rough sets and genetic algorithms (Voges et al., 2002a, 2002b). However, these studies do not use unsupervised learning to create interval sets. A modification of the K-means algorithm to create interval of clusters will provide an efficient method for representing clusters with vague and imprecise boundaries.

2. Mathematical Review

2.1 Rough set Topology

T is the topology on X
\[ T = \{ A \subseteq X : \text{either } A = \emptyset \text{ or } X-A \} \]

Let X denote the universe, and let \( X R X \) be an equivalence relation on X. The pair \( (T, X) \) is called an approximation space or topological space. The equivalence relation \( R \) partitions the set X into disjoint subsets. Such a partition of the universe is denoted by

\[ A = \bigcup A_i \quad A_i \in T \]

Lower bound of \( A(X) = \emptyset \) (null)
Upper bound of \( A(X) = X \) (space it self)
Upper and lower bound belong to T

If two elements A, B belong to T

and \( A \cap B = \emptyset \) then A \( \cap \) B belong to T.

A \( \cap \) B \( \neq \emptyset \) then also A \( \cap \) B belong to T.

2.2 Review of K-means approach

K-means clustering is one of the most popular statistical clustering techniques (Hartigan, 1979; MacQueen, 1967). The name K-means originates from the means of the \( k \) clusters that are created from \( n \) objects. Let us assume that the objects are represented by \( m \)-dimensional vectors. The objective is to assign these \( n \) objects to \( k \) clusters. Each of the clusters is also represented by an \( m \)-dimensional vector, which is the centroid or mean vector for that cluster. The process begins by randomly choosing \( k \) objects as the centroids of the \( k \) clusters. The objects are assigned to one of the \( k \) clusters based on the minimum value of the distance \( d(v, x) \) between the object vector \( v = (v_1, \ldots, v_j, \ldots, v_m) \) and the cluster vector

\[ x = (x_1, \ldots, x_j, \ldots, x_m). \]

The distance \( d(v, x) \) is given by:

\[
d(v, x) = \sqrt{\left( \sum_{i=1}^{m} (v_j-x_j)^2 \right) / M} \tag{1}
\]

After the assignment of all the objects to various clusters, the new centroid vectors of the clusters are calculated as:

\[
x_j = \frac{\sum_{x \in x} \sum_{j=1}^{m} v_j}{\text{Size of cluster } x} \tag{2}
\]

where \( 1 \leq j \leq m \).

The process stops when the centroids of clusters stabilize, i.e. the centroid vectors from the previous iteration are identical to those generated in the current iteration.

2.3 Adaptation of K-means to rough set topology

\( d(x, y) \) measure the difference or dissimilarity between the object \( (x, y) \).

\[
d(x, y) = \|x-y\| \quad \text{for all } X
\]

\[
d(x, y) = 0 \quad \text{if } x=y
\]

\[
d(x, y) = 1 \quad \text{if } x \neq y
\]

\( d \) is the distance

\[
d(x, y) \geq 0
\]

\[
d(x, y) = 0 \quad \text{iff } x=y
\]

\[
d(x, y) = d(y,x)
\]

\[
d(x, y) \leq d(x,z) + d(z,y)
\]

for all \( x, y, z \) belong to \( X \).

A cluster is a collection of data objects that are similar to one another with the same cluster and dissimilar to the objects in other cluster. Cluster of data object can be treated collectively as one group and so be considered as a form of data compression.

A data set, to be clustered, contains \( N \) object with \( M \) variable may be represented as

\[
\begin{array}{cccc}
0 & d(1,2) & \ldots & d(1,m) \\
0 & d(2,1) & \ldots & d(2,m) \\
\vdots & \vdots & \ddots & \vdots \\
0 & d(n,1) & \ldots & d(n,m)
\end{array}
\]

where \( d(x,y) \) measure the difference or dissimilarity between the object \( (x,y) \).

The modified centroid calculation is give by

\[
d(x, y) = \sum_{x, y} d_n(X_n, Y_n) \tag{3}
\]
the Euclidean distance defined
\[ d_n(x_n, y_n) = \sqrt{(x_1-y_1)^2 + (x_2-y_2)^2 + \ldots + (x_n-y_n)^2} \]  
\[ \text{--- (4)} \]

This can be verified with the following example in which stabilized centroid is found by using the K-mean approach and by using the rough set topology.

<table>
<thead>
<tr>
<th>V</th>
<th>X</th>
<th>d(v, x)</th>
<th>X centroid</th>
<th>d(v, y)</th>
<th>d(v, y) set topology</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>8</td>
<td>3.162</td>
<td>3.8238</td>
<td>10.099</td>
<td>0.315</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>2.645</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>3.872</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>3.872</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>4</td>
<td>5.568</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

From the above table it is clear that the centroid from the set topology is more stable.

3. Summary and conclusions
This paper proposed an adaptation of the K-means algorithm to develop interval clusters of web visitors using rough set topology. In order to develop interval clusters, the K-means algorithm was modified based on the concept of lower and upper bounds using the concept of topology and find the stabilized centroid.

References


