A Spatial Overlapping Based Similarity Measure Applied to Hierarchical Clustering

Hong Chen¹ ², Gongde Guo¹ ², Yu Huang¹ ² and Tianqiang Huang¹ ²
¹ School of Mathematics and Computer Science, Fujian Normal University
² Key Laboratory of Network Security and Cryptography, Fujian Normal University
Fuzhou, Fujian, China, 350007
{hong_chen_01, yellowfish2001}@hotmail.com, {ggd, fjtq}@fjnu.edu.cn

Abstract

A novel similarity measure based on spatial overlapping relation is proposed in this paper, which calculates the similarity between a pair of data points by using the mutual overlapping relation between them in a multi-dimensional space. A spatial overlapping based hierarchical clustering method SOHC was also developed and implemented aimed to justify the effectiveness of the proposed similarity measure. SOHC works well both in low-dimensional and high-dimensional datasets, and is able to cluster arbitrary shape of clusters. Moreover, it can work for both numerical and categorical attributes in a uniform way. Experimental results carried out on some public datasets collected from the UCI machine learning repository and predictive toxicology domain show that SOHC is a promising clustering method in data mining.

1. Introduction

Clustering can be considered as one of the most important unsupervised learning problems. It is a descriptive task that seeks to identify homogeneous groups of objects based on the values of their attributes [1]. A cluster is a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters [2]. Several kinds of clustering methods have been developed in the past few years such as Partitioning Clustering, Hierarchical Clustering, Density-based Clustering, Grid-based Clustering and Model-based Clustering [2]. Given a set of objects and a clustering criterion, Partitioning clustering obtains a partition of the objects into clusters such that the objects in a cluster are more similar to each other than to objects in different clusters [1]. The K-means and K-medoid methods are of this kind. Hierarchical Clustering creates a hierarchical decomposition of the set of data using some criterion. Generally speaking, there are two kinds of Hierarchical Clustering: Divisive and Agglomerative. Agglomerative is more popular, such as CURE and BIRCH [3, 4]. Single linkage, complete linkage and average linkage are the three most representative traditional HC algorithms due to different similarity measures between clusters. Density-based Clustering such as DBSCAN [5], are based on connectivity and density functions. STING [6] and CLIQUE [1] are the two most representative Grid-based methods. This method divides the spatial area into rectangle cells and every manipulation is based on the cell that has much less computational cost. In model-based Clustering such as COBWEB etc., a model is hypothesized for each of the clusters and tries to find the best fit of that model to each other [2].

Most of traditional clustering algorithms use distance to measure the similarity/proximity between a pair of data points. A number of similarity measures based on distance have been proposed and are widely used such as Minkowski Distance, Euclidean Distance, Mahalanobis Distance, Manhattan Distance and Cosine Angle Distance [4]. However, these similarity measures which need to convert/transform different types of data to a uniform format may bring some distortions or cause information lose. Moreover, in high-dimensional datasets, traditional clustering algorithms tend to break down both in terms of accuracy, as well as efficiency, due to the so-called “curse of dimensionality” [5]. It is necessary to develop a new method which can deal with all types of data without any measures of similarity. Wang, et al [7] proposed a data reduction method, called DR, based on lattices and hyper relations. The process of DR can be regarded as a merging process of simple tuples and hyper tuples [7] that have been generated in the same class to generate new hyper tuples. The success of each merging operation depends on whether the new hyper tuple generated from this merging operation covers in same sense a simple tuple of another class. If the new hyper tuple does cover a simple tuple of another class, the operation is cancelled.
The merging operation repeats recursively until all the data including hyper tuples and same class simple tuples cannot be merged again. However, datasets from real-world applications often contain a number of irrelevant or useless attributes and noisy data which directly affect the performance of DR. In such cases, fully coverage-based methods, i.e. DR [7], P&C [8], cannot discover well the relationship of tuples in a multi-dimensional data space. Huang, et al [9] proposed a PCC classification method by considering partial coverage relationship of attributes of tuples with the same and different class labels and class distribution of dataset in a multidimensional data space, whilst without considering distance metric and conversion for categorical attribute.

Unlike a supervised method as PCC, in this paper we present a novel Spatial Overlapping Based Hierarchical Clustering method SOHC. The success of each merging operation depends on spatial overlapping relationship of attributes of tuples. SOHC works well on high-dimensional datasets, and can cluster arbitrary shape of clusters. It is also capable to cope with data with numerical and categorical attributes.

2. Preliminary knowledge

Assume $D$ is a finite set, denoted as $D=\{d_1, d_2, \ldots, d_n\}$, where $d_i$ is a simple tuple with $m$ attributes, denoted as $d_i = (a_{i1}, a_{i2}, \ldots, a_{im})$. The values of attributes can be categorical, numerical or binary. To describe the idea of SOHC, some operators [9] need to be introduced here.

Merge operator ‘$\cup$’ - The ‘$\cup$’ operator is defined as the set union operation for categorical data, and the interval merging operation for numerical data respectively.

Intersection operator ‘$\cap$’ - The ‘$\cap$’ operator is defined as the set intersection operation for categorical data, and the interval overlapping operation for numerical data respectively.

Overlapping operator - Suppose $Y$ and $Z$ are two hyper tuples denoted as $Y = (Y_1, Y_2, \ldots, Y_m)$ and $Z = (Z_1, Z_2, \ldots, Z_m)$ respectively, if $Z$ is an interval, it is represented as $Z_j = [Z_{j1}, Z_{j2}]$, and otherwise, it is represented as a set. $\lambda$ is a predefined threshold whose value is between 0 and 1. It is used to control the ratio of attribute coverage. $\text{Olpi}(Y_j, Z_j)$ function whose value is between 0 and 1, which is used to reflect the overlapping between two hyper tuples and is defined as:

$$\text{Olpi}(Y_j, Z_j) = \frac{\text{LEN}(\cap(Y_j, Z_j))}{\text{LEN}(\cup(Y_j, Z_j))},$$

where

$$\text{Olpi}(Y_j, Z_j) = \begin{cases} 
\text{LEN}(\cap(Y_j, Z_j)) & \text{for } \text{NA}, Y_j = Z_j, \text{for CA} \\
\text{LEN}(\cup(Y_j, Z_j)) & \text{for } \text{NA}, Y_j \neq Z_j, \text{for CA} 
\end{cases}$$

In above equations, $\text{LEN()}$ is a function used to calculate the length of an interval. NA stands for Numerical Attribute; CA stands for Categorical Attribute.

3. The proposed SOHC method

3.1. The basic idea of SOHC

The model constructing process of SOHC is similar to hierarchical tree constructing process in hierarchical clustering method. However, for a typical hierarchical clustering we have to merge data points in each new step - two most similar clusters are merged into a new cluster. The information of each point is needed when the similarity is computed (complete linkage, single linkage etc.). The bottleneck is the “range queries”, that is to find set Neighbors(X,T) for every data points X. Usually, exhaustively searching the whole dataset is needed [10].

To solve this problem, we proposed a SOHC method which similarity calculation between two data points/clusters is based on spatial overlapping. In each step two clusters, whose similarity are the most similar through scanning the similarity matrix instead of making use of all data’s information, are merged. For simplicity, all the data including the raw instances (leaf nodes) and merged clusters (intermediate nodes) are treated in a uniform way as hyper tuple. At beginning of clustering, each instance is a leaf node in a tree. If an instance $X$, holds a numeric attribute $X_{yi}$, this attribute is enlarged by $(\max(X_i) - \min(X_j)) \cdot \delta$ for both right and left directions, where $\delta (0<\delta<1)$ is a parameter. The coverage based measure may result in the priority of merging for some specific clusters, bringing so-called Matthew Effect [11]. A type of ten-fold cross method was used to throw clusters’ order into confusion in order to restrain Matthew Effect.

3.2. The detailed SOHC method

Let $D=\{d_1, d_2, \ldots, d_n\}$ be a dataset with $n$ instances, where $d_i = (a_{i1}, a_{i2}, \ldots, a_{im})$ is a simple tuple with $m$ attributes; Parameter $\delta$ is a predefined threshold used to control the ratio of the enlargement of the leaf nodes; Parameter $r$ is used to control the spatial overlapping rate; $k$ is the predefined number of clusters.
Procedure SOHC $(D, \delta, r, k)$

Input $D, \delta, r, k$
Output $k$ Clusters.

\{
    L = \text{InitLeaves}(D);  // * Initialize leaves
    // * Construct tree from bottom to up, generate k Clusters .
    \text{MakeTree}(L);
\}

InitLeaves(D) // * Method for generating leaves
Input $D$
Output $L$

\{
    \text{For each instance } i, i \in [1,n]$
    \text{For each dimension } j \in [1,m] \text{ do}
    \text{If the } j \text{-th dimension is numeric do}
    \quad \quad a_j = [a_j - (\text{max}(a_j) - \text{min}(a_j)) \cdot \delta, a_j + (\text{max}(a_j) - \text{min}(a_j)) \cdot \delta];
\}

MakeTree(L) // * build tree bottom-up(clustering)
\{
    \text{While } L.\text{size} > k \text{ do}
    \{
        // * method for restraining the Matthew Effect
        // * Generate the similarity matrix $N (L.\text{size} \times L.\text{size})$,
        \text{Cross}(L);
        N_{ij} = \text{Olp}(L.\text{get}(i), L.\text{get}(j)); \text{ max}=0;
        \text{For each ClusterNode } i \in [1, L.\text{size}]
        \quad \text{For each ClusterNode } j \in [i, L.\text{size}]
        \quad \quad \text{If } N_{ij} > \text{max} \& \& N_{ij} > r \text{ then}
        \quad \quad \quad \text{max} = N_{ij}, \text{ max } i = i; \text{ max } j = j;
        \quad \quad // * method for merging
        \quad \text{Merge } (L.\text{get}(\text{max } i), L.\text{get}(\text{max } j));
        \text{If no pair of ClusterNode is merged then}
        \text{reexpand the ClusterNodes}
    \}
\}

Before the merging process of each level, the priority cluster pri has to be found out. The others was ordered with respect to the similarity with it, then kindred ten-fold cross method was used to throw clusters’ order into confusion in order to restrain Matthew Effect.

3.3. A graphic illustration of SOHC

An example of the spatial overlapping based merging process is given in Figure 1. It shows two cases of merging operation.

3.4. Advantages of SOHC over other clustering methods

After the description of the SOHC method, we know that SOHC provides a novel coverage based similarity measurement which is different from traditional methods based on distance. It has following advantages:

1. It can cope with numerical and categorical attributes in a uniform way. Most of the clustering method recently can only deal with numerical and categorical attributes separately. (DBSCAN, CURE for numerical, ROCK for categorical);
2. It can be performed in a high-dimensional dataset. Due to the so-called “curse of dimensionality”, many methods tend to break down both in terms of accuracy, as well as efficiency, such as K-means, K-centroid, DBSCAN, STING;
3. It is able to cluster arbitrary shape of clusters, not to be affected by initial cluster centroids (k-means);
4. It is not to be affected by the sequence of the data input (BIRCH).

4. Experiments and Evaluation

In order to evaluate the effectiveness of the SOHC algorithm, eight public datasets collected from the UCI Machine Learning Repository [12] and seven datasets from predictive toxicology domain are used for evaluation in the experiments. Among these toxicology datasets, five of them, i.e. Trout, Dietary, Oral_Qual, Daphnia_Qual, and Bees have been developed by the DEMETRA project [13, 14]; APC dataset is provide by CSL[13, 15]; Phenols dataset comes from TETRATOX database[16].
[**Experiment 1**] Two clustering methods: SimpleKmeans, FarthestFirst and three traditional HC clustering methods: Single Linkage, Complete Linkage, Average Linkage which are based on Euclidean distance were selected to be compared with the proposed SOHC method. All the clustering methods including SOHC are implemented and integrated into WEKA software package [17]. In the experiment, the parameters for each clustering method were set to obtain the best accuracy. For SOHC, the $\lambda$ defined in definition 5 was set to 0, and the parameter $r$ was set to 0. The experimental results carried out on different datasets are presented in Table 1.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Clustering methods</th>
<th>SK</th>
<th>FF</th>
<th>SL</th>
<th>CL</th>
<th>AL</th>
<th>SOHC</th>
<th>$\delta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iris</td>
<td></td>
<td>11.33%</td>
<td>14%</td>
<td>63.33%</td>
<td>30.67%</td>
<td>7.33%</td>
<td>8.00%</td>
<td>0.05</td>
</tr>
<tr>
<td>Labor</td>
<td></td>
<td>22.81%</td>
<td>26%</td>
<td>45.61%</td>
<td>31.58%</td>
<td>29.80%</td>
<td>24.56%</td>
<td>0.03</td>
</tr>
<tr>
<td>Hepatitis</td>
<td></td>
<td>26.45%</td>
<td>26%</td>
<td>45.81%</td>
<td>21.29%</td>
<td>23.00%</td>
<td>24.56%</td>
<td>0.0038</td>
</tr>
<tr>
<td>Echo</td>
<td></td>
<td>26.45%</td>
<td>26%</td>
<td>40.91%</td>
<td>18.18%</td>
<td>18.94%</td>
<td>18.18%</td>
<td>0.0001</td>
</tr>
<tr>
<td>Ionosphere</td>
<td></td>
<td>28.77%</td>
<td>29.34%</td>
<td>47.29%</td>
<td>34.19%</td>
<td>35.90%</td>
<td>22.22%</td>
<td>0.15</td>
</tr>
<tr>
<td>Sonar</td>
<td></td>
<td>45.67%</td>
<td>46.15%</td>
<td>47.60%</td>
<td>46.63%</td>
<td>38.94%</td>
<td>38.46%</td>
<td>0.12</td>
</tr>
<tr>
<td>Anneal</td>
<td></td>
<td>39.76%</td>
<td>48.44%</td>
<td>71.83%</td>
<td>67.71%</td>
<td>36.08%</td>
<td>31.51%</td>
<td>0.01</td>
</tr>
<tr>
<td>Clean1</td>
<td></td>
<td>45.67%</td>
<td>46.15%</td>
<td>42.86%</td>
<td>45.59%</td>
<td>47.06%</td>
<td>38.87%</td>
<td>0.2</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td>30.05%</td>
<td>31.78%</td>
<td>30.65%</td>
<td>36.98%</td>
<td>29.58%</td>
<td>25.00%</td>
<td></td>
</tr>
<tr>
<td>Trout</td>
<td></td>
<td>54.61%</td>
<td>54.26%</td>
<td>54.96%</td>
<td>55.32%</td>
<td>59.22%</td>
<td>50.00%</td>
<td>0.175</td>
</tr>
<tr>
<td>Oral_Quail</td>
<td></td>
<td>51.72%</td>
<td>45.69%</td>
<td>60.34%</td>
<td>48.28%</td>
<td>44.83%</td>
<td>40.52%</td>
<td>0.15</td>
</tr>
<tr>
<td>Dietary_Quail</td>
<td></td>
<td>56.91%</td>
<td>60.04%</td>
<td>70.73%</td>
<td>65.85%</td>
<td>64.23%</td>
<td>57.72%</td>
<td>0.009</td>
</tr>
<tr>
<td>Daphnia</td>
<td></td>
<td>56.06%</td>
<td>54.17%</td>
<td>62.88%</td>
<td>54.92%</td>
<td>62.5%</td>
<td>52.65%</td>
<td>0.005</td>
</tr>
<tr>
<td>Bees</td>
<td></td>
<td>60.95%</td>
<td>56.19%</td>
<td>71.43%</td>
<td>60.95%</td>
<td>67.62%</td>
<td>51.43%</td>
<td>0.152</td>
</tr>
<tr>
<td>APC</td>
<td></td>
<td>58.33%</td>
<td>60%</td>
<td>63.33%</td>
<td>65%</td>
<td>68.33%</td>
<td>56.67%</td>
<td>0.0018</td>
</tr>
<tr>
<td>Phenols</td>
<td></td>
<td>44.8%</td>
<td>34.8%</td>
<td>60%</td>
<td>39.2%</td>
<td>55.2%</td>
<td>36.4%</td>
<td>0.0022</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td>54.77%</td>
<td>52.88%</td>
<td>63.38%</td>
<td>55.65%</td>
<td>59.91%</td>
<td>49.34%</td>
<td></td>
</tr>
</tbody>
</table>

In Table 1, the meaning of the title in each column is follows: SK-SimpleKmeans; FF-FarthestFirst; SL-Single Linkage; CL-Complete Linkage; AL-Average Linkage; SOHC-A Spatial overlapping Based Hierarchical Clustering method.

From Table 1, it is clear that the SOHC method obtains the lowest error rate in most of the datasets from UCI except for iris and labor datasets. SOHC also gets the lowest average error rate (25.00%) among 6 clustering methods over 8 datasets. Compared with AL method which obtains the lowest error rate among other five clustering methods: SK, FF, SL, CL, AL, the error rate of SOHC has reduced 4.58%.

Moreover, Table 1 shows that SOHC also gets the lowest average error rate (49.34%) among 6 clustering methods over 7 datasets. Compared with FF method which obtains the lowest error rate among other five clustering methods: SK, FF, SL, CL, AL, the error rate of SOHC has reduced 3.54%.

[**Experiment 2**] In this experiment, we observe the different performance of different clustering methods SK, FF, SL, CL, AL and SOHC along with the increasing of the dimensions in high-dimensional dataset clean1. Six subsets of the original clean1 dataset with the size of 11, 16, 41, 61, 81, 111 attributes separately were selected through randomized attributes selection. The experimental results are presented in Figure 2.

**Figure 2. Rates of incorrectly clustered instances**

From Figure 2, it is clear that the error rates of SK, FF, CL, AL and SOHC increase along with dimensions increasing except for SL. Moreover, the SOHC method obtained the lowest error rate among 6 methods in comparison.

In order to compare them in the same environment,
the SK, FF algorithms used in the experiments are from the Weka Software package. The SL, CL, AL and SOHC have been implemented and integrated into the WEKA software package and its source code can be downloaded from following link:
http://mcs.fjnu.edu.cn/datamining/download/SOHC.zip

5. Conclusions

In this paper, we proposed a novel spatial overlapping based hierarchical clustering method SOHC. It is based on a novel coverage relation measurement which is different from the distance-based metric. SOHC is suitable and performs well for high-dimensional dataset. It is able to cluster the arbitrary shape of clusters and can work for both numerical and categorical attributes in a uniform way. Experimental results carried out on some public datasets collected from the UCI machine learning repository and predictive toxicology domain show the effectiveness of the proposed method. Further research is required into how to improve the robustness of SOHC.

Acknowledgements

The project was supported by the Natural Science Foundation of Fujian Province of China (No. 2007J0016). The authors acknowledge also the support of the research project “Research on Spatial Minimal Coverage-based Classification Method” from Fujian Provincial Department of Education under Grant No. JA07051.

References

[12] UCI machine learning repository:
http://www.ics.uci.edu/~mlearn/ MLRepository.html