Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2013, Article ID 323750, 8 pages
http://dx.doi.org/10.1155/2013/323750

Research Article

A Multidimensional and Multimembership Clustering Method for Social Networks and Its Application in Customer Relationship Management

Peixin Zhao,[1] Cun-Quan Zhang,[2] Di Wan,[3] and Xin Zhang[4]

[1] *School of Management, Shandong University, Jinan, Shandong 250100, China*
[2] *Department of Mathematics, West Virginia University, Morgantown, WV 26506, USA*
[3] *Department of Physics and Astronomy, University of Victoria, Victoria, BC, Canada V8W 2Y2*
[4] *Foundation Department, Shandong College of Electronic Technology, Jinan, Shandong 250200, China*

Correspondence should be addressed to Peixin Zhao; pxzhao@126.com

Received 15 July 2013; Accepted 7 August 2013

Academic Editor: Yoshinori Hayafuji

Community detection in social networks plays an important role in cluster analysis. Many traditional techniques for one-dimensional problems have been proven inadequate for high-dimensional or mixed type datasets due to the data sparseness and attribute redundancy. In this paper we propose a graph-based clustering method for multidimensional datasets. This novel method has two distinguished features: nonbinary hierarchical tree and the multi-membership clusters. The nonbinary hierarchical tree clearly highlights meaningful clusters, while the multimembership feature may provide more useful service strategies. Experimental results on the customer relationship management confirm the effectiveness of the new method.

1. Introduction

A social network is a set of people or groups each of which has connections of some kind to some or all of the others. Although the general concept of social networks seems simple, the underlying structure of a network implies a set of characteristics which are typical to all complex systems. Social network plays an extremely important role in many systems and processes and has been intensively studied over the past few years in order to understand both local phenomena, such as clique formation and their dynamics, and network-wide processes, for example, flow of data in computer networks [1], energy flow in food webs [2], customer relation management [3–6], and so forth. Modern information and communication technology has offered new interaction modes between individuals, like mobile phone communications and online interactions. Such new social exchanges can be accurately monitored for very large systems, including millions of individuals, representing a huge opportunity for the study of social science.

Clustering analysis is a data mining technique developed for the purpose of identifying groups of entities that are similar to each other with respect to certain similarity measures. Many different clustering methods have been proposed and used in a variety of fields. Jain [7] broadly divided these methods into two groups: hierarchical clustering and partitioned clustering. Hierarchical clustering is the grouping of objects of interest according to their similarity into a hierarchy, with different levels reflecting the degree of inter-object resemblance. The most well-known hierarchical methods are singlelink and completelink. In singlelink hierarchical methods, the two clusters whose two closest members have the smallest distance are merged in each step; in completelink cases, the two clusters whose merger has the smallest diameter are merged in each step. Compared to hierarchical clustering methods, partitioned clustering methods find all the clusters simultaneously as a partition of the data lK-means, which is widely used for the ease of implementation, simplicity, and efficiency where a certain data point cannot be simultaneously included in

more than one cluster [8]. Based on the difference of their capabilities, applicability, and computational requirements, clustering methods can be categorized into several different approaches: partitioning, hierarchical, density-based, grid-based, and model-based. No particular clustering method has been shown to be superior to all its competitors in all aspects [9].

In recent years, community detection based on clustering has become a growing research field partly as a result of the increasing availability of a huge number of networks in the real world. The most intuitive and common definition of community structure is that such network seems to have communities in them: subsets of vertices within which vertex-vertex connections are dense, but between which connections are relatively sparse. Yang and Luo [10] show that community structure has close relationship with some functionality such as robustness and fast diffusion. It is an important network property and is able to reveal many hidden features of the given network [11]. The detection and analysis of communities in social networks have played an important role in the mining of different kinds of networks, including the World Wide Web [12, 13], communication networks [14], and biological networks [15].

Most traditional community detection algorithms based on clustering are limited to handling one-dimensional datasets [16, 17]. However, the datasets to be mined in real life often contain millions of objects described by many various types of attributes or variables. For example, in customer relation management, a customer can be depicted by multidimensional data or mixed type data such as gender, age, income, education level, and so forth. In such cases, data mining operations and methods are required to be scalable as well as capable of dealing datasets' complex structures and dimensions. Previous researches were mainly focused on the representation of a set of items with a single attribute, which is apparently unsuitable for the scenarios described above: (i) a single attribute can not accurately represent all the dimensions of items; (ii) clustering according to a single attribute often fails to capture the inherent dependency among multiple attributes and leads to meaningless cluster.

Under such considerations, in this paper we firstly introduce two pretreatment methods for multi-dimensional and mixed type data, followed by a new clustering approach for community detection in social networks. In this approach, individuals and their relationships are denoted by weighted graphs, and the graph density we defined gives a better quantity depict of the overall correlation among individuals in a community, so that a reasonable clustering output can be presented. In particular, our method produces "trees" of simple hierarchy and allows for fuzzy (overlapping) clusters, which distinguishes it from other methods. In order to verify the utility/effectiveness of our method, we did a (preliminary) evaluation against a mobile customer segmentation use case. The numerical output of which shows supporting evidence for further (improvement) application.

The rest of the paper is organized as follows. In Section 2 we summarize the related works of community detections in social networks. In Section 3, we introduce the details of the novel clustering approach for multiattribute data sets. As an application in customer relationship management, this approach is used to analyze mobile customer segmentation problem in Section 4. Finally, a summary and conclusions are given in Section 5.

2. Related Works

The detection for communities has brought about significant advances to the understanding of many real-world complex networks. Plenty of detection algorithms and techniques have been proposed drawing on methods and principles from many different areas, including physics, artificial intelligence, graph theory, and even electrical circuits [11]. The spectral bisection methods [18] and the Kernighan-Lin [19] algorithm are early solutions to this problem in computer society. The spectral approach bisects graph iteratively, which is unsuitable to general networks. For the Kernighan-Lin algorithm, it requires a priori knowledge about the sizes of the initial divisions. In 2002, Girvan and Newman [20] proposed a divisive hierarchical clustering algorithm referred to as GN, which can generate optimizion of the division of a network by iteratively cutting the edge with the greatest betweenness value. However, a disadvantage of GN is that its time complexity is $O(m^2n)$ on a network of n nodes and m edges or $O(m^3)$ on a sparse network; then Newman [21] proposed a faster algorithm, referred to as NM, with time complexity $O(n^2)$ or $O((m+n)n)$ on a sparse network. A lot of works have been done to improve GN and NM; for example, Radicchi et al. [22] proposed a similar algorithm with GN by using the edge-clustering coefficient as a new metric with a smaller time complexity $O(m^2)$; Clauset et al. [23] have also proposed a fast clustering algorithm with $O(n\log^2 n)$ time complexity on sparse graph. Especially in 2007, Ou and Zhang [24] proposed a new clustering method with the feature of hierarchical tree and overlapping clusters, the complexity of this method is $O(hn^2 \log n)$ where h denotes the height of the hierarchical structure. This method was, respectively, used to cluster extremist web pages [25] and some classic social networks [26] with single weighted edges.

Random walk has also been successfully used in finding network communities [27, 28]. The idea of this method is that the walk tends to be trapped in dense parts of a network corresponding to communities. Pons and Latapy [27] proposed a measure of similarity between vertices based on random walks which has several important advantages: it captures well the community structure in a network, it can be computed efficiently, and it can be used in an agglomerative algorithm to compute efficiently the community structure of a network. The algorithm called Walktrap runs in time $O(mn^2)$ and space $O(n^2)$ in the worst case and in time $O(n^2 \log n)$ and space $O(n^2)$ in most real-world cases; Hu et al. [29] proposed a method for the identification of community structure based on a signaling process of complex networks. Each node is taken as the initial signal source to excite the whole network one time, and the source node is associated with an n-dimensional vector which records the effects of the signaling process. By this process, the topological relationship of nodes on the network could be transferred into a geometrical

structure of vectors in n-dimensional Euclidean space. Then the best partition of groups is determined by F statistics, and the final community structure is given by the K-means clustering method.

Spectral clustering techniques have seen an explosive development and proliferation over the past few years [30–32]. Previous work indicated that a robust approach to community detection is the maximization of the benefit function known as "modularity" over possible divisions of a network, but Newman and Girvan [30] showed that the maximization process can be written in terms of the modularity matrix, which plays a role in community detection similar to that played by the graph Laplacian in graph partitioning calculations, and the time complexity of this algorithm is $O(n^2)$. They also proposed an objective function for graph clustering called the Q function, which allows for automatic selection of the number of clusters, and then higher values of the Q function were proven to correlate well with good graph clustering. White and Smyth [31] showed how the Q function can be reformulated as a spectral relaxation problem and proposed two new spectral clustering algorithms that seek to maximize Q. Capocci et al. [32] developed some spectral-based algorithm to reveal the structure of a complex network, which could be blurred by the bias artificially overimposed by the iterative bisection constraint. Such a method should be able to conjugate the power of spectral analysis to the caution needed to reveal an underlying structure when there is no clear cut partitioning, as is often the case in real networks.

Lots of other community detection algorithms have also been proposed in the recent literatures. For example, Wu and Huberman [33] proposed a method which partitions a network into two communities, where the network is viewed as an electric circuit, and a battery is attached to two random nodes that are supposed to be within two communities. Shi et al. [11] proposed a new genetic algorithm for community detection, using the fundamental measure criterion modularity Q as the fitness function. A special locus-based adjacency encoding scheme is applied to represent the community partition; Shi et al. [34] proposed a novel method based on particle swarm optimization to detect community structures by optimizing network modularity.

3. Multidimensional and Multimembership Clustering Method for Social Networks

3.1. Similarity of Multidimensional Data. Traditional distance functions include Euclidean distance, Chebyshev distance, Manhattan distance, Mahalanobis distance, Weighted Minkowski distance, and Cosine distance. Among these distance functions, Mahalanobis distance is based on correlations between variables by which different patterns can be identified and analyzed. It gauges similarity of an unknown sample set to a known one. It differs from Euclidean distance in that it takes into account the correlations of the data set and is scale-invariant. In other words, it is a multivariate effect size.

All these distance functions have their own advantages and disadvantages in practical applications. Some research results shows that Euclidean distance has better performance in vector models, while some other numerical examples in high dimensional spaces show that the farthest and nearest distance are almost equal, although Euclidean distance is used to measure the similarity between data points. That is in high-dimensional data, traditional similarity measures as used in conventional clustering algorithms are usually not meaningful. This problem and related phenomena require adaptations of clustering approaches to the nature of high-dimensional data. This area of research has been a highly active one in recent years. Common approaches are known as, for example, subspace clustering, projected clustering, pattern-based clustering, or correlation clustering. Subspace clustering is the task of detecting all clusters in all subspaces, which means that a point might be a member of multiple clusters, each existing in a different subspace. Subspaces can either be axis parallel or affine. Projected clustering seeks to assign each point to a unique cluster, but clusters may exist in different subspaces. The general approach is to use a special distance function together with a regular clustering algorithm. Correlation clustering provides a method for clustering a set of objects into the optimum number of clusters without specifying that number in advance.

In 2011, A new function "Close()" is presented based on the improvement of traditional algorithm to compensate their inadequacy for high-dimensional space [35]. Let

$$X = (x_1, x_2, \ldots, x_n), \quad Y = (y_1, y_2, \ldots, y_n) \tag{1}$$

denote two points in n-dimensional space. The function "Close()" is defined as

$$\text{Close}(X, Y) = \frac{\sum_{i=1}^{n} e^{-|x_i - y_i|}}{n}. \tag{2}$$

It depicts the similarity degree between two data points and has the following properties.

(a) The minimum value of the function is 0, which means that the similarity degree between X and Y is smallest since the difference comes closest to infinity in each dimension.

(b) The maximum value of the function is 1, which means that the similarity degree between X and Y is largest since they come closest to coinciding in each dimension.

Similar to the weighted operator in traditional distance functions, the close function can be corrected as

$$\text{Close}(X, Y) = \frac{\sum_{i=1}^{n} \omega_i e^{-|x_i - y_i|}}{n}, \tag{3}$$

where $\omega_i \in [0, 1]$ denotes the importance degree of data in the ith dimension. Advantages of the new function are obvious in high-dimensional similarity measurement according to the comparison in [35]. Quantitative analysis also proved that this function can avoid the effects of noise and the curse of high-dimension.

3.2. Similarity of Mixed Type Data. For clustering multiattributes datasets, we first introduce a method for the measurement of similarity between items as follows [36]. The multiattribute datasets can be separated into two parts: the pure numeric datasets and pure categorical datasets. Some existing efficiency clustering methods designed for these two types of data sets are employed to produce corresponding clusters. For the similarity matrix, we define $S(i, j)$ as the number of times the given sample pair x_i and x_j has co-occurred in a cluster [37].

Consider

$$S(i,j) = S\left(x_i, x_j\right) = \frac{1}{H}\sum_{k=1}^{H}\delta\left(\pi_k\left(x_i\right), \pi_k\left(x_j\right)\right), \qquad \delta(a,b) \equiv \begin{cases} 1, & a = b, \\ 0, & a \neq b, \end{cases} \tag{4}$$

where H denotes the number of the clustering. $\pi_k(x_i)$ and $\pi_k(x_j)$ denote the cluster label of items x_i and x_j, respectively. Then for the pure numerical datasets, the similarity can be defined as

$$S_1(i,j) = \frac{n}{N} = \frac{\sum_{i=1}^{N} C(i,j)}{N}, \tag{5}$$

where N is the number of clustering and n is the number of times the pattern pair (x_i, x_j) is assigned to the same cluster among the N clustering. If (x_i, x_j) is assigned to the same cluster, $C(i, j) = 1$, otherwise $C(i, j) = 0$.

For the pure categorical datasets, the similarity can be defined as

$$S_2(i,j) = \frac{n}{m} = \frac{\sum_{i=1}^{m} C(i,j)}{m}, \tag{6}$$

where m denotes the number of attributes. Then the similarity of multiattribute datasets can be denoted by

$$S = S_1 + \alpha S_2, \tag{7}$$

where α is a user-defined parameter. If $\alpha > 1$, the categorical datasets is more important than the numerical datasets; if $\alpha < 1$, numerical datasets is more important. $S_1(i, j)$ and $S_2(i, j)$ can also be used as two-dimensional (or multidimensional) datasets to represent the similarities between items x_i and x_j.

3.3. Multidimensional and Multimembership Clustering Method for Social Networks. A graph or network is one of the most commonly used models to represent real-valued relationships of a set of input items. Since many traditional techniques for one-dimensional problems have been proven inadequate for high-dimensional or mixed type datasets due to the data sparseness and attribute redundancy, the graph-based clustering method for single dimensional datasets proposed in [24–26] can be extended as follows to directly cluster multidimensional datasets.

Let $G = (V, E)$ be a graph with the vertex set V and associated with r weights:

$$\omega_k : E(G) \longmapsto [0,1], \quad k = 1, 2, \ldots, r. \tag{8}$$

For a subgraph $C(|V(C)| > 1)$ of G, we define the kth density of C by

$$d_k(C) = \frac{2\sum_{e\in E(C)} \omega_k(e)}{|V(C)|\left(|V(C)| - 1\right)}. \tag{9}$$

In single weighted graph C, if $\omega(e) = 1$ and $d(C) = 1$ for every edge e in C, the subgraph C induces a clique. For a multiweighted graph $(G; \omega_1, \omega_2, \ldots, \omega_r)$, a subgraph C is called a Δ-quasiclique if $d_k(C) \geq \Delta$ for some positive real number Δ and for every $k \in \{1, 2, \ldots, r\}$ (r is the number of weights on the edge).

Clustering is a process that detects all dense subgraphs in G and constructs a hierarchically nested system to illustrate their inclusion relation.

A heuristic process is applied here for finding all quasicliques with density of various levels. The core of the algorithm is deciding whether or not to add a vertex to an already selected dense subgraph C. For a vertex $v \notin V(C)$, we define the contribution of v to C by

$$c_k(v, C) = \frac{\sum_{u\in V(C)} \omega_k(uv)}{|V(C)|}. \tag{10}$$

A vertex v is added into C if $c_k(v, C) > \alpha d(C)$ where α is a user specified parameter.

In short, the main steps of our algorithm can be described as shown in Algorithm 1.

Trace the process of each vertex, and obtain the hierarchic tree.

Our detailed community detection algorithm that can find Δ-quasicliques in G with various levels of Δ is as follows. A hierarchically nested system is constructed to illustrate their inclusion relation.

Step 0. $l \leftarrow 1$ where l is the indicator of the levels in the hierarchical system:

$$M_0 \longleftarrow \gamma \max\left\{\omega_k(e) : \forall e \in E(G), \forall k\right\}, \tag{11}$$

where γ $(0 < \gamma < 1)$ is a user specified parameter (γ is a cut-off threshold).

Step 1 (the initial step). Let F be the set of all edges e of G with

$$\min\left\{\omega_k(e) : k = 1, 2, \ldots, r\right\} \geq M_0. \tag{12}$$

Let $m = |F|$. Sort the edges of the set F as a sequence $S = e_1, \ldots, e_m$ such that

$$\sum_{k=1}^{r}\omega_k\left(e_1\right) \geq \sum_{k=1}^{r}\omega_k\left(e_2\right) \geq \cdots \geq \sum_{k=1}^{r}\omega_k\left(e_m\right), \tag{13}$$

$\mu \leftarrow 1$, $p \leftarrow 0$, and $L_l \leftarrow \emptyset$ where L_l is the community sets in the lth hierarchical level.

Step 2 (One has starting a new search).

$$p \longleftarrow p + 1, \qquad C_p \longleftarrow V\left(e_\mu\right). \qquad L_l \longleftarrow L_l \cup \left\{C_p\right\}. \tag{14}$$

Input: A graph $G = (V; \omega_1, \omega_2, \ldots, \omega_r)$ is a multi-weighted graph with ω_k: $E(G) \mapsto [0, 1]$.
Output: Meaningful community sets in G.

Algorithm: Detect Δ-quasi-cliques in G with various levels of Δ, and construct a hierarchically nested system to illustrate their inclusion relation.

While $E(G) \neq \emptyset$
begin
determine the value of M_0
Decompose(G, M_0)
$E_0 = \{e \in E(G)$: $\omega_k(e) \geq M_0, k = 1, 2, \ldots, r\}$
for each edge in E_0 in decreasing order of weights, if the two vertexes of edge are not in any community, create a new empty community C Choose v in the rest vertex sets that have maximum contribution to C and add v in it.
Merging (G)
Merge two communities according to their common vertexes;
Contract each community to a vertex and redefine the weight of the corresponding edges.
Store the resulted graph to G.
End.

Algorithm 1

Step 3 (*growing*)

Substep 3.1. $U \leftarrow V(G) - V(C_p)$; if $U = \emptyset$, go to Step 4; otherwise continue.

$(*)$ Pick $v \in U$ such that $\Pi_{k=1}^{r} c_k(v, C_p)$ is a maximum.

If, for every k,

$$c_k\left(v, C_p\right) \geq \alpha_n d_k\left(C_p\right), \tag{15}$$

where $n = |V(C_p)|$ and $\alpha_n = 1 - (1/2)\lambda(n + t)$ with $\lambda \geq 1$, $t \geq 1$ as user specified parameters, then $C_p \leftarrow C_p \cup \{v\}$, and go back to Substep 3.1.

If Inequality (15) is not satisfied, then

$$U \longleftarrow U - \{v\}. \tag{16}$$

If $U \neq \emptyset$, repeat $(*)$. If $U = \emptyset$, go to Substep 3.2.

Substep 3.2. $\mu \leftarrow \mu + 1$. If $\mu > m$ go to Step 4; otherwise continue.

Substep 3.3. Suppose $e_\mu = xy$. If at least one of $x, y \notin \cup_{i=1}^{p-1} V(C_i)$, then go to Step 2; otherwise go to Substep 3.2.

Step 4 (*merging*).

Substep 4.1. List all members of L_l as a sequence $C_1, \ldots, C_s$ such that

$$\left|V\left(C_1\right)\right| \geq \left|V\left(C_2\right)\right| \geq \cdots \geq \left|V\left(C_s\right)\right|, \tag{17}$$

where $s = |L_l|$. $h \leftarrow 2$, $j \leftarrow 1$.

Substep 4.2. If

$$\left|C_j \cap C_h\right| > \beta \min\left(\left|C_j\right|, \left|C_h\right|\right), \tag{18}$$

(where $0 < \beta < 1$ is a user specified parameter), then $C_{s+1} \leftarrow C_j \cup C_h$, and the sequence L_l is rearranged as follows:
$C_1, \ldots, C_{s-1} \leftarrow$ deleting C_j, C_h from $C_1, \ldots, C_{s+1}$.
$s \leftarrow s - 1$, $h \leftarrow \max\{h - 2, 1\}$, and go to Substep 4.4.

Substep 4.3. $j \leftarrow j + 1$. If $j < h$, go to Substep 4.2.

Substep 4.4. $h \leftarrow h + 1$, $j \leftarrow 1$. If $h \leq s$, go to Substep 4.2.

Step 5. Contract each $C_p \in L_l$ as a vertex:

$$V(G) \longleftarrow \left[V(G) - \bigcup_{p=1}^{s} V\left(C_p\right)\right] \cup \{C_1, \ldots, C_s\},$$

$$\omega_k(uv) \longleftarrow \omega_k\left(C_i, C_j\right) = \frac{\sum_{e \in E_{i,j}} \omega_k(e)}{E_{i,j}}, \quad k = 1, 2, \ldots, r. \tag{19}$$

The vertex u is obtained by contracting C_i and v is obtained by contracting C_j where E_{ij} is the set of crossing edges which is defined as

$$E_{i,j} = \left\{xy : x \in C_i, y \in C_j, x = y\right\}. \tag{20}$$

For

$$q \in V(G) - \{C_1, \ldots, C_s\}, \tag{21}$$

define $\omega_k(q, C_i) = \omega_k(\{q\}, C_i)$. Other cases are defined similarly.

If $|V(G)| \geq 2$, then go to Step 6; otherwise go to End.

Step 6. One has

$$\begin{gathered} l \longleftarrow l + 1, \qquad L_l \longleftarrow \emptyset, \\ \omega_0 \longleftarrow \gamma \max\{\omega(e) : \forall e \in E(G), \forall k\}, \end{gathered} \tag{22}$$

where γ $(0 < \gamma < 1)$ is a user specified parameter, and go to Step 1 (to start a new search in a higher level of the hierarchical system).

End.

Trace the movement of each vertex and generate the hierarchic tree.

Table 1: Some information of 3000 mobile customers.

Customer number	Local call fee (Yuan)	Long distance call fee (Yuan)	Roaming fee (Yuan)	Text message and WAP fee (Yuan)	Package type
1	55.3	13.7	120.6	14.2	D
2	132	44.8	36.2	5.6	B
3	47.1	233.6	79.4	6.2	B
4	173	19.3	87.5	19.3	C
5	23.7	80.5	21	9	A
6	62.3	62.9	77.8	10.6	E
7	242.5	21.8	23.5	24.2	A
8	166.2	34.5	8	19.5	C
...	...	...	...	...	...
3000	77.6	67	21.2	24.7	D

If the input data is an unweighted graph G, the adjacency information is used for establishing the similarity matrix of G. Let $A = (a_{ij})$ be the adjacency matrix of G where

$$a_{ij} = \begin{cases} 1, & \text{there is an edge between } i \text{ and } j \\ 0, & \text{otherwise} \end{cases} \quad (23)$$

and the inner product of the ith and the jth row of A is used to describe the similarity between nodes i and j and stored as $G(i, j)$ in the similarity matrix G.

4. Simulation Examples

In order to validate the feasibility of the proposed novel approach to cluster multi-dimensional data sets, we randomly took 3000 customers' consumption lists of August 2012 from Shandong Mobile Corporation and use our new approach to divide these customers into distinguishing clusters according to 4 evaluation indices: local call fee, long distance call fee, roaming fee and text message and WAP fee. The original data of 3000 customers are listed in Table 1.

We have applied our approach to this problem, and the results of segmentation and their average consumption are listed in Table 2 and Figure 1.

As we can see from the clustering result, the long distance fee of group 1 has a high proportion of their total expenses; Groups 3 and 4 have high roaming fees; Group 8 has lower cost in each index; Groups 2, 3, and 4 have higher text message and WAP fees. Mobile corporations can initiate corresponding policies according to the clustering results. For example, for the customers in Groups 2, 3, and 4, mobile corporation should provide them with some discount text message package; for the customers in Groups 3, 4, and 6, some discount package of roaming will also help to increase customer loyalty and stability.

On the other hand, we noticed that the sum of the last column of Table 2 is larger than 3000. This is because our method allows multimembership clustering; thus some customers can belong to more than one group. For instance, Groups 8 and 1 are low value customer and high value customer respectively, and some special policies should be recommended for the 39 customers, who belong to either Group 1 or 8 to help them become loyal higher value customers.

Table 2: The customer segmentation of mobile network.

Cluster number	Average local call fee (Yuan)	Average long distance call fee (Yuan)	Average roaming fee (Yuan)	Average text message and WAP fee (Yuan)	Number of customer
1	156.9	172.8	39.8	58.5	121
2	299.1	43.2	38.7	46.9	64
3	42.6	32.9	174.7	36.2	168
4	212.8	103.3	574.3	39.7	13
5	187.9	871.5	35.3	28.7	9
6	162.1	262.3	354.8	21.2	12
7	43.0	25.8	13.7	21.2	2077
8	19.2	7.5	4.8	13.5	792

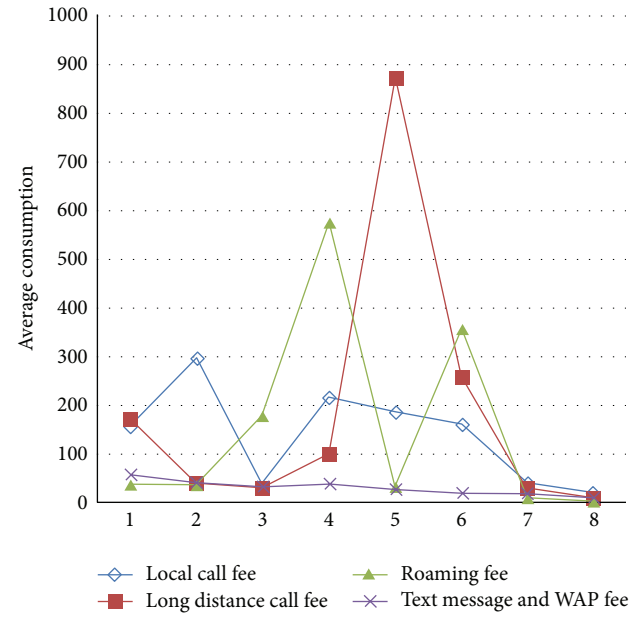

Figure 1: Average consumption list of 8 Groups.

5. Conclusions

In this paper, a graph-based new clustering method for multi-dimensional datasets is proposed. Due to the inherent sparsity of data points, most existing clustering algorithms do not work efficiently for multi-dimensional datasets, and it is not feasible to find interesting clusters in the original full space of all dimensions. These researches were mainly focused on the representation of a set of items with a single attribute, which cannot accurately represent all the attributes and capture the inherent dependency among multiple attributes. The new clustering method we proposed in this paper overcomes

this problem by directly clustering items according to the multidimensional information. Since it does not need data preprocessing, this new method may significantly improves clustering efficiency. It also has two-distinguished features: nonbinary hierarchical tree and multimembership clusters. The application in customer relationship management has proved the efficiency and feasibility of the new clustering method.

Conflict of Interests

Peixin Zhao, Cun-Quan Zhang, Di Wan, and Xin Zhang certify that there is no actual or potential conflict of interests in relation to this paper.

Acknowledgments

The first author is partially supported by the China Postdoctoral Science Foundation funded Project (2011M501149), the Humanity and Social Science Foundation of Ministry of Education of China (12YJCZH303), the Special Fund Project for Postdoctoral Innovation of Shandong Province (201103061), the Informationization Research Project of Shandong Province (2013EI153), and Independent Innovation Foundation of Shandong University, IIFSDU (IFW12109). The second is author partially supported by an NSA Grant H98230-12-1-0233 and an NSF Grant DMS-1264800.

References

[1] X. Jin, C. M. K. Cheung, M. K. O. Lee, and H. Chen, "How to keep members using the information in a computer-supported social network," *Computers in Human Behavior*, vol. 25, no. 5, pp. 1172–1181, 2009.

[2] A. Bodini, "The qualitative analysis of community food webs: implications for wildlife management and conservation," *Journal of Environmental Management*, vol. 41, no. 1, pp. 49–65, 1994.

[3] P. C. Verhoef and K. N. Lemon, "Successful customer value management: key lessons and emerging trends," *European Management Journal*, vol. 31, no. 1, pp. 1–15, 2013.

[4] C. Kiss and M. Bichler, "Identification of influencers—measuring influence in customer networks," *Decision Support Systems*, vol. 46, no. 1, pp. 233–253, 2008.

[5] E. S. Bernardes and G. A. Zsidisin, "An examination of strategic supply management benefits and performance implications," *Journal of Purchasing and Supply Management*, vol. 14, no. 4, pp. 209–219, 2008.

[6] D. Li, W. Dai, and W. Tseng, "A two-stage clustering method to analyze customer characteristics to build discriminative customer management: a case of textile manufacturing business," *Expert Systems with Applications*, vol. 38, no. 6, pp. 7186–7191, 2011.

[7] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, 2010.

[8] S. Guha, A. Meyerson, N. Mishra, R. Motwani, and L. O'Callaghan, "Clustering data streams: theory and practice," *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, no. 3, pp. 515–528, 2003.

[9] F. Cao, "A weighting K-modes algorithm for subspace clustering of categorical data," *Neurocomputing*, vol. 108, pp. 23–30, 2012.

[10] S. Yang and S. Luo, "A local quantitative measure for community detection in networks," *International Journal of Intelligent Engineering Informatics*, vol. 1, no. 1, pp. 38–52, 2010.

[11] C. Shi, Y. Wang, B. Wu, and C. Zhong, "A new genetic algorithm for community detection," in *Complex Sciences, Part II*, vol. 5 of *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, pp. 1298–1309, 2009.

[12] M. Hoerdt and U. Louis, "Completeness of the internet core topology collected by a fast mapping software," in *Proceedings of the 11th International Conference on Software, Telecommunications and Computer Networks*, pp. 257–261, 2003.

[13] A. Broder, P. Kumar, F. Maghoul et al., "Graph structure in the web," in *Proceedings of the 9th International Conference on the World Wide Web*, pp. 15–19, 2003.

[14] J. Scott, *Social Network Analysis: A Handbook*, Sage, 2000.

[15] D. A. Fell and A. Wagner, "The small world of metabolism," *Nature Biotechnology*, vol. 18, no. 11, pp. 1121–1122, 2000.

[16] T. Chen, N. L. Zhang, T. Liu, K. M. Poon, and Y. Wang, "Model-based multidimensional clustering of categorical data," *Artificial Intelligence*, vol. 176, pp. 2246–2269, 2012.

[17] L. Poon, N. L. Zhang, T. Liu, and A. H. Liu, "Model-based clustering of high-dimensional data: variable selection versus facet determination," *International Journal of Approximate Reasoning*, vol. 54, no. 1, pp. 196–215, 2012.

[18] A. Pothen, H. D. Simon, and K.-P. Liou, "Partitioning sparse matrices with eigenvectors of graphs," *SIAM Journal on Matrix Analysis and Applications*, vol. 11, no. 3, pp. 430–452, 1990, Sparse matrices (Gleneden Beach, OR, 1989).

[19] B. W. Kernighan and S. Lin, "A efficient heuristic procedure for partitioning graphs," *Bell System Technical Journal*, vol. 49, pp. 291–307, 1970.

[20] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 12, pp. 7821–7826, 2002.

[21] M. E. J. Newman, "Fast algorithm for detecting community structure in networks," *Physical Review E*, vol. 69, no. 6, Article ID 066133, pp. 1–66133, 2004.

[22] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Paris, "Defining and identifying communities in networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 9, pp. 2658–2663, 2004.

[23] A. Clauset, M. E. J. Newman, and C. Moore, "Finding community structure in very large networks," *Physical Review E*, vol. 70, no. 6, Article ID 066111, 6 pages, 2004.

[24] Y. Ou and C.-Q. Zhang, "A new multimembership clustering method," *Journal of Industrial and Management Optimization*, vol. 3, no. 4, pp. 619–624, 2007.

[25] X. Qi, K. Christensen, R. Duval et al., "A hierarchical algorithm for clustering extremist web pages," in *Proceedings of the International Conference on Advances in Social Network Analysis and Mining (ASONAM '10)*, pp. 458–463, August 2010.

[26] P. Zhao and C. Zhang, "A new clustering method and its application in social networks," *Pattern Recognition Letters*, vol. 32, no. 15, pp. 2109–2118, 2011.

[27] P. Pons and M. Latapy, "Computing communities in large networks using random walks," *Journal of Graph Algorithms and Applications*, vol. 10, no. 2, pp. 191–218, 2006.

[28] S. V. Dongen, *Graph clustering by flow simulation [Ph.D. dissertation]*, University of Utrecht, 2000.

[29] Y. Hu, M. Li, P. Zhang, Y. Fan, and Z. Di, "Community detection by signaling on complex networks," *Physical Review E*, vol. 78, no. 1, Article ID 016115, 2008.

[30] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical Review E*, vol. 69, no. 2, Article ID 026113, 2004.

[31] S. White and P. Smyth, "A spectral clustering approach to finding communities in graphs," in *Proceedings of SIAM International Conference on Data Mining*, pp. 76–84, 2005.

[32] A. Capocci, V. D. P. Servedio, G. Caldarelli, and F. Colaiori, "Detecting communities in large networks," *Physica A*, vol. 352, no. 2-4, pp. 669–676, 2005.

[33] F. Wu and B. A. Huberman, "Finding communities in linear time: a physics approach," *European Physical Journal B*, vol. 38, no. 2, pp. 331–338, 2004.

[34] Z. Shi, Y. Liu, and J. Liang, "PSO-based community detection in complex networks," in *Proceedings of the 2nd International Symposium on Knowledge Acquisition and Modeling (KAM '09)*, pp. 114–119, December 2009.

[35] C. Shao, W. Lou, and L. Yan, "Optimization of algorithm of similarity measurement in high dimensional data," *Computer Technology and Development*, vol. 20, no. 2, pp. 1–4, 2011.

[36] H. Luo and H. Wei, "Clustering algorithm for mixed data based on clustering ensemble technique," *Computer Science*, vol. 37, no. 11, pp. 234–238, 2010.

[37] A. Fred, "Finding consistent clusters in data partitions," in *Multiple Classifier Systems*, vol. 2096 of *Lecture Notes in Computer Science*, pp. 309–318, 2001.