http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 12 (23): 7849-7853, 2013
ISSN 1812-5638 / DOI: 10.3923/itj.2013.7849.7853

Wnbac: A Weighted Network Based Adaptive Clustering Algorithm for Spatial Objects

Pan Xu-Wei and Jin Min
Department of Management Science and Engineering, Zhejiang Science Technology University,
Hangzhou, 310018, Zhejiang, China

Abstract: To overcome shortcomings such as slowness of the convergence, sensitive to initial value and pre-awareness of dataset in most clustering algorithm, a Weighted Network Based Adaptive Clustering (WNBAC) algorithm is put forward. The WNBAC algorithm is to build the weighted network for spatial objects in term of the similarity among objects, then to partition nodes in the weighted network by nodes' strength and edges' weight. The core idea, main process, building procedure and parameter setting for the WNBAC algorithm are described and discussed in details. Experiment results indicate that the proposed WNBAC algorithm is both effective and efficient.

Key words: Clustering algorithm, weighted network, network partition

INTRODUCTION

Clustering is a popular data analysis and data mining technique. It means the act of partitioning a multi-attribute datasets into homogeneous groups of similar objects. The goal of a clustering algorithm is to group sets of objects into classes such that similar objects are placed in the same cluster while dissimilar objects are in separate clusters. As a data processing technique, clustering has become an active research topic in pattern recognition, data mining, statistics and machine learning with different applications such as information retrieval, image analysis, bioinformatics, medicine and web mining (Vega-Pons and Ruiz-Shulcloper, 2011).

Many clustering algorithms have been developed in the literature, some of them are (Xu and Wunsch, 2005) BRICH, CLARANS, CURE, CHAMELEON, DBSCAN, K-means, etc. These traditional clustering algorithms have well known shortcomings such as such as slowness of the convergence, sensitive to initial value and preset classed in large scale dataset etc. (Xu and Wunsch, 2005). Thereby, many efforts have been taken to get across these shortcomings. For example, some algorithms that imitate certain natural principles and phenomena of biological evolution, have been applied in various fields (Abul Hasan and Ramakrishnan, 2011; Khan and Ahmad, 2013) addresses the randomized center initialization problem of K-modes algorithm by proposing a cluster center initialization algorithm, Liao *et al.* (2013) proposed weighted fuzzy kernel-clustering algorithm with adaptive evolution algorithm to solve the problem of multi-attribute and multi-stage fuzzy synthetically evaluation.

Networks provide us with a powerful and versatile tool for processing and analyzing complex problems where nodes represent individuals and links denote the relations between them (Zhang and Liu, 2010). In clustering spatial objects, if the spatial object is taken as the node and the similarity between two objects as the weight of the edge between them, spatial objects are connected together and formed a network. Thus the clustering of spatial objects will be transformed to network partition. Therefore, a Weighted Network Based Adaptive Clustering (WNBAC) algorithm is put forward in this study trying to overcome the shortcomings stated in above.

WNBAC ALGORITHM

Basic concepts

Spatial object dataset: A spatial object x_i has m attributes $x_{i1}, x_{i2}, \ldots, x_{im}$, it is expressed as $x_i = \{x_{i1}, x_{i2}, \ldots, x_{im}\}$. If the number of objects in the dataset is n, then the spatial object dataset $X = \{x_1, x_2, \ldots, x_n\}$ is a n×m matrix.

Similarity matrix: $s(x_i, x_j)$ refers to the similarity between two spatial objects x_i and x_j, $s(x_i, x_j) \in [0, 1]$. When two spatial objects are very dissimilar, the value of s tends to 0. The greater of $s(x_i, x_j)$ means that x_i and x_j are more similar, $s(x_i, x_i) = 1$. Similarity matrix S is a n×n matrix.

Corresponding Author: Pan Xu-Wei, Department of Management Science and Engineering, Zhejiang Sci-Tech University, Hangzhou, 310018, Zhejiang, China

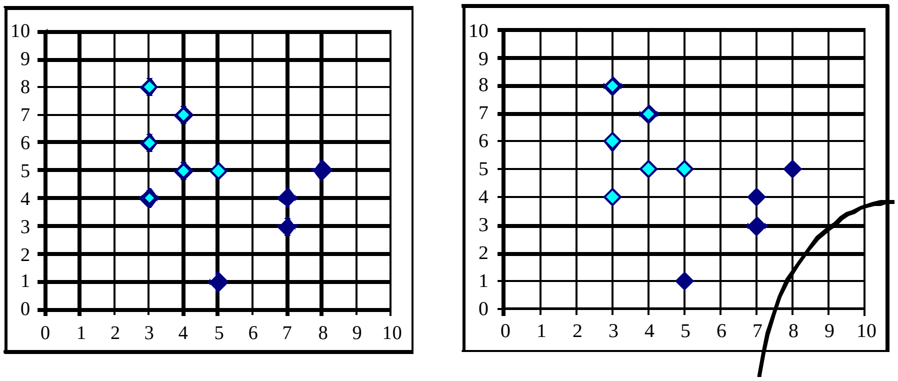

Fig. 1: Features of weighted network for spatial objects clustering

Weighted network: The weighted network means that an edge between two nodes is undirected and weighted. The weighted network is expressed as G = (V, E, W), where V is the set of nodes in the network, E is the set of edges connecting nodes and W is the set of weight of each edge. If any two nodes in the network are connected by an edge, such a network is called as full coupling network. The full coupling network is the densest network.

Node strength: The strength of a node i in weighted network is defined as follows:

$$F(i) = \sum_{j\in N_i} w_{ij}$$

Where N_i is the set of neighbors of node i and w_{ij} is the weight of the edge e_{ij} between node i and j.

Representative node: The representative node is used to represent a group (or cluster) in clustering.

Core idea and process of the algorithm: The idea of WNBAC algorithm comes from the naïve understanding of spatial objects clustering.

As shown in Fig. 1, the left is the clustering result for ten spatial objects by K-means. If the reciprocal of the distance (Euclidean distance) for two nodes is taken as the similarity between them and the nodes are connected with the similarity weighted edge when the similarity between two nodes are in the front 50% of the set of similarities by ascending order, the weighted network for these spatial objects is shown in the right of Fig. 1. Where, the size of nodes and edges indicate the strength value of nodes and the similarity between two nodes respectively. It can be seen that nodes in the same cluster are connected by bigger size edges and there is a bigger node than others in each cluster. These features indicate that similarity weighted network for spatial objects may be used to partition them, so the WNBAC algorithm is put forward.

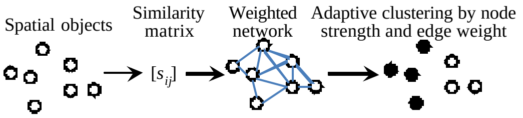

Fig. 2: Main process of WNBAC algorithm

The core idea of WNBAC algorithm is to use different roles of nodes and edges in the similarity weighted network to partition spatial objects. Main process of WNBAC algorithm is shown in Fig. 2. Firstly, the similarity among spatial objects is calculated and the similarity matrix is constructed. Then the weighted network for spatial objects is built in term of the similarity matrix. Finally, the weighted network is adaptively partitioned into groups (clusters) by exploiting the strength of nodes and the weight of edges and final clusters is obtained. Thereby, the clustering of spatial objects is transformed to the partition of nodes in the similarity weighted network.

Building the WNBAC algorithm:

- **Step 1: Build weighted network:** In this step, the similarity among spatial objects is calculated firstly according to characteristics of objects' attributes and the similarity matrix is formed. Then, the weighted network G = (V, E, W) for spatial objects is built, in which spatial objects are taken as nodes and the similarity between two objects is taken as the weight of the connected edge for them. In the building weighted network, the weighted parameter α is introduced. α is the threshold which is used to determine whether two objects are connected. If $s(x_i, x_j)<\alpha$, there is no edge built between the object x_i and x_j; if $s(x_i, x_j)\geq\alpha$, there is an edge built between x_i and x_j and the weight is $s(x_i, x_j)$. When $\alpha = 0$, the weighted network is a full coupling network

- **Step 2: Adaptive clustering:** In this step, nodes' strength and edges' weight in the similarity weighted network are used to partition nodes into groups to achieve spatial objects clustering. Firstly, the strength of every node is calculated. The node with the most strength is taken as the representative node for the first cluster and it is marked as visited. Then, it is determined whether adjacent nodes of this node belong to its represented cluster. The rule for determination is shown as follows: a is the representative node, b is an adjacent node of a, W_b is the set of weights for the edges between b and its adjacent nodes, if the weight of the edge between a and b (w_{ab}) belongs to the front θ% of W_b (W_b is ordered by ascending), the node b is partitioned into the a represented cluster and the node b is marked as visited. The other adjacent nodes of the node a are traversed in turns and the cluster represented by a is formed. After a cluster is formed, the nodes marked as visited in the weighted network are removed. In the rest of weighted network, the same process is conducted in returns until there are no remaining nodes. Thus, all nodes in the weighted network are partitioned into certain groups and the clustering of spatial objects is achieved

    The pseudocode of WNBAC algorithm is as follows:

```
Algorithm: WNBAC algorithm
  Input: X={$x_1$, $x_2$, ..., $x_n$}-spatial objects
           and their attributes;
         $\alpha$-the weighted parameter;
         $\theta$-the partition parameter.
Output:
     Clusters for spatial objects
//Calculate the similarity
S=calculateSim(X)
  //Build the similarity weighted network
G(V, E, W)=buildNet(S, $\alpha$)
  //Adaptive clustering
while(V is not empty) do
  for each node v in V do
  //Compute the strength of nodes
    computeStr(v)
  end for
  insert all computeStr (v) (v in V) by
             ascending order into List L
  select the first vertex $v_i$ in L
  mark $v_i$ as visited
  $C_j$= next cluster
  assign $v_i$ to $C_j$
  for each adjacent node $v_k$ of $v_i$ do
  //Obtain the set of weights for $v_k$
    $W_k$ = getWeight($v_k$, $\theta$)
    if $w_{ik} \in W_k$
      assign $v_k$ to $C_j$
      mark $v_k$ as visited
    end if
  end for
  delete $v \in C_j$ from V
  delete all from L
  obtain a cluster $C_j$ with label = $v_i$
end while
```

Main strategies to implement four functions of calculateSim(X), buildNet (S, α), computeStr (v) and getWeight (v_k, θ) in the algorithm have been stated in above, so details about them are not given here.

Parameters discussion: There are two parameters in WNBAC algorithm, i.e., the weighted parameter α and the partition parameter θ. α is used to determine whether two spatial objects are connected according to the similarity between them. It is suitable from instinct that about half of spatial objects in the dataset are connected by α. So, the probable value of α is set as the similarity value by which the set of similarities is divided into two sets with about the same size. θ is used to determine whether an adjacent node is partitioned into the representative node cluster. Setting θ as 50 maybe suitable from naïve instinct for the reason that if a node grouped into the representative node cluster, the weight of the edge between it and the representative node should be over the half of weights of all its connected edges.

EXPERIMENTAL RESULTS

Collections used in the experiments: Generally, the standard dataset is used to examine the effectiveness of clustering algorithm. So the "iris" dataset and "glass" dataset in UCI standard datasets are used in experimental studies. The "iris" dataset includes 150 spatial objects and each object has four attributes. These 150 objects have been divided into 3 categories. The "glass" dataset includes 214 spatial objects and each object has six attributes. These 214 objects have been divided into 6 categories.

Evaluation measure: The Jaccard coefficient (Halkidi *et al*., 2002) which is an external evaluation measure is adopted in our studies for the reason that two datasets used in the experiment have definite categories. The Jaccard coefficient is bigger, the clustering results is more similar with definite categories, which means the clustering algorithm is more effective.

Parameters examination: Since values of attributes of objects in two experimental datasets are numeric, the Euclidean distance is adopted and the similarity is calculated as follows:

$$s(x_i, x_j) = 1 - \frac{d(x_i, x_j) - \min D}{\max D - \min D}$$

where, $s(x_i, x_j)$ and d (x_i, x_j) are respectively the similarity and Euclidean distance between x_i and x_j, minD and maxD

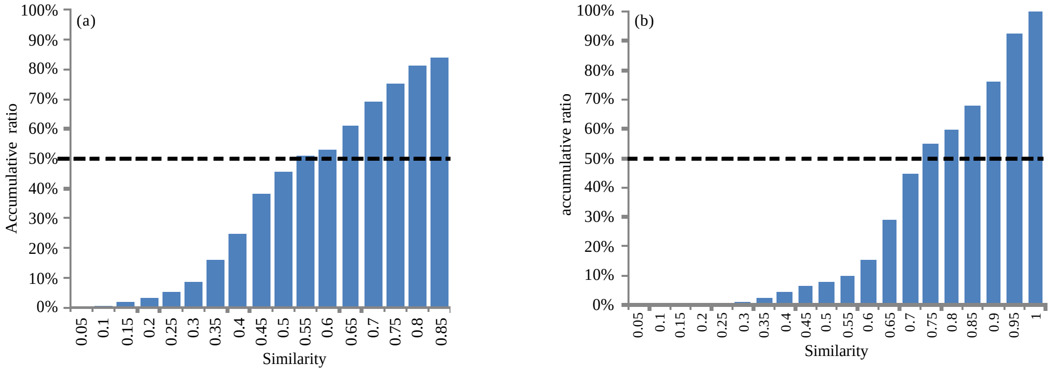

Fig. 3: Similarity accumulative distribution for two experimental datasets, (a) "iris" dataset and (b) "glass" dataset

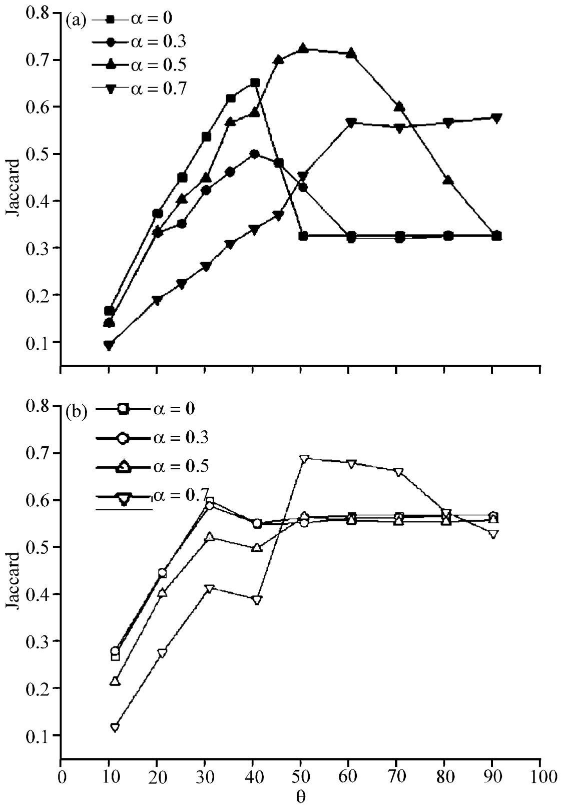

Fig. 4: Experimental results for examining the parameter setting approaches, (a) The "iris" dataset and (b) The "glass" dataset

Table 1: Compared evaluated result

	"iris" dataset (α = 0.5, θ = 0.5) (k = 3)	"glass" dataset (α = 0.7, θ = 0.5) (k = 6)
K-means	0.6823	0.6453
WNBAC	0.7217	0.6915
Improved	5.77%	7.16%

are respectively the minimum and maximum of all distances.

According to the above equation, the similarity accumulative distribution for two dataset is shown in Fig. 3. It is seen that the similarity values are in 0.5-0.55 and 0.7-0.75 for the "iris" dataset and the "glass" dataset, respectively when 50% line is first crossed with the pillars. That means the probable value of the weighted parameter α for the "iris" dataset may be in 0.5-0.55, for the "glass" dataset may be in 0.7-0.75 in term of the discussion of parameters in above.

As discussed in above, the probable value of α is set as the similarity value by which the set of similarities is divided into two sets with about the same size and è is suitable for setting 50. To examine whether such approaches for setting parameters is desirable, different values of α and θ are used in experiment. The examination results are shown in Fig. 4. It can be seen that the best experimental result is occurred in α = 0.5, θ = 50 for the "iris" dataset and α = 0.7, θ = 50 for the "glass" dataset. The results meet with the expectant parameter values, i.e., α is in 0.5-0.55 for the "iris" dataset, in 0.7-0.55 for the "glass" dataset and θ is 50 for both datasets. Therefore, although the approach with little knowledge about dataset for setting α and θ is simple and efficient, it is very effective. Thus, the WNBAC algorithm overcomes the pre-awareness of large scale dataset in many clustering algorithms.

Effectiveness evaluation: To verify the effectiveness of the proposed algorithm for clustering, the comparison between K-means and WNBAC algorithm is conducted. The compared result is shown in Table 1, which shows that the clustering result obtained from the proposed WNBAC algorithm (parameters are set with little knowledge about dataset) is better than from the K-means algorithm (parameters are set with in-depth understanding the dataset). This result implies the effectiveness of the proposed WNBAC algorithm.

CONCLUSION

This study puts forward a Weighted Network Based Adaptive Clustering (WNBAC) algorithm, in which the clustering of spatial objects is transformed to the partition of similarity weighted network. The core idea, main process, building procedure and parameter setting for the WNBAC algorithm are discussed in details. Experimental results indicate that both the parameter setting is efficient and the proposed algorithm is effective. In the future work, we plan to conduct more experiments with different datasets to further verify the effectiveness of the proposed algorithm and improve the performance of the WNBAC algorithm.

ACKNOWLEDGMENTS

We appreciate the support of the National Natural Science Foundation of China under the grant 71071144, the Zhejiang Provincial Natural Science Foundation of China under the grant Z6110334 and LY12F02044 and the '521' program of Zhejiang Sci-Tech University.

REFERENCES

Abul Hasan, M.J. and S. Ramakrishnan, 2011. A survey: Hybrid evolutionary algorithms for cluster analysis. Artif. Intell. Rev., 36: 179-204.

Halkidi, M., Y. Batistakis and M. Vazirgiannis, 2002. Cluster validity methods: Part II. ACM SIGMOD Rec., 13: 40-45.

Khan, S.S. and A. Ahmad, 2013. Cluster center initialization algorithm for *K*-modes clustering. Exp. Syst. Appl., 40: 7444-7456.

Liao, L., J.Z. Zhou and Q. Zou, 2013. Weighted fuzzy kernel-clustering algorithm with adaptive differential evolution and its application on flood classification. Nat. Hazards, 69: 279-293.

Vega-Pons, S. and J. Ruiz-Shulcloper, 2011. A survey of clustering ensemble algorithms. Int. J. Patt. Recognit. Artif. Intell., 25: 337-372.

Xu, R. and D. Wunsch II, 2005. Survey of clustering algorithms. IEEE Trans. Neural Networks, 16: 645-678.

Zhang, Z.K. and C.A. Liu, 2010. A hypergraph model of social tagging networks. J. Stat. Mech. Theory Exp., Vol. 2010. 10.1088/1742-5468/2010/10/P10005