International Journal of Fuzzy Logic and Intelligent Systems, vol. 11, no. 3, September 2011, pp. 178−183
http://dx.doi.org/10.5391/IJFIS.2011.11.3.178

Medoid Determination in Deterministic Annealing-based Pairwise Clustering

Kyung Mi Lee and Keon Myung Lee

Dept. of Computer Science, Chungbuk National University, and PT-ERC
Cheongju, Korea

Abstract

The deterministic annealing-based clustering algorithm is an EM-based algorithm which behaves like simulated annealing method, yet less sensitive to the initialization of parameters. Pairwise clustering is a kind of clustering technique to perform clustering with inter-entity distance information but not enforcing to have detailed attribute information. The pairwise deterministic annealing-based clustering algorithm repeatedly alternates the steps of estimation of mean-fields and the update of membership degrees of data objects to clusters until termination condition holds. Lacking of attribute value information, pairwise clustering algorithms do not explicitly determine the centroids or medoids of clusters in the course of clustering process or at the end of the process. This paper proposes a method to identify the medoids as the centers of formed clusters for the pairwise deterministic annealing-based clustering algorithm. Experimental results show that the proposed method locate meaningful medoids.

Keywords: clustering, data analysis, deterministic annealing, pairwise clustering

1. Introduction

Data clustering is one of exploratory data analysis techniques that groups a data set into partitions so as that each partition contains similar entities but the entities from different partitions have low similarities[1]. Various clustering algorithms have been developed in such domains as statistics, pattern recognition, machine learning, image processing. Some algorithms produce hierarchical structures for clusters and others just form partitions. Clustering results are affected by the representation of data and the adopted similarity measures. The attributes of data can be either discrete ones like categorical values and ordinal values or continuous ones. Hence the similarity measures sometimes need to take into account both discrete and continuous ones together. In some applications, the clustering needs to be performed only when the distance information among data is available, yet the attribute values of data are not given. Such a clustering problem is called the pairwise clustering. On the contrary, the traditional clustering problem for data with attribute values is called the vector-based clustering[3].

Once a cluster formation is done, the representatives of each cluster come to be determined. Those representatives play the role of the reference exemplars for classifying new data, and they also can be used as code vectors for data compression[4]. In vector-based clustering for data with attribute values, the centroid of a cluster can be determined by the attribute-wise weighted means of attribute values with respect to membership degrees of data toward a specific cluster. The pairwise clustering does not allow to get the centroids of clusters because there are no specified attributes and their values. Instead, the medoids can be elected as the representatives of clusters, which are the entities believed to be located in the central region of the clusters.

The distance information between entities in a cluster is used in medoid selection. This paper is concerned with the medoid identification in the deterministic annealing(DA)-based pairwise clustering which neither specifies nor uses the meidoids of clusters. The remainder of this paper is organized as follows: Section 2 briefly presents the clustering methods and the basic ideas of deterministic annealing. Section 3 explains the DA-based pairwise clustering algorithm and Section 4 introduces the proposed medoid identification method for the DA-based pairwise clustering algorithm. Section 5 shows some experiment results of the proposed method and in final Section 6 draws the conclusions.

2. Related Works

2.1 Clustering Methods

Various clustering algorithms have been developed and put in practice for real applications. Clustering algorithms

Manuscript received Aug. 12, 2011; revised Sep. 9, 2011; accepted Sep. 14, 2011.

This work was partially supported through PT-ERC by the Korea Science and Engineering Foundation (KOSEF) grant funded by the Korea government (MEST) in 2011.

can be categorized into the followings: partitioning methods, hierarchical methods, density-based methods, grid-based method, model-based methods[1, 3, 10, 11]. Partition methods construct k partitions of the data where k is the number of partitions usually specified earlier on. It begins with an initial partitioning and then uses an iterative procedure that attempts to improve the partitioning by moving objects from one partition to another. The optimality in partitioning clustering requires the exhaustive search for all possible combinations, hence heuristic approaches have been applied, which can be largely categorized into k-means method and k-medoid method. In k-means methods, a cluster is represented by the mean value of objects belonging to the cluster and the clusters are gradually adjusted by alternating the adjustment of means and the reassignment of objects to clusters with respect to the adjusted means. The k-medoid methods represent each cluster by one of objects that are near to the center of the cluster. The partitioning methods include k-means algorithm, PAM, CLARA, CLARANS, and so on[3].

A hierarchical clustering method creates a hierarchical decomposition of the given set of data objects[3]. According to how the hierarchical decomposition is made, a hierarchal clustering method is classified into either agglomerative or divisive approach. The agglomerative approach starts with clusters with only one object and successively merges closer clusters until the termination condition such as a maximum number of clusters meets. In the divisive approach, the initial cluster is the one that contains the whole data objects. At the successive iteration, a cluster is split into smaller clusters until a termination condition holds. In this category of clustering algorithms, there are AGNES, DIANA, BIRCH, ROCK, Chamelon, and so on[3].

The density-based method forms clusters by continuing to grow the given cluster as long as the density in the neighborhood exceeds some threshold. In order to apply this approach, the notions of neighborhood and density are defined. The distance between data points are defined and the data points within a specific distance, called radius, from a specific data are considered as neighbors. The number of data objects within a given radius is used to define the density. This method can find arbitrarily shaped clusters. DBSCAN, OPTICS, and DENCLUE are typical algorithms following the method[4].

Grid-based methods quantize the data space into a finite number of cells that build a grid structure[3]. All of the clustering operations are carried out on the quantized space. This approach allows fast processing since it works on the grids rather than data themselves. STING and WaveCluster are the representative clustering algorithms to use grid structure for clustering.

Model-based clustering algorithms find clusters by fitting the given data to some mathematical models like mixture models[3]. They are usually based on the assumption that the data are generated from a mixture of underlying probability distributions. EM(Expectation-Maximization) is a typical algorithm to construct clusters by estimating a mixture model[4].

2.2 Deterministic Annealing

Deterministic annealing (DA) is a kind of search technique that behaves like simulated annealing without probabilistic random works[3]. In clustering, a cost criterion called *distortion measure* is considered as a way to evaluate the cluster formation quality. The expected distortion shown is defined as in Eq.(1), where x denotes a data entity, y the exemplar of the cluster to which x is clustered, $p(x, y)$ the joint probability distribution, and $p(y|x)$ the association probability, and $d(x, y)$ the distance or the dissimilarity between x and y.

$$D = \sum_x \sum_y p(x, y)d(x, y) = \sum_x p(x) \sum_y p(y|x)d(x, y) \tag{1}$$

DA approach tries to seek the distribution which minimizes D subject to a specified level of randomness that can be measured by the entropy $H(X, Y)$ [3].

$$H(X, Y) = -\sum_{x \in X} \sum_{y \in Y} p(x, y)\log p(x, y) \tag{2}$$

The search problem can be formulated as minimization of the Lagrangian F, where T is the Lagrangian multiplier, D the expected distortion, H the entropy for the considered distribution.

$$F = D - TH \tag{3}$$

F can be regarded as the free energy of a candidate probability distribution. The distribution to minimize F is the one which minimizes the expected distortion D and maximizes the randomness of the distribution. T is a parameter to control the relative importance of D and H. It begins with a large value and gradually decreases, and thus gradually emphasizes the expected distortion. Minimizing F with respect to the conditional probabilities $p(y|x)$ gives Gibb's distribution to $p(y|x)$, and minimizing F with respect to y produces the formula for the centroid of the corresponding cluster. The DA algorithm for clustering consists of minimizing F with respect to ys, starting at high values of T and tracking the minimum while lowering T and determining $p(y|x)$s and ys [3]. Simulated annealing has been proved to find a solution to follow a Gibbs distribution[3], and DA also finds a Gibbs distribution deterministically by differentiating the function of an expected value of the distortion and the randomness. In that sense, DA shows a similar behavior of simulated annealing with probabilistic random walk.

3. Pairwise DA-based Clustering Algorithm

The above-mentioned DA-based clustering algorithm cannot be applied to the pairwise clustering problem because data do not have any attribute information. To handle this problem, Hofmann and Buchmann[5] proposed a pairwise DA-based clustering algorithm as follows. The algorithm makes use of the notion *mean-field* which account for partial cost to assign an entity to a cluster.

Suppose that there are a data set $X = \{x_i \in R^d | i = 1, \cdots, n\}$ and K clusters. The assignment variable M is a boolean matrix to tell which data element belongs to which cluster.

$$M = (M_{iv})|_{i=1\cdots N, v=1\cdots K} \in \{0,1\}^{N\times K} \tag{4}$$

The cost function for pairwise clustering with K clusters with respect to the cluster assignment M is defined as follows:

$$H^{pc}(M) = \frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{N}\frac{D_{ik}}{N}\left(\sum_{v=1}^{K}\frac{M_{iv}M_{kv}}{p_v} - 1\right) \tag{5}$$

In the above equation, D_{ik} indicates the distance between data d_i and $d-k$, and p_v the percentage of data in that cluster, i.e., $p_v = \sum_{i=1}^{N} M_{iv}/N$.

Let E_{iv} denote the mean-field of data d_i to cluster C_v, $E = E_{iv}|_{i=1\cdots N, v=1\cdots K}$ the mean-field values of N data to K clusters. Assignments M of data to clusters are supposed to be randomly drawn from the set of admissible configurations with the constraint of Eq.(4) according to the Gibbs distribution.

$$P^{Gb}(H^{pc}(M)) = \frac{exp(-H^{pc}(M)/T)}{\sum_{M'\in M} exp(-H^{pc}(M')/T)} \tag{6}$$

Although each assignment variable M_{iv} interacts with all other assignment variables, the average interaction of M_{iv} with other assignment variables are approximated by a mean-field $\mathcal{E}_{iv}$. A mean-field approximation of the Gibbs distribution $P^{Gb}(H^{pc})$ neglects the correlations between the stochastic variables and determines the most similar factorized distribution within an $\mathcal{E}$-parameterized family of distributions $P^0(\mathcal{E})$.

$$H^0(M,\mathcal{E}) = \sum_{i=1}^{N}\sum_{v=1}^{K} M_{iv}\mathcal{E}_{iv} \tag{7}$$

$$P^0(\mathcal{E}) \equiv P^{Gb}(H^0) \tag{8}$$

The distribution $P^0(\mathcal{E}^*)$ which represents most accurately the statistics of the original problem is determined by the minimum of the Kullback-Leibler divergence $\mathcal{I}$ of the original Gibbs distribution as shown in Eq.(9).

$$\mathcal{E}^* = \arg\min_{\mathcal{E}} \mathcal{I}(P^0(\mathcal{E})||P^{Gb}(H^{pc})) \tag{9}$$

The mathematical handling of Eq.(9) gives the optimal potential $\mathcal{E}^*_{iv}$ and the cluster assignments $\langle M_{i\alpha}\rangle$ as follows:

$$\mathcal{E}^*_{iv} = \frac{1}{\sum_{j=1,j\neq i}^{N}\langle M_{jv}\rangle + 1}\sum_{k=1}^{N}\langle M_{kv}\rangle \left(D_{ik} - \frac{1}{2\sum_{j=1,j\neq i}^{N}\langle M_{jv}\rangle}\sum_{j=1}^{N}\langle M_{jv}\rangle D_{jk}\right) \tag{10}$$

$$\langle M_{i\alpha}\rangle = \frac{exp(-\mathcal{E}^*_{i\alpha}/T)}{\sum_{v=1}^{K} exp(-\mathcal{E}^*_{iv}/T)} \tag{11}$$

The averaging brackets $\langle\cdot\rangle$ denote the average with respect to $P^{Gb}(H^0)$. The Hofmann and Buhmann's pairwise DA-based clustering algorithm is an EM algorithm which iterates the steps of Expectation and Maximization as follows: In the procedure, α indicates the cooling rate for the temperature parameter T, T_0 the starting temperature, and T_{final} the final temperature at which annealing stops.

```
procedure : Pairwise DA-based clustering
  Initialize E*(0)_iv and ⟨M_iv⟩^(0) at random.
  T ← T_0.
  while (T > T_final)
     t ← 0
     do
        estimate ⟨M_iv⟩^(t+1) as a function
          of E*(t)_iv.
        calculate E*(t+1)_iv for given
          ⟨M_iv⟩^(t+1).
        t ← t + 1.
     until all parameters are converged.
     T ← αT; ⟨M_iv⟩^0 ← ⟨M_iv⟩^(t); E*(0)_iv ← E*(t)_iv
  end while.
```

4. Medoid Determination in Deterministic Annealing-based Pairwise Clustering

Clustering is to form clusters with similar data entities which are separated from other clusters.The representatives of each cluster are, sometimes, needed to classify unseen data or to characterize clusters. In vector-based clustering, the centroids of clusters can be easily determined by averaging their members.Due to lack of attribute value information in pairwise clustering, a medoid which is one of members for the cluster, is selected as the representative of the cluster.

The pairwise DA-based clustering algorithm itself neither finds the centroids or the medoids of clusters in the course of cluster formation. We are concerned with the medoids of cluster results by the algorithm because it is inherently impossible to get the centroids. We pay attention to the mean-field values $\mathcal{E}_{iv}$ which expresses the partial cost

of assigning a data d_i to cluster v. The data with the minimum mean-field can be regarded as the closest one to the center of the cluster.

Once the clustering is done by the pairwise DA-based clustering algorithm, the clusters C_v with the membership degree M_{iv} of data d_i and the mean-field values $\mathcal{E}_{iv}$ are given. From the above observations, the medoid μ_v can be defined as follows:

$$\mu_v = \arg\min_i \{\mathcal{E}^*_{iv} | \langle M_{iv} \rangle \geq \theta\} \tag{12}$$

$$\mu_v = \arg\min_i \{\frac{\sum_{\langle M_{kv} \rangle \geq \theta} \mathcal{E}^*_{ik}}{|\{k | \mathcal{E}^*_{ik} \geq \theta|}\} \tag{13}$$

$$\mu_v = \arg\min_i \{max_{\langle M_{kv} \rangle \geq \theta} \mathcal{E}^*_{ik}\} \tag{14}$$

In the above definition, θ indicates the minimum membership degree of a data to be regarded as a member of a cluster. The threshold allows to control the outliers not to affect the determination of medoids. When it is doubtful to have a single data as a medoid, it is possible to select multiple ones to represent a cluster. The medoid set M_v of a cluster C_v is defined as follows: Here θ_{rv} denotes the radius threshold from the cluster center within which the data are considered as medoids.

$$M_v = \{d_i | \mathcal{E}^*_{iv} \leq \theta_{rv}, \min\{\mathcal{E}^*_{kv} | \langle M_{kv} \rangle \geq \theta\} < \theta_{rv}\} \tag{15}$$

The next definition chooses a medoid which shows biggest difference with respect to other clusters' medoid sets.

$$\mu_v = \arg\min_i \{\frac{\mathcal{E}^*_{iv}}{\min\{\mathcal{E}^*_{ik} | d_k \in M_{v'}, v' \neq v\}} | \langle M_{iv} \rangle \geq \theta\} \tag{16}$$

The formula of Eq.(12)-(14) and (16) can be used to determine medoids of clusters formed by the pairwise DA-based clustering algorithm. The choice of one of them depends on which aspect the analysts emphasize.

In order to determine the medoids for the DA-based pairwise clustering, the following procedure can be applied: First, make the cluster formation using the procedure `Pairwise DA-based clustering` and get the parameter values $\mathcal{E}^*_{iv}$ and $\langle M_{iv} \rangle$. Next, for each cluster, its medoid is determined by the selected medoid selection method.

The distance between cluster can be determined in a various way like minimum distance, maximum distance, mean distance, average distance, and so on. Once medioids are available, we can assign mean distances among clusters as follows: Here $CD(i, j)$ denotes the distance between clusters i and j, μ_i and μ_j their medoid, and $d(a, b)$ the distance between objects a and b.

$$CD(i, j) = d(\mu_i, \mu_j) \tag{17}$$

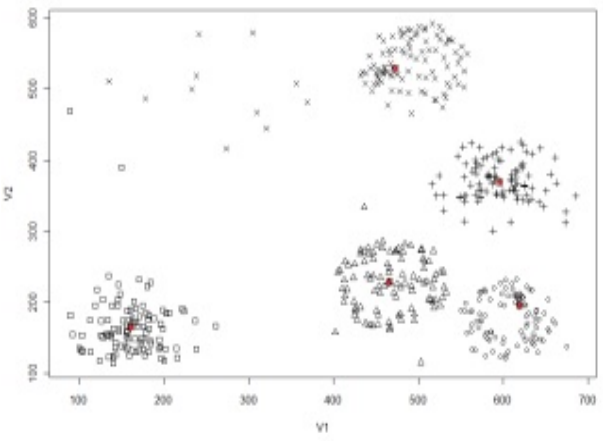

Figure 1: An aritificially generated data set used in the first experiment and the cluster medoids identified by the proposed method, on which medoids are marked by asterisks

5. Experiments

In order to show that the proposed method for DA-based pairwise clustering is capable of identifying medoids, two experiments have been carried out. The first data set is an artificially generated data set of size 500 with five clusters in two dimensional space, illustrated in Fig. 1 on which the identified medoid locations are marked by asterisks.

The second data set was the IRIS data that consists in 150 data having 4 attributes along with class label. Fig. 2 illustrates the clustering results obtained by the pairwise clustering algorithm and the medoid determination method. The positions marked by the star indicates the selected medoid for the clusters. For comparison, the k-means algorithm was applied to the IRIS data. In Fig. 3, the asterisks indicates the centroids found by the k-means algorithm. We could observe that the medoid points by the proposed method are located in comparable to mean positions by k-means algorithm.

6. Conclusions

The DA-based clustering algorithm is a fast EM-based algorithm which behaves like simulated annealing method, yet less sensitive to the initialization of parameters. The pairwise DA-based clustering algorithm repeatedly alternates the steps of estimation of mean-fields and the update of membership degrees of data objects to clusters until termination condition holds. Then the algorithm does not explicitly determine the centroids or medoids of clusters in the course of clustering process or at the end of the process. This paper proposed a method to identify the medoids as the centers of formed clusters for the pairwise DA-based

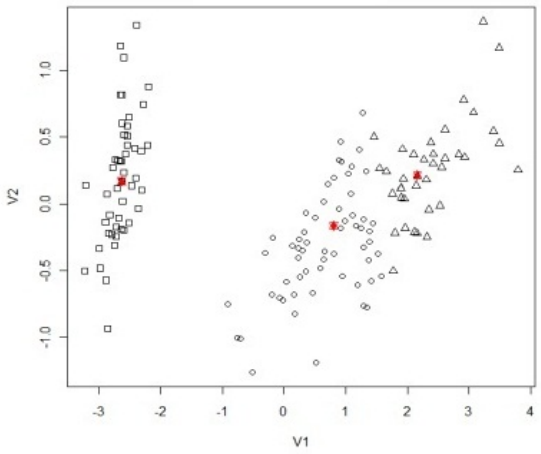

Figure 2: The medoids for the IRIS data which are obtained by the pairwise clustering and the proposed medoid identification method, marked by asterisks

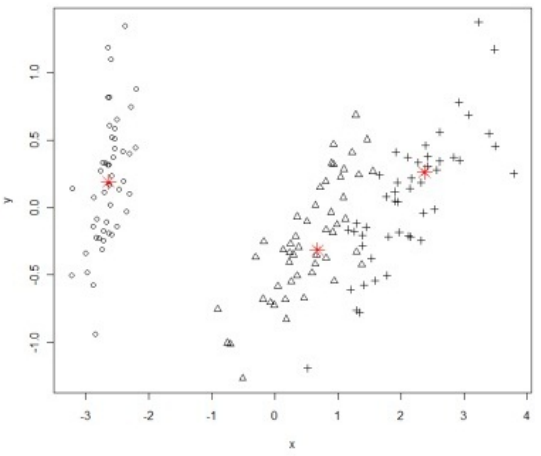

Figure 3: The centroids for the IRIS data which are determined by the k-means algorithm, marked by asterisks

clustering algorithm. In order to show the applicability of the proposed method, it has applied to a synthetic data set of 500 data objects and the IRIS data set. It has been observed that the locations of the identified medoids are close to those of the means obtained by the k-means algorithm. This implies that the proposed method could find the meaning medoids for the pairwise DA-based clustering algorithm that is not equipped to provide them.

References

[1] A. K. Jain, M. N. Murty, P. J. Flynn, "Data Clustering: A Review, *ACM Computing Surveys*," vol.31, no.3, pp.264–323, 1999.

[2] P. -N.Tan, M. Steinbach, V. Kumar, *Introduction to Data Mining*, Addison-Wesley, 2006.

[3] J. Han, M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, 2006.

[4] K. Rose, "Deterministic Annealing for Clustering, Compression, Classification, Regression, and Related Optimization Problems, *Proc. of the IEEE*," vol.86, no.11, pp.2210–2239, 1998.

[5] G. Fung, *A Comprehensive Overview of Basic Clustering Algorithms*, June 2001.

[6] T. Hofmann, J. M. Buhmann, "Pairwise Data Clustering by Deterministic Annealing," *IEEE Trans. on PAMI*, vpl.19, no.1, pp.1–14, 1997.

[7] Z. Yanjie, W. Shuanhu, "A Pairwise Clustering based Biclustering Method," *Proc. of 2nd Int. Conf. on Signal Processing Systems*, pp.V1.311–314, 2010.

[8] K. A. Arai, R. Barakbah, "Hierarchical K-means: an algorithm for centroids initiliazation for K-means," *Reports of the Faculty of Science and Engineering, Saga University*, vol.36, no.1, pp.25–31, 2007.

[9] X. Yang, Q. Song, A. Cao, "A Weighted Deterministic Annealing Algorithm for Data Clustering," *Int. J. of Computational Intelligence Research*, vol.2, no.1, pp.81–85, 2006.

[10] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[11] B. Clarke, E. Fokoue, H. H. Zhang, *Principles and Theory for Data Mining and Machine Learning*, Springer, 2009.

Kyung Mi Lee

PhD course student, Department of Computer Science of Chungbuk National University
Research Area: soft computing, artificial intelligence applications
E-mail : kmlee07@cbnu.ac.kr

Keon Myung Lee

Professor, Department of Computer Science of Chungbuk National University
Research Area: machine learning, data mining, bioinformatics, cloud computing
E-mail : kmlee@cbnu.ac.kr