

A Weighted Deterministic Annealing Algorithm for Data Clustering

Xulei Yang¹, Qing Song¹ and Aize Cao²

¹School of Electrical and Electronic Engineering,
Nanyang Technological University, Singapore 329978
{yangxulei, eqsong}@pmail.ntu.edu.sg

²Medical Center, Vanderbilt University,
VU Station B, #351631, Nashville, TN, 37235
aizecao@vanderbilt.edu

Abstract: The deterministic annealing (DA) approach to clustering and its extensions has demonstrated substantial performance improvement over standard supervised and unsupervised learning methods. However, its performance will be severely distorted if there exists noise (or outliers) in the given data set. This paper proposes a new robust clustering method -- weighted deterministic annealing (WDA) algorithm, which attempts to solve the noise sensitivity problem by reformulating the source distribution of conventional DA approach. The proposed method generates different weight for different pattern: bigger weight for good (non-noisy) pattern and smaller weight for noisy pattern, which effectively reduced the effect of noise on the final partitions. The superiority of the proposed method is supported by simulation results.

Keywords: Deterministic Annealing, Robust Clustering, Supervised and Unsupervised Learning, Noise Sensitivity.

I. Introduction

Interest in clustering has increased recently because of new areas of application, such as data mining, image and speech processing, and bio-informatics. A very rich literature on clustering has developed in the past three decades. In general, fuzzy clustering has been shown to have advantages over hard clustering in that total commitment of a vector to a given class is not required. In the fuzzy clustering literature, the fuzzy c-means clustering algorithm, proposed by Dunn [1] and extended by Bezdek [2] has been successfully applied to a wide variety of problems [2]. However, one of the serious disadvantages of this method is its sensitivity to noise (or outliers) in data set. In this case, partitioned cluster centers can be placed far away from the true positions.

Many researchers have attempted to overcome this problem by modifying the constraint of probability in the memberships of FCM. The noise clustering (NC) [3] defines an extra noise cluster, which attempts to include all the noise with high membership value, such that the good (true) clusters

are mostly composed of good (non-noisy) patterns. The possibilistic c-means (PCM) [4] is a possibilistic clustering method in which one input pattern slightly affects good clusters if its possibility to be noise is high, such that the partitioned results are slightly distorted by noise. Both methods work satisfactorily only provided appropriate values of the related resolution parameters are specified [5]: δ for NC and η for PCM. But the problem is the suitable values of the resolution parameters are difficult to specify for a certain clustering problem. Another serious problem with these two methods is their sensitivity to the initialization of the cluster prototypes, which makes them unrealistic for practical applications. Actually, these problems also occur in other robust clustering methods, such as least biased fuzzy clustering algorithm (LBFC) [6], fuzzy possibilistic c-means algorithm (FPCM) [7] and credibilistic fuzzy c-means algorithm (CFCM) [8], and so on.

In this paper, we presented a novel robust clustering method -- weighted deterministic annealing (WDA) algorithm, which attempts to solve the noise sensitivity problem by reformulating the source distribution of conventional DA approach. The proposed method generates different weight for different pattern: bigger weight for good (non-noisy) pattern and smaller weight for noisy pattern, which effectively reduced the effect of noise on the final partitions. As a result, the proposed method performs better than existing clustering approaches in term of independence of initialization of cluster prototypes, robustness against noise, and non-requirement of any resolution parameter. The rest of this paper is organized as follows. In section 2, we first review deterministic annealing clustering algorithm, then develop the proposed novel robust clustering method. The experimental results of the proposed clustering method are reported in section 3. Finally, conclusion and discussion are presented.

II. The Proposed Clustering Algorithm

A. Review of Conventional DA Clustering Algorithm

Let the input data set be $X = \{x_1, x_2, \dots, x_L\} \subset R^n$, where L is the number of input patterns, and n is the dimension of input space. Based on a measurement of similarity (most used one is the squared Euclidean distance), the data set is partitioned into K clusters denoted by X_1, X_2, \dots, X_K . Let $V = \{v_1, v_2, \dots, v_K\} \subset R^n$ be the set of according cluster centers. The conventional deterministic annealing (DA) algorithm [9] [10] aims to minimize the following Lagrangian formulation

$$F = J_e - TH_s \quad (1)$$

Where T is the Lagrange multiplier, which is analogical to the temperature in statistic physic, J_e is the cost function defined by

$$J_e = \sum_{j=1}^L \sum_{k=1}^K p(x_j) p(v_k | x_j) d(x_j, v_k) \quad (2)$$

And H_s is the conditional Shannon entropy defined by

$$H_s = - \sum_{j=1}^L \sum_{k=1}^K p(x_j) p(v_k | x_j) \log p(v_k | x_j) \quad (3)$$

Where $d(x_j, v_k)$ is the squared Euclidian distance between x_j and v_k , and $p(v_k | x_j)$ is the association probability relating input pattern x_j with cluster center v_k , which is similar to the membership u_{kj} in FCM algorithm. According to the maximum entropy principle [11], the probability distribution that minimizing F is Gibbs distribution

$$p(v_k | x_j) = \frac{p(x_j) e^{-\frac{d(x_j, v_k)}{T}}}{\sum_{i=1}^K p(x_j) e^{-\frac{d(x_j, v_i)}{T}}} \quad (4)$$

The denominator is normalization function, which is analogical to partition function in statistic physic. Minimizing F we get the expression of cluster center as

$$v_k = \frac{\sum_{j=1}^L p(x_j) p(v_k | x_j) x_j}{\sum_{j=1}^L p(x_j) p(v_k | x_j)} \quad (5)$$

The equations (4) and (5) are just the conventional DA clustering algorithm. Detailed discussions about derivative and implementation of DA can be found in [10].

B. Noise Sensitivity Problem of DA

The deterministic annealing (DA) approach to clustering and its extensions has demonstrated substantial performance improvement over standard supervised and unsupervised

learning methods [10]. But its performance will be distorted severely when there exists noise in the data set. In this case, partitioned cluster centers can be placed far away from the true positions. The reason is that the constraint on the association probability of clusters is equivalent to one

(namely $\sum_{k=1}^K p(v_k | x_j) = 1$), such that the noisy patterns

will inevitably affect the cluster centers during the clustering procedure as discussed in [5]. In practical implementation, the source probability $p(x_j)$ is supposed to be identically

distributed (namely, $p(x_j) = 1/L$) because we don't have

any prior information about the data set beforehand. From (5), we found there may be two ways to improve the robustness of conventional DA against noise: one method is to relax the constraint on the association probability to make

$0 \leq \sum_{k=1}^K p(v_k | x_j) \leq 1$ instead of $\sum_{k=1}^K p(v_k | x_j) = 1$. The

other method is to reformulate the source probability such that the good (non-noisy) patterns will have stronger effects (bigger weights) on cluster centers than the noisy patterns. In [12], the basic idea of NC is used for DA to relax the constraint on the association probability of clusters so that the developed RDA algorithm is robust against noise, which belonged to the first method. This paper focuses on the second method: we reformulate the source probability $p(x_j)$ to propose the so called weighted deterministic annealing (WDA) clustering algorithm, which is robust against noise but does not require to specify any resolution parameter.

C. The Proposed WDA Clustering Algorithm

We attempt to reduce the noise sensitivity of DA clustering algorithm by modifying the source distribution $p(x_j)$ of data set such that the algorithm generates different weight for different pattern: big weight for the good pattern which is normally near the cluster centers and small weight for noisy pattern which is normally far from the cluster centers. If a pattern has a low value of weight, it is atypical to the data set and considered as a noisy pattern. To do this, we reformulate the source distribution by

$$p(x_j) = \frac{e^{-\frac{\min_k d(x_j, v_k)}{T}}}{\sum_{i=1}^L e^{-\frac{\min_k d(x_i, v_k)}{T}}} \quad (6)$$

instead of the following in the conventional DA clustering algorithm [9][10]

$$\sum_{k=1}^K p(v_k | x_j) = 1 \quad (7)$$

Equation (6) assigns different weight to different pattern according the minimum distance between the data pattern and all the cluster centers. The smaller the distance is, the bigger the weight is. Such that the good pattern near the cluster

centers is assigned bigger weight so that has bigger effect on the cluster centers and the noisy pattern far from cluster centers is assigned smaller weight so that has smaller effect on the cluster centers. From (6), it is clear that at limited high T , these are uniform source distribution, each pattern equally affects all the clusters, i.e., all the weights are equal. As T is lowered, the source distribution becomes more discriminating. The good patterns get more and more weights and the noise lose more and more weights. And at limited low T , the weights of noise become nearly zero, such that the effect of noise can be eliminated. This is the basic idea of the proposed clustering method against noise.

The proposed weighted deterministic annealing (WDA) is reformulated as minimization of the following Lagrangian

$$F_w = \sum_{j=1}^L \sum_{k=1}^K p(x_j) p(v_k | x_j) d(x_j, v_k) + T \sum_{j=1}^L \sum_{k=1}^K p(x_j) p(v_k | x_j) \log p(v_k | x_j) \quad (8)$$

Note (8) has the same expression as conventional DA, but in this formulation the source distribution is now defined by (6) instead of (7). Similar to conventional DA, the probability distribution $p(v_k | x_j)$ in above equation minimizing F_w is Gibbs distribution defined by (4) (according to maximum entropy principle [12]). The corresponding minimum of F_w is obtained by plugging (4) into (8), which is

$$F_w^* = \min_{p(v_k | x_j)} F_w = -T \sum_{j=1}^L p(x_j) \log \sum_{k=1}^K e^{-\frac{d(x_j, v_k)}{T}} \quad (9)$$

Minimize (9) against v_k we have

$$\frac{\partial}{\partial v_k} F_w^* = 0 \Rightarrow \sum_{j=1}^L p(x_j) e^{-\frac{d(x_j, v_k)}{T}} (x_j - v_k) = 0 \quad (10)$$

Divide the above equation by the normalization factor

$$Z_{x_j} = \sum_{k=1}^K p(x_j) e^{-\frac{d(x_j, v_k)}{T}} \quad (11)$$

which leads to

$$\sum_{j=1}^L \frac{p(x_j) e^{-\frac{d(x_j, v_k)}{T}}}{Z_{x_j}} x_j = \sum_{j=1}^L \frac{p(x_j) e^{-\frac{d(x_j, v_k)}{T}}}{Z_{x_j}} v_k \quad (12)$$

Use (4) the above equation can be rewritten as

$$\sum_{j=1}^L p(x_j) p(v_k | x_j) x_j = \sum_{j=1}^L p(x_j) p(v_k | x_j) v_k \quad (13)$$

then we get the expression of cluster centers as

$$v_k = \frac{\sum_{j=1}^L p(x_j) p(v_k | x_j) x_j}{\sum_{j=1}^L p(x_j) p(v_k | x_j)} \quad (14)$$

Note (14) has the same expression as conventional DA (see (5)), but in this expression the source distribution $p(x_j)$ is now defined by (6) instead of (7).

Equations (4), (6) and (14) are the proposed WDA clustering algorithm. The advantages of WDA over existing clustering approaches will be shown in the next section. Here we mention a few important properties of WDA algorithm: First, the cluster center is obtained through annealing procedure and independent of the initialization. Second, the characteristic of the entropy maximization is to gain probability relationship among all the training data through the soft clustering procedure, which is similar in philosophy with the fuzzy clustering. Third, the source distribution defined by (6) assigns different weight to different pattern in data set, which is the basic idea of proposed clustering method against noise.

III. Simulation Results

The effectiveness of the proposed WDA clustering algorithm is supported by several artificial data sets. In all simulation results, the original data set and partitioned results are displayed by using different marks: the original patterns are marked by “circle”, the centers of good clusters (assume the noise are artificially removed) are marked by “diamond”, the partitioned cluster patterns are marked by “circle” and “cross”, the partitioned cluster centers are marked by “star”.

Fig.1 shows the robustness of WDA against noise through the comparison of performances between conventional DA and the proposed WDA. The two original noisy data sets are shown in (a) and (d). In each case, there are two assumed clusters and some noisy patterns. We can see the partitioned

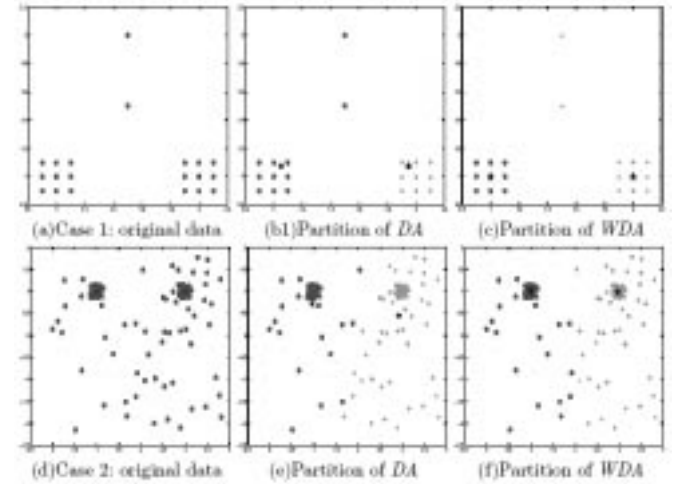


Figure 1. The robustness of the proposed WDA algorithm against noise compared to conventional DA algorithm.

cluster centers of DA are obviously distorted by noisy patterns as shown in (b) and (e). While the proposed WDA has much better performance than DA, the partitioned cluster centers of WDA are hardly distorted by the noisy patterns as

shown in (c) and (f). It is also observed from simulations (which we do not show in this paper for space limitation) that for clear data set (supposed noisy patterns are artificially removed) DA and WDA have similar performances and the partitioned cluster centers of these two methods are almost coincided.

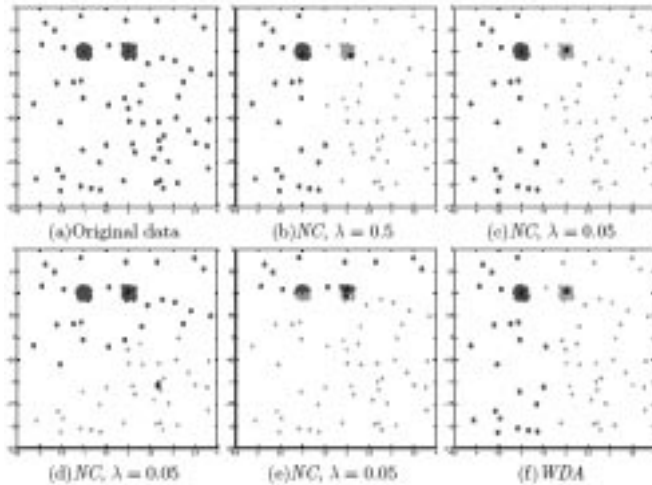


Figure 2. The advantage of the proposed WDA algorithm over conventional NC algorithm.

Fig.2 shows the advantages of the proposed WDA algorithm over other robust clustering algorithms mentioned in section 1. For simplification, here we just compare the performance between WDA and NC. We can see the clustering results of NC depend on the value of λ (which directly determines the value of the resolution parameter δ [3]). Different values may lead to different results, which can be obtained from the different positions of the cluster centers at different values of λ , as shown by (b)(c). Note that the value of λ is difficult to specify for a certain clustering problem. Even a suitable value is specified ($\lambda = 0.05$ for this case), the clustering results of NC may be different in different runs due to the sensitivity to the initialization of cluster prototypes, as shown by (c)(d)(e). While our proposed WDA is independent of such problems and the clustering result is much better as shown in (f).

IV. Conclusion and Discussion

In this paper, we have proposed a novel robust clustering method --weighted deterministic annealing (WDA) algorithm, which performs better than existing clustering approaches in term of independence of initialization of cluster centers, robustness against noise, and non-requirement of resolution parameters. The superiority of the proposed clustering method is supported by simulation results. Compared to DA, WDA is less sensitive to noise (or outliers). Compared to NC, WDA is hardly dependent on initialization of cluster prototypes, and does not need to specify any resolution parameter. One mayor issue needed further

investigation is how to determine the right number of clusters for a certain data set. Though it is not discussed in this paper, we believe this problem can be solved on the basis of entropy changing, as discussed in [6].

Acknowledgment

The authors would like to thank the anonymous reviewers for their valuable comments on an early version of this paper.

References

- [1] J.C. Dunn. "A Fuzzy Relative of the ISODATA Process and its Use in Detecting Compact Well-Separated Cluster", *Journal of Cybernetics*, 3(3), pp. 32-57, 1973.
- [2] J.C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, New York, 1981.
- [3] R.N. Dave. "Characterization and Detection of Noise in Clustering", *Pattern Recognition Letters*, 12(11), pp. 657-664, 1991.
- [4] R. Krishnapuram, J.M., Keller. "A Possibilistic Approach to Clustering", *IEEE Trans. on Fuzzy Systems*, 1(2), pp. 98-110, 1993.
- [5] R.N. Dave, R. Krishnapuram. "Robust Clustering Methods: A Unified View", *IEEE Trans. on Fuzzy Systems*, 5(2), pp. 270-293, 1997.
- [6] G. Beni, X. Liu. "A Least Biased Fuzzy Clustering Method", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 16(9), pp. 954-960, 1994.
- [7] N.R. Pal, K. Pal, J.C. Bezdek. "A Mixed C-Means Clustering Model". In *Proceedings of the IEEE International Conference on Fuzzy Systems*, pp. 11-21, 1997.
- [8] K.K. Chintalapudi, M. Kam. "The Credibilistic Fuzzy C-Means Clustering Algorithm". In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*, pp. 2034-2039, 1998.
- [9] K. Rose, E. Gurewitz, G.C. Fox. "Statistical Mechanics and Phase Transitions in Clustering", *Physical Review Letters*, 65(8), pp. 945-948, 1990.
- [10] K. Rose. "Deterministic Annealing for Clustering, Compression, Classification, Regression, and Related Optimization Problems", *Processings of the EEE*, 86(11), pp. 2210-2239, 1998.
- [11] E.T. Jaynes. "Information Theory and Statistical Mechanics", *Physical Review*, 106(4), pp. 620-630, 1957.
- [12] X.L. Yang, Q. Song, S. Liu. "A Robust Deterministic Annealing Algorithm for Data Clustering". In *Proceedings of the IEEE International Joint Conference on Neural Networks*, In Press, 2005.

Author Biographies

Xulei Yang born in Zhejiang, China, 1977. He received the B.E. degree and M.E. degree in EE School, Xi'an Jiaotong University in 1999 and 2002 respectively. He just finished the PhD study in EEE School, NTU in 2005. His current research interests include pattern recognition, image processing, and machine vision.

Qing Song received the B.S. and the M.S. degrees from Harbin Shipbuilding Engineering Institute and Dalian Maritime University, China in 1982 and 1986, respectively. He obtained the PhD degree from the Industrial Control

Center at Strathclyde University, U.K in 1992. He is currently an associate professor and an active industrial consultant at the school of EEE, NTU. His current research interests focus on a few computational intelligence related research programs targeted for practical applications.

Aize Cao received the B.S. degree in Beijing Institute of Technology in 1993, M.S. degree in Changchun Institute of Optics and Fine Mechanics, Chinese Acad. SCS. in 1996, and PhD degree in EEE School, NTU in 2005, respectively. She is now a research fellow in medical center, Vanderbilt University. Her current research interests include data clustering, image segmentation, and medical image analysis.