An Alternative to Center-Based Clustering Algorithm Via Statistical Learning Analysis

Rui Nian[1,2], Guangrong Ji[2], and Michel Verleysen[1]

[1] Machine Learning Group, DICE, Université catholique de Louvain, Place du Levant, 3-B-1348 Louvain-la-Neuve, Belgium
[2] College of Information Science and Engineering, Ocean University of China, Qingdao, China, 266003
nianrui_80@163.com, grji@mail.ouc.edu.cn,
michel.verleysen@uclouvain.be

Abstract. This paper presents an alternative for center-based clustering algorithms, in particular the k-means algorithm, via statistical learning analysis. The essence of statistical learning principle, i.e., both the empirical risk and structural assessment, is taken into particular consideration for the clustering algorithm so as to derive and develop the relevant minimization mathematical criterion with automatic parameter learning and model selection in parallel. The proposed algorithm roughly decides on the number of clusters, by earning activation for the winners and assigning penalty for the rivals, so that the most competitive center wins for possible prediction and the extra ones are driven far away when starting the algorithm from a too large number of clusters without any prior knowledge. Simulation experiments prove the feasibility of the algorithm and show good performances of the double learning tasks during clustering.

1 Introduction

Clustering, an unsupervised learning process to pursue natural groups among unlabeled data, is one of the most important tasks in Intelligent Computing and Machine Learning. Typically, there is still no perfect solution for the generation and evaluation of clusters. Clustering algorithms proposed in the literature try to seek the minimization of certain mathematical criterion as good as possible [1-4].

In practice, center-based clustering has shown its generality, maneuverability and effectiveness for many applications [1-3]. Usually assuming that each cluster adheres to a unimodal distribution, center-based clustering algorithms try to make each center describe the truth in a single cluster drawn from one mode. However, there still remain some problems not yet completely solved in center-based clustering. First, the best value for the number of clusters is not always clear; it is usually required to specify the number of clusters beforehand in most cases, which is often an ad hoc decision based on prior knowledge, assumptions, and practical experience, and becomes more difficult in a high dimensional space. Second, in mixture distribution cases, it is still a NP hard problem to select and partition data into approximately original clusters which capture and reflect the natural attributes among data.

D.-S. Huang et al. (Eds.): ICIC 2008, LNAI 5227, pp. 693–700, 2008.

In virtue of statistical tools, problems existing in clustering could be to some extent discovered, analyzed and solved. Classical statistics typically focuses on sufficient statistic cases, while statistical learning theory is a machine learning principle to explore the inherent distribution, dependence structure, and generalization ability as good as possible from a finite sample size [5, 6]. Vapnik first put forward statistical learning theory for model complexity based on a minimal capacity measure - VC dimension confidence [5]. Lei Xu also proposed a general statistical learning framework, Bayesian Ying-Yang harmony learning theory, for simultaneous parameter learning and model selection [6].

In this paper, with the help of the general statistical learning principle from a direct perspective, we heuristically explore an alternative conceptually equivalent to the previous work [4, 7], for center-based clustering algorithms, in particular for k-means algorithm, by taking both the empirical risk and structural assessment into consideration. We derive the relevant minimization mathematical criterion with joint parameter learning and model selection, try to seek one solution to learn about both the range and the number of the clusters in mixture distribution cases simultaneously, and accomplish the double learning tasks during clustering procedure. Simulation experiments prove the feasibility of the algorithm and show good performances for both the clustering itself and the estimation on the number of clusters.

2 Center-Based Clustering Algorithms

Center-based clustering algorithms consider that each cluster follows a unimodal distribution and attempt to seek centers from natural clusters [3]. Given an input data set $X=\left\{x_t\right\}_{t=1}^{N}$ drawn from K^* true clusters, the task of center-based clustering is to partition X into K categories, each being represented by an inner center y_ℓ in the representation domain $Y=\left\{y_\ell\right\}_{\ell=1}^{K}$ in a machine learning system. The k-means algorithm is one of the most popular center-based clustering algorithms. The basic idea of k-means algorithm is to partition data into clusters with the objective that tries to achieve the minimization of the total intra-cluster variance, or, the Mean Square Error (MSE) function [1]. Similar to the k-means algorithm, the Expectation-Maximization (EM) algorithm for mixtures of Gaussians is another widely studied method in center-based clustering algorithms, which maximizes the likelihood estimation in probabilistic models that depends on unobserved latent variables [2].

In this paper, for the center-based clustering algorithms, we lay emphasis on the k-means algorithm and explore some improvements. In general, the membership for k-means algorithm is:

$$M(y_\ell|x_t)=\begin{cases}1 & if\ \ell=\ell_t \\ 0 & otherwise\end{cases}, \qquad (1)$$

$$\ell_t=\arg\min_j d^2(x_t,y_j)$$

where $d(x_t, y_j)$ is the similarity measure between x_t and y_j, and each input is assigned to the nearest cluster label ℓ_t. The objective function for optimization in the k-means algorithm is the Mean Square Error function as follows:

$$E_{MSE} = \frac{1}{N}\sum_{\ell=1}^{K}\sum_{t=1}^{N} M(y_\ell | x_t) d^2(x_t, y_\ell) = \frac{1}{N}\sum_{t=1}^{N} d^2(x_t, y_{\ell_t}) \ . \qquad (2)$$

Here Y is obtained by minimizing this objective function $\min_Y E_{MSE}$. Only one winner ℓ_t is activated and its corresponding inner center y_{ℓ_t} is modified, while the rest remain all the same. This basic clustering algorithm takes the conventional competitive learning of winner-take-all (WTA) learning [1].

3 Clustering Via Statistical Learning Analysis

In most cases, the performance of the above classical center-based clustering algorithm greatly depends on the number of clusters fixed in advance and contributes to good clustering results only if the number of clusters has already been known as prior knowledge. However, when the number of clusters is unknown beforehand, it will be quite difficult to achieve a reasonable solution.

In order to tackle this problem, an alternative clustering mechanism is directly inspired from statistical learning analysis. The essence of statistical learning analysis is to achieve Structural Risk Minimization instead of Empirical Risk Minimization, as a sound statistical basis for the assessment of model adequacy [5]. Given that the learning model is completely unknown, the goal for clustering here not only concerns the issue of parameter learning, but also attaches great importance to the construction of the predictive models from the data to be learned.

3.1 Membership Hypothesis

One typical membership hypothesis is first specifically considered [7], so that not only the winner would be modified to adapt to the input, but also its rival will receive some penalty:

$$\begin{aligned} M(y_\ell | x_t) &= \begin{cases} 1 & \text{if } \ell = \ell_t \\ -1 & \text{if } \ell = \ell_r \\ 0 & \text{otherwise} \end{cases} \\ \ell_t &= \arg\min_j \gamma_j d^2(x_t, y_j) \qquad , \\ \ell_r &= \arg\min_{j, j \neq \ell_t} \gamma_j d^2(x_t, y_j) \end{aligned} \qquad (3)$$

where $\gamma_j = n_j \big/ \sum_{\ell=1}^{K} n_\ell$ is the relative winning frequency of the inner representation y_j, as a conscience strategy to reduce the winning probability of certain frequent winners to some extent, and n_j is the cumulative number of the occurrences of $M(y_j | x_t) = 1$ during the past learning. When starting from a number of clusters that is larger than the natural number of groups in the dataset, the aim is to automatically adjust the effective number of clusters.

3.2 Objective Function

On the basis of the above membership hypothesis, the clustering procedure here will not only be determined by the winner, but also by its rival. In other words, for each input, both the corresponding inner centers of the winner and the rival are modified by feed-back, with one for pure learning and the other for penalty.

Heuristically, when taking the membership hypothesis into the objective function of the k-means algorithm, we can update the objective function for optimization as:

$$E = \frac{1}{N}\sum_{\ell=1}^{K}\sum_{t=1}^{N} M(y_\ell | x_t) d^2(x_t, y_\ell) = \frac{1}{N}\sum_{t=1}^{N} d^2(x_t, y_{\ell_t}) - \frac{1}{N}\sum_{t=1}^{N} d^2(x_t, y_{\ell_r}) \; . \tag{4}$$

The above function is made up of two parts in the sense of statistical learning, one for empirical risk calculation, the other one for structural assessment. Let the objective function E be decomposed into the plain part E_{MSE} and the additional part E_{SR}, $E = E_{MSE} + E_{SR}$. Minimizing the objective function E will lead to x_t partitioned into the direction of both the minimal similarity with y_{ℓ_t} and maximal dissimilarity with y_{ℓ_r} simultaneously.

In detail, for the benefit of more compatibility in high-dimensional space, here we replace the commonly used Euclidean norm by one higher order metric as the similarity measure so that one of the problems encountered in high-dimensional space, i.e., the "concentration of measure" phenomenon, will be diminished to some degree [8]. And in order to seek a simpler solution as well as to avoid a negative or infinite objective function coming from the structural assessment, we take the Cityblock distance for an easy realization in the second part with only additions, subtractions, and arithmetic comparisons, and turn it into a power fraction expression instead. Inspired by the Minkowski distance metric, the objective function E then becomes

$$\begin{aligned} E_{MSE} &= \frac{1}{Np}\sum_{t=1}^{N}\Big(\sum_{m=1}^{d} \left| x_{t,m} - y_{\ell_t,m} \right|^p\Big) \\ E_{SR} &= \frac{\alpha}{Np}\sum_{t=1}^{N}\Big(\sum_{m=1}^{d} \left| x_{t,m} - y_{\ell_r,m} \right|\Big)^{-p} \end{aligned} \; . \tag{5}$$

Here d refers as the dimension in the space, $p \geq 2$ defines the order of the average error in the objective function, and α is introduced as a constant factor, $0 < \alpha \leq 1$, to control the influence of the structural assessment. Although E_{SR} is an indispensable part in the model construction of the proposed clustering method, E_{MSE} still plays the most essential role in the whole learning process, which should be attached greater importance to.

3.3 Adaptive Algorithm

With the above selection and modification of mathematical criterion, the derivatives of the objective function E with respect to the center y_ℓ can be computed and an iterative procedure for clustering can then be derived:

$$\frac{\partial E}{\partial y_\ell} = \begin{cases} \frac{\partial E_{MSE}}{\partial y_\ell} = -\frac{1}{N}\sum_{t=1}^{N} \left| x_t - y_\ell \right|^{p-2} (x_t - y_\ell) & if \ \ell = \ell_t \\ \frac{\partial E_{SR}}{\partial y_\ell} = \frac{\alpha}{N}\sum_{t=1}^{N} \operatorname{sgn}(x_t - y_\ell)(\sum_{m=1}^{d} \left| x_{t,m} - y_{\ell,m} \right|)^{-p-1} & if \ \ell = \ell_r \\ 0 & otherwise \end{cases}, \quad (6)$$

where $\operatorname{sgn}(x_t - y_\ell) = (x_t - y_\ell) / \left| x_t - y_\ell \right|$, refers to the sign function that extracts the sign from the difference between the input x_t and the center y_ℓ.

For each input x_t, the adaptive update algorithm for the center y_ℓ is:

$$y_\ell^{new} = y_\ell^{old} + \Delta y_\ell$$

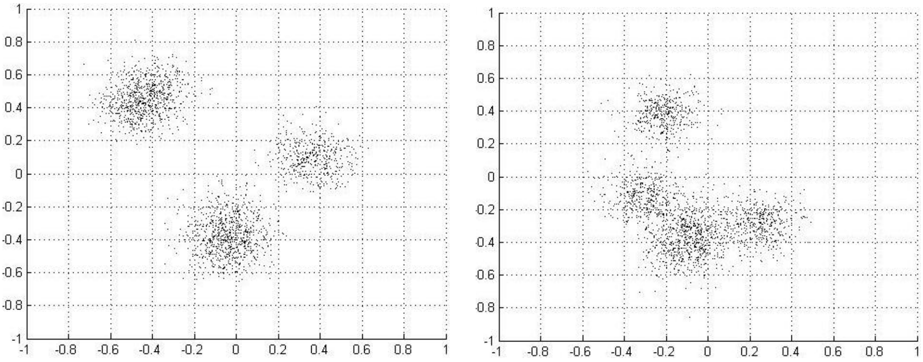
$$\Delta y_\ell = \begin{cases} \eta_t \left| x_t - y_\ell^{old} \right|^{p-2} (x_t - y_\ell^{old}) & if \ \ell = \ell_t \\ -\eta_r \operatorname{sgn}(x_t - y_\ell^{old})(\sum_{m=1}^{d} \left| x_{t,m} - y_{\ell,m}^{old} \right|)^{-p-1} & if \ \ell = \ell_r \\ 0 & otherwise \end{cases}, \quad (7)$$

where η_t and η_r both are constant rates for learning, with $0 < \eta_r \leq \eta_t < 1$. These iterative steps are repeated until one of the two following conditions is fulfilled: either each extra center y_ℓ is pushed far away from the data, or if the clustering results remain roughly fixed for all inputs. When the above rates are appropriately selected, the clustering algorithm has the capacity to not only assign a suitable cluster position to each input, but also to automatically allocate a proper number of clusters for the input dataset.

4 Simulation Experiment

Simulation experiments on a sample database (Gaussian mixture) were carried out to verify the performance of the proposed clustering algorithm. The experimental dataset consists of a set of samples following a mixture of no more than five Gaussian distributions with different location, mixture proportion and degree of overlap among clusters inside the [-1, 1] domain in a 2-dimensional space. Some examples of datasets are shown in Fig.1.

Given a hypothetical number of clusters larger than the original number of mixtures, both k-means and the proposed algorithm were respectively employed for clustering.



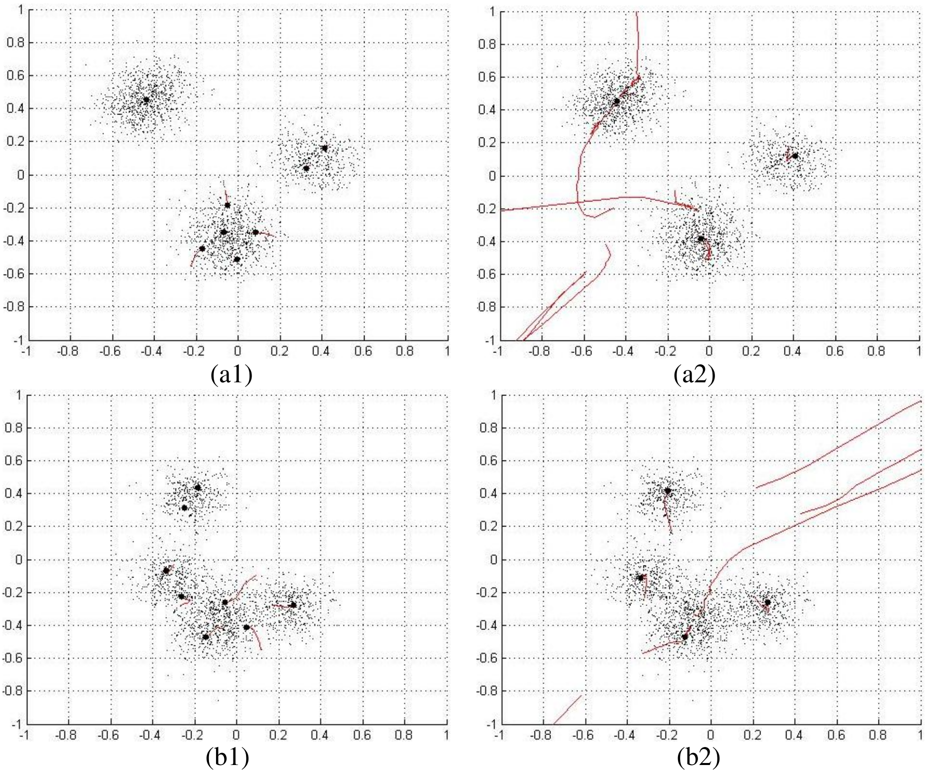
(a)

(b)

Fig. 1. Dataset examples

5 Result Analysis

Starting from a too large number of clusters (set to eight here), the clustering performances as well as the paths of centers in both k-means and the proposed algorithm for the above example databases are shown as Fig.2. A comparison could be made accordingly between their clustering results. Fig. 2 (a1) and (b1) are the results of the k-means algorithm, and (a2) and (b2) are the results of the clustering algorithm proposed in this paper with the learning rate $\eta_t = 0.005$ and $\eta_r = 0.0005$; (a1) and (a2) refer to dataset (a), and (b1) and (b2) to dataset (b). The clustering algorithm proposed in this paper earned activation for the winners and assigned penalty for their rivals, so that the winners concentrate more around the natural centers of the clusters and their rivals are driven far away from the datasets. Samples from unknown clusters are then assigned to the most competitive clusters, whose centers are representative of the datasets. With adequate parameters, the effective number of clusters can be easily observed, while the extra ones can be identified and removed after or even during



(a1) (a2)

(b1) (b2)

Fig. 2. Clustering performances and paths of cluster centers during learning. The final position is identified by a dot. Fig.2 (a1) and (b1) show the results of the k-means algorithm; (a2) and (b2) show the results of the clustering algorithm proposed in this paper. (a1) and (a2) correspond to dataset (a), and (b1) and (b2) to dataset (b) in Fig.1.

learning. On the contrary, the k-means algorithm maintains the originally given number of clusters, some of them turning out to be meaningless at the end for the correct number was not guessed before learning.

6 Conclusions

In this paper, an alternative for center-based clustering algorithms, in particular the k-means algorithm, is presented via statistical learning analysis. The essence of statistical learning principle, i.e., both the empirical risk and structural assessment, is taken into particular account for the clustering algorithm so as to derive and develop the relevant minimization mathematical criterion with automatic parameter learning and model selection in parallel. The proposed clustering algorithm roughly decides on the number of real clusters, prompts the winner by activation and obstructs its rival by penalty, so that the most competitive center wins for possible prediction and the extra ones are driven far away from the distribution. The only prerequisite is to start with a number of clusters that exceeds the natural number of clusters in the dataset. Simulation experiments achieve good performances of the double learning tasks in clustering, and show how the number of effective clusters is automatically extracted during learning.

Acknowledgments. Rui Nian is sponsored by the China Scholarship Council for her overseas study in Belgium. This research was fully supported by the Natural Science Foundation of China (60572064).

References

1. Selim, S.Z., Ismail, M.A.: K-Means-Type Algorithms: A Generalized Convergence Theorem and Characterization of Local Optimality. IEEE Trans. On PAMI-6 1, 81–87 (1984)
2. Bilmes, J.A.: A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models. Technical Report, University of Berkeley, ICSI-TR-97–021 (1997)
3. Zhang, B.: Comparison of the Performance of Center-Based Clustering Algorithms. In: Whang, K.-Y., Jeon, J., Shim, K., Srivastava, J. (eds.) PAKDD 2003. LNCS (LNAI), vol. 2637, pp. 63–74. Springer, Heidelberg (2003)
4. Ma, J.W., Cao, B.: The Mahalanobis Distance Based Rival Penalized Competitive Learning Algorithm. In: Wang, J., Yi, Z., Żurada, J.M., Lu, B.-L., Yin, H. (eds.) ISNN 2006. LNCS, vol. 3971, pp. 442–447. Springer, Heidelberg (2006)
5. Vapnik, V.N.: The Nature of Statistical Learning Theory. Springer, Berlin (1995)
6. Xu, L.: Bayesian Ying Yang harmony learning. The handbook of brain theory and Neural Networks. MIT Press, Cambridge (2002)
7. Xu, L., Krzyzak, A., Oja, E.: Rival Penalized Competitive Learning for Clustering Analysis, RBF Net, and Curve Detection Bayesian Ying Yang harmony learning. IEEE Trans. Neural Networks 4, 636–649 (1993)
8. Verleysen, M.: Learning High-dimensional Data. In: Ablameyko, S., et al. (eds.) Limitations and Future Trends in Neural Computation, pp. 141–162. IOS Press, Amsterdam (2003)