FUZZY C-MEANS CLUSTERING USING TRANSFORMATIONS INTO HIGH DIMENSIONAL SPACES

Sadaaki Miyamoto

Institute of Engineering Mechanics and Systems
University of Tsukuba
Ibaraki 305-8573, Japan

Daisuke Suizu

Graduate School of Systems and
Information Engineering
University of Tsukuba
Ibaraki 305-8573, Japan

ABSTRACT

Algorithms of fuzzy c-means clustering with kernels employed in nonlinear transformations into high dimensional spaces in the support vector machines are studied. The objective functions in the standard method and the entropy based method are considered and iterative solutions in the alternate optimization algorithm are derived. Explicit cluster centers in the data space are not obtained by this method in general but fuzzy classification functions are useful which have much more information than crisp clusters in the hard c-means. Numerical examples using radial basis kernel functions are given.

1. INTRODUCTION

Recently Support Vector Machines (SVM) [11, 12] have been focused upon by many researchers in pattern classification. A characteristic of the SVM is the use of nonlinear transformations into high dimensional spaces whereby classifiers may have high nonlinearities.

In the use of such nonlinear transformations, it is unnecessary to obtain an explicit representation of the function of the transformation, but to have the function of the scalar product in the high dimensional space is sufficient, and the latter is called a kernel function [12].

Apart from the SVM itself, the use of the kernel functions have also been considered. In particular, we remark the study by Girolami [4] where crisp clustering based on the minimization of the trace of a matrix in the high dimensional space is proposed.

No doubt the work by Girolami is important, but there are much more to be studied along his discussions. First, there are technical problems in the minimization of the trace and hence the problem should be reformulated; second, fuzzy clustering has to be studied which is said to have advantages over crisp clustering with regard to robustness.

In this paper the second problem of fuzzy c-means clustering using transformations into high dimensional spaces is considered. The standard objective function as well as entropy based objective function are used and solutions of iterative minimizations are derived.

Numerical examples show that clusters are obtained that cannot be generated without the use of a kernel function.

2. PRELIMINARIES

Assume that objects to be clustered are points in the p-dimensional Euclidean space $\boldsymbol{R}^p$ and are denoted by $x_k = (x_k^1, \ldots, x_k^p)$, $k = 1, \ldots, n$. The Euclid norm in $\boldsymbol{R}^p$ is denoted by $\|\cdot\|$.

The membership by which object x_k belongs to cluster i is denoted by $c \times n$ matrix $U = (u_{ik})$ and the constraint for U is

$$M_{prb} = \left\{ (u_{ik}) \,:\, u_{ik} \in [0,1], \sum_{i=1}^{c} u_{ik} = 1 \text{ for all } k \right\}$$

as usual; cluster centers are $v_i = (v_i^1, \ldots, v_i^p)$. Moreover we put $V = (v_1, \ldots, v_c)$.

The objective function

$$J_m(U,V) = \sum_{k=1}^{n} \sum_{i=1}^{c} (u_{ik})^m \|x_k - v_i\|^2 \quad (m > 1) \qquad (1)$$

has been considered [3, 2]. In the standard alternate minimization algorithm **FCM** [3, 2], we put $J(U,V) = J_m(U,V)$ and $M = M_{prb}$.

Procedure FCM

FCM1. Set initial value for $\bar{V}$.

FCM2. Solve

$$\min_{U \in M} J(U, \bar{V})$$

and let the optimal solution be $\hat{U}$. Put $\bar{U} = \hat{U}$.

FCM3. Solve

$$\min_{V} J(\bar{U}, V)$$

and let the optimal solution be $\hat{V}$. Put $\bar{V} = \hat{V}$.

FCM4. If the solution $(\bar{U}, \bar{V})$ is convergent, stop. (A convergence criterion is omitted here for simplicity.) Otherwise, go to **FCM2**.

Remark: It is possible to exchange the order of the steps **FCM2** and **FCM3**. In that case an initial value for $\bar{U}$ should be given in **FCM1**.

Optimal solutions in **FCM2** and **FCM3** are as follows. Notice that we write u_{ik} and v_i instead of $\bar{u}_{ik}$ and $\bar{v}_i$ for simplicity.

$$u_{ik} = \left[\sum_{j=1}^{c} \left(\frac{\|x_k - \bar{v}_i\|^2}{\|x_k - \bar{v}_j\|^2} \right)^{\frac{1}{m-1}} \right]^{-1}, \tag{2}$$

$$v_i = \frac{\sum_{k=1}^{n} (\bar{u}_{ik})^m x_k}{\sum_{k=1}^{n} (\bar{u}_{ik})^m} \tag{3}$$

Remark: When we put $m = 1$ in J_m and

$$M = M_c = \left\{ (u_{ik}) \,:\, u_{ik} \in \{0,1\}, \sum_{i=1}^{c} u_{ik} = 1 \text{ for all } k \right\}$$

then the algorithm is reduced to hard c-means.

Entropy based fuzzy c-means

Li and Mukaidono [7] have proposed entropy maximization in fuzzy clustering. Moreover Miyamoto *et al.* [8, 9] have reformulated the method of entropy in the framework of the alternate optimization algorithm **FCM**. We use the latter method here which we call the entropy based method or entropy method.

In this method the objective function is

$$J^\lambda = \sum_{k=1}^{n} \sum_{i=1}^{c} u_{ik} \|x_k - v_i\|^2 + \lambda^{-1} \sum_{k=1}^{n} \sum_{i=1}^{c} u_{ik} \log u_{ik}.$$

where λ is a positive parameter.

The same algorithm **FCM** is used for $J = J^\lambda$ and $M = M_{prb}$. Optimal solutions in **FCM2** and **FCM3** are as follows.

$$u_{ik} = \frac{e^{-\lambda \|x_k - \bar{v}_i\|^2}}{\sum_{j=1}^{c} e^{-\lambda \|x_k - \bar{v}_j\|^2}}, \tag{4}$$

$$v_i = \frac{\sum_{k=1}^{n} \bar{u}_{ik} x_k}{\sum_{k=1}^{n} \bar{u}_{ik}} \tag{5}$$

3. HIGH DIMENSIONAL SPACES AND KERNEL FUNCTIONS

A high dimensional space used in SVM is denoted by S here which is called a feature space, whereas the original space $\boldsymbol{R}^p$ is called the data space. S is in general an infinite dimensional inner product space. Its inner product is denoted by $\langle \cdot, \cdot \rangle$. The norm of S is denoted by $\| \cdot \|_S$.

Notice that in SVM, a mapping $\Phi : \boldsymbol{R}^p \to S$ is employed and x_k is transformed into $\Phi(x_k)$. Explicit representation of $\Phi(x_k)$ is not usable in general but the inner product $\langle \Phi(x_k), \Phi(x_\ell) \rangle$ is expressed by a kernel:

$$K(x_k, x_\ell) = \langle \Phi(x_k), \Phi(x_\ell) \rangle.$$

The next objective function based on entropy is first studied and the standard objective function is later mentioned:

$$J^\lambda(U, W) = \sum_{k=1}^{n} \sum_{i=1}^{c} u_{ik} \|\Phi(x_k) - W_i\|_S^2 + \lambda^{-1} \sum_{k=1}^{n} \sum_{i=1}^{c} u_{ik} \log u_{ik} \tag{6}$$

where W_i is the cluster center in the high dimensional feature space and we put $W = (W_1, \ldots, W_c)$. Moreover we put

$$d_{ik} = \|\Phi(x_k) - W_i\|_S^2$$

for simplicity, whence

$$J^\lambda(U, W) = \sum_{k=1}^{n} \sum_{i=1}^{c} u_{ik} d_{ik} + \lambda^{-1} \sum_{k=1}^{n} \sum_{i=1}^{c} u_{ik} \log u_{ik}.$$

For deriving the solution in **FCM3**, let $\phi = (\phi_1, \ldots, \phi_c)$ be an arbitrary element in S and ε be a small positive number. From

$$J(U, W + \varepsilon\phi) - J(U, W) = 2\varepsilon \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik} \langle \phi_i, \Phi(x_k) - W_i \rangle + o(\varepsilon^2)$$

the Euler equation is

$$\sum_{k=1}^{n} u_{ik} (\Phi(x_k) - W_i) = 0$$

Put

$$U_i = \sum_{k=1}^{n} u_{ik},$$

then we have

$$W_i = \frac{1}{U_i} \sum_{k=1}^{n} u_{ik} \Phi(x_k), \quad i = 1, \ldots, c. \tag{7}$$

We do not have an explicit form of W_i, since it uses $\Phi(x_k)$. We therefore substitute (7) into d_{ik} to obtain

$$\begin{aligned} d_{ik} &= \langle \Phi(x_k) - W_i, \Phi(x_k) - W_i \rangle \\ &= \langle \Phi(x_k), \Phi(x_k) \rangle - 2\langle W_i, \Phi(x_k) \rangle + \langle W_i, W_i \rangle \\ &= \langle \Phi(x_k), \Phi(x_k) \rangle - \frac{2}{U_i} \sum_{j=1}^{n} u_{ij} \langle \Phi(x_j), \Phi(x_k) \rangle \\ &\quad + \frac{1}{U_i^2} \sum_{j=1}^{n} \sum_{\ell=1}^{n} u_{ij} u_{i\ell} \langle \Phi(x_j), \Phi(x_\ell) \rangle \end{aligned}$$

Let

$$K(x, y) = \langle \Phi(x), \Phi(y) \rangle$$

be a kernel function and put

$$K_{j\ell} = K(x_j, x_\ell) = \langle \Phi(x_j), \Phi(x_\ell) \rangle.$$

Using the kernel function we have

$$d_{ik} = K_{kk} - \frac{2}{U_i} \sum_{j=1}^{n} u_{ij} K_{jk} + \frac{1}{U_i^2} \sum_{j=1}^{n} \sum_{\ell=1}^{n} u_{ij} u_{i\ell} K_{j\ell}, \quad (8)$$

whereby the solution in **FCM2** becomes

$$u_{ik} = \frac{e^{-\lambda d_{ik}}}{\sum_{j=1}^{c} e^{-\lambda d_{jk}}} \quad (9)$$

Notice that in **FCM**, W_i is not explicitly calculated but calculations of (9) and (8) are repeated until convergence. Hence the convergence in **FCM4** should be based on U, and not W.

Various kernel functions have been proposed from which the Radial Basis Function(RBF) kernel

$$K_{j\ell} = \exp(-c\|x_j - x_\ell\|^2)$$

is used in the next section for numerical examples.

Fuzzy classification functions [9, 10] are available in fuzzy c-means which show how prototypical an arbitrary point in the data space is to a cluster by extending the membership u_{ik} to the whole space. The next formula calculates a fuzzy classification function

$$u_i(x) = \frac{e^{-\lambda D_i(x)}}{\sum_{j=1}^{c} e^{-\lambda D_j(x)}} \quad (10)$$

where $D_i(x) = \|\Phi(x) - W_i\|_S^2$. We have

$$\begin{aligned} D_i(x) &= K(x, x) - \frac{2}{U_i} \sum_{j=1}^{n} u_{ij} K(x, x_j) \\ &\quad + \frac{1}{U_i^2} \sum_{j=1}^{n} \sum_{\ell=1}^{n} u_{ij} u_{i\ell} K_{j\ell} \end{aligned} \quad (11)$$

using u_{ik} after convergence. Note also that

$$K(x, x_j) = \exp(-c\|x - x_j\|^2)$$

when RBF kernel is used.

Standard fuzzy c-means using the high dimensional space

A similar algorithm is derived for the standard method of fuzzy c-means. The objective function is thus assumed to be

$$\begin{aligned} J_m(U, W) &= \sum_{k=1}^{n} \sum_{i=1}^{c} (u_{ik})^m \|\Phi(x_k) - W_i\|_S^2 \\ &= \sum_{k=1}^{n} \sum_{i=1}^{c} (u_{ik})^m d_{ik} \end{aligned}$$

where $d_{ik} = \|\Phi(x_k) - W_i\|_S^2$.

It is easy to see the Euler equation is

$$\sum_{k=1}^{n} (u_{ik})^m (\Phi(x_k) - W_i) = 0$$

Put

$$U_i' = \sum_{k=1}^{n} (u_{ik})^m$$

and we obtain

$$W_i = \frac{1}{U_i'} \sum_{k=1}^{n} (u_{ik})^m \Phi(x_k), \quad i = 1, \ldots, c. \quad (12)$$

Instead of (8), we use

$$\begin{aligned} d_{ik} &= K_{kk} - \frac{2}{U_i'} \sum_{j=1}^{n} (u_{ij})^m K_{jk} \\ &\quad + \frac{1}{(U_i')^2} \sum_{j=1}^{n} \sum_{\ell=1}^{n} (u_{ij} u_{i\ell})^m K_{j\ell} \end{aligned} \quad (13)$$

for the solution of **FCM2**:

$$u_{ik} = \left[\sum_{j=1}^{c} \left(\frac{d_{ik}}{d_{jk}} \right)^{\frac{1}{m-1}} \right]^{-1} \quad (14)$$

By substituting this u_{ik} into (13), we have the iterative algorithm.

For the fuzzy classification function in the standard method, u_{ik} after convergence is used in

$$\begin{aligned} D_i(x) &= K(x, x) - \frac{2}{U_i'} \sum_{j=1}^{n} (u_{ij})^m K(x, x_j) \\ &\quad + \frac{1}{(U_i')^2} \sum_{j=1}^{n} \sum_{\ell=1}^{n} (u_{ij} u_{i\ell})^m K_{j\ell} \end{aligned} \quad (15)$$

and calculate

$$u_i(x) = \left[\sum_{j=1}^{c} \left(\frac{D_i(x)}{D_j(x)} \right)^{\frac{1}{m-1}} \right]^{-1} \quad (16)$$

In particular, prototypical regions close to the cluster centers have memberships near unity. Hence we can distinguish prototypical regions from regions near cluster boundaries.

4. NUMERICAL EXAMPLES

Throughout all numerical experiments the number of clusters are two and the initial value in the algorithm **FCM** has been two initial cluster centers randomly chosen from the data points. Ten trials for each method and each example with different initial centers have been tested and the solution that has attained minimum of the objective function values has been selected as the best result. The kernel function is the RBF kernel as above.

Figure 1 shows a three-dimensional representation of a classification function for a cluster. The data for clustering is seen on the plane where two groups are found by sight:

one ring and a sphere in the ring. A similar data set has been experimented by Girolami [4] and he reported that two clusters have successfully been separated. In this figure the standard FCM with $m = 2$ has been used and the parameter in the RBF kernel is $c = 10$ which has been used for all examples except the last. Figure 2 depicts another result obtained from the entropy FCM with $\lambda = 10.0$ and the same parameter in the kernel. Apparently both figures exhibit successful separation of the two clusters. Remark that such separation of two clusters such that one is inside the other is impossible by an ordinary c-means type methods, since a c-means algorithm will provide a Voronoi region for a cluster [5, 9, 10], while the data in these figures cannot be separated to Voronoi regions.

Another example is more difficult to handle. In Figure 3 the data points on the plane forms two 'crescents'. The problem is to separate the two groups. This figure has been obtained from the standard method with $m = 2$ but without a kernel: an ordinary FCM. The result shows there are misclassifications. When we use the entropy based FCM, we have similar results with misclassifications, although we omit the latter result. Unfortunately, even if we use the RBF kernel for the standard FCM as in Figure 4, we still have misclassified points. A good result has been obtained from the entropy FCM as we see in Figure 5 where we use $\lambda = 15.0$ and $c = 50.0$. However, we still have a problem, that is, the cluster is unstable and sensitive to parameters in FCM and the kernel function.

5. CONCLUSION

Fuzzy c-means clustering using the high dimensional spaces and associated kernel functions has been studied. Two methods of the standard FCM and the entropy based FCM have been considered. They have shown different clusters and classification functions in the numerical examples.

Girolami [4] uses stochastic approximation algorithm for crisp clustering, whereas in the case of fuzzy clustering direct solutions are derived without a combination of different methods.

Fuzzy clustering has a number of advantages over crisp clustering. For example, fuzzy classification functions can be employed to attach memberships to all points in the data space. Moreover, the two objective functions can be used whereby we obtain different clusters.

We have remarked in the introduction that in Girolami's formulation there is a technical problem. We have no space to describe the problem in detail and the way to solve it. We will show these in a forthcoming paper.

There are many possible studies in the future, since many variations of fuzzy c-means are available. For example, possibilistic clustering [6] using such kernels seems promising. Moreover other type of clustering algorithms in which the idea of SVM is used (e.g. [1]) should further be studied using fuzzy classification and compared with the approach herein.

This research has partially been supported by the Grant-in-Aid for Scientific Research (C), Japan Society for the Promotion of Science, No.13680475.

6. REFERENCES

[1] A.Ben-Hur, D.Horn, H.T.Siegelmann, V.Vapnik, A support vector clustering method, *Proc. of ICPR2000*, pp. 724–727, 2000.

[2] J.C.Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum, New York, 1981.

[3] J.C.Dunn, A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, *J. of Cybernetics*, Vol.3, pp. 32–57, 1974.

[4] M.Girolami, Mercer kernel based clustering in feature space, *IEEE Trans. on Neural Networks*, to appear. Available from http://cis.paisley.ac.uk/giroci0/publications.html

[5] T.Kohonen, *Self-Organization and Associative Memory*, Springer-Verlag, Heiderberg, 1989.

[6] R.Krishnapuram, J.M.Keller, A possibilistic approach to clustering, *IEEE Trans. on Fuzzy Syst.*, Vol.1, No.2, pp. 98–110, 1993.

[7] R.-P.Li, M.Mukaidono, A maximum entropy approach to fuzzy clustering, *Proc. of the 4th IEEE Intern. Conf. on Fuzzy Systems (FUZZ-IEEE/IFES'95)*, Yokohama, Japan, March 20-24, 1995, pp. 2227–2232.

[8] S.Miyamoto, M.Mukaidono, Fuzzy c - means as a regularization and maximum entropy approach, *Proc. of the 7th International Fuzzy Systems Association World Congress (IFSA'97)*, June 25-30, 1997, Prague, Chech, Vol.II, pp.86–92, 1997.

[9] S.Miyamoto, *Introduction to Cluster Analysis*, Morikita-Shuppan, Tokyo, 1999 (in Japanese).

[10] S.Miyamoto, Methods in hard and fuzzy clustering, In: Z.Q.Liu and S.Miyamoto, eds., *Soft Computing and Human Centered Machines*, Springer, Tokyo, 2000, pp. 85–129.

[11] V.N.Vapnik, *Statistical Learning Theory*, Wiley, New York, 1998.

[12] V.N.Vapnik, *The Nature of the Statistical Learning Theory, 2nd Ed.*, Springer, New York, 2000.

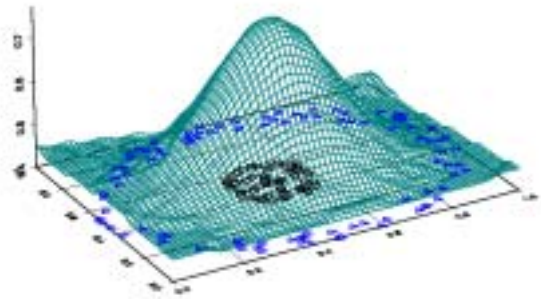

Figure 1: Two clusters and a classification function from a ball and ring data; standard FCM with $m = 2$ is used.

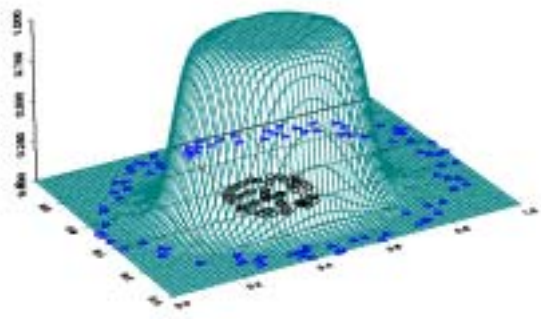

Figure 2: Two clusters and a classification function from a ball and ring data; entropy based FCM with $\lambda = 10.0$ is used.

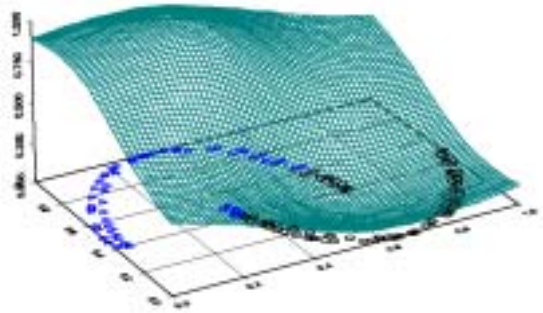

Figure 3: Two clusters and a classification function from two 'crescents'data; standard FCM with $m = 2$ *without* a kernel is used.

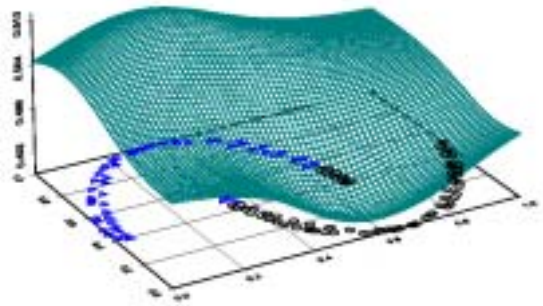

Figure 4: Two clusters and a classification function from two 'crescents'data; standard FCM with $m = 2$ *with* the RBF kernel is used.

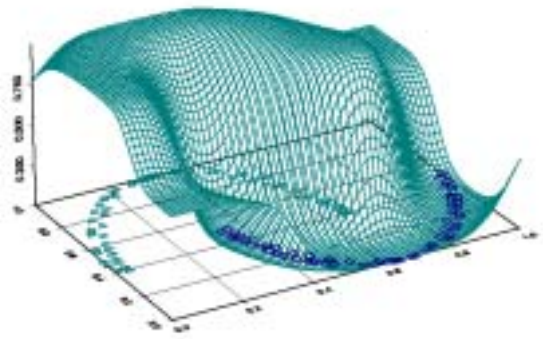

Figure 5: Two clusters and a classification function from two 'crescents'data; entropy based FCM with $\lambda = 15.0$ *with* the RBF kernel is used.