2009 International Symposium on Nonlinear Theory and its Applications
NOLTA'09, Sapporo, Japan, October 18-21, 2009

An Explict Mapping for Kernel Data Analysis and Application to c-Means Clustering

Sadaaki Miyamoto[1] Keisuke Sawazaki[2]

1.Department of Risk Engineering, University of Tsukuba
Tsukuba, Japan
2.Graduate School of Systems and Information Engineering, University of Tsukuba
Tsukuba, Japan
Email: miyamoto@risk.tuskuba.ac.jp, sawazaki@soft.risk.tsukuba.ac.jp

Abstract— *Kernel data analysis is now becoming standard in many applications of data analysis. An implicit mapping into a high-dimensional feature space is first assumed, in other words, an explicit form of the mapping is unknown but their inner product should be known. Contrary to this common assumption, we introduce an explicit mapping which is standard in a sense. The reason why we use this mapping is as follows. (1) the use of this mapping does not lose any fundamental information in kernel data analysis. (2) We have the same formulas in every kernel methods with and without this explicit mapping. (2) Usually the derivation becomes simpler by using this mapping. (3) New applications of the kernel methods become possible using this mapping. As an application we consider fuzzy c-means clustering and principal component analysis. A typical numerical example is shown to observe the effectiveness of the present method.*

Keywords— Kernel data analysis, fuzzy clustering, explicit mapping

1 Introduction

Kernel functions [6] noted in support vector machines[7] now is a well-known technique and becoming standard in many applications of data analysis. An important point in the kernel trick is that although we consider a mapping from a data space into a high-dimensional feature space, we need not to know its explicit form but we should know the inner product of the feature space. Generally, the feature space is not uniquely determined. Accordingly every formulas in data analysis using kernel functions should be described in terms of the inner product.

Although kernel functions are really useful, but the derivation is sometimes complicated when original formulas should be rewritten by the inner product forms. A typical example is kernel fuzzy c-means clustering [3] and kernel SOM [4].

Here is a question: can we have a useful and explicit mapping and explicit representation of a high-dimensional feature space? If we have such a mapping, we can use many existing tools of multivariate analysis. The answer is YES as far as we do not need a function with a variable $x \in \boldsymbol{R}^p$. Note that some methods in data analysis such as the principal component and cluster analysis does not need such a function, while classification rules such as support vector machines and discriminant analysis need functions with a variable.

This means that as far as we are concerned with clustering and principal component analysis, we have a simpler method. This paper shows the way how we have such an explicit mapping and how to use this mapping. This mapping is simple enough and useful in the sense that it leads to the same formulas when transformed into the inner product forms. In short, this explicit mapping and associated feature space have all information that is used in kernel functions for data analysis.

To illustrate the effectiveness of the present method, a typical result of kernel fuzzy c-means clustering with cluster centers will be shown using kernel principal component analysis.

2 An Explicit Mapping for Kernel Functions

2.1 Preliminary consideration

A set of data $X = \{x_1, \ldots, x_n\} \subset \boldsymbol{R}^p$ is assumed to be given. Each data unit is also called an object or an individual and it is a point in the p-dimensional real space $x_k = (x_k^1, \ldots, x_k^p)^T \in \boldsymbol{R}^p$. We consider a mapping into a high-dimensional feature space $\Phi\colon \boldsymbol{R}^p \to H$ and associated kernel function

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

where $\langle \cdot, \cdot \rangle$ is an inner product of H. We also assume $\| \cdot \|_H$ is a norm of H. Thus H is an inner product space. In this section we first suppose, as usual, we do not know function $\Phi(\cdot)$ explicitly but we know $K(x, y)$. Specifically, the Gaussian kernel is frequently used:

$$K(x, y) = \exp(-\lambda \|x - y\|^2)$$

where $\lambda > 0$ and $\|x\|$ is the norm of $\boldsymbol{R}^p$.

An objective function of fuzzy c-means using the feature space H is the following.

$$J_H(U, W) = \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ki})^m \|\Phi(x_k) - W_i\|_H^2, \quad (m > 1)$$

where $U = (u_{ki})$ is $n \times c$ matrix representing membership of x_k to cluster i. U has the next constraint when it is optimized.

$$M = \{U = (u_{ki}) : \sum_{i=1}^{c} u_{ki} = 1,\ u_{\ell j} \geq 0, \forall \ell, j\},$$

while $W = (W_1, \ldots, W_c)$ is a collection of cluster centers in H.

The iterative algorithm of fuzzy c-means clustering [1] is basically an alternative minimization of $J_H(U, W)$ with re-

spect to U and W until convergence. We have the next solutions.

$$u_{ki} = \left[\sum_{j=1}^{c} \left(\frac{D(x_k, W_i)}{D(x_k, W_j)} \right)^{\frac{1}{m-1}} \right]^{-1}, \quad (1)$$

$$W_i = \frac{\sum_{k=1}^{n} (u_{ki})^m \Phi(x_k)}{\sum_{k=1}^{n} (u_{ki})^m} \quad (2)$$

where we put

$$D(x_k, W_i) = \|\Phi(x_k) - W_i\|_H^2. \quad (3)$$

Equations (1) and (2) should be repeated until convergence, but since $\Phi(x_k)$ is unknown, we should use another formula for kernel fuzzy c-means clustering.

The formula is derived by eliminating (2) from iteration, i.e., we substitute (2) into (3) to have an updating formula for $D(x_k, W_i)$[3]:

$$D(x_k, W_i) = K(x_k, x_k) - \frac{2}{\sum_{k=1}^{n} (u_{ki})^m} \sum_{j=1}^{n} (u_{ji})^m K(x_j, x_k) + \frac{1}{\{\sum_{k=1}^{n} (u_{ki})^m\}^2} \sum_{j=1}^{n} \sum_{\ell=1}^{n} (u_{ji} u_{\ell i})^m K(x_j, x_\ell). \quad (4)$$

We thus repeat (1) and (4) until convergence when kernel fuzzy c-means clustering should be used.

2.2 An explicit mapping and its properties

The above derivation needs consideration to transform an original algorithm into another one, due to the fact that no explicit mapping is available.

This means that if we have an explicit mapping, then no derivation is needed and we just use an existing algorithm.

We show two explicit mappings of which the latter is more useful and will be used throughout applications. First mapping is very simple:

$$\Phi_1(x_k) = e_k \quad (k = 1, 2, \ldots, n) \quad (5)$$

where $H = \boldsymbol{R}^n$ and e_k is the k-th unit vector that has unity as ith component and all other components are zero. Note that $\Phi_1 \colon X \to \boldsymbol{R}^n$, i.e., $\Phi(\cdot)$ is not defined on $\boldsymbol{R}^p$ but is limited to the finite set X. We also assume that the inner product of $\boldsymbol{R}^n$ is

$$\langle e_k, e_\ell \rangle = K(x_k, x_\ell) \quad (6)$$

instead of the standard inner product $\langle e_k, e_\ell \rangle = \delta_{k\ell}$ ($\delta_{k\ell}$ is the Kronecker delta).

We have

Proposition 1 *If kernel $K(x, y)$ is positive definite, $\langle e_k, e_\ell \rangle$ defined by (6) satisfy the axioms of the inner product of $\boldsymbol{R}^n$, that is, $\boldsymbol{R}^n$ with (6) is an inner product space.*

Note 1 *The detailed proof is given in standard textbooks [6]. As a rough sketch of the proof, note that the Mercer condition[7]*

$$\int \int K(x, y) \eta(x) \eta(y) dx dy \geq 0$$

for all $\eta(x)$ guarantees

$$\sum_{i,j} K(x_i, x_j) \zeta_i \zeta_j \geq 0$$

$\forall \zeta_i \in \boldsymbol{R}$, by putting $\eta(x) = \sum_i \zeta_i \delta(x - x_i)$.

Thus the matrix $K = (K(x_i, x_j))$ is positive semi-definite. The kernel function generally does not distinguish positive-semidefiniteness and positive-definiteness, while positive-definiteness is required for the definition of an inner product.

To solve this problem, we introduce another explicit mapping. For this purpose note that a positive semi-definite matrix K can be divided into $K = K^{\frac{1}{2}} K^{\frac{1}{2}}$. We thus define the second mapping:

$$\Phi_2(x_k) = K^{\frac{1}{2}} e_k \quad (k = 1, 2, \ldots, n) \quad (7)$$

In this case, $\Phi_2 \colon X \to \boldsymbol{R}^n$, i.e., the domain and the codomain are the same as Φ_1, but now the inner product of $\boldsymbol{R}^n$ is standard:

$$\langle e_k, e_\ell \rangle = \delta_{k\ell}. \quad (8)$$

We now apply mapping Φ_2 to fuzzy c-means clustering. It is sufficient to show the optimal solution of W_i.

Proposition 2 *For all positive semi-definite kernel $K(x, y)$ and mapping Φ_2 by (7), the cluster centers are the same and are given by*

$$W_i = \left(\frac{(u_{1i})^m}{\sum_{k=1}^{n} (u_{ki})^m}, \ldots, \frac{(u_{ni})^m}{\sum_{k=1}^{n} (u_{ki})^m} \right)^T, \quad i = 1, \ldots, c \quad (9)$$

Note 2 *The proof of this proposition is not difficult by observing closely (2). Note that W_i given by (2) is the solution of*

$$\min_{W_i} \sum_{k} (u_{ki})^m \|x_k - W_i\|^2$$

where the space can be an arbitrary inner product space, since the derivation of (2) uses a general variational principle valid for any Hilbert space. Note moreover that we substitute (9) into (1) to have the optimal solution of U. It should be noted that although optimal W_i is the same for all positive definite kernel, optimal U differs because (6) give different values for different kernels.

The proof of the next proposition is almost trivial and is omitted.

Proposition 3 *Substituting (9) into (3), we have (4).*

That is, (9) derived from the single mapping (5) has all necessary and sufficient information for kernel fuzzy c-means clustering.

As noted above, formulas in kernel principal component analysis are derived without any difficulty by using Φ_2, since the inner product is just standard. The mapping (7) is moreover useful for application to LVQ and SOM [2], where vectors for updating quantization vectors should be based on learning. The explicit mapping enables vector representations in $\boldsymbol{R}^n$, we can use every formulas in LVQ and SOM, while usual kernel methods should eliminate quantization vectors [5].

3 A Typical Numerical Example

A set of data shown in Fig. 1 was used to show the effectiveness of the proposed method. This example is well-known in the sense that a usual fuzzy c-means (abbreviated as FCM) algorithm cannot separate the outer circle and the inner ball, while a kernel-based c-means algorithm can. Figure 2 depicts the result of an ordinary FCM, while Figure 3 successfully separated the two groups. The ordinary FCM algorithm with Φ_2 has been used.

Moreover Figure 4 shows two major principal components with mapping Φ_2, where the method of the ordinary principal component analysis has been applied.

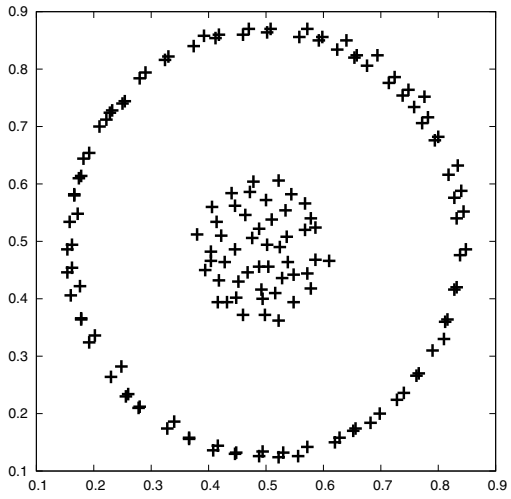

Figure 1: A ball in a circle with 150 data points.

4 Conclusions

We have proposed the use of an explicit mapping for kernel based data analysis. The second mapping seems to be more useful in general but either one of Φ_1 and Φ_2 can be used, as there is no fundamental difference between the two, except that the second can be applied even when the kernel is positive semi-definite. In summary, we note the following advantages of the present method.

1. Using this mapping, we do not lose any fundamental information in kernel data analysis.
2. Generally the derivation becomes simpler using this mapping.
3. New applications of the kernel methods become easier using this mapping.

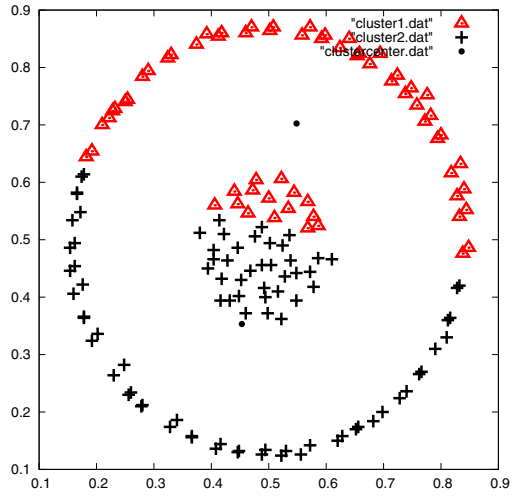

Figure 2: The result of ordinary FCM applied to data in Fig. 1.

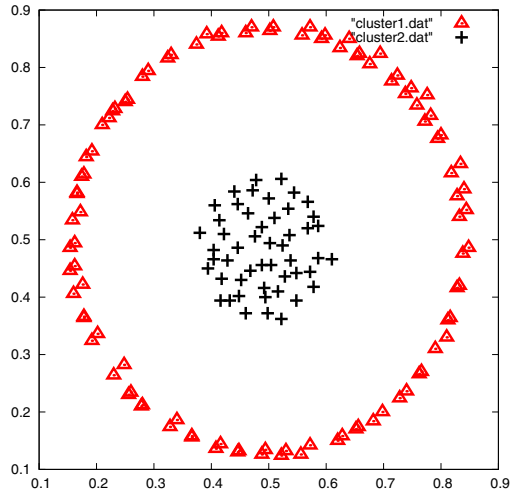

Figure 3: The result of kernel-based FCM applied to data in Fig. 1.

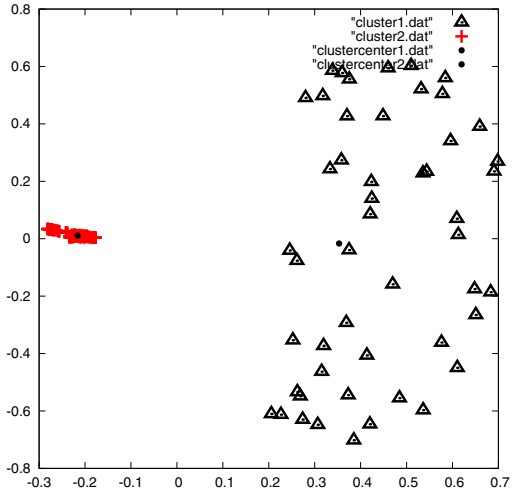

Figure 4: The display of two major principal components to data in Fig. 1 with kernel Φ_2.

The last statement should be further studied. In relation to fuzzy clustering, the method of fuzzy c-varieties should be studied. We moreover have many research possibilities related to SOM.

Moreover we have another possibility to use nonlinear methods of data analysis to the transformed data, where the above explicit mapping is essential. Without such a mapping, we cannot apply a nonlinear tool of data analysis.

Acknowledgment

This research has partly been supported by the Grant-in-Aid for Scientific Research, JSPS, Japan, No.19300074.

References

[1] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Functions, Plenum Press, New York, 1981.

[2] T. Kohonen, Self-Organizing Maps, 2nd Ed., Springer, Berlin, 1995.

[3] S. Miyamoto, D. Suizu, Fuzzy c-means clustering using transformations into high dimensional spaces, Proc. of FSKD'02: 1st International Conference on Fuzzy Systems and Knowledge Discovery, Nov. 18-22, 2002, Singapore, Vol.2, pp. 656–660.

[4] S. Miyamoto, H. Ichihashi, K. Honda, Algorithms for Fuzzy Clustering, Springer, Berlin, 2008.

[5] K. Mizutani, S. Miyamoto, Fuzzy multiset space and c-means clustering using kernels with application to information retrieval. in T.Bilgiç et al. Eds.: IFSA 2003, LNAI2715, Springer, Berlin, pp.387-395, 2003.

[6] B. Schölkopf, A.J. Smola, Learning with Kernels, the MIT Press, Cambridge, 2002.

[7] V.N. Vapnik, Statistical Learning Theory, Wiley, 1998.