An Explicit Mapping for Kernel Data Analysis and Application to Text Analysis

Sadaaki Miyamoto[1] Yuichi Kawasaki[2] 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, {kawasaki,sawazaki}@soft.risk.tsukuba.ac.jp

Abstract— *Kernel data analysis is now becoming standard in every application of data analysis and mining. Kernels are used to represent a mapping into a high-dimensional feature space, where an explicit form of the mapping is unknown. Contrary to this common understanding, we introduce an explicit mapping which we consider standard. 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 and we have the same formulas in every kernel methods. (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 an example of text mining where we use fuzzy c-means clustering and cluster centers in the high-dimensional space and visualize the centers using kernel principal component analysis.*

Keywords— Kernel data analysis, fuzzy clustering, explicit mapping, text mining

1 Introduction

Kernel functions [9] which are not very new but was noted in support vector machines[10] and now is a well-known technique and becoming standard in many applications of data analysis and data mining. An important point in the use of kernel functions 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 formula 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] in which the cluster centers in a feature space should be eliminated using inner product forms.

Here is a question: can we have a useful and explicit mapping and explicit representation of a high-dimensional feature space? The answer is YES, and we will show the mapping that is simple enough and useful in the sense that it leads to the same formulas when transformed into the inner product forms. To summarize, this explicit mapping and associated feature space have all information that is used in kernel functions for data analysis.

Another question arises: what is useful in this explicit mapping? We will show a real application of an example of text analysis. A result of kernel fuzzy c-means clustering and cluster centers will be shown using kernel principal component analysis.

Proofs of propositions are mostly omitted to save the space but brief notes are given as appendix.

2 Kernel Functions and an Explicit Mapping

2.1 Preliminary consideration

Although we consider kernel fuzzy c-means (FCM) clustering [3] and kernel principal component analysis (KPCA)[9] later, we describe FCM and omit KPCA for simplicity.

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

$$\boldsymbol{K}(x, y) = \langle \Phi(x), \Phi(y) \rangle \tag{1}$$

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 suppose we do not know function $\Phi(\cdot)$ explicitly but we know $\boldsymbol{K}(x, y)$. Specifically, the Gaussian kernel is used frequently:

$$\boldsymbol{K}(x, y) = \exp(-\lambda \|x - y\|^2) \tag{2}$$

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.

$$J_H(U, W) = \sum_{i=1}^{c} \sum_{k=1}^{n} (u_{ki})^m \|\Phi(x_k) - W_i\|_H^2, \quad (m > 1) \tag{3}$$

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.

$$M = \{U = (u_{ki}) : \sum_{i=1}^{c} u_{ki} = 1,\ u_{ki} \geq 0, \forall k, i\}. \tag{4}$$

Moreover $W = (W_1, \dots, W_c)$ shows cluster centers.

The iterative algorithm of fuzzy c-means clustering is basically an alternative minimization of $J_H(U, W)$ with respect to

ISBN: 978-989-95079-6-8

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

$$u_{ki} = \left[\sum_{j=1}^{c} \left(\frac{D(x_k, W_i)}{D(x_k, W_j)} \right)^{\frac{1}{m-1}} \right]^{-1}, \tag{5}$$

$$W_i = \frac{\sum_{k=1}^{n} (u_{ki})^m \Phi(x_k)}{\sum_{k=1}^{n} (u_{ki})^m} \tag{6}$$

where we put

$$D(x_k, W_i) = \|\Phi(x_k) - W_i\|_H^2. \tag{7}$$

Equations (5) and (6) are 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 (6) from iteration, i.e., we substitute (6) into (7) to have an updating formula for $D(x_k, W_i)$[3]:

$$D(x_k, W_i) = \boldsymbol{K}(x_k, x_k) - \frac{2}{\sum_{k=1}^{n} (u_{ki})^m} \sum_{j=1}^{n} (u_{ji})^m \boldsymbol{K}(x_j, x_k) + \frac{1}{\{\sum_{k=1}^{n} (u_{ki})^m\}^2} \sum_{j=1}^{n} \sum_{\ell=1}^{n} (u_{ji} u_{\ell i})^m \boldsymbol{K}(x_j, x_\ell). \tag{8}$$

We hence repeat (5) and (8) until convergence when kernel fuzzy c-means clustering should be used.

2.2 An explicit mapping and its properties

We sometimes have applications in which we wish to have explicit cluster centers. Even if we cannot show them in a low-dimensional data space, there are ways to have approximate representations or visualizations, one of which is to show kernel principal components and another is to use SOM mapping.

To this end we use the following explicit mapping that is simple enough but seems unnoticed.

$$\Phi(x_k) = e_k \quad (k = 1, 2, \ldots, n) \tag{9}$$

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 $\Phi \colon X \to \boldsymbol{R}^n$, i.e., $\Phi(\cdot)$ is not defined on $\boldsymbol{R}^p$ but is limited to the finite set X. Moreover we assume that the inner product of $\boldsymbol{R}^n$ is

$$\langle e_k, e_\ell \rangle = \boldsymbol{K}(x_k, x_\ell) \tag{10}$$

instead of the usual $\langle e_k, e_\ell \rangle = \delta_{k\ell}$.

We have the next proposition.

Proposition 1. *If kernel $\boldsymbol{K}(x, y)$ is positive definite, $\langle e_k, e_\ell \rangle$ defined by (10) satisfy the axioms of the inner product of $\boldsymbol{R}^n$, that is, $\boldsymbol{R}^n$ with (10) is an inner product space.*

We now apply this mapping to fuzzy c-means clustering. It is sufficient to show the optimal solution of W_i. We then have the next proposition of which the proof is easy and omitted.

Proposition 2. *For all positive definite kernel $\boldsymbol{K}(x, y)$ and mapping Φ by (9), the cluster centers are the same and are given by*

$$W_i = \left(\frac{(u_{1i})^m}{\sum_{k=1}^{n} (u_{ki})^m}, \ldots, \frac{(u_{ni})^m}{\sum_{k=1}^{n} (u_{ki})^m} \right)^T, \quad i = 1, \ldots, c \tag{11}$$

The next proposition of which the proof is almost trivial shows an important result of the equivalence between the usual technique and the explicit mapping method.

Proposition 3. *Using the explicit mapping (11) to (7), we have (8).*

That is, (11) derived from the single mapping (9) has all necessary and sufficient information for kernel fuzzy c-means clustering.

As noted above, formulas in kernel principal component analysis are derived likewise. Since the derivation repeats what is described in textbooks [9], we omit the details. The mapping (9) 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 [6].

3 Analysis of Terms from a Set of Texts

We hereafter show an application in which mapping (9) has a natural interpretation. A typical example is analysis of terms from s set of texts. A well-known model for term analysis uses a vector-space model [8] which is also called a bag model. Mizutani and Miyamoto[6] generalize the model into fuzzy multiset model and use kernel based learning and clustering. Note that a term may occur many times in a text which we call here term occurrences, or simply occurrences. Term occurrences are handled as term frequencies or count in bags in the vector space model, or it is treated as fuzzy multisets in [6] when memberships are attached to each term occurrence. When c-means clustering is applied, cluster centers have implications as they are representatives of clusters.

In these models, however, a structure inside a text such as distances or topology between occurrences is not considered. Miyamoto and Kawasaki[5] propose a kernel-based model that handles fuzzy neighborhood. As a result, kernel-based data analysis can be applied but a drawback is that we cannot have cluster centers. We consider this model of fuzzy neighborhood here and uses the above explicit mapping and overcome the last drawback.

3.1 Fuzzy neighborhood model

Let us consider a model that consists of a quintuple:

$$< T, O, d, R, N >$$

where each element is as follows.

1. $T = \{t_1, \ldots, t_n\}$: a finite set of *terms* or *keywords*;
2. O and d: a metric space; O is a finite set of *term occurrences* where a metric d is defined;
3. R : a fuzzy relation of $T \times O$;
4. N : a fuzzy relation of $O \times O$ called *fuzzy neighborhood*.

ISBN: 978-989-95079-6-8

The relation $R(t,o)$ shows correspondence between a term and its occurrence. For simplicity we assume, for every $t \in T$, $R(t,\cdot) = \{o \in O : R(t,o) > 0\}$ is not empty. The most simple relation is

$$R(t,o) = \begin{cases} 1 & (o \text{ is an occurrence of } t), \\ 0 & (\text{otherwise}), \end{cases}$$

and this relation is assumed hereafter in this paper.

Moreover we impose the following conditions to N:

(i) $N(o,o) = 1$ for all $o \in O$;

(ii) $d(o,o') > d(o,o'') \Rightarrow N(o,o') \le N(o,o'')$;

(iii) $N(o,o') \to 0$ as $d(o,o') \to \infty$;

(iv) $N(o,o') = N(o',o)$.

Our purpose is to define a natural inner product space on T using the structure of occurrence space. We define a proximity relation $p(t,t')$, $t,t' \in T$:

$$p(t,t') = \sum_{o \in O} \sum_{o' \in O} R(t,o)N(o,o')R(t',o'). \tag{12}$$

If the matrix $P = (p(t,t'))$ is positive-definite, we can define an inner product space. If P is positive definite, we can use a normalized relation

$$s(t,t') = \frac{p(t,t')}{\sqrt{p(t,t)p(t',t')}}. \tag{13}$$

which is also positive-definite.

Generally, the above equation (12) does not guarantee the positive definiteness. We hence consider conditions for the positive-definiteness. To obtain a general condition is not practically useful. We therefore describe a specific, but broad enough, application example.

3.2 A model of a text set

A set of texts where each term may occur many times is represented as a sequence. For example, let a term set is $T = \{a,b,c\}$. Then an example of O can be represented by

$$O = aabcbbcccabbccabab$$

or more precisely, we put suffixes:

$$O = a_1\, a_2\, b_1\, c_1\, b_2\, b_3\, c_2\, c_3\, c_4\, a_3\, b_4\, b_5\, c_5\, c_6\, a_4\, b_6\, a_5\, b_7 \tag{14}$$

in order to distinguish occurrences. Obviously, $R(a,\cdot) = \{a_1,\ldots,a_5\}$. What if two or more texts should be handled? We can connect those texts into a sequence using dummy terms. Hence a sequence is sufficient to represent a set of texts. A natural distance d on O is given by

$$d(o,o') = \{\text{the number of occurrences between } o \text{ and } o'\}+1 \tag{15}$$

Thus, $d(a_1,c_1) = 3$, $d(c_6,c_2) = 7$, *etc* in the above example.

Next, two specific examples of fuzzy neighborhood are considered.

Fuzzy neighborhood using a monotone function

Let us show and example of a fuzzy neighborhood. For this purpose we introduce a monotone-decreasing function $f\colon \boldsymbol{R} \to [0,1]$ such that

(I) $f(0) = 1$,

(II) $f(-x) = f(x)$,

(III) $\lim_{x\to\infty} f(x) = 0$.

We define

$$N(o,o') = f(d(o,o')). \tag{16}$$

We have

Proposition 4. *Fuzzy relation $N(o,o')$ defined by (16) satisfies the conditions (i)–(iv) for a fuzzy neighborhood.*

The next proposition is crucial to this model.

Proposition 5. *If function $f(x)$ is convex on $[0,\infty)$, then $P = (p(t,t'))$ is positive-definite. Namely, we can define an inner product space using P.*

Let us consider a simple example where a convex function

$$f(x) = \begin{cases} 1 - \frac{x}{3} & (0 \le x \le 3) \\ 0 & (x > 3) \end{cases}$$

is used and $f(-x) = f(x)$. For the example of (14), $N(a_1,b_1) = 1/3$, $N(a_2,b_1) = 2/3$, *etc.* we hence have

$$p(a,b) = \sum_{i=1}^{5} \sum_{j=1}^{7} N(a_i,b_j) = 4.$$

3.3 Fuzzy neighborhood using a hierarchical structure

Another class for fuzzy neighborhood uses a hierarchical structure of a text, that is, a text frequently has chapters, sections, subsections, and paragraphs. Hence the next distance can be induced using $0 < \alpha < \beta < \gamma$.

- If o and o' occurs in a same paragraph, $d(o,o') = \alpha$;
- else if o and o' occurs in a same subsection, $d(o,o') = \beta$;
- else if o and o' occurs in a same section, $d(o,o') = \gamma$;
- else $d(o,o') = \infty$.

Such a distance induced from a hierarchical classification satisfies

$$d(o,o'') \le \max\{d(o,o'), d(o',o'')\}. \tag{17}$$

which is called ultra-metric property. It is also known that this property is equivalent to fuzzy equivalence that is also called fuzzy similarity. That is,

$$N(o,o') = f(d(o,o')) \tag{18}$$

is a fuzzy equivalence.

We have the next proposition.

Proposition 6. *If $d(o,o')$ satisfies (17) and $N(o,o')$ is defined by (18), then $P = (p(t,t'))$ is positive-definite. Namely, we can define an inner product space using this P.*

ISBN: 978-989-95079-6-8

Let us moreover consider a particular case of d when O consists of documents $D_1, \ldots, D_m$:

$$d(o, o') = \begin{cases} 0 & (o \text{ and } o' \text{ are in a same document}), \\ \infty & (o \text{ and } o' \text{ are in different documents}). \end{cases} \quad (19)$$

Suppose also that the frequency of occurrences of t in D_i is $F_i(t)$.

We have the next proposition.

Proposition 7. *If $d(o, o')$ is given by (19) and $N(o, o')$ is defined by (18). then*

$$p(t, t') = \sum_{i=1}^{m} F_i(t) F_i(t'), \quad (20)$$

which is used in the vector space model.

We thus note that the model in this section generalizes the vector space model.

Finally, we note that the above model can directly be applied to a document set $D = \{d_1, \ldots, d_m\}$ using

$$R_D(d, t) = \begin{cases} 1 & (\ t \text{ occurs in } d), \\ 0 & (\ t \text{ doe no not occur in } d). \end{cases}$$

Then we have a positive definite measure between two documents:

$$p_D(d, d') = \sum_{t, t' \in T} R_D(d, t) p(t, t') R_D(d', t'), \quad (21)$$

and a normalized measure

$$s_D(d, d') = \frac{p_D(d, d')}{\sqrt{p_D(d, d) p_D(d', d')}}, \quad (22)$$

which can be used as inner products of the document space.

4 Numerical Examples

We briefly show the results of kernel fuzzy c-means clustering (KFCM) with the kernel principal components (KPCA) applied to a set of documents extracted from a Japanese newspaper articles. Sixty articles from ASAHI Shinbun (ASAHI News) have been used. They are categorized by the publisher into three categories of 'Economics', 'Politics', and 'Social Affairs'; each category has 20 articles. Each of these categories are subdivided into subcategories of 'economic statistics', 'finance', and so on. Terms occurred more than twice were used and the number of terms is 556. The above measure (22) has been used throughout. The method of fuzzy c-means with $m = 1.2$ and with $c = 3$ has been applied. The following two neighborhood functions have been used.

- First type uses a convex function $f(x) = 1.2^{-d(o,o')}$.
- Second type uses a fuzzy equivalence: If two term occurrences are the same term, $N(o, o') = 1$; if they are in a same paragraph, $N(o, o') = 0.9$; if they are in a same document, $N(o, o') = 0.7$; if they are in a same subcategory: $N(o, o') = 0.5$; if they are in a same category: $N(o, o') = 0.3$.

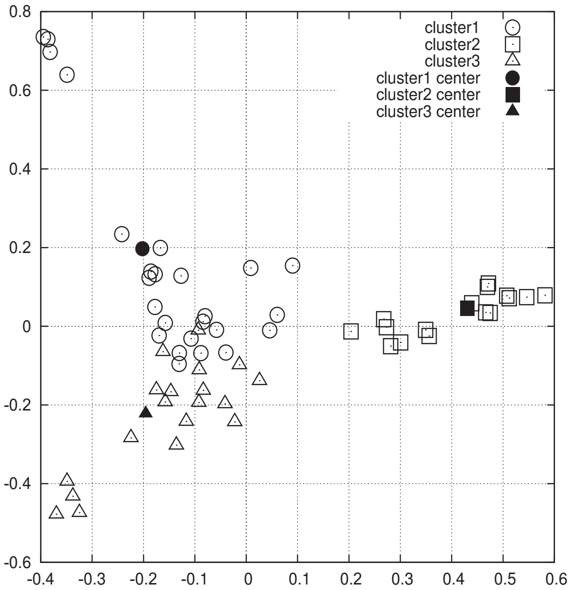

Figure 1: Two-dimensional display from KPCA with three clusters and cluster centers, where the first type of neighborhood function $f(x) = 1.2^{-d(o,o')}$ has been used.

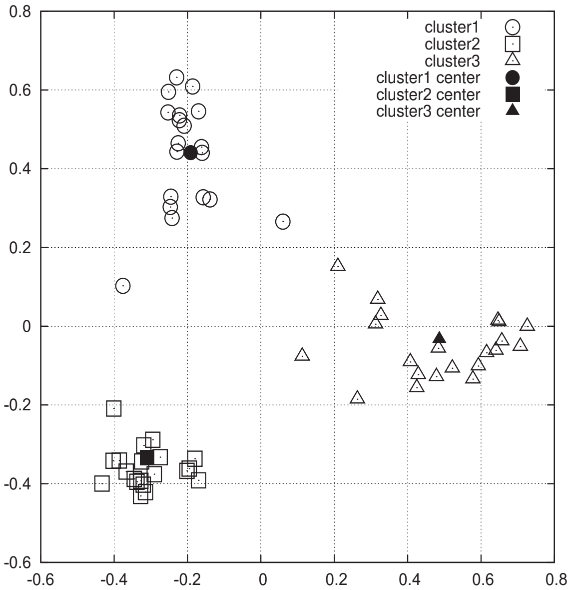

Figure 2: Two-dimensional display from KPCA with three clusters and cluster centers, where the second type of fuzzy equivalence has been used.

Figures 1 and 2 respectively show the two-dimensional figures of two-major axes from KPCA using the first and the second types of neighborhoods. The three symbols of white squares, triangles, and circles show the three clusters obtained from KFCM: fuzzy clusters have been made crisp by the maximum membership rule

$$x_k \to \textit{cluster } i \iff u_{ki} = \max_{1 \le j \le c} u_{kj}.$$

The black square, triangle, and circle are the cluster centers of the corresponding clusters. It seems that the second neigh-

ISBN: 978-989-95079-6-8

borhood of fuzzy equivalence divides more clearly the three clusters.

We have checked the correspondence of the obtained clusters with the actual classification provided from the publisher. The six correspondences have all been checked and the maximum correspondence percentages are given next. In the calculation 1000 trials with random initial values were used and the average numbers are shown below. That is, when we used the first type of neighborhood, the correct classification ratio was 73.5%, while the second type of fuzzy equivalence produced 97.9%.

We also tested a subclass of the above articles: 20 'Economics' and 20 'Politics' articles with the same conditions except that the number of clusters $c = 2$. The number of terms which occurred more than twice is 390. The results are shown as Figures 3 and 4. The separation was better than the previous case of three clusters. The correct classification ratio was 93.2% for the first type of neighborhood and 99.8 for the second neighborhood.

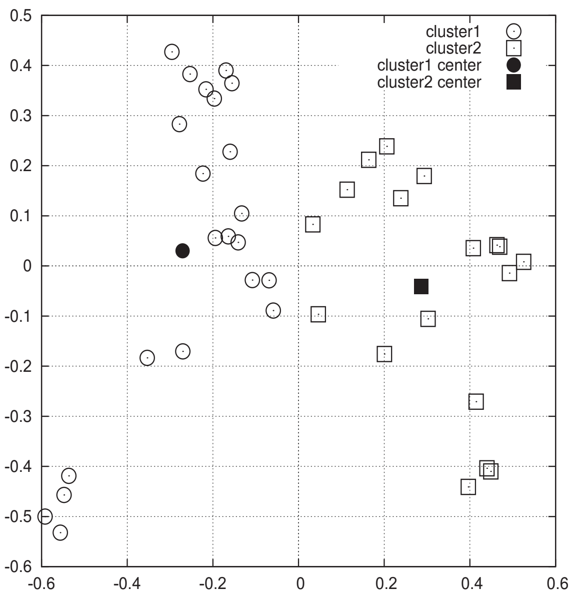

Figure 3: Two-dimensional display from KPCA with two clusters and cluster centers, where the first type of neighborhood function $f(x) = 1.2^{-d(o,o')}$ has been used.

5 Conclusion

We have proposed the use of an explicit mapping for kernel based data analysis. To summarize, 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.

The last statement should be put into practice. In relation to fuzzy clustering, the method of fuzzy c-varieties should be

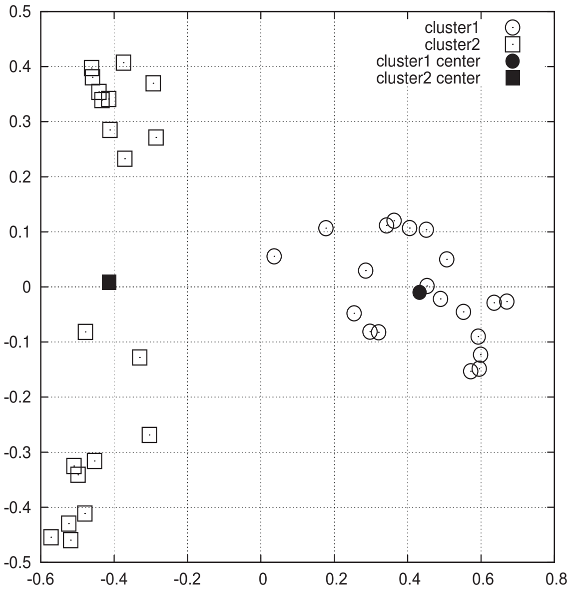

Figure 4: Two-dimensional display from KPCA with two clusters and cluster centers, where the second type of fuzzy equivalence has been used.

studied. We moreover have many research possibilities related to SOM.

This mapping invokes several problems to be solved. For example, when the number of objects are large, we have a problem of many dimensions which should be overcome using some handling large matrix techniques. We also should consider when and where such an explicit mapping become useless. Clarification of such a boundary between usefulness and uselessness is the ultimate objective of the present study.

Acknowledgment

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

Appendix: Notes on Proofs

Since detailed proofs are too lengthy, we give notes on how to prove the propositions given here. The proofs are mostly straightforward.

Note on proof of Proposition 1

The detailed proof is given in standard textbooks [9]. As a rough sketch of the proof, note that the Mercer condition[10]

$$\int\int \boldsymbol{K}(x,y)\eta(x)\eta(y)dxdy \geq 0$$

for all $\eta(x)$ guarantees

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

$\forall \zeta_i \in \boldsymbol{R}$, by putting $\eta(x) = \sum_i \zeta_i \delta(x - x_i)$. Thus the matrix $K = (\boldsymbol{K}(x_i,x_j))$ is positive semi-definite. The kernel function generally does not distinguish positive semi-definiteness

ISBN: 978-989-95079-6-8

and positive definiteness, while positive-definiteness is required for the definition of an inner product. For this purpose a simple trick is to use a regularization which means that we take

$$K \to K + \epsilon I \quad (\epsilon > 0)$$

where ϵ is sufficiently small. The matrix $K + \epsilon I$ is positive definite and approximates K when K is positive semi-definite. In this paper we thus assume positive-definiteness throughout, by using such a regularization when needed.

Note on proof of Proposition 2

The proof of this proposition is not difficult by observing closely (6). Note that W_i given by (6) 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 (6) uses a general variational principle valid for any Hilbert space. Note moreover that we substitute (11) into (5) 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 (10) give different values for different kernels.

Note on proof of Proposition 3

A straightforward calculation shows this proposition is valid.

Note on proof of Proposition 4

The proof is immediate and omitted here.

Note on proof of Proposition 5

Take an arbitrary $c \in O$. The conclusion is immediately obtained from the Pólya's theorem [7] which states that

$$\sum_{a,b} z_a z_b f(|x(a) - x(b)|) \geq 0$$

when f is convex on $[0, +\infty)$ and $f(x) \to 0$ as $x \to +\infty$. Note that $x(a)$ is the real value defined by $x(a) = D(a, c)$ when a is the left hand of c; $x(a) = -D(a, c)$ when a is right hand of c.

Note on proof of Proposition 6

To prove this proposition, we consider a partition matrix $U = (u_{ij})$. Namely, an $n \times n$ real matrix U is called a partition matrix iff there exists a partition $K_1, \ldots, K_c$ of $\mathbf{n} = \{1, 2, \ldots, n\}$ ($\bigcup_j K_j = \mathbf{n}$ and $K_i \cap K_j = \emptyset$, for $i \neq j$) such that

$$u_{ij} = 1 \quad \forall i, j \in K_h$$

for some h and

$$u_{ij} = 0 \quad \forall i \in K_\ell, j \in K_h$$

for $h \neq \ell$.

We have

Lemma: *A partition matrix U is positive semi-definite. It is positive definite if and only if U is identity matrix ($U = I$).*

The proof of the lemma is immediate by observing

$$x^t U x = \sum_{i=1}^{c} \left(\sum_{x_j \in K_i} x_j \right)^2 .$$

The proof of this proposition is now straightforward. We assume that O is a finite set for simplicity. Then an equivalence relation is represented by a partition matrix. Moreover a fuzzy equivalence relation F is represented by a finite collection $U_1, \ldots, U_k$ of partition matrix and positive $\beta_1, \ldots, \beta_k$:

$$F = \sum_{j=1}^{k} \beta_j U_j .$$

Consequently we have

$$x^t F x = \sum_{j=1}^{k} \beta_j x^t U_j x \geq 0.$$

Hence $N(a, b)$ is positive definite. From Proposition 6, $p(t, t')$ is also positive definite. The proposition is thus proved.

Note on proof of Proposition 7

By straightforward calculation, we see this proposition holds.

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] S. Miyamoto, Y. Kawasaki, Kernel Space for Text Analysis Based on Fuzzy Neighborhoods, *WCCI 2008 Proceedings, 2008 IEEE World Congress on Computational Intelligence*, June 1-6, 2008, Hong Kong, pp.738-743, 2008.

[6] 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.

[7] G. Pólya, Remarks on characteristic functions, Proc. First Berkeley Conf. on Math. Stat. and Prob., pp.115–123, Univ. of Calf Press, Berkeley, 1949.

[8] G. Salton, M.J. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, New York, 1985.

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

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

ISBN: 978-989-95079-6-8