University of Nebraska at Omaha
DigitalCommons@UNO

Computer Science Faculty Publications | Department of Computer Science

9-1998

Minimum Cross-Entropy Approximation for Modeling of Highly Intertwining Data Sets at Subclass Levels

Qiuming Zhu
University of Nebraska at Omaha, qzhu@unomaha.edu

Follow this and additional works at: https://digitalcommons.unomaha.edu/compscifacpub

Part of the Computer Sciences Commons

Recommended Citation
Zhu, Qiuming, "Minimum Cross-Entropy Approximation for Modeling of Highly Intertwining Data Sets at Subclass Levels" (1998). *Computer Science Faculty Publications*. 46.
https://digitalcommons.unomaha.edu/compscifacpub/46

This Article is brought to you for free and open access by the Department of Computer Science at DigitalCommons@UNO. It has been accepted for inclusion in Computer Science Faculty Publications by an authorized administrator of DigitalCommons@UNO. For more information, please contact unodigitalcommons@unomaha.edu.

Minimum Cross-Entropy Approximation for Modeling of Highly Intertwining Data Sets at Subclass Levels

QIUMING ZHU zhuq@unomaha.edu
Department of Computer Science, University of Nebraska at Omaha, Omaha, NE 68182-0050

Abstract. We study the problem of how to accurately model the data sets that contain a number of highly intertwining sets in terms of their spatial distributions. Applying the Minimum Cross-Entropy minimization technique, the data sets are placed into a minimum number of subclass clusters according to their high intraclass and low interclass similarities. The method leads to a derivation of the probability density functions for the data sets at the subclass levels. These functions then, in combination, serve as an approximation to the underlying functions that describe the statistical features of each data set.

Keywords: cross-entropy, intertwining data sets, probability distribution, subclasses, cross-entropy minimization

1. Introduction

One of the problems often encountered in a data analysis system is to derive an intrinsic model description on a set (or sets) of data in terms of their inherent properties, such as their membership categories and/or their statistical distribution characteristics. For example, in a database mining process it is necessary to extract the information from a large set of data points (records) and model the data in terms of their uniformity and regularities. This is often done by first obtaining the statistical distributions of the data sets that are grouped in terms of one or more designated key fields, regarded as labels, of the data points, and then mapping them to a set of objective functions. A speech recognition system also needs to have a data model be developed from a large set of experimental data before it can distinguish words and phrases spoken by different people. In these modeling processes, the system typically deals with the problems of (1) the relations between a set of known labels (also named as categories or classes), denoted as $\Omega = \{\omega_1, \omega_2, \ldots, \omega_c\}$, and a set of data points, denoted as $S = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n\}$; and (2) the derivation of a set of descriptive functions, often statistical distributions, denoted as $\{\pi(\mathbf{X}, \omega_i), i = 1, 2, \ldots, c\}$, that depicts the membership characteristics for each of the extracted (or recognized) data-label, denoted as $\mathbf{X}$-ω, relations.

The problem can also be expressed in this way: Given a data set $S = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n\}$, it forms a multidimensional space $R(\mathbf{X})$, where $\mathbf{X}$ represents a member of S. A categorization made according to the labels of the $\mathbf{X}$'s partitions the $R(\mathbf{X})$ into a number of subspaces $R(\omega_i)$, $i = 1, 2, \ldots, c$, where usually we have

$$R(\omega_i) \subseteq R(\mathbf{X}), \bigcup_i R(\omega_i) = R(\mathbf{X}), \text{ and } R(\omega_i) \cap R(\omega_j) = \emptyset; \quad \forall j \neq i.$$

The $R(\omega_i)$'s represent the data sets of $\mathbf{X}$'s with high intraclass and low interclass similarities based on the characteristics specified in each of the ω_i's. It is typical in such a system to assume the existence of a set of functions $\{\pi(\mathbf{X}, \omega_i), i = 1, 2, \ldots, c\}$ that makes:

$$R(\omega_i) = \{\mathbf{X} \mid \forall(j \neq i)[\pi(\mathbf{X}, \omega_i) > \pi(\mathbf{X}, \omega_j)]\}.$$

We call the $\pi(\mathbf{X}, \omega_i)$'s the objective functions of the data modeling. In the statistics domain, the $\pi(\mathbf{X}, \omega_i)$ is expressed as

$$\pi(\mathbf{X}, \omega_i) = p(\mathbf{X} \mid \omega_i)p(\omega_i)$$

where $p(\mathbf{X} \mid \omega_i)$, a *conditional density* function of $\mathbf{X}$, specifies the probability for the data point $\mathbf{X}$ being in category (label) ω_i, and $p(\omega_i)$, the a priori probability, represents the likelihood for the appearance of the label ω_i in the data set.

The $\pi(\mathbf{X}, \omega_i)$'s are in linear or piece-wise linear functions (Nath et al., 1992; Juang and Katagiri, 1992; Ney, 1995) when the $R(\omega_i)$ subspaces are in convex and continual regions. However, there are cases that the $R(\omega_i)$'s do not possess the linearity feature because of the irregular and complex natures on the $\mathbf{X}$-ω relations of the data sets. Figure 1 shows an example where the labeled data set o1 has a concave distribution region and o2 has two discontinuous distribution regions. The data points of these two data sets together form an intertwining distribution in the $R(\mathbf{X})$ space. For the description of the data sets as shown in figure 1, the $\pi(\mathbf{X}, \omega_i)$'s are to be in high-order nonlinear functions. These functions, while not impossible, are often complex to describe and computationally expensive to obtain.

A data analysis system is usually built upon two paradigms. (1) A set of mathematical functions is acquired first by utilization of the statistical distributions of the data sets or an algorithmic computation of the data sets (Nath et al., 1992; Juang and Katagiri, 1992; Ney, 1995). The mathematical functions are then used to partition the data into groups such that each corresponds to a data category. This approach is generally referred to as "discriminant analysis." (2) A data space is first partitioned into a number of subspaces based on algorithmic computation of certain relational or statistical properties of the data

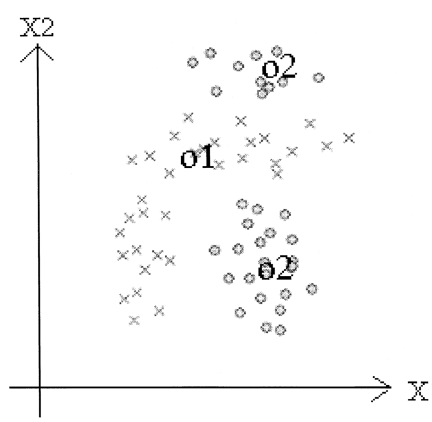

Figure 1. Data points in intertwining distribution.

sets. These subspaces are then modeled in mathematical functions that distinguish the data groups according to their intrinsic properties (Chan and Cheung, 1992; Ishibuchi et al., 1993). This process is generally referred to as "clustering analysis." In both of these approaches, distributions of the data sets are modeled, continuously or discontinuously, at the class level determined by the associated labels of the data sets (Banfield and Raftery, 1993; Bennett and Mangasarian, 1992; Man and Gath, 1994).

In this paper a different approach from the above is taken. We model each data set by a minimum number of subspaces and seek the description of the data set at the subclass levels. The overall description of the data set will then be derived as a combination of the descriptions at the subclass levels. In this approach, the spaces for each data group as a whole may be discontinuous. However, the subspaces within each data group will be continual. Section 2 presents the principles of the cross-entropy minimization technique for the data modeling. Section 3 discusses the foundation of the minimum-set subclass modeling approach. Section 4 describes a computational model by applying the minimum cross-entropy approach to the data modeling at the subclass level. Section 5 presents the algorithms for the construction of subclasses for given data sets. In Section 6, we present the experiment results of applying the minimum-set subclass modeling technique to some intertwining data sets. Section 7 contains concluding remarks.

2. Cross-entropy minimization

The principle of cross-entropy minimization has been a subject of study by Shore and Gray (1982), Rao and Nayak (1985), and Jones and Byrne (1990). Derived from a set of axioms of consistent inference, the technique considered generally a minimum distance approach for the reconstruction of a real function from finitely many linear function values. The problem is expressed as a reconstruction of the positive function $\pi(\mathbf{X})$, $\pi(\mathbf{X}) = \{\pi(\mathbf{X}, \omega_i), i = 1, 2, \ldots, c\}$, defined on the measurable set Π of positive measures, subject to the constraints

$$r_k = \int \pi(\mathbf{X}) g_k(\mathbf{X})\, d\mathbf{X}, \quad k = 0, 1, \ldots, M; \tag{2.1}$$

where the integral is over Π. M is the dimension of the data point (vector) $\mathbf{X}$. The constraint set $\{r_k\}$ consists of known measurable, locally bounded and linearly independent function value $g_k(\mathbf{X})$'s that could be the data point itself or certain transformations of the data points. In the context of data analysis, the $\{r_k\}$ correspond to the expected values of the data points, and $\pi(\mathbf{X})$ the unknown probability density functions of the data sets.

Let $\{Q\}$ be the collection of all admissible functions defined on the data sets $S = \{\mathbf{X}\}$, that is, $\pi(\mathbf{X})$ is a member of $\{Q\}$. For the given sets of data, the problem of reconstructing $\pi(\mathbf{X})$ is to find, as output of the system, an admissible data-consistent reconstruction $\pi(\mathbf{X}) \in \{Q\}$, that is "optimal" in some appropriate sense.

Let $P(\mathbf{X})$ be an a prior estimate of $\pi(\mathbf{X})$ and $Q(\mathbf{X})$ be a posterior estimation of $\pi(\mathbf{X})$, both $P(\mathbf{X})$ and $Q(\mathbf{X})$ are members of $\{Q\}$. The general approach of optimization is to select the posterior estimate $Q(\mathbf{X})$ such that a distortion (discrepancy, or distance) measurement

of the form

$$D[Q(\mathbf{X}),\, P(\mathbf{X})] = \int f(Q(\mathbf{X}),\, P(\mathbf{X}))\, d\mathbf{X}, \tag{2.2}$$

is minimum.

A careful study of the various conditions for the $f(\cdot,\cdot)$ function leads to one measurement that holds the property of directed orthogonality. This measurement, known as cross-entropy between two functions $Q(\mathbf{X})$ and $P(\mathbf{X})$, is expressed as

$$H[Q(\mathbf{X}),\, P(\mathbf{X})] = \int Q(\mathbf{X}) \log\left(\frac{Q(\mathbf{X})}{P(\mathbf{X})}\right) d\mathbf{X}, \tag{2.3}$$

which is also called Kullback distortion (Jones, 1992). The principle states that, of all the distributions that satisfy the constraints, the posterior $Q(\mathbf{X})$ with the least cross-entropy with respect to the prior $P(\mathbf{X})$ should be chosen to properly approximate the $\pi(\mathbf{X})$.

An important fact that makes the above $Q(\mathbf{X})$ the best estimate of $\pi(\mathbf{X})$ rests on the cross-entropy's well-known and unique property as an information measure. For example, cross-entropy satisfies $H[Q(\mathbf{X}),\, P(\mathbf{X})] \geq 0$ with equality only if $Q(\mathbf{X}) = P(\mathbf{X})$ almost everywhere. The general concept of cross-entropy minimization can be stated as the following: Given a positive prior probability density $P(\mathbf{X})$, if there exists a posterior that satisfies the constraints (2.1) and

$$\int Q(\mathbf{X})\, d\mathbf{X} = 1, \tag{2.4}$$

and minimizes the cross-entropy (1.3), then it has the form

$$Q(\mathbf{X}) = P(\mathbf{X}) \exp\left(-\lambda - \sum_{k=0}^{M} \beta_k\, g_k(\mathbf{X})\right), \tag{2.5}$$

where λ and β_k are Lagrangian multipliers whose values are determined by the constraints (2.1) and (2.4). The cross-entropy at the minimum, therefore, can be expressed in terms of the Lagrangian multipliers and the r_k as follows:

$$H[Q(\mathbf{X}),\, P(\mathbf{X})] = -\lambda - \sum_{k=0}^{M} \beta_k\, r_k. \tag{2.6}$$

It is necessary to choose λ and β_k so that the constraints are satisfied. Conversely, if one can find the values for λ and β_k in (2.5) such that the constraints (2.1) and (2.4) are satisfied, then the solution for the objective function exists and is given by (2.5). Unfortunately, it is usually impossible to obtain a closed-form solution expressed directly in terms of the known expected values r_k rather than in terms of the Lagrangian multipliers. Computational methods for finding approximate solutions are, however, available (Shore, 1982).

The minimum cross-entropy method fits nicely into the paradigm of data clustering problems. When a mathematical function, that is, $\pi(\mathbf{X})$, of a data set distribution is sought,

it can be readily modeled as an $Q(\mathbf{X})$ function in the cross-entropy minimization. A positive a prior function $P(\mathbf{X})$ for the description of $\pi(\mathbf{X})$ could be always assumed in this process. Informally speaking, $H[Q(\mathbf{X}), P(\mathbf{X})]$ is a measure of the "information divergence" or "information dissimilarity" between $Q(\mathbf{X})$ and $P(\mathbf{X})$. Shore and Gray (1982) showed the application of this approach to the problems of classifying an input vector of measurements to a fixed set of data centers by a nearest neighbor rule. However, the application of this method to multiple subclass modeling of the data sets is a new attempt.

3. Subclass modeling of intertwining data sets

We consider a paradigm in which a complexly distributed data set can be models as consisting of a number of subsets, each with a relatively simple distribution. Under this modeling approach, for example, the labeled data sets of figure 1 would be reconstructed in four subsets as shown in figure 2, where each subset of the data point is enclosed in a convex distribution region.

A question often asked about is, what constraints should be applied to these subsets to make the data model a valid and accurate one. By investigating the data set clustering techniques and their relationship with the cross-entropy minimization approach, it reveals that it is necessary to construct a minimum set of these subgroups to properly represent the intrinsic properties of the data set. We therefore introduce the minimum-set subclass modeling technique that is to be used to make an accurate description of the distribution of a complex intertwining data set.

To describe the minimum-set subclass modeling, let's start from the description of the data set S, in which each data point $\mathbf{X}$ is associated with a specific label ω_i, $\omega_i \in \Omega$. Let S_i be used to denote the set of the data points that have been labeled by ω_i, i.e.,

$$S = \bigcup_{i=1}^{c} S_i, \quad S_i \cap S_j = \emptyset, \quad \forall i \neq j.$$

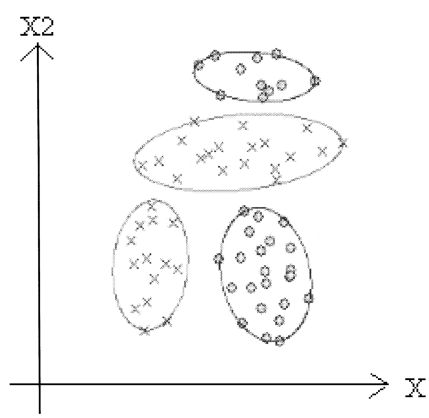

Figure 2. Subclass model on data sets of figure 1.

That is, for each $\mathbf{X} \in S$, there exists an i, $i = 1, 2, \ldots, c$, such that $[(\mathbf{X} \in S_i) \Rightarrow (\mathbf{X} \propto \omega_i)]$. Here, $\mathbf{X} \propto \omega_i$ stands for $\mathbf{X}$ being labeled by ω_i. We then have the following definitions.

Definition (Subclass clusters (SC)). Let S_i be a set of data points labeled ω_i, $S_i \subseteq S$ and $\omega_i \in \Omega$. Let ε_{ik} be the kth subset of S_i. That is, $\varepsilon_{ik} \subseteq S_i$, where $k = 1, 2, \ldots, d_i$, and d_i is the number of subset in S_i. Let $p(\mathbf{X} \mid \varepsilon_{ik})$ be a probability density function for the data point $\mathbf{X}$'s in ε_{ik}. The *subclass clusters* of S_i are defined as the set $\{\varepsilon_{ik}\}$ that satisfies the following conditions:

$$\bigcup_{k=1}^{d_i} \varepsilon_{ik} = S_i, \tag{3.1}$$

$$\forall l \neq k\, [\varepsilon_{ik} \cap \varepsilon_{il} = \emptyset], \tag{3.2}$$

$$\forall l \neq k\, [(\mathbf{X} \in \varepsilon_{ik}) \Rightarrow (p(\mathbf{X} \mid \varepsilon_{ik}) > p(\mathbf{X} \mid \varepsilon_{il}))], \tag{3.3}$$

$$\forall (j \neq i)[(\mathbf{X} \in \varepsilon_{ik}) \Rightarrow (p(\mathbf{X} \mid \varepsilon_{ik}) \geq p(\mathbf{X} \mid \varepsilon_{jl}))]. \tag{3.4}$$

Definition (Minimum-set subclass clusters (MSSC)). Let ε_{ik} and ε_{il} be two subclass clusters of the data points in S_i, $k \neq 1$ and $\varepsilon_{il} \neq \emptyset$. Let $\varepsilon_i = \varepsilon_{ik} \cup \varepsilon_{il}$ and $p(\mathbf{X} \mid \varepsilon_i)$ be the probability density function defined on ε_i. We say that the subclass cluster set $\{\varepsilon_{ik}; k = 1, 2, \ldots, d_i\}$ is a *minimum-set subclass clusters* of S_i, if for any $\varepsilon_i = \varepsilon_{ik} \cup \varepsilon_{il}$ we would have:

$$\exists (j \neq i)\, \exists m\, \exists (\mathbf{X} \in \varepsilon_i)\, [(p(\mathbf{X} \mid \varepsilon_i) < p(\mathbf{X} \mid \varepsilon_{jm})], \tag{3.5}$$

or

$$\exists (j \neq i)\, \exists m\, \exists (\mathbf{X} \in \varepsilon_{jm})\, [(p(\mathbf{X} \mid \varepsilon_i) > p(\mathbf{X} \mid \varepsilon_{jm})], \tag{3.6}$$

where ε_{jm} is the mth subclass cluster for a data point set S_j. The above definition means that a subclass cluster of a data set must be large enough such that any joint set of them would then violate the subclass definition (condition (3.4)).

We consider the construction of the subclass clusters ε_{ik} as a step-by-step process that determines the members of the subset by a sequential examination of the data points in S_i. We will then consider the determination of the probability density functions, $p(\mathbf{X} \mid \varepsilon_{ik})$'s, for the subclasses constructed. As the formation of subclass clusters may rely on the utilization of $p(\mathbf{X} \mid \varepsilon_{ik})$ (according to the conditions (3.3) and (3.4)), we give a general assumption that it is to be a positive function. A computational process that applies the cross-entropy minimization technique to construct the subclass clusters and the associated probability density functions of the data sets is described in following section.

4. Cross-entropy minimization for the subclass clusters

The process associated with the subclass modeling involves (1) the selection of data points from S_i into a particular subclass cluster ε_{ik}, and (2) the determination of the probability

density function $p(\mathbf{X} \mid \varepsilon_{ik})$, and thus the $\pi(\mathbf{X}, \omega_i)$, for each data set. Let's consider the construction of the subclasses as a sequential process of cross-entropy minimization, where each data point in the data set adds a constraint to the $Q(\mathbf{X})$ function to be obtained. Shore and Gray (1982) illustrated that it is useful and convenient to view cross-entropy minimization as one implementation of an abstract information operator "o". The operator takes two arguments—the a prior function $P(\mathbf{X})$ and new information I_k—and yields a posterior function $Q(\mathbf{X})$, that is $Q(\mathbf{X}) = P(\mathbf{X})\, o\, I_k$, where I_k also stands for the known constraints on expected values:

$$I_k\colon \int Q(\mathbf{X}) g_k(\mathbf{X})\, d\mathbf{X} = r_k. \tag{4.1}$$

By requiring the operator o satisfy a set of axioms, the principle of minimum cross-entropy follows. The axioms of o are informally phrased by Shore and Gray (1982) as the following:

(1) *Uniqueness*: The results of taking new information into account should be unique.
(2) *Invariance*: It should not matter with respect to the coordinate system the data point accounts for new information.
(3) *System independence*: It should not matter whether information about systems is accounted separately in terms of different probability densities or together in terms of a joint density.
(4) *Subset independence*: It should not matter whether information about system states is accounted in terms of a separate conditional density or in terms of the full system density.

Thus, given a prior probability density $P(\mathbf{X})$ and new information in the form of constraint I_k on expected value r_k, there is essentially one posterior density function that can be chosen in a manner as the axioms stated above.

Considering two constraints I_1 and I_2 associated with the data modeling expressed as:

$$I_1\colon \int Q_1(\mathbf{X}) g_k(\mathbf{X})\, d\mathbf{X} = r_k^{(1)}, \tag{4.2}$$

$$I_2\colon \int Q_2(\mathbf{X}) g_k(\mathbf{X})\, d\mathbf{X} = r_k^{(2)}; \tag{4.3}$$

where $Q_1(\mathbf{X})$ and $Q_2(\mathbf{X})$ are the density function estimations at two different times. The $r_k^{(1)}$ and $r_k^{(2)}$ represent the expected values of the function in the consideration of different data points in S, that is, in terms of the new information about $Q(\mathbf{X})$ contained in the data points $\{\mathbf{X}\}$. Taking count of these constraints, we have (Shore and Gray, 1982)

$$(P(\mathbf{X})\, o\, I_1)\, o\, I_2 = Q_1(\mathbf{X})\, o\, I_2 \tag{4.4}$$

and

$$H[Q_2(\mathbf{X}), P(\mathbf{X})] = H[Q_2(\mathbf{X}), Q_1(\mathbf{X})] + H[Q_1(\mathbf{X}), P(\mathbf{X})] + \sum_{k=0}^{M} \beta_k^{(1)} \left(r_k^{(1)} - r_k^{(2)}\right); \tag{4.5}$$

where, $Q_1(\mathbf{X}) = P(\mathbf{X})\, o\, I_1$, $Q_2(\mathbf{X}) = P(\mathbf{X})\, o\, I_2$, and the $\beta_k^{(1)}$'s are the Lagrangian multipliers associated with $Q_1(\mathbf{X})$. From (4.5) it follows that

$$H[Q(\mathbf{X}), Q_j(\mathbf{X})] = H[Q(\mathbf{X}), P(\mathbf{X})] - H[Q_j(\mathbf{X}), P(\mathbf{X})] - \sum_{k=0}^{M} \beta_k^{(j)} \left(r_k^{(j)} - r_k\right); \tag{4.6}$$

Substitute $H[Q_j(\mathbf{X}), P(\mathbf{X})]$ by Eq. (2.6) we have

$$H[Q(\mathbf{X}), Q_j(\mathbf{X})] = H[Q(\mathbf{X}), P(\mathbf{X})] + \lambda^{(j)} + \sum_{k=0}^{M} \beta_k^{(j)} r_k. \tag{4.7}$$

The minimum $H[Q(\mathbf{X}), Q_j(\mathbf{X})]$ is computed by taking the counts of I_j, $j = 1, \ldots, n$ (where n is the total number of data points) and a value j such that $H[Q(\mathbf{X}), Q_j(\mathbf{X})] \leq H[Q(\mathbf{X}), Q_i(\mathbf{X})]$ for $i \neq j$. The process would take count of the data points one at a time, and choose the $Q_j(\mathbf{X})$ with respect to the selected data point that has the *minimum distance* (nearest neighbor) from the existing functions.

Applying the cross-entropy minimization technique to the construction of the probability density functions $p(\mathbf{X} \mid \omega_i)$ for a given data set, the technique calls for an approximation of the functions under the constraints of the expected values of the data clusters. Expressed as a computational model for the classification of data points, Shore and Gray (1982) showed that the technique resulted in taking the arithmetic mean of the member components in $\{\varepsilon_{ik}\}$ as the representation of the data set. The same result was presented by Jones and Byrne (1990). According to Jones, the best set of data to represent the sets $\{\varepsilon_{ik}\}$ is given by $\{\mu_{ik}\}$, where

$$\mu_{ik} = \frac{1}{N_{ik}} \sum_{\mathbf{X}_j \in \varepsilon_{ik}} \mathbf{X}_j, \tag{4.8}$$

where N_{ik} is the number of data points in the cluster ε_{ik}, i.e., $N_{ik} = \|\varepsilon_{ik}\|$. We call μ_{ik} a moving centroid of the cluster. That means, when data points are examined one by one and added into the subclass clusters in the construction process, the cluster centroid are adjusted to the new expectation values constantly. The covariance parameters Σ_{ik} of the clusters can be estimated by extending the results of the moving centroid and expressed as:

$$\Sigma_{ik} = \frac{1}{N_{ik}} \sum_{\mathbf{X}_j \in \varepsilon_{ik}} (\mathbf{X}_j - \mu_{ik})(\mathbf{X}_j - \mu_{ik})^T, \tag{4.9}$$

The parameters are to be continuously updated also upon the examination of additional data points $\mathbf{X}$'s and the addition of them into the selected subclass clusters.

5. Subclass clustering and the functional approximation algorithms

The following definitions are made as a preparation for the algorithms to be described. Note that they are defined on the concept of *Data clusters* (*DC*) which is a collection of data

point $\mathbf{X}$'s such that they all having the same label ω_i. A data cluster becomes a subclass cluster when it satisfies the conditions (3.1)–(3.4).

Definition (Distance between two data clusters). Let $p(\mathbf{X} \mid \varepsilon_{ik}) \sim G(\mu_{ik}, \Sigma_{ik})$ and $p(\mathbf{X} \mid \varepsilon_{jl}) \sim G(\mu_{jl}, \Sigma_{jl})$ be the approximations of the probability density functions for two data clusters ε_{ik} and ε_{jl}, respectively; where $[p(\mathbf{X} \mid \varepsilon_{ik}) \sim G(\mu_{ik}, \Sigma_{ik})]$ means $p(\mathbf{X} \mid \varepsilon_{ik})$ is a Gaussian density function with parameters μ_{ik} and Σ_{ik}. The *distance* between the two data clusters ε_{ik} and ε_{jl} is defined as

$$\frac{\|\mu_{ik} - \mu_{jl}\|}{|\Sigma_{ik}| + |\Sigma_{jl}|}. \tag{5.1}$$

A function *Distance*$(\varepsilon_1, \varepsilon_2)$ will return a value of the above for two data clusters ε_1 and ε_2.

Definition (Merge of two data clusters). Let ε_{ik} and ε_{il} be two data clusters in the data set S_i. The *Merge* of ε_{ik} and ε_{il} is defined as a data cluster ε such that

$$\varepsilon = \varepsilon_{ik} \cup \varepsilon_{il},$$

and

$$p(\mathbf{X} \mid \varepsilon) \sim G(\mu, \Sigma), \tag{5.2}$$

where $\mu = \frac{1}{\|\varepsilon\|} \sum_{\mathbf{X}_j \in \varepsilon} \mathbf{X}_j$, $\Sigma = \frac{1}{\|\varepsilon\|} \sum_{\mathbf{X}_j \in \varepsilon} (\mathbf{X}_j - \mu)\,(\mathbf{X}_j - \mu)^T$, and $\|\varepsilon\|$ is the total number of data points in the cluster ε.

A function *Merge*$(\varepsilon_1, \varepsilon_2)$ will do the above computation and return ε and $p(\mathbf{X} \mid \varepsilon)$ which is the approximation of the probability density function on the merged data cluster.

Definition (Intersection of two data clusters). Let ε_{ik} and ε_{jl} be two data clusters. Let $p(\mathbf{X} \mid \varepsilon_{ik})$ and $p(\mathbf{X} \mid \varepsilon_{jl})]$ be two approximations of the probability density functions defined on ε_{ik} and ε_{jl}, respectively. We say that the two data clusters intersect if

$$\exists \mathbf{X} \in \varepsilon_{ik} [p(\mathbf{X} \mid \varepsilon_{ik}) < p(\mathbf{X} \mid \varepsilon_{jl})] \forall_j \neq i, \tag{5.3}$$

or

$$\exists \mathbf{X} \in \varepsilon_{jl} \, [p(\mathbf{X} \mid \varepsilon_{jl}) < p(\mathbf{X} \mid \varepsilon_{ik})] \, \forall_j \neq i. \tag{5.4}$$

A function *Intersect*$(\varepsilon_1, \varepsilon_2)$ will return true if ε_1 and ε_2 intersect; otherwise the function will return false.

The subclass clustering algorithm to be described inherits the basic control mechanism from the agglomerative hierarchical clustering of Duda and Hart (1973). The major difference of our algorithm from the original one is the embedding of the cross-entropy minimization technique in the processes of assigning data point to clusters and the computation of

the parameters of the clusters. The algorithm computes an approximation of the probability density function (PDF) for each cluster and uses it to rectify and adjust the clusters. The PDFs are updated constantly as new data points are examined, in applying the cross-entropy minimization principle to explore the data set characteristics. To describe the algorithm, we first make (or restate) a set of specifications of the symbolic notations:

c	The total number of classes in data set S.
S_i	A subset of data set S; S_i contains the data points in class ω_i, $i = 1, 2, \ldots, c$.
$\mathbf{X}$	A data point in n-dimensional space, $\mathbf{X} \in S$.
ε	A subclass cluster; when subscripts are used, ε_{ik} means the kth cluster of S_i.
E_i	The subclass clusters for data set S_i.
$\|E_i\|$	The number of subclass clusters in set E_i.
μ_i	The expectation vector for the PDF of a subclass cluster ε_i.
Σ_i	The covariance matrix for the PDF of a subclass cluster ε_i.

Algorithm. Minimum set subclass clustering and PDF estimation

Input: $\{S_i\}$, $i = 1, 2, \ldots, c$.
Output: $\{E_i\}$, $i = 1, 2, \ldots, c$.

Step 1: **for** each S_i ($i = 1, 2, \ldots, c$) **do** /* Initialize subclass clusters */
Step 1.1: $E_i \leftarrow \emptyset$, $\|E_i\| \leftarrow 0$;
Step 1.2: **for** each $\mathbf{X} \in S_i$ **do**
Step 1.2.1: $k \leftarrow \|E_i\|$; $\varepsilon_{ik} \leftarrow \mathbf{X}$; initialize($\mu_{ik}$, Σ_{ik}),
Step 1.2.2: $E_i \leftarrow E_i \cup \{\varepsilon_{ik}\}$; $\|E_i\|$++;

Step 2: **Repeat:** /* form minimum number, nonintersecting clusters */
Step 2.1: find a pair (ε_{ik}, ε_{il}) such that (ε_{ik}, $\varepsilon_{il} \in E_i$) and ($k \neq l$) and *Distance*(ε_{ik}, ε_{il}) is the minimum among all pairs of (ε_{ik}, ε_{il}) in E_i, $i = 1, 2, \ldots, c$;
Step 2.2: $\varepsilon \leftarrow$ *Merge*(ε_{ik}, ε_{il}),
Step 2.3: **if** NOT(*Intersect*(ε, ε_{jm}) $\forall_j \neq i$ and $\forall m$) **then**
Step 2.3.1: $\varepsilon_{ik} \leftarrow \varepsilon$; compute ($\mu_{ik}$, Σ_{ik}); remove ε_{il} from E_i
Step 2.3.2: $E_i \leftarrow E_i \cup \{\varepsilon_{ik}\}$, remove ε_{il} from E_i, $\|E_i\|$--;
Step 2.4: **Until** no change is made on every $\|E_i\|$.

Step 3: **Return** $\{E_i\}$, $i = 1, 2, \ldots, c$.

This algorithm converges in a finite number of operations. Without loss of generality, we assume that the number of data points, $\|S_i\| = n, i = 1, 2, \ldots, c$. At the Step 1 of the algorithm, the assignment of each data point to a trivial cluster takes an $O(cn)$ time complexity. In Step 2.1, the selection and merge of clusters takes at most $O(n^2)$ computation for every data set S_i. That is a total of $O(cn^2)$ computation for the entire data set S. Step 2.2 takes a maximum of $O(n)$ time complexity. The operation of intersection checking at Step 2.3 $O(cn^2)$ computation at the worst case. Considering the entire Step 2 is to be done

in the maximum of n times, the overall computation in Step 2 then takes a $O(c^2n^3)$ time complexity. Let N be the total number of data points in the training set S, $N = cn$, the worst case time complexity of this algorithm can then be expressed as $O(N^3/c)$.

Since the subclass clusters are already labeled by their belonging classes, the mapping of the data points from their subclasses to their parent classes is straightforward. Therefore, after the execution of the algorithm, we have the data clusters $E_i = \{\varepsilon_{ik}\}, i = 1, 2, \ldots, c$, constructed from the original data set S with a PDF function $p(\mathbf{X} \mid \varepsilon_{ik})$ associated with each ε_{ik}.

6. Experimental results

We conducted the experiment on the data clustering and the functional approximation of the data sets by simulation. The simulation uses randomly generated data sets that exhibit complex and intertwining distributions. Some of the experimental data sets are shown in figure 3, where figures in (a) show the plots of the data sets and (b) show the subclass clusters constructed on the data sets. Symbols "×", "○" and "△" are used to indicate data points of

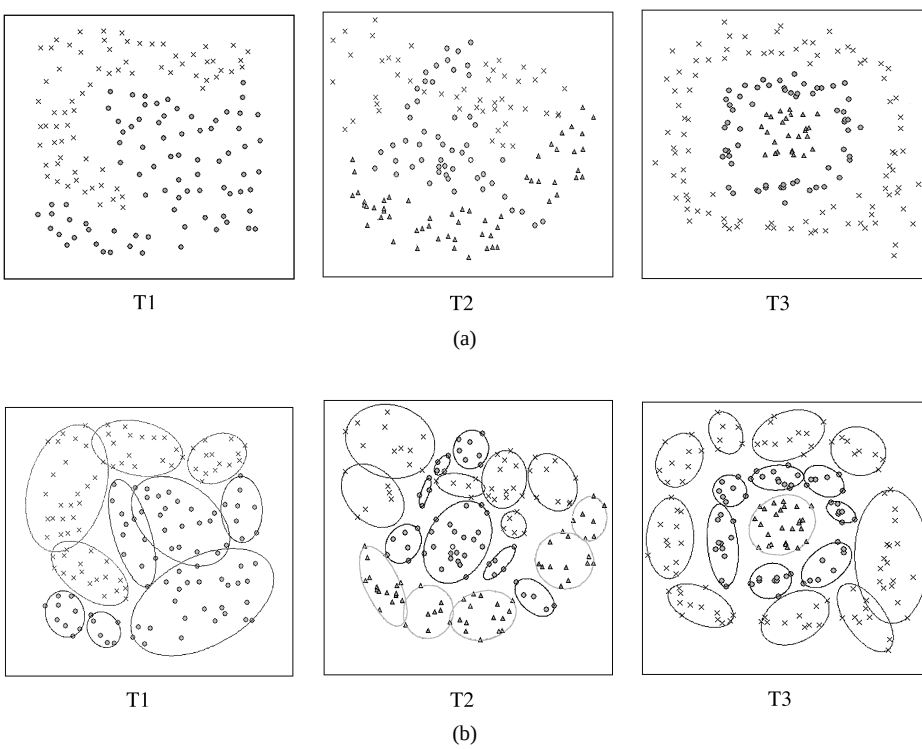

Figure 3. Data sets and their subclass clusters of the test cases.

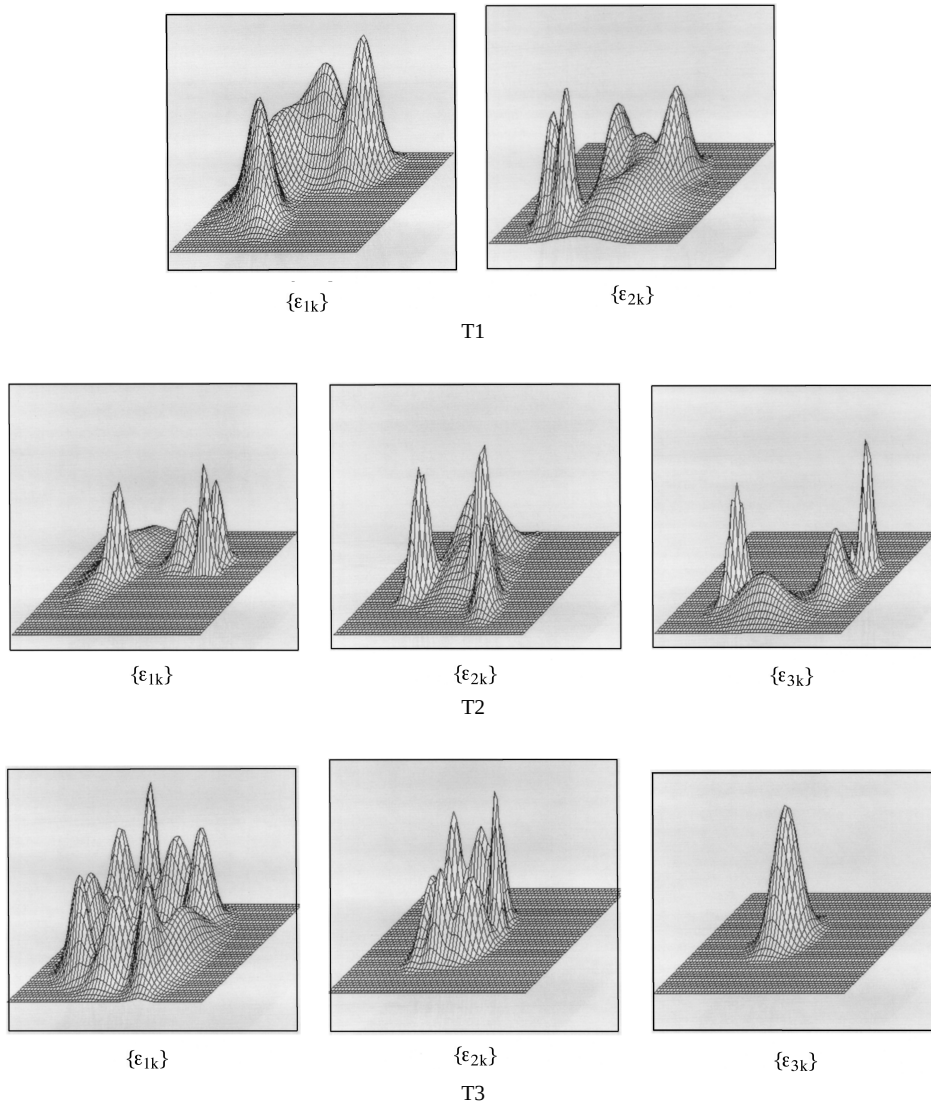

Figure 4. The $p(\mathbf{X} \mid \varepsilon_{ik})$ functions of the subclass data models of the test cases.

different classes. For the illustration purpose, only two-dimensional data are shown in the examples. Every data point in the example is correctly categorized into its labeled group, as the technique being designed for, though the points may be in different subclass clusters. A description of the data sets is obtained by the combination of the probability density functions resulted from the subclass clusters. These functions are shown in figure 4, where we show the functional descriptions of the $p(\mathbf{X} \mid \varepsilon_{ik})$'s for the data points in the same label category.

7. Conclusions

One of the tasks of data analysis is to describe the partitions of a multidimensional space of given data sets in a number of separated subspaces, each corresponding to a specific data category. Many data analysis techniques attempted a linear or piece-wise linear solutions for approximating nonlinearly distributed data sets, thus sacrificed certain degree of the nonlinear precision. In this research, we developed a data clustering technique based on a subclass modeling that is able to provide a solution for modeling the highly intertwining data sets. The process is conducted in terms of the minimization of the cross-entropy of the resulting data models in a multiple subclass space. The technique derives an approximation of the probability density functions of the data sets in the subspace partitions by considering both the interclass and intraclass properties of the data points. Though the distributions of the subclasses are in simple convex form, which renders to simplicity of computation, the overall distribution of the data sets remain the properties of a nonlinear, complex spatial distribution. The technique does not require human interaction to predetermine or tune up the system parameters, thus can be conveniently applied, with adequate generality and accuracy, to many practical data analysis problems.

Acknowledgment

The authors are thankful to the reviewers who provided very detailed and useful comments on the earlier version of this paper. The support of the University Committee on Research at the University of Nebraska at Omaha is also acknowledged.

References

Avi-Itzhak, H.I., Van Mieghem, J.A., and Rub, L. (1995). Multiple Subclass Pattern Recognition: A Maximin Correlation Approach, *IEEE Trans. Pattern Anal., Machine Intell.*, 17(4), 418–431.

Banfield, J.D. and Raftery, A.E. (1993). Model-Based Gaussian and Non-Gaussian Clustering, *Biometrics*, 49, 803–821.

Bennett, K.P. and Mangasarian, O.L. (1992). Robust Linear Programming Discrimination of Two Linearly Inseparable Sets, *Optimization Methods and Software*, 1, 23–34.

Chan, K.P. and Cheung, Y.S. (1992). Clustering of Clusters, *Pattern Recognition*, 25(2), 211–217.

Duda, R.O. and Hart, P.E. (1973). *Pattern Classification and Scene Analysis*, John Wiley & Sons.

Ishibuchi, H., Nozaki, H.K., and Tanaka, H. (1993). Efficient Fuzzy Partition of Pattern Space for Classification Problems, *Fuzzy Sets and Systems*, 59, 295–304.

Jones, L.K. and Byrne, C.L. (1990). General Entropy Criteria for Inverse Problems, with Applications to Data Compression, Pattern Classification, and Cluster Analysis, *IEEE Transactions on Information Theory*, 36(1), 23–30.

Juang, B.H. and Katagiri, S. (1992). Discriminative Learning for Minimum Error Classification, *IEEE Transactions on Signal Processing*, 40, 3043–3054.

Man, Y. and Gath, I. (1994). Detection and Separation of Ring-Shaped Clusters using Fuzzy Clustering, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(8), 855–861.

Nath, R., Jackson, W., and Jones, T.W. (1992). A Comparison of the Classical and the Linear Programming Approaches to the Classification Problem in Discriminant Analysis, *Journal of Statistical Computation and Simulation*, 41(1), 73–93.

Ney, H. (1995). On the Probablistic Interpretation of Neural Network Classifiers and Discriminative Training Criteria, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(2), 107–119.

Rao, C.R. and Nayak, T.K. (1985). Cross-Entropy, Dissimilarity Measures, and Characterizations of Quadratic Entropy, *IEEE Transactions on Information Theory*, 31(5).

Shore, J.E. and Gray, R.M. (1982). Minimum Cross-Entropy Pattern Classification and Cluster Analysis, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 4(1), 11–17.