Proceedings of the International MultiConference of Engineers and Computer Scientists 2008 Vol I
IMECS 2008, 19-21 March, 2008, Hong Kong

# Similarity and Pattern Recognition

Chun-Hung Tzeng, Computer Science Department
Ball State University, Muncie, IN 47306, U.S.A., tzeng@bsu.edu

*Abstract*—**This paper formally defines similarities as tolerance relations, which are reflexive and symmetric binary relations. An abstract set with a similarity is called a tolerance space. The training data set in a learning task is a given database of independent identically distributed random pairs $(X_i, Y_i)$, where each $X_i$ is a record and $Y_i$ is its label: $Y_i \in \{0, 1\}$. The goal of the learning is to design a classifier of which the error probability is near to the theoretical limitation, the Bayes error. The learning process consists of finding a similarity of feature vectors $\psi(X_i)$'s and the learning result is a representative data clustering on the tolerance space of feature vectors. The information about a record $X$ derived from the representative clustering is the set of representatives similar to the feature vector $\psi(X)$. The percentage of the records of class 1 in the intersection of these representative clusters is used to estimate the conditional probability of $Y = 1$. This paper defines a $\theta$-classifier, which assigns the record to class 1 if the conditional probability is larger than the threshold $\theta$. If the clustering is a partition, the threshold $\theta = \frac{1}{2}$ minimizes error probability in the training data set. In general, an optimal $\theta$-classifier has a different threshold. The experiments show the trade-off between the number of clusters and the error probabilities of the optimal $\theta$-classifiers.**

***Keywords: similarity, tolerance-space, representative-clustering, pattern-recognition, classification***

## 1 Introduction

Pattern recognition is about guessing or predicting the unknown nature of an observation, a discrete quantity such as black or white, abnormal or normal, one or zero, sick or healthy, real or fake. Usually, an observation is a collection of numerical measurements such as an image or a vector of weather data ([3, 9]). Formally, an observation is a $d$-dimensional vector $x$. The unknown nature is a *class*, which is denoted by $y$ and takes values in a finite set $\mathcal{C} = \{0, 1, 2, ..., M\}$. The task is to create a function $g : \mathcal{R}^d \longrightarrow \mathcal{C}$. The value $g(x)$ represents the guess of $y$, given $x$. The function $g$ is called a *classifier*. The classifier errs on $x$ if $g(x) \neq y$. In a probabilistic setting (e.g., [3]), we consider a random pair $(X, Y)$ on $\mathcal{R}^d \times \mathcal{C}$, of which a distribution describes the frequency of encountering particular pairs in practice. The probability of error for a classifier $g$ is $L(g) = P(g(X) \neq Y)$. A best possible classifier $g^*$ which has the minimal probability of error is called *Bayes classifier* or *Bayes rule*. The minimal probability of error is called the *Bayes error*.

In the two-class problem $\mathcal{C} = \{0, 1\}$, the Bayes classifier is defined by $g^*(x) = 1$ iff $\eta(x) > 1/2$, where $\eta(x)$ is the posterior probability $\eta(x) = P(Y = 1|X = x)$. In most cases, the distribution of $(X, Y)$ and the Bayes classifier $g^*$ are unknown. To design a classifier is based on a given database of pairs $(X_i, Y_j), 1 \leq i \leq n$. Such a task is called a *supervised learning*. Many classification rules have been proposed (e.g., [3, 9]). For example, the $k$-nearest neighbor rule $g_n(x)$ takes a majority vote over the $Y_i$'s in subset of $k$ pairs $(X_i, Y_i)$ from the given database that have the smallest values $\|X_i - x\|$.

Similarities or dissimilarities play a central role in the pattern recognition, implicitly or explicitly. For example, the $k$-nearest neighbor rule uses the Euclidean distance to measure similarities. Most similarity measures are reflexive and symmetric. That is, each $x$ is not dissimilar to itself and the similarity measure of two $x$'s is independent of the order of $x$.

This paper introduces a probabilistic model for the two-class pattern recognition on an abstract space $\Omega$ based on a formally defined similarity, called a *tolerance relation* and denoted by $\xi$. The similarity $\xi$ is a reflexive and symmetric binary relation on $\Omega$. The pair $\Omega^\xi = (\Omega, \xi)$ is called a *tolerance space*. Zeeman [19] introduced tolerance relation to describe the imperfection of human sight. A representative system of a tolerance space was studied in [17]. More references on tolerance relation are in [13].

In the tolerance space, the neighborhood of an element $x$ is $\xi(x) = \{u \in \Omega : \xi(x, u)\}$, the set of all elements similar to $x$. Let $\mathcal{F}_\xi$ be the Borel field generated by all such neighborhoods. The probabilistic model is constructed on the measurable space $(\Omega^\xi, \mathcal{F}_\xi)$. The pattern recognition problem is represented by a random pair $(X, Y)$, $X \in \Omega$ and $Y \in \{0, 1\}$. A classifier is a measurable function $g : \Omega \longrightarrow \{0, 1\}$, which errs on $X$ if $g(X) \neq Y$. Bayes rule is extended to this model, which is based on the posterior probability $\eta(x) = P(Y = 1|X \in \overline{x})$, where $\overline{x}$ is the minimal measurable set containing $x$. Bayes error is also defined.

This paper designs classifiers based on a training data set, a database of pairs $(X_i, Y_i)$, where $1 \leq i \leq n$, the record $X_i \in \Omega$, and the class label $Y_i \in \{0, 1\}$. We assume that

ISBN: 978-988-98671-8-8 

the sequence $(X_1, Y_1), ..., (X_n, Y_n)$ is a sequence of independent identically distributed (i.i.d.) random pairs. Let $\Omega_n = \{X_i | 1 \leq i \leq n\}$. A data pre-processing is formally represented by a function $\psi : \Omega_n \longrightarrow \Phi$. The image of the function $\Phi_m = \psi(\Omega_n)$ is the feature vector set of the training data set. The cardinality $|\Phi_m|$ is usually much smaller than $|\Omega_n|$ for computational purpose.

For each feature vector $x \in \Phi_m$, we store the frequencies of $Y$ in the training data set in $n_0(x)$ and $n_1(x)$, which are the numbers of records of class 0 and class 1 in $\psi^{-1}(x) \subseteq \Omega_n$, respectively. On the training data set, the conditional probability $Y = 1$, given $\psi(X) = x$, is $\eta_{\Phi_m}(x) = \frac{n_1(x)}{n_0(x)+n_1(x)}$. The Bayes error based on this conditional probability should near the original Bayes error. Otherwise, the data pre-processing should be redesigned.

On $\Phi_m$, we define a similarity $\xi$ and consider the tolerance space $(\Phi_m)^\xi$, on which we apply the data clustering introduced in [16] and find a representative system $\mathcal{R}_\xi = \{R_1, R_2, ..., R_k\}$ of the space.

Each feature vector in $\Phi_m$ is similar to at least one $R_i$. For each $i$, the cluster represented by $R_i$ is its neighborhood $\xi(R_i)$. Such a data clustering is not a partition in general. Two clusters may have non-empty intersection. $\mathcal{R}_\xi$ is minimal if there is no smaller representative system. Maak [12] introduced the concept of minimal representative system to approximate the mean value of some functions on abstract groups. In this paper, we use a minimal representative system to represent a pattern learned from the training data set.

Consider a random pair $(X, Y)$. Given the value of $X$, we derive information about $X$ from the representative system $\mathcal{R}_\xi$. We find the representatives similar to the feature vector $\psi(X)$, denoted by $N_X = \{R'_1, ..., R'_{k'}\}$, and consider their intersection, denoted by $F_X$. Let $t_0$ and $t_1$ be the record numbers of class 0 and class 1 in the subset $\psi^{-1}(F_X)$ of the training data set, respectively. On the training data set, the conditional probability of $Y = 1$, given $\psi(X) \in F_X$, is $\eta(X) = \frac{t_0}{t_0+t_1}$. We use this conditional probability to estimate the posterior probability of any record $X$.

The *$\theta$-classifier* ($\theta \in [0, 1]$) is defined by $g_\theta(X) = 1$ iff $\eta(X) > \theta$. If the clustering $\mathcal{R}_\xi$ is a partition, then $\theta = \frac{1}{2}$ minimizes the error probability (e.g., [15]). In general, we need to find an optimal threshold $\theta_0$ for the minimal error probability. The error probability of $g_{\theta_0}$ on the training data set is a performance indicator. The similarity $\xi$ need to be adjusted if the error probability does not near the Bayes error based on conditional probability $\eta_{\Phi_m}$. This paper does not consider other criteria, such as false positive. The experiments in the paper show that the error probability of the $\theta_0$-classifier for the testing data set can be well predicted for proper similarities.

The remainder of this paper is organized as follows: Section 2 reviews tolerance space and data clustering. Section 3 introduces the probabilistic model of pattern recognition. Section 4 introduces the supervised learning and $\theta$-classifiers. Section 5 describes the experiments. Section 6 is the conclusion.

## 2 Tolerance Space and Representative Clustering

In this section, we define a similarity formally. Let $\Omega$ be an abstract set of elements.

**Definition 2.1** A *tolerance relation* $\xi$ on $\Omega$ is a binary relation with two conditions: $(x, x) \in \xi$ for any $x \in \Omega$ (reflexivity), and $(x, y) \in \xi \Rightarrow (y, x) \in \xi$ (symmetry).

The pair $(\Omega, \xi)$ is called a *tolerance space*, denoted by $\Omega^\xi$. We also use $\xi$ as a predicate: $\xi(x, y)$ iff $(x, y) \in \xi$. If $\xi(x, y)$, we say that $x$ is *$\xi$-similar* to $y$, or $x$ and $y$ are $\xi$-similar. We omit $\xi$ when there is no ambiguity. On a metric space $(M, d)$, for example, each positive real number $\varepsilon > 0$ defines a tolerance relation $d_\varepsilon$: $d_\varepsilon(x, y)$ if $d(x, y) < \varepsilon$, $x, y \in M$. Any undirected graph can be treated as a tolerance space (and vice versa), where $\Omega$ is the set of all vertices, and two vertices are similar if they are the same vertex or they are adjacent (e.g., Figure 1). Any equivalent relation $\xi$ on $\Omega$ is a tolerance relation having the transitive property.

The set inclusion $\subseteq$ is a partial ordering on tolerance relations on $\Omega$. The collection of all tolerance relations forms a lattice with a unique lower bound and a unique upper bound. The lower bound is called the *discrete tolerance relation*, which is $\{(x, x) \mid x \in \Omega\}$ (i.e., each element $x$ is similar only to itself). We call the corresponding tolerance space the *discrete tolerance space*, denoted by $\Omega^0$. The upper bound is called the *trivial tolerance relation*, which is $\Omega \times \Omega$ (i.e., any two elements are similar). The corresponding tolerance space is called the *trivial tolerance space*, denoted by $\Omega^\infty$.

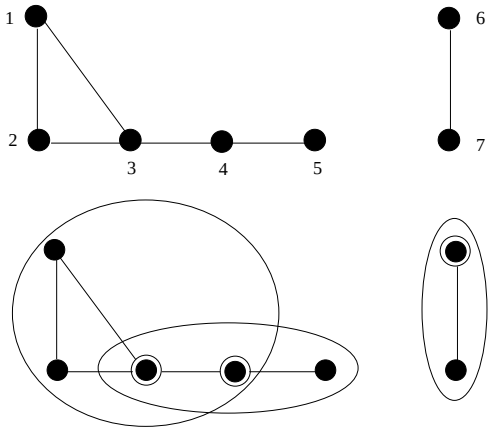


Figure 1: A tolerance space and a minimal clustering.

For each $x \in \Omega$, the set $\xi(x) = \{u \in \Omega : \xi(x, u)\}$, consisting of all elements similar to $x$, is called the *neighborhood* of $x$. Let $\mathcal{F}_\xi$ be the Borel field generated by all neighborhoods. The pair $(\Omega^\xi, \mathcal{F}_\xi)$ is a measurable space and each element of $\mathcal{F}_\xi$ is called a $\xi$-*measurable* set. For example, consider $\Omega = \mathcal{R}^n$, the $n$-dimensional Euclidean space, and the tolerance relation $\xi(x, y)$ iff $\|x - y\| < \varepsilon$ for a fixed given $\varepsilon > 0$, where $\|x - y\|$ is the Euclidean distance. Then the corresponding $\mathcal{F}_\xi$ is the collection of all Borel sets in $\mathcal{R}^n$.

For any $x \in \Omega^\xi$, the singleton $\{x\}$ is not always $\xi$-measurable (i.e., in $\mathcal{F}_\xi$). In Figure 1, for example, $\{1\}$, $\{2\}$, $\{6\}$, and $\{7\}$ are not $\xi$-measurable, and the singletons $\{3\}$, $\{4\}$, and $\{5\}$ are $\xi$-measurable. We define indistinguishable elements as follows.

**Definition 2.2.** Two elements $x$ and $u$ in $\Omega^\xi$ are *indistinguishable*, denoted by $x \sim u$, if they have the same neighborhood (i.e., $\xi(x) = \xi(u)$).

Note that $\sim$ is an equivalent relation. Let $\overline{x} = \{u \mid u \sim x\}$, the set of all indistinguishable elements of $x$, which is called the *indistinguishable set* of $x$ and satisfies $\overline{x} = \bigcap_{u \in \xi(x)} \xi(u) - \bigcup_{u \notin \xi(x)} \xi(u)$. In Figure 1, for example, $\overline{1} = \overline{2} = \{1, 2\}$ and $\overline{6} = \overline{7} = \{6, 7\}$. To avoid advanced measure theory (e.g., [10]), we assume $\overline{x}$ is $\xi$-measurable for any $x \in \Omega$ and is the smallest $\xi$-measurable set containing $x$. This assumption is always true if $\Omega$ is a countable set.

A function from a tolerance space to another tolerance space $f : \Omega^\xi \longrightarrow \Phi^\zeta$ is called *measurable* if $f^{-1}(A) \in \mathcal{F}_\xi$ for all $A \in \mathcal{F}_\zeta$. A measurable function maps indistinguishable sets into indistinguishable sets.

**Theorem 2.3.** If $x \sim y$ in $\Omega^\xi$ and the function $f : \Omega^\xi \longrightarrow \Phi^\zeta$ is measurable, then $f(x) \sim f(y)$ in $\Phi^\zeta$.

*Proof.* Let $a = f(x)$. Since $\overline{a}$ is $\zeta$-measurable and $\overline{x}$ is the smallest $\xi$-measurable set containing $x$, the inverse $f^{-1}(\overline{a})$ is $\xi$-measurable and contains $\overline{x}$. Therefore, $f(y) \in \overline{a}$ and $f(x) \sim f(y)$.

For any real-valued measurable function $f$ on $\Omega^\xi$, $f(x) = f(u)$ if $x \sim u$ (considering the Borel sets of real numbers). Especially, every real-valued measurable function on the trivial tolerance space $\Omega^\infty$ is a constant function.

Any function from a set to a tolerance space defines a tolerance relation on the domain. Suppose that $f : \Omega \longrightarrow \Phi^\zeta$ is a function from an arbitrary set $\Omega$ to a tolerance space $\Phi^\zeta$. We define a tolerance relation $\xi_\zeta$ on $\Omega$: $\xi_\zeta(x, u)$ for $x, u \in \Omega$ if $f(x)$ and $f(u)$ are $\zeta$-similar in $\Phi^\zeta$. Then the function $f : \Omega^{\xi_\zeta} \longrightarrow \Phi^\zeta$ is measurable. For example, in a data pre-processing $f : \Omega \longrightarrow \Phi$, $f(x)$ is the feature vector of $x$. Each similarity on the feature vectors defines a similarity on the original data set $\Omega$.

We review the data clustering in a tolerance space $\Omega^\xi$ [16]. Here we assume that $\Omega^\xi$ is finite. Each $x \in \Omega$ is called a *representative* of its neighborhood $\xi(x)$. A set of elements $\{r_1, ..., r_k\}$ is called a *representative system* of the tolerance space $\Omega^\xi$ if the corresponding neighborhoods cover the whole space $\Omega$. A representative system $\{r_1, ..., r_k\}$ is *minimal* if there is no other representative system with less than $k$ members. In this paper, we call such a minimal representative system a *minimal representative clustering*. Note that the clustering in this paper is not crisp in general and two clusters may have nonempty intersection. Such a property is important in the uncertainty study of AI [18] and is different from many clustering algorithms (e.g., [8]).

To search for a minimal representative clustering is intractable in general. We introduce a heuristic search [16] which uses a concept of density. The *density* function on $\Omega$ is the number of elements in $\xi(x)$: $den(x) = |\xi(x)|$. According to $den(x)$, we sort the elements of $\Omega$: $x_1, x_2, ..., x_m$, so that $den(x_1) \geq den(x_2) \geq ... \geq den(x_m)$. To search for a representative system, choose the neighborhood $\xi(x_1)$ first. Then choose the next neighborhood $\xi(x_i)$ (according to the density order) which is not covered by previously chosen neighborhoods. Repeat the process until the whole space is covered. Finally, scan the chosen neighborhoods backwards and delete the neighborhood which is contained in the union of other already collected neighborhoods. In this paper, we use such a *sub-minimal* representative system to approximate a minimal representative clustering. Figure 1 shows a minimal representative clustering computed by the heuristic method.

## 3 Bayes Classifier and Bayes Error

Pattern recognition is about guessing or predicting the unknown nature of an observation, a discrete quantity such as black or white, one or zero, sick or healthy, abnormal or normal. An observation is a collection of information about an object, such as an image, a vector of weather data, or an internet message. Note that about the object the information of an observation is not always complete and has some degree of uncertainty. The observation usually is represented by a vector of several components and each component may be numerical or categorical. Formally, we use $x$ to denote an observation and $\Omega$ to denote the space of all possible observations. The unknown nature of the observation is called a *class*. It is denoted by $y$ and takes values in a finite set. For simplicity, this paper considers only two possible classes (e.g., normal and anomaly), denoted by 0 and 1. In pattern recognition, one creates a function $g : \Omega \longrightarrow \{0, 1\}$ which represents one's guess of $y$ given $x$. The mapping $g$ is called a *classifier*. The classifier errs on $x$ when $g(x) \neq y$. In the following, we introduce a probabilistic model. First we define a probability measure on a tolerance space as follows.

**Definition 3.1** A *probability measure* on a tolerance space $\Omega^\xi$ is a probability measure $\mu$ on the measurable space $(\Omega^\xi, \mathcal{F}_\xi)$. The triple $(\Omega^\xi, \mathcal{F}_\xi, \mu)$ is a called a probability space of $\Omega^\xi$.

Given a probability space $(\Omega^\xi, \mathcal{F}_\xi, \mu)$, each measurable function $f : \Omega^\xi \longrightarrow \Phi^\zeta$ defines a probability measure $\mu_\zeta$ on $(\Phi^\zeta, \mathcal{F}_\zeta)$ as follows. For any measurable set $A \in \mathcal{F}_\zeta$, $\mu_\zeta(A) = \mu(f^{-1}(A))$. The triple $(\Phi^\zeta, \mathcal{F}_\zeta, \mu_\zeta)$ is a probability space. For example, the function $f$ is a feature selection and any probability measure on the original data is transformed to the feature vectors.

Let $\xi$ be a tolerance relation on $\Omega$ and $(\Omega^\xi, \mathcal{F}_\xi, \mu)$ a probability space. If there is no ambiguity, we omit the $\xi$ in our notation. Let $(X, Y)$ be a random pair taking their respective values from $\Omega$ and $\{0, 1\}$. The random pair may be defined by a pair $(\mu, \eta)$, where $\mu$ is the probability measure of $X$ and $\eta$ is the regression of $Y$ on $X$. That is, for any $A \in \mathcal{F}$, $P(X \in A) = \mu(A)$, and for any $x \in \Omega$, $\eta(x) = P(Y = 1|X \in \overline{x})$. Note that $\eta$ is measurable and $\eta(x) = \eta(y)$ if $x$ and $y$ are indistinguishable.

A *classifier* is a measurable function $g : \Omega \longrightarrow \{0, 1\}$. Here we treat $\{0, 1\}$ as a discrete tolerance space. The probability of error $L(g) = P(g(X) \neq Y)$ is called the *error* of $g$, which is the integration of a conditional probability:

$$L(g) = P(g(X) \neq Y) = \int_\Omega P(g(X) \neq Y|X \in \overline{x}) d\mu(x)$$

**Definition 3.2** The *Bayes classifier* is defined as follows

$$g^*(x) = \begin{cases} 1 \text{ if } \eta(x) > 1/2, \\ 0 \text{ otherwise.} \end{cases}$$

The error $L(g^*)$ is called the *Bayes Error*. Similarly to the classical Bayes rule [3], we can prove

**Theorem 3.3** The Bayes classifier minimizes the error; that is, for any classifier $g$, $L(g^*) \leq L(g)$.

The Bayes error $L(g^*)$ is the theoretical limitation of the performance of any designed classifier $g$. Note that $L(g*) = 0$ if $\eta(x) \in \{0, 1\}$ for all $x \in \Omega$. The goal of this paper is to construct a classifier to approximate the Bayes error as near as possible.

## 4 Optimal $\theta$-Classifier

**Training Data Set** The Bayes classifier $g^*$ depends upon the tolerance space $\Omega^\xi$ and the distribution of $(X, Y)$. In most cases, both the tolerance relation $\xi$ and the distribution are not given so that both Bayes classifier and Bayes error are unknown. To design a classifier is usually based on a given database of pairs $(X_i, Y_i)$, $1 \leq i \leq n$. The database may be the result of experimental observations. It could also be obtained through an expert or teacher who filled in the $Y_i$'s after having seen the $X_i$'s. To find a classifier with a small error is hopeless unless there is some assurance that the $(X_i, Y_i)$'s jointly are some representatives of the unknown distribution. Here, we assume the data (i.e., $(X_1, Y_1), ..., (X_n, Y_n)$) is a sequence of independent identically distributed (i.i.d.) random pairs with the same distribution as that of $(X, Y)$. To construct a classifier on the basis of $X_1, Y_1, ..., X_n, Y_n$ is called *learning*, *supervised learning*, or *learning with a teacher*. The given database is called the *training data set*. In this section, we introduce a process of constructing such a classifier using similarities on the training data set.

Each $X_i$ in the training data set usually consists of several components, each of which may be numerical or categorical. We also assume that the class of $X_i$ is binary; that is, $Y_i \in \{0, 1\}$. Let $\Omega_n$ be the set of all $X_i$'s: $\Omega_n = \{X_i | 1 \leq i \leq n\}$. Note that $|\Omega_n| \leq n$ because it is possible that $X_i = X_j$ for different $i$ and $j$. For each $t \in \Omega_n$, consider two counts: $f_0(t) = |\{i \mid 1 \leq i \leq n, X_i = t, Y_i = 0\}|$ and $f_1(t) = |\{i \mid 1 \leq i \leq n, X_i = t, Y_i = 1\}|$. Note that $\sum_{t \in \Omega_n} (f_0(t) + f_1(t)) = n$. Then on the training data set (as a discrete space), the conditional probability of $Y = 1$, given $X = t$, is $\eta_n(t) = \frac{f_1(t)}{f_0(t)+f_1(t)}$. Let the corresponding Bayes classifier and Bayes error (on $\Omega_n$) be denoted by $g_n^*$ and $L_n(g_n^*)$, respectively. Note that $L_n(g_n^*) = 0$ if $X_i \neq X_j$ for any $i \neq j$. The Bayes error $L_n(g_n^*)$ is the theoretical limitation of the learning. If the Bayes error is too large, the performance of the learning result will be poor. Usually, larger data set or more information for each $X$ is needed.

**Data Pre-processing** Before designing a classification scheme, a data pre-processing is commonly applied to the training data set, such as data cleaning, data integration, data granulation, feature extraction, data transformation, and data reduction (e..g, [8, 9]). Formally, we use a function to represent the data preprocessing, $\psi : \Omega_n \longrightarrow \Phi$. For convenience, we call each sample $X$ a *record* and $\psi(X)$ the *feature vector* of $X$. All possible feature vectors form the set $\Phi$. Usually the function $\psi$ is not one-to-one. For the frequency of $Y_i$, we store two integers for each feature vector $x$: $n_0(x) = |\{i \mid 1 \leq i \leq n, \psi(X_i) = x, Y_i = 0\}|$ and $n_1(x) = |\{i \mid 1 \leq i \leq n, \psi(X_i) = x, Y_i = 1\}|$. That is, $n_0(x)$ and $n_1(x)$ are the numbers of records of class 0 and class 1 in $\psi^{-1}(x) \subseteq \Omega_n$, respectively. Let the image of $\Omega_n$ be denoted by $\Phi_m = \psi(\Omega_n)$ $(\subset \Phi)$. Usually, the size of $\Phi_m$ is much smaller than that of $\Omega_n$ for the computational purpose. Note that $\sum_{x \in \Phi_m} (n_0(x) + n_1(x)) = n$. Consider $\Phi_m$ as a discrete space. For each $x \in \Phi_m$, the conditional probability of $Y = 1$ on the training data set, given $\psi(X) = x$, is $\eta_{\Phi_m}(x) = \frac{n_1(x)}{n_0(x)+n_1(x)}$. Let the corresponding Bayes classifier and Bayes error (on $\Omega_n$) be denoted by $g_{\Phi_m}^*$ and $L_n(g_{\Phi_m}^*)$, respectively. The feature vector $\psi(X)$ may not contain all information of $X$ and it is always true that $L_n(g_n^*) \leq L_n(g_{\Phi_m}^*)$. The data

pre-processing simplifies the computation, still the Bayes error $L_n(g^*_{\Phi_m})$ should be near to $L_n(g^*_n)$. If the difference between the two errors is too large, the pre-processing $\psi$ should be re-designed.

**Representative Classification Scheme** On $\Phi$ we define a suitable tolerance relation (i.e., similarity), denoted by $\xi$. To define similarity depends on each individual task and there are many methods to measure similarities or dissimilarities (e.g., [9]). Here we simply assume $\xi$ is given. Consider representative clustering on the tolerance space $(\Phi_m)^\xi$ and let $\mathcal{R}_\xi = \{R_1, R_2, ..., R_k\}$ be a sub-minimal representative system computed by the algorithm introduced in Section 2.

Deciding whether an event is surprising is one of the tasks in statistics [6, 7]. Here we also define a concept of surprising records with respective to the tolerance space $(\Phi_m)^\xi$. The tolerance relation $\xi$ is first transformed to the training data set. Two records $X_i$ and $X_j$ in the training data set are similar if the corresponding feature vectors $\psi(X_i)$ and $\psi(X_i)$ are similar. A record $t \in \Omega_n$ is called a *surprise* if $t$ is not similar to any other record in $\Omega_n$ and there is only one $i$ for which $X_i = t$. Let $x = \psi(t)$. Then $t$ is a surprise iff $\psi^{-1}(x) = \{t\}$, the neighborhood $\xi(x) = \{x\}$ is a singleton in the tolerance space $(\Phi_m)^\xi$, and $n_0(x) + n_1(x) = 1$. Therefore, the feature vector of a surprise is always in $\mathcal{R}_\xi$.

Let the numbers of the surprises of class 0 and class 1 be denoted by $\alpha_\xi$ and $\beta_\xi$, respectively, which are independent of the choice of $\mathcal{R}_\xi$. Finally, we use $\mathcal{L} = (\Phi_m,\ \mathcal{R}_\xi,\ \alpha_\xi,\ \beta_\xi)$ to represent the result computed above, which is called a *representative classification scheme*.

**$\theta$-Classifiers** Consider the random pair $(X, Y)$. Given the value of $X$, we design classifiers of $X$ based on a representative classification scheme $\mathcal{L} = (\Phi_m,\ \mathcal{R}_\xi,\ \alpha_\xi,\ \beta_\xi)$ as follows. We find the representatives (in $\mathcal{R}_\xi$) similar to the feature vector $\psi(X)$, denoted by $N_X = \{R'_1, ..., R'_{k'}\}$. First consider the case that $N_X$ is not empty. Let the intersection of the clusters of these representatives be $F_X = \bigcap_{i=1}^{k'} \xi(R'_i)$, which is the set of all feature vectors (in $\Phi_m$) similar to all representatives in $N_X$. If $F_X = \emptyset$, then it is replaced by the union of the clusters: $F_X = \bigcup_{i=1}^{k'} \xi(R'_i)$. The set $F_X$ is a piece of information about $X$ derived from $\mathcal{L}$, from which we compute posterior probability of $Y$.

In the subset $\psi^{-1}(F_X)$ of $\Omega_n$, the number of records of class 0 is $t_0 = \sum_{x \in F_X} n_0(x)$ and the number of records of class 1 is $t_1 = \sum_{x \in F_X} n_1(x)$. On the training data set, the conditional probability of $Y = 1$, given $\psi(X) \in F_X$, is $\frac{t_1}{t_0+t_1}$. For any random sample $(X, Y)$, we use this value to predict the conditional probability, which is denoted by $\eta_\mathcal{L}(X) = P(Y = 1|\ \mathcal{L}) = \frac{t_1}{t_0+t_1}$. Next if $N_X = \emptyset$, that is, $\psi(X)$ is not similar to any of the representatives in $\mathcal{R}_\xi$, then we say that the record $X$ is an *unknown surprise* (w.r.t. $\mathcal{L}$). If $\alpha_\xi + \beta_\xi > 0$, then the conditional probability is estimated as $\eta_\mathcal{L}(X) = P(Y = 1|\ \mathcal{L}) = \frac{\beta_\xi}{\alpha_\xi+\beta_\xi}$. If $\alpha_\xi + \beta_\xi = 0$, then we say that the conditional probability of unknown surprises is not predictable, that is, $\eta_\mathcal{L}(X)$ is undefined. Note that all unknown surprises have the same predicted conditional probability and will be classified to one class; therefore, the similarity should be modified if there are too many surprises in the training data set.

In the following we introduce a simple non-randomized classifier using a threshold, without considering any loss function. Let $\theta$, $0 < \theta < 1$, be a real number, then we define the following classifier:

**Definition 4.1** The *$\theta$-classifier* is defined as follows

$$g_\theta(X) = \begin{cases} 1 & \text{if } \eta_\mathcal{L}(X) > \theta, \\ 0 & \text{if } \eta_\mathcal{L}(X) \le \theta, \\ \text{unknown} & \text{if } \eta_\mathcal{L}(X) \text{ is undefined.} \end{cases}$$

The error of $g_\theta$ on the training data set can by computed, denoted by $L_\mathcal{L}(\theta) = L_n(g_\theta)$, which satisfies $L_n(g^*_n) \le L_n(g^*_{\Phi_m}) \le L_\mathcal{L}(\theta)$. The error $L_\mathcal{L}(\theta)$ is an indicator of the performance of the classifier $g_\theta$. If the representative clustering $\mathcal{R}_\xi$ is a partition of $\Phi_m$, the error $L_\mathcal{L}(\theta)$ reaches its minimum at $\theta = \frac{1}{2}$. In the special case that $\xi$ is the discrete tolerance relation (i.e., each element in $\Phi_m$ is a representative in $\mathcal{R}_\xi$), this error is the Bayes error: $L_n(g^*_{\Phi_m}) = L_\mathcal{L}(\frac{1}{2})$. In the other extreme case that $\xi$ is the trivial tolerance relation (i.e., there is only one cluster in $\mathcal{R}_\xi$), the information derived from $\mathcal{L}$ is the numbers of records of class 1 and class 0 in the training data set. Therefore, the classifier will classifies all records to class 1 or 0, depending on the fact that the record number of class 1 in the training data set is larger or not. In general, $\mathcal{R}_\xi$ is not a partition and $g_{\frac{1}{2}}$ does not have the minimal error.

**Definition 4.2** The $g_{\theta_0}$ is called an *optimal $\theta$-classifier* if

$$L_\mathcal{L}(\theta_0) \le L_\mathcal{L}(\theta), \quad 0 < \theta < 1.$$

Since $\Phi_m$ is finite, an optimal $\theta$-classifier can be computed. If the error $L_\mathcal{L}(\theta_0)$ (on the training data set) is too far away from the Bayes error $L_n(g^*_{\Phi_m})$, new similarities should be considered. This paper considers only the probability of error and does not consider other criteria, such as false positive. The experiments in the next section show that the error of the $\theta$-classifier on testing data set can be well predicted for proper similarities.

## 5 Experiments

In our experiments, we use the KDD-99 cup data set [11]. There are 212971 records in total. There are 41 attributes, categorical (e.g., type of protocol) or continuous (e.g., length of the connection). Each record is la-

beled with normal or certain attack. The 42nd attribute is the label, of which there are 23 different types: one is normal and others are attacks. All attacks are classified as anomalies in the experiments. We use 0 to represent a normal and 1 to represent an anomaly. Our task is to classify records based on the 41 attributes into two classes: anomaly or normal.

**Pre-Processing and Training Data Set** Not all 41 features are significant in the classification task. The traditional statistical test of the equality of two distributions is used in this process. We select an attribute to the feature vector if the distributions of the value domains are significantly different for the two types of records. We have selected 9 attributes. Each categorical value is labeled by an integer. The values of continuous attributes are granulated into a linearly ordered discrete values according to their distribution. Each record $t$ is mapped into a feature vector $\psi(t)$. By a few feature vectors, 85.08% (139213 records in total) of total records are classified perfectly. Those records are not used in the experiments in order to make the error more visible.

From the rest of the records, we choose randomly about half of normals and half of anomalies as the training data set, and the rest as the testing data set. Let the training data set be denoted by $D_{train}$ and the testing data set by $D_{test}$. The $D_{test}$ consists of 29603 normals and 7310 anomalies, 36913 in total. Let $\psi(D_{train})$ be denoted by $F_{train}$. The numbers of records in $D_{train}$ and $F_{train}$ and the corresponding Bayes errors are shown in Table 1. Note that the pre-processing reduces the number of the records from 36845 to 1160 (3.1% of 36845). It also reduces the number of features from 41 to 9. The Bayes error is raised from 0 to 0.0091.

| | |
|---|---|
| normal | 29550 |
| anomaly | 7295 |
| total | 36845 |
| Bayes error on $D_{train}$ | 0 |
| feature vectors | 1160 |
| Bayes error on $F_{train}$ | 0.0091 |

**Table 1: Information about the training data set.**

**Tolerance Relations and $\theta$-classifiers** On features vectors we define a generalized Hamming distance, denote by $d$. For each $\varepsilon > 0$, we define a tolerance relation on feature vectors: $d_\varepsilon(f, g)$ if $d(f, g) \leq \varepsilon$.

Consider the training process on the tolerance space $(F_{train})^{d_\varepsilon}$. For a representative classification scheme $\mathcal{L}_\varepsilon = (F_{train},\ \mathcal{R}_{d_\varepsilon},\ \alpha_{d_\varepsilon},\ \beta_{d_\varepsilon})$, we study the $\theta$-classifier $g_\theta$. In this paper, we introduce some results for different $\varepsilon$'s: 0.01, 0.5, 2.0, and 9.0. For $\varepsilon = 0.01$, $(F_{train})^{d_{0.01}}$ is the discrete space, in which all 1160 feature vectors are representatives and the clustering is a partition. Therefore, the Bayes classifier (on $D_{train}$) is an optimal $\theta$-classifier (i.e., $\theta_0 = 0.5$). For other two cases, the clusterings are not partitions. The cluster numbers and computations are significantly reduced, but errors increase. The experiment result is summarized in Table 2, which shows the trade off between the computations and errors. Furthermore, the error functions $L_{\mathcal{L}_\varepsilon}(g_\theta)$ are depicted in Figure 2 for $\varepsilon = 0.5$. The top one is for $D_{train}$ and the bottom one is for $D_{test}$.

If $\varepsilon$ is sufficiently large (e.g. $\varepsilon \geq 9.0$), then all records are indistinguishable. That is, the only information used in this case about $D_{train}$ is that there are 29550 normals and 7295 anomalies and $\eta_{\mathcal{L}}(X) = 0.1959$ for any $X$. Therefore, each record will be classified as a normal. The error of this classifier is 0.1980, but no anomalies are classified at all.

| $\varepsilon$ | $\lvert\mathcal{R}_{d_\varepsilon}\rvert$ | $\alpha$ | $\beta$ | $\theta_0$ | train-error |
|---|---|---|---|---|---|
| 0.01 | 1160 | 236 | 92 | 0.50 | 0.0091 |
| 0.5 | 240 | 12 | 17 | 0.11 | 0.0202 |
| 2.0 | 38 | 0 | 0 | 0.60 | 0.1012 |
| 9.0 | 1 | 0 | 0 | 0.50 | 0.1980 |

| $\varepsilon$ | unknown surprises | test-error |
|---|---|---|
| 0.01 | 339 | 0.0132 |
| 0.5 | 121 | 0.0236 |
| 2.0 | 19 | 0.1048 |
| 9.0 | 0 | 0.1980 |

**Table 2: Information about some classification models and optimal $\theta$-classifiers.**

## 6 Conclusion

This paper introduces a probabilistic classification model on a tolerance space. Similarities are formally defined as reflexive and symmetric binary relations. Probability measures are constructed on the measure space generated by neighborhoods in the space. Training data set is a database of pairs $(X_i, Y_i)$, where each $X_i$ is a record and $Y_i$ is the label (0 or 1) of the record. The data preprocessing maps the records in the training data set to feature vectors.

The learning process includes search for similarities of feature vectors. For each similarity, the learning result is a representative clustering of the tolerance space of feature vectors. The information about a given record is the set of representatives similar to the feature vector of the record. In the intersection of the corresponding clusters, the percentage of records of class 1 is used to predict the posterior probability of the given record. The $\theta$-classifier use $\theta$ as a threshold and classifies the record

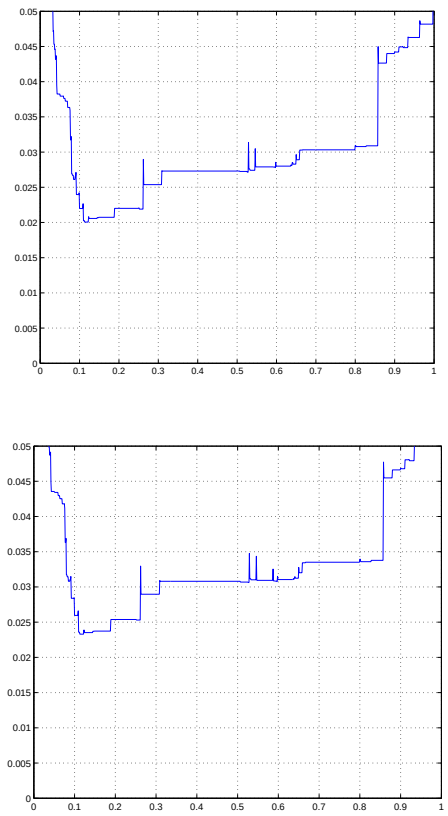

Figure 2: Error of $\theta$-classifiers for $\varepsilon = 0.5$.

to class 1 if the posterior probability is larger than $\theta$. An optimal threshold is not necessarily $\frac{1}{2}$ in general.

A concept of surprising record is introduced. The model assigns the same posterior probability to all unknown surprises. Therefore, the learning process should avoid similarities with too many surprises. The experiments demonstrate the trade-off between computations and classification errors.

Future works include more experiments and more investigation on the mathematical theory of the model, such as the handling of surprises and other criteria of classifiers.

## References

[1] Berger, J. O., *Statistical Decision Theory and Baysian Analysis,* 2nd Edition, Springer-Verlag, 1985.

[2] Brun, M., Sima, C., Hua, J., Lowey, J., Carroll, E. S., Shu, E., Dougherty, E. R., "Model-based evaluation of clustering validation measures," *Pattern Recognition* 40, pp. 807-824, 2007.

[3] Devroye, L, and Györffi, G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Springer, 1996.

[4] Dougherty, E. R., and Brun, M., "A probabilistic theory of clustering," *Pattern Recognition* 37, pp. 917-925, 2004.

[5] Dunham, M. H., *Data Mining, Introductory and Advanced Topics*, Prentice Hall, 2003.

[6] Good, I. J., *THE ESTIMATION OF PROBABILITIES, An Essay on Modern Bayesian Methods*, Research Monograph No.30, MIT Press, Cambridge, Mass., 1965.

[7] Good, I. J., *The Foundations of Probability and Its Applications* University of Minnesota Press, Minneapolis, 1983.

[8] Han, J. and Kamber, M., *Data Mining Concepts and Techniques*, 2nd Edition, Morgan Kaufmann, 2006.

[9] Hastie, T, Tibshirami, R. and Friedman, J., *The Elements of Statistical Learning*, Springer, 2001.

[10] Jacobs, K., *Measure and Integral*, Academic Press, 1978.

[11] KDD-99 Cup Dataset
http://kdd.ics.uci.edu/databases/kddcup99.html

[12] Maak, W., *Fastperiodische Funktionen*, Springer-Verlag, 1967.

[13] Schroeder, M. J. and Wright, M. H., "Tolerance and Weak Tolerance Relations," *Journal of Combinatorial Mathematics and Combinatorial Computing* 11, pp. 123-160, 1992.

[14] Sun, F.-S. and Tzeng C.-H., "Classification and Anomaly Detection in Tolerance Space," *Proc. of the Ninth IASTED International Conference on Intelligent Systems and Control*, pp. 206 - 211, Honolulu, Hawaii, 2006.

[15] Tzeng, C.-H. *A Theory of Heuristic Information in Game-Tree Search* Springer-Verlag, 1988.

[16] Tzeng, C.-H. and Sun, F.-S., "Data Clustering in Tolerance Space," in Berthold, Lenz, Bradley, Kruse, and Borgelt, editors, *Advances in Intelligent Data Analysis V*, Springer-Verlag pp. 297-306, 2003.

[17] Tzeng, C.-H. and Tzeng, C.-S. O., "Tolerance spaces and almost periodic functions," *Bull. Inst. Math. Acad. Sinica* 6, pp.159-173, 1978.

[18] Zadeh, L., "A New Direction in AI Toward a Computational Theory of Perceptions," *AI Magazine* Spring 2001, pp. 73-84.

[19] Zeeman, E. C., "The Topology of the Brain and Visual Perception," *Topology of 3-Manifolds and related Topics*, Prentice-Hall, pp. 240-256, 1962.