International Journal of
Electronic Commerce Studies
Vol.1, No.2 , pp.117-138, 2010

A PATTERN SEARCH IN DATA ANALYSIS

Chun-Hung Tzeng,
Ball State University, Indiana, USA
tzeng@bsu.edu

Fu-Shing Sun,
Ball State University, Indiana, USA
fsun@bsu.edu

ABSTRACT

This paper introduces a probabilistic model of two-class pattern recognition. The measurable sets are defined by a similarity, which is a reflexive and symmetric binary relation. The heuristic information model is formulated by a type of data clustering called representative clustering. The heuristic information about a data record is a data subset containing the record, which is computed by comparing the record with all representative records. For the corresponding classifiers, both Bayes and Neyman-Pearson Theorems are proved in this paper. In application, the knowledge discovering process searches for similarity and representative clustering in a training data set. The evaluation is extended to records in a testing data set. The experiment shows the trade-off between the number of representatives and classifier performance.

Keyword: Pattern-recognition, Similarity, Representative, Heuristic-information

1. INTRODUCTION

Pattern recognition involves guessing or predicting the unknown nature of an observation. In a classical formulation[1,2], 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 $C = \{0, 1, 2, ..., M\}$. The task is to create a *classifier*, which is a function $g : \Re^d \to C$. The value $g(x)$ represents the guess of y, given x. A probabilistic setting [1] considers a random pair (X, Y) on $\Re^d \times C$, of which a distribution describes the frequency of encountering particular pairs in practice. The *error rate* of g is $L(g) = P(g(X) \neq Y)$.

In the two-class problem $C = \{0, 1\}$, there are other types of errors[3]: the *false positive rate* $L^{(0)}(g)$ and the *false negative rate* $L^{(1)}(g)$. The basic information about a given $X = x$ is the posterior probability: $\eta(x) = P(Y = 1 \mid X = x)$. For a real $\theta (0 \le \theta < 1)$, the classifier g_θ is defined as $g_\theta(x) = 1$ iff $\eta(x) > \theta$. In classical theory, Bayes Theorem shows that the *Bayes classifier* $g_{0.5}$ has the minimal error rate. Neyman-Pearson Theorem shows that g_θ for some θ minimizes the false negative rate if the false positive rate is required to be kept under a certain level.

In most cases, the distribution of (X, Y) is unknown. The design of a classifier is based on a training data set $\{(X_i, Y_i) \mid 1 \le i \le n\}$. Many classification rules have been proposed[1,2]. For example, the *k*-nearest neighbor rule takes a majority vote over the Y_i's in the subset of *k* pairs (X_i, Y_i) that have the smallest values $\| X_i - x \|$.

To classify a given observation $X = x$, a classifier usually searches first for information about *x* from certain knowledge discovered from the training data set and then makes the decision. The searched information is often incomplete. This paper calls such information *heuristic information*. A formal probabilistic formulation of heuristic information has been introduced[4], which is represented by a partition $\{P_1, P_2, ...\}$ of the random space *X*. For an observation $X = x$, the search computes the particular P_k for which $x \in P_k$. Then, the decision is based on the posterior probability $P(Y = 1 \mid X \in P_k)$. In the rough set terminology[5, 6, 7], the set of the members of the same class (1 or 0) is a rough set, and the partition is an approximation. The boundary consists of the partition members with the posterior probability strictly in the open interval $(0,1)$. That is, the members consist of both classes of records.

In data mining, similarities play a central role. For example, the *k*-nearest neighbor rule uses Euclidean distance to measure similarities. Most similarity measures are reflexive and symmetric. Using the two properties as a formal similarity, this paper introduces a probabilistic pattern recognition model. The first step of knowledge discovery is to search for a similarity, so that similar X_i's are likely of the same class. The second step is to compute a data cluster based on the similarity. There are two goals in this process: each cluster should contain as many X_i's as possible, and a

high percentage of X_i's in each cluster should belong to one class. That is, each cluster should exhibit a typical pattern of X_i's.

This paper develops a mathematical theory of this approach for a general data set, not restricted to the real $\Re^d$. The similarity is called a tolerance relation[8]. A space Ω with a tolerance relation ξ is called a *t*olerance space, denoted by Ω^{ξ}. The neighborhood of an element x is the set of all elements similar to x, denoted by $\xi(x) = \{u \in \Omega \mid \xi(x,u)\}$. All neighborhoods generate the Borel field $\mathcal{F}_{\xi}$. The probabilistic pattern recognition is based on the measurable space (Ω^{ξ}, $\mathcal{F}_{\xi}$). The probability is represented by a random pair (X,Y), $X \in \Omega$ and $Y \in \{0, 1\}$. A classifier is a measurable function $g : \Omega \to \{0, 1\}$. About a given $X = x$, the full information is the posterior probability $\eta(x) = P(Y = 1 \mid X \in \overline{x})$, where $\overline{x}$ is the minimal measurable set containing x. Both Bayes and Neyman-Pearson Theorems are true in this abstract model.

Heuristic information is derived from a representative clustering $\boldsymbol{R}_{\xi} = \{r_1, r_2, \ldots, r_k\}$, of which the neighborhoods $\xi(r_i)$'s cover the whole space Ω. These neighborhoods do not form a partition in general and generate a Borel subfield, which is the heuristic information model in this theory. About a given value $X = x$, the heuristic information in this subfield is the smallest measurable set A_i containing x, which is computed by comparing x with the representatives. The heuristic decision is then based on the posterior probability $P(Y = 1 \mid X \in A_i)$.

In application, the similarity and representative clustering are searched in a *training data set* $\{(X_i, Y_i)\}$. To measure the choice of similarity, this paper computes the number of surprising X_i's and the conditional probability of two X_i's being of the same class, given that they are similar. There should be as few surprising records as possible, and the conditional probability should be as large as possible. To measure performance, this paper uses error rates and ROC (i.e., Receiver operating characteristic) analysis. A good prediction is expected if the similarity ξ is suitable and the training data set contains all typical patterns. The experiment demonstrates the trade-off between computations and classifier performance.

The rest of this paper is organized as follows. Section 2 introduces the mathematical model of the pattern recognition on a tolerance space. Section

3 introduces the application in supervised learning. Section 4 describes the experiment. Section 5 provides the conclusion.

2. MATHEMATICAL THEORY

2.1 Tolerance space

Let Ω be an abstract finite set. A binary relation ξ on Ω is a subset of the Cartesian product $\xi \subseteq \Omega \times \Omega$.

Definition 2.1 A *tolerance relation* ξ on Ω is a reflexive and symmetric binary relation; that is, $(x, x) \in \xi$ for any $\xi \in \Omega$, and $(x, y) \in \xi \Rightarrow (y, x) \in \xi$.

The set Ω with a tolerance relation ξ is called a *tolerance space*, denoted by Ω^{ξ}. We also use ξ as a predicate: $\xi(x, y)$ iff $(x, y) \in \xi$. If $\xi(x, y)$, we say that x is ξ-*similar* to y, or x and y are ξ-similar. We may omit ξ when there is no ambiguity. Any undirected graph is a tolerance space (and vice versa), where Ω is the set of all vertices, and two vertices are similar if they are the same vertex or if they are adjacent (e.g., Figure 1).

The minimal tolerance relation is the *discrete tolerance relation* $\{(x, x) \mid x \in \Omega\}$, where each x is similar only to itself. The corresponding tolerance space is called the *discrete tolerance space*, denoted by Ω^0. The maximal one is $\Omega \times \Omega$ and is called the *trivial tolerance relation*, where each x is similar to all members in Ω. The corresponding tolerance space is the *trivial tolerance space*, denoted by Ω^{∞}.

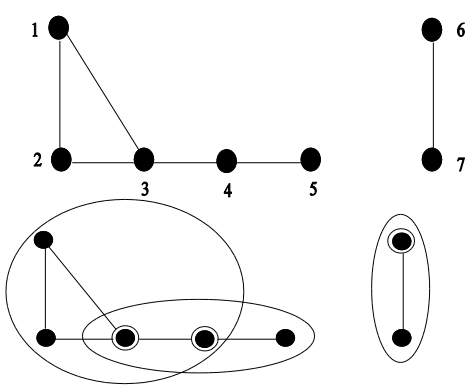

Figure 1. A tolerance space and a minimal representative clustering.

For each $x \in \Omega^{\xi}$, the set $\xi(x) = \{u \in \Omega \mid \xi(x,u)\}$ is called the *neighborhood* of x. Let $\mathcal{F}_{\xi}$ be the Borel field generated by all neighborhoods. The pair $(\Omega, \mathcal{F}_{\xi})$ is a measurable space, and each element of $\mathcal{F}_{\xi}$ is called a *ξ-measurable* set. For example, consider $\Omega = \Re^n$ (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 between x and y. Then the corresponding $\mathcal{F}_{\xi}$ is the collection of all Borel sets in $\Re^n$.

For any $x \in \Omega^{\xi}$, the singleton $\{x\}$ is not always ξ-measurable (i.e., in $\mathcal{F}_{\xi}$). In Figure 1, for example, the singletons {3}, {4}, and {5} are ξ-measurable, but {1}, {2}, {6}, and {7} are not ξ-measurable. Two elements x and u in Ω^{ξ} 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 $\overline{6} = \overline{7} = \{6, 7\}$. The set $\overline{x}$ is the smallest ξ-measurable set containing x.

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

Theorem 2.1 If $x \sim y$ in Ω^{ξ} and the function $f : \Omega^{\xi} \to \Phi^{\zeta}$ is measurable, then $f(x) \sim f(y)$ in Φ^{ζ}.

Proof. Let $a = f(x)$. Since $\overline{a}$ is ζ-measurable and $\overline{x}$ is the smallest ξ-measurable set containing x, the inverse $f^{-1}(\overline{a})$ is ξ-measurable and hence contains $\overline{x}$. Therefore, $y \in \overline{x} \subseteq f^{-1}(\overline{a})$, that is, $f(y) \in \overline{a}$ and $f(x) \sim f(y)$. □

A function $f : \Omega \to \Phi^{\zeta}$ from an arbitrary set Ω to a tolerance space Φ^{ζ} defines a tolerance relation ξ_{ζ} on Ω: $\xi_{\zeta}(x, u)$ for $x, u \in \Omega$ if $f(x)$ and $f(u)$ are ζ-similar in Φ^{ζ}. The function f is measurable. For example, in a data

pre-processing $f:\Omega \to \Phi$, each similarity in the feature vectors (i.e., Φ) defines a similarity in the original data set Ω.

2.2. Pattern recognition and classifiers

Pattern recognition involves about guessing or predicting the unknown nature of an observation. Formally, we use x to denote an observation and Ω to denote the space of all possible observations. This paper considers only two classes, normal and anomaly, denoted by 0 and 1. In the following, we introduce a probabilistic model of pattern recognition in a tolerance space.

A *probability measure* of a tolerance space Ω^{ξ} is a probability measure μ of the measurable space $(\Omega, \mathcal{F}_{\xi})$. Given the probability space $(\Omega, \mathcal{F}_{\xi}, \mu)$, each measurable function $f: \Omega^{\xi} \to \Phi^{\zeta}$ defines a probability measure μ_{ζ} in $(\Phi^{\zeta}, \mathcal{F}_{\zeta})$ as follows. For any measurable set $A \in \mathcal{F}_{\zeta}$, $\mu_{\zeta}(A) = \mu(f^{-1}(A))$.

Let $(\Omega, \mathcal{F}_{\xi}, \mu)$ be a probability space. Let (X, Y) be a random pair taking their respective values from Ω and $\{0, 1\}$. The random pair is defined by a pair (μ, η), where μ is the probability measure of X and η is the posterior probability, given the value of X. That is, for any $A \in \mathcal{F}_{\xi}$, $P(X \in A) = \mu(A)$, and for any $x \in \Omega$, $\eta(x) = P(Y = 1 \mid X \in \overline{x})$. Note that η is measurable (relative to $\mathcal{F}_{\xi}$), and $\eta(x) = \eta(y)$ if x and y are indistinguishable. Formally, we have the following definition:

Definition 2.2 The quadruplet $(\Omega, \mathcal{F}_{\xi}, \mu, \eta)$ is called a *pattern recognition* in the tolerance space Ω^{ξ}.

In pattern recognition, one creates a classifier g: $\Omega \to \{0, 1\}$ to represent one's guess of y, the class of the given observation x. The classifier errs on x when $g(x) \neq y$. In this paper, we only consider measurable classifiers, which always assign indistinguishable observations (i.e., elements) to the same class. In the following, we omit the similarity notation ξ.

Definition 2.3 A *classifier* in the tolerance space Ω is an $\mathcal{F}$-measurable function $g:\Omega \to \{0,1\}$, where $\{0, 1\}$ is treated as the discrete tolerance space.

For a *classifier* g, the *error rate* $L(g)$ is the probability of error:

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

The *false positive rate* $L^{(0)}(g)$ and the *false negative rate* $L^{(1)}(g)$ are

$$L^{(0)}(g) = P(g(X) = 1 \mid Y = 0),$$
$$L^{(1)}(g) = P(g(X) = 0 \mid Y = 1).$$

Their relation is

$$L(g) = L^{(0)}(g)P(Y = 0) + L^{(1)}(g)P(Y = 1).$$

Definition 2.4 Let $0 \leq \theta < 1$. The classifier g_θ is defined as follows: $g_\theta(x) = 0$ if $\eta(x) > \theta$; otherwise, $g_\theta(x) = 0$.

Note that g_θ is $\mathcal{F}$-measurable and that if $\eta(x) \in \{0, 1\}$ (i.e., any indistinguishable elements are of the same class), then $L(g_\theta) = L^{(0)}(g_\theta) = L^{(1)}(g_\theta) = 0$ for $0 \leq \theta < 1$. The classifier $g_{0.5}$ is the *Bayes classifier,* and $L(g_{0.5})$ is the *Bayes Error*. Similar to classical theory [1], we present following two theorems:

Theorem 2.2 (**Bayes**) For any classifier g, $L(g_{0.5}) \leq L(g)$.

Proof. Let $A^* = \{x \mid g_{0.5}(x) = 1\}$ and $A = \{x \mid g(x) = 1\}$. For any given $x \in \Omega$, we have the conditional probabilities:

$$P(g(X) \neq Y \mid X \in \overline{x}) = P(g(X) = 1, Y = 0 \mid X \in \overline{x}) + P(g(X) = 0, Y = 1 \mid X \in \overline{x})$$

$$= \mathbf{1}_A(x)P(Y = 0 \mid X \in \overline{x}) + (1 - \mathbf{1}_A(x))P(Y = 1 \mid X \in \overline{x})$$

$$= \mathbf{1}_A(x)(1 - \eta(x)) + (1 - \mathbf{1}_A(x))\eta(x)$$

$$= \mathbf{1}_A(x)(1 - 2\eta(x)) + \eta(x),$$

similarly

$$P(g_{0.5}(X) \neq Y \mid X \in \overline{x}) = \mathbf{1}_{A^*}(x)(1 - 2\eta(x)) + \eta(x),$$

where $\mathbf{1}_A$ and $\mathbf{1}_{A^*}$ are the indicators of the sets A and A^*, respectively. Thus,

$$P(g(X) \neq Y \mid X \in \bar{x}) - P(g_{0.5}(X) \neq Y \mid X \in \bar{x})$$

$$= (2\eta(x) - 1)(\mathbf{1}_{A^*}(x) - \mathbf{1}_A(x)) \geq 0.$$

Therefore,

$$L(g) - L(g_{0.5}) = \int_\Omega [P(g(X) \neq Y \mid X \in \bar{x}) - P(g_{0.5}(X) \neq Y \mid X \in \bar{x})] d\mu(x) \geq 0.$$

Theorem 2.3 (**Neyman-Pearson**) For any classifier g and θ ($0 \leq \theta < 1$), if $L^{(0)}(g) \leq L^{(0)}(g_\theta)$, then $L^{(1)}(g) \geq L^{(1)}(g_\theta)$.

Proof. Assume that $L^{(0)}(g) \leq L^{(0)}(g_\theta)$. That is,

$$P(g(x) = 1, Y = 0) \leq P(g_\theta(x) = 1, Y = 0).$$

Let $A = \{x \mid g(x) = 1\}$, $A_\theta = \{x \mid \eta(x) > \theta\}$, $W = A - A_\theta$, $Z = A_\theta - A$, and $Y = A \cap A_\theta$.

Note that all of these sets are $\mathcal{F}$-measurable, $\eta(x) > \theta$ on Z, and $\eta(x) \leq \theta$ on W.

We have

$$P(g(X) = 1, Y = 0) = \int_\Omega P(g(X) = 1, Y = 0 \mid X \in \bar{x}) d\mu(x),$$

and

$$P(g(X) = 1, Y = 0 \mid X \in \bar{x}) = \mathbf{1}_A(x)(1 - \eta(x)).$$

Therefore,

$$P(g(X) = 1, Y = 0) = \mu(A) - \int_A \eta(x) d\mu(x).$$

Similarly,

$$P(g_\theta(X)=1, Y=0) = \mu(A_\theta) - \int_{A_\theta} \eta(x) d\mu(x).$$

The assumption implies that

$$\mu(A) - \int_A \eta(x) d\mu(x) \le \mu(A_\theta) - \int_{A_\theta} \eta(x) d\mu(x).$$

Since $\int_W \eta(x) d\mu(x) \le \theta\mu(W)$ and $\int_Z \eta(x) d\mu(x) > \theta\mu(Z)$, we have

$\mu(W) \le \mu(Z)$.

Now show that $L^{(1)}(g) \ge L^{(1)}(g_\theta)$:

$$P(g_\theta(x)=0, Y=1) \le P(g(x)=0, Y=1).$$

Since $P(g(X)=0, Y=1 \mid X \in \overline{x}) = (1-\mathbf{1}_A(x))\eta(x)$,

$$\begin{aligned} P(g(X)=0, Y=1) &= 1 - \int_A \eta(x) d\mu(x) \\ &= 1 - \int_W \eta(x) d\mu(x) - \int_Y \eta(x) d\mu(x). \end{aligned}$$

Similarly,

$$P(g_\theta(X)=0, Y=1) = 1 - \int_Z \eta(x) d\mu(x) - \int_Y \eta(x) d\mu(x).$$

Therefore,

$$P(g(X)=0, Y=1) - P(g_\theta(X)=0, Y=1)$$

$$= \int_Z \eta(x) d\mu(x) - \int_W \eta(x) d\mu(x) \ge \theta(\mu(Z) - \mu(W)) \ge 0.$$

The Bayes classifier is an optimal classifier if we are required to minimize the error rate. If the false positive rate $L^{(0)}$ is required to be kept under a certain level, then g_θ minimizes the false negative rate $L^{(1)}$ for some θ.

2.3 Representative data clustering and heuristic information

In the pattern recognition defined above, the posterior probability η is usually unavailable or is difficult to compute[4].In this section, we introduce an estimation of the posterior probability based on a heuristic information model, which is derived by data clustering. We first introduce a formal model of the heuristic information.

We introduce data clustering in a tolerance space Ω^{ξ}.[9, 10, 11] Each $x \in \Omega$ is called a *representative* of its neighborhood $\xi(x)$. A set of elements $\{r_1,...,r_k\}$ is a *representative system* of the tolerance space Ω^{ξ} if the corresponding neighborhoods cover the whole space Ω^{ξ}. A representative system $\{r_1,...,r_k\}$ is *minimal* if there is no other representative system with fewer than k members. The concept of minimal representative system comes from Maak.[12]. Tzeng[13] applies the minimal representative system in tolerance space.

A representative system forms a clustering of the tolerance space in which the clusters are the corresponding neighborhoods. We call this a *representative clustering*. Note that the clustering is generally not a partition (e.g., Figure 1). Let $\boldsymbol{R}_{\xi} = \{r_1,...,r_k\}$ be a *minimal* representative system. The neighborhoods of r_i's generate a partition by set intersection:

$$\{A_i \mid i = 1,...,n\}.$$

The value n is normally greater than k. For example, the partition in Figure 1 is {{1, 2}, {3, 4}, {5}, {6, 7}}. For an $x \in \Omega$, the member A_i containing x is computed as follows. We compare x with the k representatives. Let S_x be the set of all representatives similar to x and $N_x = \boldsymbol{R}_{\xi} - S_x$. Then

$$A_i = \bigcap_{r \in S_x} \xi(r) - \bigcup_{r \in N_x} \xi(r).$$

Given x, the information $x \in A_i$ will be used to compute the estimation of η: $P(Y = 1 \mid X \in A_i)$. Since each A_i is $\mathcal{F}$-measurable, the Borel field $\mathcal{D}$ generated by the partition is a subfield of $\mathcal{F}$: $\mathcal{D} \subseteq \mathcal{F}$. The conditional expectation of η, $\eta_{\mathcal{D}} = E(\eta \mid \mathcal{D})$, is constant on all elements of A_i :

$$\eta_{\mathcal{D}}(x) = \frac{1}{P(A_i)} \int_{A_i} \eta(t) d\mu(t) = P(Y = 1 \mid X \in A_i).$$

We use the following definition:

Definition 2.5 Given a representative system $\boldsymbol{R}_\xi$, the Borel subfield $\mathcal{D}$ is called the *heuristic information model* relative to $\boldsymbol{R}_\xi$, and $\eta_{\mathcal{D}} = E(\eta \mid \mathcal{D})$ is the *estimation* of η based on $\mathcal{D}$.

Each representative in $\boldsymbol{R}_\xi$ is called a *typical pattern* of Ω relative to the similarity ξ. For each x, the partition member A_i containing x is the heuristic information about x. We consider classifiers based on the representative system $\boldsymbol{R}_\xi$.

Definition 2.6 A $\boldsymbol{R}_\xi$*-classifier* is a $\mathcal{D}$-measurable classifier.

Note that a $\boldsymbol{R}_\xi$*-classifier* $g^{\boldsymbol{R}_\xi}$ is a constant function on each A_i: $g^{\boldsymbol{R}_\xi}(x) = \sum_i c_i \mathbf{1}_{A_i}(x)$, where $c_i \in \{0, 1\}$ is the value of $g^{\boldsymbol{R}_\xi}$ on A_i and $\mathbf{1}_{A_i}$ is the indicator of A_i for each i. Since any $\boldsymbol{R}_\xi$-classifier $g^{\boldsymbol{R}_\xi}$ is a classifier in the tolerance space Ω, $L(g^{\boldsymbol{R}_\xi}) \geq L(g_{0.5})$. For $\boldsymbol{R}_\xi$-classifiers, we have the following definitions and theorems:

Definition 2.7 For each θ, $0 \leq \theta < 1$, the $\boldsymbol{R}_\xi$-classifier $g_\theta^{\boldsymbol{R}_\xi}(x) = 1$ iff $\eta_{\mathcal{D}}(x) > \theta$.

Theorem 2.4 The Bayes $\boldsymbol{R}_\xi$-classifier $g_{0.5}^{\boldsymbol{R}_\xi}$ minimizes the error rate of all $\boldsymbol{R}_\xi$-classifiers.

Theorem 2.5 (**Neyman-Pearson Theorem**) For any $\boldsymbol{R}_\xi$-classifier $g^{\boldsymbol{R}_\xi}$, if $L^{(0)}(g^{\boldsymbol{R}_\xi}) \leq L^{(0)}(g_\theta^{\boldsymbol{R}_\xi})$, then $L^{(1)}(g^{\boldsymbol{R}_\xi}) \geq L^{(1)}(g_\theta^{\boldsymbol{R}_\xi})$.

In application, the Bayes error $L(g_{0.5}^{\boldsymbol{R}_\xi})$ should be as close to the original Bayes error $L(g_{0.5})$ as possible. To be a typical pattern, each representative r_i should be similar to as many records as possible.

In the rough set terminology [6], the set of the members of Class 1 is a rough set, and the partition $\{A_i\}$ is an approximation. The boundary is the set of the members in which $0 < \eta_{\mathcal{D}}(x) < 1$.

3. APPLICATION: SUPERVISED LEARNING

3.1 Training Data Set

This section introduces a learning process from a given training data set $\{(X_i, Y_i) \mid 1 \le i \le n\}$. We assume that the data (i.e., (X_1, Y_1), ... , (X_n, Y_n)) comprise a sequence of independent identically distributed random pairs.

Each X_i in the training data set is called a record, which usually consists of several numerical or categorical components. The class Y_i is binary; that is, $Y_i \in \{0,1\}$. Let $\Omega = \{X_i \mid 1 \le i \le n\}$. Note that $|\Omega| \le n$, because it is possible that $X_i = X_j$ for different i and j. For each $x \in \Omega$, consider the following two counts:

$$f_0(x) = |\{i \mid 1 \le i \le n, X_i = x, Y_i = 0\}|, \text{and}$$
$$f_1(x) = |\{i \mid 1 \le i \le n, X_i = x, Y_i = 1\}|.$$

Note that $\sum_{x \in \Omega} (f_0(x) + f_1(x)) = n$. Let μ be the frequency on Ω. Then

$$\mu(x) = P(X = x) = \frac{f_0(x) + f_1(x)}{n}.$$

The conditional probability of Y=1, given $X = x$, is

$$\eta(x) = \frac{f_1(x)}{f_0(x) + f_1(x)}.$$

First, we treat Ω as a discrete tolerance space. For a real θ, $0 \le \theta < 1$, we define the classifier $g_\theta(x) = 1$ iff $\eta(x) > \theta$. The Bayes error is

$$L(g_{0.5}) = \frac{1}{n}\left(\sum_{f_0(x) < f_1(x)} f_0(x) + \sum_{f_0(x) \ge f_1(x)} f_1(x)\right).$$

Other errors can be computed similarly. Note that the error rates $L(g_\theta) = L^{(0)}(g_\theta) = L^{(1)}(g_\theta) = 0$ if $X_i \neq X_j$ for any $i \neq j$. The Bayes error $L(g_{0.5})$ is the theoretical limitation of the learning. If the Bayes error is too large, then the performance of the learning result will be poor. In this case, a larger training data set and/or more attributes of the records are needed.

3.2 Data Pre-Processing

Formally, we use a function to represent data pre-processing[2,14] $\phi : \Omega \to \Phi$. We call $\phi(X)$ the *feature vector* of X. All feature vectors form the set Φ. Usually, the function ϕ is not one-to-one. For the frequency of Y_i, we store two numbers for each feature vector $z \in \Phi$:

$$n_0(z) = |\{i \mid 1 \le i \le n, \phi(X_i) = z, Y_i = 0\}| \quad \text{and}$$
$$n_1(z) = |\{i \mid 1 \le i \le n, \phi(X_i) = z, Y_i = 1\}|.$$

That is, $n_0(z)$ and $n_1(z)$ are the numbers of records for Classes 0 and 1 in $\phi^{-1}(z) \subseteq \Omega$, respectively. Usually, the size of Φ is much smaller than that of Ω for computational purposes. Note that $\sum_{z \in \Phi} (n_0(z) + n_1(z)) = n$. For each $z \in \Phi$,

$$\mu(z) = \frac{n_0(z) + n_1(z)}{n}.$$

Let the partition $\{\phi^{-1}(z) \mid z \in \Phi\}$ of Ω generate the Borel field $\mathcal{F}$. For each $x \in \phi^{-1}(z)$, consider the conditional probability

$$\eta_{\mathcal{F}}(x) = P(Y = 1 \mid x \in \phi^{-1}(z)) = \frac{n_1(z)}{n_0(z) + n_1(z)}.$$

Then $\eta_{\mathcal{F}} = E(\eta \mid \mathcal{F})$, the conditional expectation of η, which can also be treated as a function on Φ. Based on $\eta_{\mathcal{F}}$, we consider the $\mathcal{F}$ -measurable classifiers $g_\theta^{\mathcal{F}}$ for $0 \le \theta < 1$. We have $L(g_{0.5}) \le L(g_{0.5}^{\mathcal{F}})$. The data pre-processing simplifies the computation but may increase the Bayes error. Therefore, in selecting the feature vectors $\phi : \Omega \to \Phi$, the Bayes error $L(g_{0.5}^{\mathcal{F}})$ should be as small as possible.

3.3 Representative clustering and classifiers

In the training process, a similarity ξ is searched in the feature vectors Φ. The similarity ξ also defines a similarity (also denoted by ξ) on the records in the training data set Ω. To measure the suitability of ξ, we compute the conditional probability $P(Y_i = Y_j \mid \xi(X_i, X_j))$, which is called the *similarity effect* in this paper. The effect should be as close to 1 as possible. That is, two similar vectors are likely to be of the same class.

To search for a minimal representative clustering of Φ^{ξ} is intractable in general. In the following, we use a heuristic method.[9] The goal of the heuristic search is to find a clustering with as few clusters as possible. The idea is to collect large clusters first. For this purpose, we use a density function: *den*(x) = the number of records (in Ω) of which the feature vectors are similar to x. According to the density function, we sort all feature vectors: $x_1, x_2, ..., x_m$, so that

$$den(x_1) \geq den(x_2) \geq ... \geq den(x_m).$$

To search for a sub-minimal representative clustering, choose the cluster $\xi(x_1)$ first. Then, according to the above order, choose the next cluster $\xi(x_i)$ which is not in the union of previously chosen clusters. Repeat the process until the whole space Φ is covered. Finally, scan the chosen clusters backwards and delete the ones that are contained in the union of other collected clusters. The result is a sub-minimal representative clustering; the set of representatives is called a *sub-minimal* representative system. Note that the input order has a minimal effect in this process. For example, the representative clustering in Figure 1 is computed by this heuristic method.

Let $\boldsymbol{R}_{\xi} = \{R_1, R_2, ..., R_k\}$ be a sub-minimal representative system of Φ^{ξ}. Let $\mathcal{D}$ be the heuristic information model generated by the representative system in Φ. We can also treat $\mathcal{D}$ as a Borel subfield of $\mathcal{F}$ in the training data set Ω. Then the conditional expectation of $\eta_{\mathcal{F}}$

$$\eta_{\mathcal{D}} = E(\eta_{\mathcal{F}} \mid \mathcal{D}) = E(\eta \mid \mathcal{F}).$$

For each record x in the training data set, $x \in \Omega$, we compare its feature vector $\phi(x)$ to all representatives R_i's. Let the representatives similar to the feature vector $\phi(x)$ form the set S_x, and $N_x = \boldsymbol{R}_\xi - S_x$. Then the partition component of $\mathcal{D}$ containing $\phi(x)$ is

$$F_x = \bigcap_{R \in S_x} \xi(R) - \bigcup_{R \in N_x} \xi(R)$$

The number of records of Class 0 in F_x is $t_0 = \sum_{z \in F_x} n_0(z)$ and the number of records of Class 1 is $t_1 = \sum_{z \in F_x} n_1(z)$.

Therefore,

$$\eta_{\mathcal{D}}(x) = P(Y = 1 \mid \phi(X) \in F_x) = \frac{t_1}{t_0 + t_1}.$$

Note that the computation is on the feature vectors only. Based on $\eta_{\mathcal{D}}$, the Bayes error of $g_{0.5}^{\boldsymbol{R}_\xi}$ satisfies $L(g_{0.5}^{\boldsymbol{R}_\xi}) \geq L\left(g_{0.5}^{\mathcal{F}}\right) \geq L\left(g_{0.5}\right)$. Let $g_\theta^{\boldsymbol{R}_\xi}$ be defined similarly based on $\eta_{\mathcal{D}}$. In the training process, $L(g_{0.5}^{\boldsymbol{R}_\xi})$ should be as close to the Bayes error in the feature vectors as possible.

A record $x \in \Omega$ is called a surprise[15, 16] if x is not similar to any other record and if there is only one i for which $X_i = x$. Therefore, the feature vector of a surprise is always in $\boldsymbol{R}_\xi$. Let the numbers of the surprise records of Classes 0 and 1 be denoted by α_ξ and β_ξ, respectively, which are independent of the choice of $\boldsymbol{R}_\xi$. The similarity ξ does not provide any useful information about surprises; therefore, a learning process should eliminate or reduce the number of surprises by adjusting feature vectors and similarities.

Receiver operating characteristic (ROC) analysis can also be used to study the posterior probability $\eta_{\mathcal{D}}(x)$. The area under the curve (AUC) should be as close to 1 as possible.

3.4 Testing Data Set

Let x be any given record which has the same structure as the record in the training data set. The heuristic information F_x (in $\mathcal{D}$) about x relative to $\boldsymbol{R}_\xi$ is computed similarly by comparing its feature vector $\phi(x)$ to all representatives in $\boldsymbol{R}_\xi$. We use the conditional expectation $\eta_{\mathcal{D}}$ in the training data set to approximate the posterior probability: $\eta_{\mathcal{D}}(x) = \frac{t_1}{t_0 + t_1}$. If $\phi(x)$ is not similar to any representative in $\boldsymbol{R}_\xi$, we call such a record an *unknown surprise* (w.r.t. ξ). In this case, we use the surprises in the training data set. If $\alpha_\xi + \beta_\xi > 0$, then the posterior probability is estimated as $\eta_{\mathcal{D}}(x) = \frac{\beta_\xi}{\alpha_\xi + \beta_\xi}$. If there are no surprises in the training data set (i.e., $\alpha_\xi + \beta_\xi = 0$), then $\eta_{\mathcal{D}}(x)$ is undefined. That is, the representative system $\boldsymbol{R}_\xi$ does not provide any information about the unknown surprise.

If the training data set is large enough to contain typical records and the similarity ξ is well chosen so that there are no or very few surprises, then the function $\eta_{\mathcal{D}}$ will be a good estimation of posterior probability in the testing data set, and the errors of $g_\theta^{\boldsymbol{R}_\xi}$ can be well predicted by the errors in the training data set, as shown by the experiment in the next section.

4. EXPERIMENTS

In our experiments, we use the KDD-99 cup data set,[17] which is adapted from the data set of the 1998 DARPA Intrusion Detection Evaluation Program.[18] We have loaded 494,020 records in our database with 42 attributes, both categorical and continuous. Each record is labeled as a normal or certain attack. The 42nd attribute is the label. The label has 23 different types: one is normal, and the others are attacks. All attacks are classified as anomalies in this experiment. 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.

4.1 Training Data Set and Pre-Processing

In order to make the errors more visible in the experiment, we exclude 420,262 records that can be easily classified. From the rest of the records, we randomly choose about half normals and half anomalies as the

training data set and the rest as the testing data set. Let the training data set be D_{train} and the testing data set be D_{test}. The testing data set D_{test} consists of 29,603 normals and 7,310 anomalies (36,913 total). We have selected 9 attributes, and each record t is mapped into a feature vector $\phi(t)$ with 9 components. Let the set of the feature vectors $\phi(D_{train})$ be 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 records in D_{train} from 36,845 to 1,160 (3.1% of 36,845) in $\phi(D_{train})$. It also reduces the number of features from 41 to 9. The Bayes error is raised from 0 to 0.0091.

Table 1. Information about the training data set

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

4.2 Tolerance Relations and g_θ-Classifiers

On features vectors we define a Hamming distance d, in which the absolute differences between numerical values are scaled down to the unit interval [0,1]. For each $\varepsilon > 0$, we define a tolerance relation in the set of feature vectors: $d_\varepsilon(f,g)$ if $d(f,g) \le \varepsilon$.

Consider the training process in the tolerance space $(F_{train})^{d_\varepsilon}$. Let $\boldsymbol{R}_\varepsilon$ be a corresponding representative clustering, and let α_ε and β_ε be the numbers of surprising normals and anomalies, respectively. We study the $\boldsymbol{R}_\varepsilon$-classifier $g_\theta^{\boldsymbol{R}_\varepsilon}$. The following are results for some ε's: 0.01, 0.25, 0.5, 2.0, 3.0, and 9.0. For $\varepsilon = 0.01$, $(F_{train})^{d_{0.01}}$ is the discrete space, in which all 1,160 feature vectors are representatives, and the clustering is a partition. For other cases, the clusterings are not partitions. The cluster numbers and computations are significantly reduced, but errors increase. The experimental result is summarized in Tables 2 and 3 (where B- is Bayes and u- is unknown). The result shows the trade-off between the computations and performances. The error functions L, $L^{(0)}$, and $L^{(1)}$ on

D_{train} for $\varepsilon = 0.5$ are depicted in Figure 2. The ROC curves of D_{train} are depicted in Figure 3. For D_{test}, we have very similar results.

Table 2. Experimental result for D_{train}

ε	$\lvert \boldsymbol{R}_{\varepsilon} \rvert$	α	β	effect	AUC	B-error
0.01	1160	236	92	0.9865	0.9995	0.0091
0.25	437	30	34	0.9816	0.9991	0.0122
0.5	240	12	17	0.9732	0.9983	0.0188
3.0	9	0	0	0.7887	0.9671	0.0852
4.0	4	0	0	0.7114	0.8527	0.1769
9.0	1	0	0	0.6824	0.5000	0.1980

If ε is sufficiently large (e.g. $\varepsilon \geq 9.0$), then all records are indistinguishable. That is, the only information about D_{train} is that there are 29,550 normals and 7,295 anomalies; the posterior probability $\eta(x) =$ 0.1980 for any x. Therefore, the Bayes rule classifies each record as a normal. The error rate is 0.1980, but no anomalies are classified at all; that is, the false negative rate is 1. The similarity effect can also be computed directly from the number of normals and anomalies. The AUC is 0.5, which means that this similarity is worthless.

Table 3. Experimental result for D_{test}

ε	u-surprises	effect	AUC	B-error
0.01	339	0.9851	0.9969	0.0108
0.25	238	0.9794	0.9963	0.0149
0.50	121	0.9709	0.9958	0.0213
3.00	1	0.7769	0.9640	0.0879
4.00	0	0.7110	0.8522	0.1767
9.00	0	0.6824	0.5000	0.1980

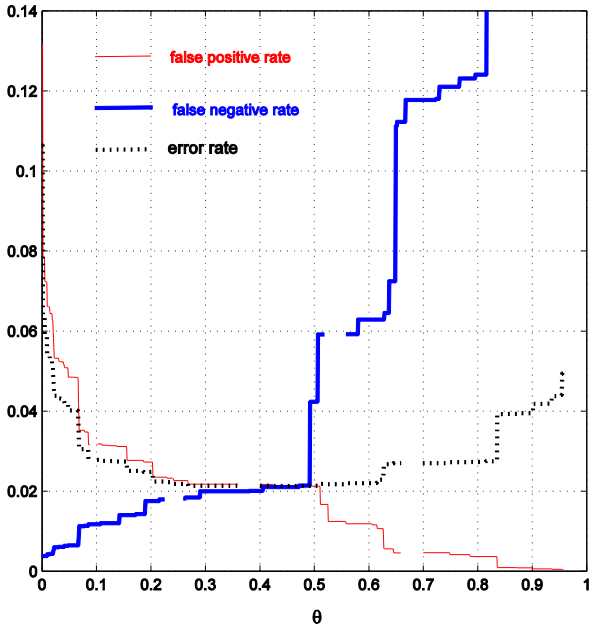

Figure 2. Error functions in D_{train} for $\varepsilon = 0.5$

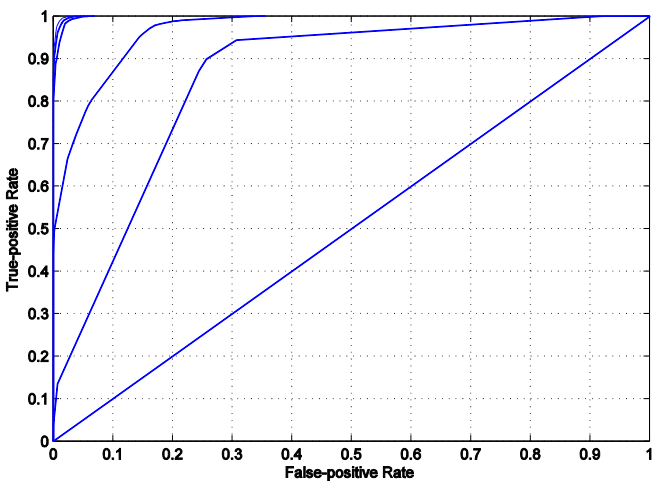

Figure 3. ROC curves in D_{train} for $\varepsilon = 0.01, 0.25, 0.5, 3, 4, 9$

5. CONCLUSION

This paper introduces a probabilistic model of pattern recognition in tolerance space. The learning process includes searching for a similarity in the feature vectors and computing a representative clustering. The partition generated by the clustering provides the heuristic information model. Based on the posterior probability, both Bayes and Neyman-Pearson Theorems are true for the training data set. For a record in the testing data set, the set of representatives similar to the feature vector of the record is used to predict the posterior probability. The experiment demonstrates the trade-off between computations and classification performances. And the prediction is not effective for surprise records. The future work includes combination of different similarities and reduction of surprise records.

6. REFERENCES

[1] L. Devroye and G. L. Gyoerffi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

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

[3] M. A. E. Maloof. *Machine Learning and Data Mining for Computer Security*. Springer-Verlag, 2006.

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

[5] S. Greco, B. Matarazzo, R. Slowinski, Rough set theory for multicriteria decision analysis, *European Journal of Operational Research,* 129(1), p 1-47, 2001, Elsevier.

[6] L. Polkowski. *Rough Sets, Mathematical Foundations*, Physica-Verlag, Heidelberg, 2002.

[7] J. Stepaniuk, Knowledge discovery by application of rough set models, in *L. Polkowski, S. Tsumoto, T. Y. Lin, editors, Rough Set Methods and Applications : new developments in knowledge discovery in information systems,* p 137-233, Physica-Verlag, Heidelberg, Germany, 2000.

[8] E. C. Zeeman. The topology of the brain and visual perception. In *Topology of 3-Manifolds and related Topics*, p 240–256. Proc. The Univ. of Georgia Institue, 1962.

[9] C.-H. Tzeng and F.-S. Sun. Data clustering in tolerance space. In *Berthold, Lenz, Bradley, Kruse, and Borgelt, editors, Advances in Intelligent Data Analysis V*, p 297–306. Springer-Verlag, 2003.

[10] C.-H. Tzeng. Similarity and pattern recognition. In *Proc. Intern. Conf. on Data Mining and Appl.* ICDMA'08, March 2008.

[11] C.-H. Tzeng. A Probabilistic Model of Pattern Recognition on Abstract Data. In *Proc. of The 2009 Intern. Conf. on Data Mining.* DMIN 2009, p 319-325.
[12] W. Maak. *Fastperiodische Funktionen*. Springer-Verlag, 1967.
[13] C.-H. Tzeng, C.-H. and C.-S. O. Tzeng, Tolerance spaces and almost periodic functions, *Bull. Inst. Math. Acad. Sinica* 6, p159-173, 1978.
[14] J. Han and M. Kamber. *Data Mining Concepts and Techniques*. 2nd Edition, Morgan Kaufmann, 2006.
[15] I. J. Good. *THE ESTIMATION OF PROBABILITIES, An Essay on Modern Bayesian Methods*. Research Monograph No.30, MIT Press, Cambridge, Mass., 1965.
[16] I. J. Good. *The Foundations of Probability and Its Applications*. University of Minnesota Press, Minneapolis, 1983.
[17] ACM KDD-Cup 1999, Computer Network Intrusion Detection, http://kdd.ics.uci.edu/databases/kddcup99/.
[18] DARPA Intrusion Detection Evaluation, http://www.ll.mit.edu/mission/communications/ist/CST/index.html.