Supervised Pattern Recognition with Heterogeneous Features

Ventzeslav Valev

Saint Louis University
College of Art and Sciences
Department of Mathematics and Computer Science
St. Louis, MO 63103, USA
valevv@slu.edu

Abstract. In this paper, we address the supervised pattern recognition problem with heterogeneous features, where the mathematical model is based on construction of thresholds. Non-Reducible Descriptors (NRDs) for fuzzy features are obtained through the use of a threshold value, which is calculated based on the distance between patterns. For solving the problem with real features the mathematical model for construction of thresholds is based on parallel feature partitioning. Boolean formulas are used to represent NRDs.

1 Introduction

We assume that a phenomenon to be studied using the available information is in the form of *patterns*. Let us denote by M, the set of all such patterns Q. M is viewed as a union of a finite number of subsets $K_1, K_2, \ldots, K_l$ which are called *classes*. We assume that classes do not intersect, but they can overlap. However, the information available pertains only to the partitioning of some subset $M' \subset M$, called the *training set*. We assume that there are m patterns in M', which are divided into l classes, and each pattern Q contains n features of the pattern described. This information is organized as a table which we call the *training table*, denoted by $T_{n,m,l}$, assuming that there are m_1 patterns in class K_1, m_2 patterns in class K_2, ..., m_l patterns in class K_l. The first m_1 rows of $T_{n,m,l}$ will correspond to patterns in K_1, the next m_2 rows will correspond to patterns in K_2, and so on.

The *supervised pattern recognition problem* is formulated as follows. Using the training set, the class membership of patterns in the training set, and the description Q, assign a pattern $Q \in M \backslash M'$ to one of the classes $K_1, \ldots, K_l$.

As a rule, all models for solving pattern recognition problem use the concepts of similarity or dissimilarity. These concepts are used in the mathematical models of learning procedures and in the decision rules. Usually, for measuring similarity or dissimilarity, a metric in the pattern space is introduced. When describing complex patterns, different types of features are used. Properties of different patterns are measured and these measurements are usually performed

A. Sanfeliu et al. (Eds.): CIARP 2004, LNCS 3287, pp. 693–700, 2004.

in different scales. Generally speaking, features can take values from the following sets: $\{0,1\}$; $\{0,1,\ldots,d-1\}$, d - integer, $d>2$; R, where R is the set of all real numbers; fuzzy interval $[0,1]$.

As a rule, all models for solving pattern recognition problem are oriented towards one type of feature. Introducing a metric in a space with different types of features produces methodological difficulties. These difficulties are related to calculating the distance between two feature vectors, having components obtained as a result of the measurement of apparently incomparable quantities.

The methodological difficulties discussed above may be avoided in different ways. One possible methodological approach is to transform the feature space. Another approach is to transform some of the features from one type to another. The solution of the pattern recognition problem in both cases is constructed in the transformed space. Another group of methods are directed to search the solution of the supervised pattern recognition problem in the initial feature space. Some of them use a geometrical approach based on feature partitioning. Usually, each feature is considered separately and it is divided into segments [1]. Many models acquire decision rules, often expressed in logical form, but also in other forms like schemata. Other models use a set of representative instances [2], hyperrectangles (exemplars) [3], [4], or decision trees [5].

In the present paper binary description for fuzzy features are obtained through the use of a threshold value, which is calculated based on the distance between patterns defined in a manner similar to Hamming distance between binary features. For solving the problem with real features the mathematical model for construction of thresholds is based on feature partitioning. In contrast with the sequential methods, here all features are considered in parallel.

After transforming all the pattern descriptions into binary a mathematical model based on Non-Reducible Descriptors (NRDs) is applied. An NRD is a descriptor of minimal length. In other words, an NRD contains information on the smallest number of features that are necessary to describe a pattern uniquely.

The rest of the paper is organized as follows. In section 2, we present procedure for construction NRDs. In section 3, our model for construction NRD for fuzzy features is presented. Finally, in section 4, a method for transforming real features into binary or k-values based on parallel feature partitioning is proposed.

2 Non-reducible Descriptors

For ease of reference, we include here some definitions similar to the ones introduced in previous related work [6]. We will also describe briefly the computational procedure of [6] for the construction of all NRDs for pattern recognition problems with binary features. This will enable us to present the procedure for problems with fuzzy and real features in a simple and straightforward manner.

For pattern recognition problems with binary features, each pattern Q is represented in the training table by means of a sequence $t_1,t_2,\ldots,t_n$ with $t_i \in \{0,1\}$. Each member of this sequence corresponds to the presence or absence of the corresponding feature $x_1,x_2,\ldots,x_n$.

Definition 1. Let $Q_r = (t_{r,1}, t_{r,2}, \ldots, t_{r,n})$. The subsequence $(t_{r,j_1}, t_{r,j_2}, \ldots, t_{r,j_d})$, $j_d \leq n$ is called a *descriptor* for the pattern $Q_r \in K_i$ if there does not exist any other pattern $Q_s \in K_p$, $p = 1, 2, \ldots, i-1, i+1, \ldots, l$ in the training table with the same subsequence.

Definition 2. A given descriptor is called a *Non-Reducible Descriptor* (NRD) if none of its arbitrarily chosen proper subsequences is a descriptor.

Definition 2 means that if an arbitrarily chosen feature is removed, then this descriptor loses its property of being a descriptor. Therefore, an NRD is a descriptor of minimal length. Next, we assume that the NRD of pattern Q_r is given by $t_{j_1}, t_{j_2}, \ldots, t_{j_d}$.

Next, we need some more definitions for formulating the problem of constructing the NRDs for the patters $Q_r \in K_i$ for some i, $1 \leq i \leq l$. Let m' denote the number of patterns not belonging to K_i.

Definition 3. The *dissimilarity matrix* for a pattern $Q_r \in K_i$ is a binary matrix $L_r = [l_{v,j}]$; $v = 1, \ldots, m'$, $j = 1, \ldots, n$, with

$$l_{v,j} = \begin{cases} 1, \text{ if } t_{r,j} \neq t_{v,j}, \\ 0, \text{ otherwise,} \end{cases}$$

where $t_{r,j}$ and $t_{v,j}$ are the values of feature j of $Q_r \in K_i$ and $Q_v \notin K_i$, respectively.

Note that since the classes are pairwise disjoint, it follows that every row of the matrix L_r contains at least one unit.

Definition 4. The number of features d in an NRD is called its *rank* and is denoted by $\mathrm{NRD}^{\,d}$.

Definition 5. Columns $j_1, j_2, \ldots, j_d$ of an arbitrary $\{0,1\}$-matrix A of order $m \times n$ form a *covering* of M if there does not exist a row p, $p = 1, 2, \ldots, m$, such that $a_{p,j_q} = 0$ for $q = 1, 2, \ldots, d$.

The following procedure of [6], restated using the terminology presented above, is very useful in formulating the pattern recognition problem with binary features in terms of dissimilarity matrices.

The problem for constructing all NRDs for an arbitrary pattern Q_r is equivalent to permuting the rows and columns of the dissimilarity matrix L_r to obtain a matrix $L_r^{'}$ of order $m' \times n$ of the form

$$L_r^{'} = \begin{bmatrix} E_d & P_1 \\ P_2 & P_3 \end{bmatrix},$$

satisfying the following properties:

a) Submatrix E_d is an identity submatrix of order d, and no further permutations of rows and columns of L_r will result in a larger identity submatrix comprising E_d;
b) The columns of the submatrix P_2 form a covering of P_2; in other words, each row of P_2 has at least one unit.

Therefore, E_d is the maximal identity submatrix, where d is the rank of the constructed NRD. Note that the above problem always has a solution because each row of the dissimilarity matrix L_r must contain at least one unit due to the pairwise disjointedness of the classes.

Example. Patients are characterized as suffering from strep-throat or flu depending on the presence or absence of a combination of the following symptoms: sore throat (feature x_1), cough (x_2), cold (x_3), and fever (x_4). Let K_1 denote the class of patients suffering from strep-throat, and K_2, the class of patients suffering from flu. The following training table consists of information pertaining to 7 patients, the first two in K_1, and the last five in K_2. The information pertaining to each patient is represented as a row in the following training table. A 1 in a particular column represents the presence of the corresponding symptom, and a 0 represents the absence of that symptom. Thus, in the notation of this paper, Q_1, $Q_2 \in K_1$, and $Q_3, \ldots, Q_7 \in K_2$.

$$T = \begin{array}{c} \begin{matrix} x_1 & x_2 & x_3 & x_4 \end{matrix} \\ \left[\begin{array}{cccc} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ \hline 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \end{array}\right] \end{array}.$$

For the object Q_1, the sequence $(t_{1,1}, t_{1,2}, t_{1,4}) = (1, 1, 0)$ is a descriptor. The sequence $(t_{1,1}, t_{1,2}) = (1, 1)$ is an NRD for the object Q_1, and it is expressed by the conjunction x_1x_2. Also, the sequence $(t_3) = (0)$ is an NRD for the object Q_1, and it is expressed by the conjunction $\bar{x}_3$. Similarly, for the object Q_2, the sequence $(t_{2,1}, t_{2,3}, t_{2,4}) = (1, 1, 0)$ is descriptor. Also, the sequence $(t_{2,1}, t_{2,4}) = (1, 0)$ is an NRD for the object Q_2, and it is expressed by the conjunction $x_1\bar{x}_4$.

This would mean that, based on the first patient's symptoms and diagnosis, the presence of sore throat and cough are sufficient to diagnose a patient with strep-throat. Similarly, based on the second patient's symptoms and diagnosis, the presence of sore throat and the absence of fever are sufficient to diagnose a patient with strep-throat as well.

In the next sections we will consider how to transform fuzzy features and real features into binary.

3 Fuzzy Features

Since the training table is not a $\{0, 1\}$-matrix for pattern recognition problems with fuzzy features, the previous definition of the dissimilarity matrix (for binary features) does not apply directly. Therefore, we need to extend this concept appropriately for the case of fuzzy features [7]. Once this is accomplished, the procedure for construction of NRDs may be applied as well. For this purpose,

we make the following definition. For simplifying the notation we will assume that all features in the training table are fuzzy. Otherwise, we will apply the proposed model only to the fuzzy part of the descriptions.

Definition 6. The *distance matrix* for a pattern $Q_r \in K_i$ is the $m' \times n$ numerical matrix $D_r = [d_{v,j}]$ with $d_{v,j} = |t_{r,j} - t_{v,j}|$, where $t_{r,j}$ and $t_{v,j}$ are the values of feature j of $Q_r \in K_i$ and $Q_v \notin K_i$, respectively.

Note that since there are totally m patterns in the training table, m such distance matrices can be obtained, one for each pattern. In order to obtain binary dissimilarity matrices, we would like to determine a *threshold distance value* ε. We accomplish this as follows:

a) In each distance matrix, determine the maximal element for each row.
b) Define ε as the minimal element among all the maximal elements obtained from the distance matrices.

We have used this procedure for determining ε in order to ensure that the dissimilarity matrices will have at least one unit in each row. We are now ready to define dissimilarity matrices for fuzzy features.

Definition 7. The *dissimilarity matrix* for the pattern $Q_r \in K_i$ is the $m' \times n$ binary matrix $L_r = [l_{v,j}]$, with

$$l_{v,j} = \begin{cases} 1, \text{ if } d_{v,j} \geq \varepsilon, \\ 0, \text{ otherwise,} \end{cases}$$

where $d_{v,j}$ is the value in matrix D_r.

However, if the value corresponding to the feature x_j in a descriptor is at least ε it denotes the presence of the feature x_j, and therefore will be represented by the occurrence of x_j in the NRD. Similarly, if this value is smaller than ε, then it denotes the absence of x_j, and therefore will be represented by the occurrence of the feature $\bar{x}_j$ in the NRD.

4 Real Features

We will consider the case when all features take value from the set of real numbers R. Otherwise, we will apply the proposed model only to the real part of the descriptions. Without loss of generality, we will discuss the pattern recognition problem with two classes K_1 and K_2. Let the set M' be entirely contained in a restricted closed region D. Let $K_1 \cap K_2 = \emptyset$, $K_1 = \{Q_1, \ldots, Q_r\}$, $K_2 = \{Q_{r+1}, \ldots, Q_m\}$.

Let us assume that for each feature x_j, $j = 1, 2, \ldots, n$ the threshold set $E_j = \{\varepsilon_j^1, \ldots, \varepsilon_j^{v(j)}\}$ is given. The set element ε_j^s belongs to the region defined for the feature x_j and $\varepsilon_j^s \neq \varepsilon_j^t$ for $s \neq t$. Let $\varepsilon_j^1 < \ldots < \varepsilon_j^{v(j)}$.

Accordingly, the problem for transforming the feature space by feature partitioning is related to the construction of the sets $\tilde{E}_1, \ldots, \tilde{E}_n$ [8]. This solution means that for each coordinate axis x_j a subset $\tilde{E}_j \subseteq E_j$ is constructed, comprising of k_j elements $\{\varepsilon_j^{i_1}, \ldots, \varepsilon_j^{i_{k_j}}\}$, $\varepsilon_j^{i_1} < \ldots < \varepsilon_j^{i_{k_j}}$, $k_j \leq v(j)\}$ exists, so that the

total number of thresholds is minimal. The sets $\tilde{E}_1, \ldots, \tilde{E}_n$ are called threshold sets and their members, thresholds. The problem for construction the threshold sets $\tilde{E}_1, \ldots, \tilde{E}_n$ is discussed next.

On each coordinate axis x_j, $j = 1, \ldots, n$ the values of the corresponding features from the descriptions of the training patterns $Q_1, \ldots, Q_m$ are plotted. Among all possible open intervals, restricted by two neighboring values, only those whose ends belong to the description of patterns from different classes are considered. Let the number of intervals for the jth coordinate axis be $v(j)$. Let

$$n_1 = \sum_{j=1}^{n} v(j),$$

From each open interval considered, one arbitrary threshold value, for example, the middle of the interval is chosen. In this way, for each coordinate axis j the sequence of threshold values $\varepsilon_j^1, \ldots, \varepsilon_j^{v(j)}$ is obtained. From this sequence the set $\tilde{E}_j$ has to be constructed. Let to each threshold component ε_j^t the binary variable x_j^t be assigned, according to the rule:

$$x_j^t = \begin{cases} 1, \text{ if } \varepsilon_j^t \text{ is chosen as a threshold,} \\ 0, \text{ otherwise,} \end{cases}$$

for $t = 1, 2, \ldots, v(j)$; $j = 1, 2, \ldots, n$.

Let us consider the ordered set of all possible pairs (Q_p, Q_q), where $Q_p \in K_1$, $p = 1, 2, \ldots, r$; $Q_q \in K_2$, $q = r+1, r+2, \ldots, m$. Obviously

$$|\{(Q_p, Q_q)\}| = r(m-r).$$

Let

$$m_1 = r(m-r).$$

The matrix $C = [c_{ij}^t]_{m_1 \times n_1}$ is constructed according to the rule:

$$c_{ij}^t = \begin{cases} 1, \text{ if } i\text{th pair of patterns differs in the } j\text{th} \\ \quad \text{coordinate axis by the threshold value} \\ \quad \varepsilon_j^t, \\ 0, \text{ otherwise.} \end{cases}$$

where $i = 1, 2, \ldots, m_1$; $j = 1, 2, \ldots, n_1$; $t = 1, 2, \ldots, v(j)$.

The matrix C is arranged in the following way. The first group of $v(1)$ columns contains information for the feature x_1, next group $v(2)$ - for the feature x_2 and so on, the last group $v(n)$ - for x_n. Each group of $v(j)$ columns, $j = 1, \ldots, n$, contains information for the thresholds $\varepsilon_j^1 < \ldots < \varepsilon_j^{v(j)}$ of the set E_j. Columns in each group are ordered sequentially in ascending order of thresholds. In the sequential rows of the matrix C, is written information for the pairs (Q_1, Q_{r+1}), (Q_1, Q_{r+2}), …, (Q_1, Q_m), (Q_2, Q_{r+1}), …, (Q_r, Q_m) according to the above rule.

The matrix C has the following properties by its construction: i) each row contains at least one unit, ii) each column contains at least one unit.

Let us suppose that for each feature x_j, $j = 1, 2, \ldots, n$, k_j thresholds, $1 \leq k_j \leq v(j)$ are chosen. This condition may be written as follows:

$$\sum_{t=1}^{v(j)} x_j^t = k_j; \; j = 1, 2, \ldots, n.$$

for $k_j = 1, 2, 3, \ldots, v(j)$.

From the condition $K_1 \cap K_2 = \emptyset$ it follows that each pair (Q_p, Q_q) differs at least by one coordinate. For ith pair this condition may be written as follows:

$$\sum_{j=1}^{n} \sum_{t=1}^{v(j)} c_{ij}^t \; x_j^t \; \geq \; 1; \quad i = 1, 2, \ldots, m_1.$$

Now we can formulate the following problem. Find out the solution for the transformation of the feature space such that for each feature x_j no more than k_j, $k_j \leq v(j)$, thresholds are constructed and

$$\sum_{j=1}^{n} k_j \quad \rightarrow \; min.$$

It means that a binary vector $x^* = (x_1^1, \ldots, x_1^{v(1)}, \ldots, x_n^1, \ldots, x_n^{v(n)})$, with dimension n_1 is found, which minimizes:

$$\sum_{j=1}^{n} \sum_{t=1}^{v(j)} x_j^t \; \rightarrow \; min,$$

at the conditions:

$$1 \; \leq \; \sum_{t=1}^{v(j)} x_j^t \; \leq \; k_j, \quad j = 1, 2, \ldots, n,$$

$$\sum_{j=1}^{n} \sum_{t=1}^{v(j)} c_{ij}^t \; x_j^t \; \geq \; 1, \quad i = 1, 2, \ldots, m_1.$$

The formulated problem is an integer-valued optimization problem. Its geometrical interpretation is as follows. Let us construct the minimal n-dimensional parallelepiped, containing the region D in R^n. Let the threshold sets $\tilde{E}_1, \ldots, \tilde{E}_n$ be given. Let us construct for each threshold the $(n-1)$-dimensional plane, perpendiculiar to the corresponding coordinate axis. As a result of crossing the n-dimensional parallelepiped containing the region D is covered by the minimal number of hyperparallelepipeds. Their number is equal to $(k_1 + 1) \ldots (k_n + 1)$. Each parallelepiped either contains patterns belonging to only one of the classes or it is the empty one. From the condition $K_1 \cap K_2 = \emptyset$ it follows that the above feature partitioning problem always has a solution. If n thresholds are constructed as a result, then we can transform pattern descriptions into binary. If more than n thresholds are obtained as a result, then the NRD may be expressed using the tools of the k-valued logic.

5 Conclusions

In this paper, a mathematical models has been proposed in the case of heterogeneous features when it is difficult to introduce metric. We have shown how the dissimilarity matrix model for the pattern recognition problem with binary features may be used to construct the Non-Reducible Descriptors for patterns in a problem with fuzzy features. For real features a mathematical model based on parallel feature partitioning has been proposed. The model is based on partitioning the feature space using minimal number of nonintersecting regions. This is achieved by solving the integer-valued optimization problem, which leads to the construction of minimal covering. The proposed models may be used in a variety of fields, including medicine, molecular biology and social sciences.

References

1. H.A.Güvenir, I.Sirin, Classification by feature partitioning. Machine Learning **23** (1996) 47–67.
2. D.W.Aha, D.Kibler, M.K.Albert. Instance-based learning algorithms. Machine Learning **6** (1991) 77–66.
3. L.Rendell. A new basis for state-space learning systems and successful implementation. Artificial Intelligence **20** (1983) 369–392.
4. S.Salzberg. A nearest hyperrectangle method. Machine Learning **6** (1991) 251–276.
5. J.R.Quinlan. Induction of decision trees. Machine Learning **1** (1986) 81–106.
6. V.Valev and P.Radeva. A Method of solving pattern or image recognition problem by learning Boolean formulas, Proceedings of 11th International Conference on Pattern Recognition, Hague, Netherlands, August 30 - September 3, 1992, IEEE Computer Society Press **II** (1992) 359–362.
7. V.Valev, A.Asaithambi. Fuzzy non-reducible descriptors, International Journal on Machine GRAPHICS & VISION **12** No. 3 (2003) 353–361.
8. V.Valev. Supervised pattern recognition by parallel feature partitioning. Pattern Recognition **37** No. 3 (2004) 463–467.