Reducing the dimensionality of feature space in pattern recognition tasks

Sh Fazilov[1], N Mamatov[1], A Samijonov[2], Sh Abdullaev[1]
[1]Tashkent University Information Technologies named after Al-Kharezmi, Tashkent, Uzbekistan
[2]Bauman Moscow State technical university, Moscow, Russia

Abstract. The definition of an informative set of features is one of the important tasks in pattern recognition. Typically, the determination of informative features is carried out using two types of methods. The first type of methods is “direct methods”, they are directly aimed at identifying informative sets of attributes. And the second type is called "inverse methods", these methods serve to build informative sets of signs by eliminating uninformative signs from the attribute space. This article is devoted specifically to the development of the second type of method; it proposes an accelerated method and an algorithm for determining non-informative features based on the selected non-informative criterion.

1. Introduction

When solving practical recognition problems, it is often necessary to reduce the dimension of the initial feature space by eliminating non-informative features. This procedure improves the quality of recognition by eliminating “noisy” parameters and reduces processing time by reducing the amount of data.

The definition of non-informative features allows you to move from the original system of signs $x=\left(x^1,x^2,...,x^N\right)$ to the new system $z=\left(z^1,z^2,...,z^\ell\right)$, while in the new system of signs the number of signs is less $\left(\ell<N\right)$ than the original system [1-6]. New features can be formed as a function of $z=F\left(x\right)$ from the original features, i.e. by solving the optimization problem. The goal of the problem is to find such a system of signs z, at which

$$I(\tilde{z})=\min_{F\in\Omega}\left\{I(z)\right\}. \tag{1}$$

Here $I\left(z\right)$ is the specified measure of non-informativity of ℓ – dimensional system of signs z, and F is the class of permissible transformations of the original signs $x^1,x^2,...,x^N$, the class can be a linear, nonlinear, discrete, continuous or logical type of transformation.

Thus, the formation of a non-informative description of objects can be represented as a map of an N-dimensional vector x to ℓ – a dimensional vector z, which in the general case can be represented as $z=F\left(x\right)$, where F is a valid transformation, and $\ell\le N$.

The system of functions is used as valid transformations

$$z^i=f_i\left(x^i\right);\ i=\overline{1,N}, \tag{2}$$

$$f\left(x^i\right)=\begin{cases}0, \text{ if feature } x^i \text{ excluded;}\\ 1, \text{ if feature } x^i \text{ is left.}\end{cases}$$

where a new system of features is formed as a subset of the set of source features.

Heuristic criteria based on an assessment of the separability measure of objects of a given training set using the Euclidean metric are used as a measure of the non-informativeness of signs [3,8].

2. Statement of a problem and the concept of the problem decision

Let the training set be given by $x_{11}, x_{12}, ..., x_{1m_1}, x_{21}, x_{22}, ..., x_{2m_2}, ..., x_{r1}, x_{r2}, ..., x_{rm_r}$ objects, for which it is known that each group of objects $x_{p1}, x_{p2}, ..., x_{pm_p}$ belongs to a certain class X_p, $p=\overline{1,r}$.

$$x_{pi}=\left(x_{pi}^1, x_{pi}^2, ..., x_{pi}^N\right)$$

where each object x_{pi} is an N-dimensional vector of numerical signs.

For a given training sample of objects $x_{p1}, x_{p2}, ..., x_{pm_p} \in X_p$, $p=\overline{1,r}$, where x_{pi} is a vector in the N-dimensional attribute space, we introduce the vector $\lambda=\left(\lambda^1, \lambda^2, ..., \lambda^N\right)$, $\lambda^k \in \{0;1\}$, $k=\overline{1,N}$, which uniquely characterizes a certain subsystem of attributes. The components of the vector λ, equal to unity, indicate the presence of corresponding signs in this subsystem, and the zero components indicate the absence of corresponding signs.

The space of signs $\left\{ x=\left(x^1, x^2, ..., x^N\right)\right\}$ will be considered Euclidean and denoted by R^N.

Definition The vector λ is called ℓ – uninformative if the sum of its components is ℓ, i.e.

$$\sum_{i=1}^{N}\lambda^i=\ell.$$

For each subsystem defined by ℓ – an uninformative vector λ, its own ℓ – dimensional feature subspace is defined. In each of these subspaces we introduce the Euclidean norm with respect to truncation in λ

$$\|x\|_\lambda=\sqrt{\sum_{j=1}^{N}\lambda^j\left(x^j\right)^2}.$$

Denote

$$\overline{x}_p=\frac{1}{m_p}\sum_{i=1}^{m_p}x_{pi},\ p=\overline{1,r},$$

where $\overline{x}_p$ is the average object of class X_p.

We introduce the function

$$S_p(\lambda)=\sqrt{\frac{1}{m_p}\sum_{i=1}^{m_p}\left\|x_{pi}-\overline{x}_p\right\|_\lambda^2}.$$

Function $S_p(\lambda)$ characterizes the average scatter of objects of class X_p in a subset of the attributes defined by the vector λ. We define a criterion for the informativeness of subsystems in the form of a functional

$$I(\lambda)=\frac{\sum_{p,q=1}^{r}\left\|\overline{x}_p-\overline{x}_q\right\|_\lambda^2}{\sum_{p=1}^{r}S_p^2(\lambda)}. \quad (3)$$

We denote

$$a=\left(a^1,a^2,...,a^N\right);\ b=\left(b^1,b^2,...,b^N\right),$$
$$a^j=\sum_{p,q=1}^{r}\left(\bar{x}_p^j-\bar{x}_q^j\right)^2,\ j=\overline{1,N};$$
$$b^j=\sum_{p=1}^{r}\left(\frac{1}{m_p}\sum_{i=1}^{m_p}\left(\bar{x}_{pi}^j-\bar{x}_p^j\right)^2\right),\ j=\overline{1,N}.$$

Then functional (3) reduces to the form

$$I\ (\lambda)=\frac{(a,\lambda)}{(b,\lambda)}, \tag{4}$$

where (*,*) is the scalar product of vectors.

The coefficients a^j and b^j are independent of λ and are calculated in advance. To calculate the functional $I(\lambda)$ for each λ, about N operations are required.

The method proposed below for choosing non-informative feature sets is based on the use of non-informative criteria defined by functionals reducible to the form (4). Moreover, the optimization problem to be solved is formulated as

$$\begin{cases} I(\lambda)=\dfrac{(a,\lambda)}{(b,\lambda)}\to\min; \\ \lambda\in\Lambda^{\ell}, \end{cases} \tag{5}$$

where Λ^{ℓ} is the set of vectors $\lambda=\left(\lambda^1,\lambda^2,..,\lambda^N\right)$, i.e.

$$\Lambda^l=\left\{\lambda\,\middle|\,\lambda_i=\{0,1\},i=\overline{1,N},\sum_{i=1}^{N}\lambda_i=l\right\}.$$

Here, the set Λ^{ℓ} defines ℓ – non-informative space, and its element λ is ℓ – non-informative vector.

To solve problem (5), we introduce a vector function

$$\phi(\lambda)=a(b,\lambda)-b(a,\lambda), \tag{6}$$

which indicates the direction of the fastest decrease in the functional $I(\lambda)$ at the point λ.

Theorem 1. If λ and μ are two ℓ-non-informative vectors and $b^j>0,\ j=\overline{1,N}$, then $I(\lambda)>I(\mu)$ if and only if $(\phi(\lambda),\mu)<0$.

Proof. The proof follows from the sequence of inequalities

$$(\phi(\lambda),\mu)<0\leftarrow(a(b,\lambda)-b(a,\lambda),\mu)\leftarrow(a,\mu)(b,\lambda)-(b,\mu)(a,\lambda)<0\leftarrow$$
$$\leftarrow\frac{(a,\mu)}{(b,\mu)}<\frac{(a,\lambda)}{(b,\lambda)}\leftarrow I(\mu)<I(\lambda).$$

The theorem is proved.

We introduce the operator (follow)

$$\mu:\Lambda^{\ell}\to\Lambda^{\ell}$$

such that

$$\left(\phi(\lambda),\mu(\lambda)\right)=\min_{\eta\in\Lambda^{\ell}}\left(\phi(\lambda),\eta\right).$$

The μ operator has an obvious constructive view. If we order the components of the vector $\phi(\lambda)$, i.e. find a set of pairwise different indices $j_1, j_2,..., j_N$ such that

$$\phi^{j_1}(\lambda)\le\phi^{j_2}(\lambda)\le...\le\phi^{j_N}(\lambda),$$

then the components of the vector $\mu(\lambda)$ will be defined as

$$\mu^{j_1}(\lambda)=1,\mu^{j_2}(\lambda)=1,...,\mu^{j_\ell}(\lambda)=1,\mu^{j_{\ell+1}}(\lambda)=0,\mu^{j_{\ell+2}}(\lambda)=0,...,\mu^{j_N}(\lambda)=0.$$

In other words, the components of the vector $\mu(\lambda)$ corresponding to the first ℓ – maximal components of the vector $\phi(\lambda)$ are equal to unity, the rest to zero.

Obviously, $\mu(\lambda)$ is also a ℓ – informative vector, and

$$(\phi(\lambda),\mu(\lambda))=\min\left\{(\phi(\lambda),\eta)\middle|\eta\in\Lambda^\ell\right\}. \quad (7)$$

Property 1 For arbitrary $\lambda\left(\lambda\in\Lambda^\ell\right)$ true $(\phi(\lambda),\mu(\lambda))\le 0$.

Proof.

From the expression (7) follows

$$(\phi(\lambda),\mu(\lambda))<(\phi(\lambda),\lambda)=(a(b,\lambda)-b(a,\lambda),\lambda)=(a,\lambda)(b,\lambda)-(b,\lambda)(a,\lambda)=0.$$

The property proved.

From the above Theorem 1 and property, the main corollary follows $I(\lambda)\ge I(\mu(\lambda)),\lambda\in\Lambda^\ell$.

Theorem 2 If $I(\lambda)=I(\mu(\lambda))$, then $I(\lambda)=\min\left\{I(\eta)\middle|\eta\in\Lambda^\ell\right\}$.

Proof. It follows from Theorem 1 that $(\phi(\lambda),\mu(\lambda))=0$. Hence, according to (7), we have $(\phi(\lambda),\mu(\lambda))=0=\min\left\{(\phi(\lambda),\eta)\middle|\eta\in\Lambda^\ell\right\}$ or $(\phi(\lambda),\eta)\ge 0$ for arbitrary $\eta\in\Lambda^\ell$.

In accordance with Theorem 1, this means that $I(\eta)\ge I(\lambda)$ for any $\eta\in\Lambda^\ell$, i.e. $I(\lambda)=\min\left\{I(\eta)\middle|\eta\in\Lambda^\ell\right\}$.

The theorem is proved.

Theorem 2 guarantees the optimality of the obtained solution, i.e. the values of the functional $I(\lambda)$ with the solution found λ reach their minimum on the set Λ^ℓ.

Theorems 1 and 2 are the basis for the proposed method of minimizing functional (4), which is implemented as an iterative procedure. Moreover, at the first step, an arbitrary ℓ – little informative vector λ is selected, for example, $\lambda=\left(\overbrace{1,1,...,1}^{\ell},0,0,...,0\right)$.

Then, at each iteration, the new vector λ is determined from the previous one using the follow operator $\mu(\lambda)$ through assignment $\lambda=\mu(\lambda)$. The iterative process continues until the value of the functional $I(\lambda)$ decreases. In the event that such a reduction ceases, i.e. $I(\lambda)=I(\mu(\lambda))$, then λ is the optimal solution. Typically, this solution is achieved, as experiments have shown, in 3-4 steps.

The considered method can be formulated as the following algorithm:

Step 1 The following values are specified: ℓ - the required number of features; N is the total number of signs; a and b are N-dimensional vectors.

Step 2 Set the vector $\lambda=\left(\underbrace{\overbrace{1,1,...,1}^{\ell},0,0,...,0}_{N}\right)$.

Step 3 We calculate the values of the functional $I(\lambda)$.

Step 4 We calculate $\lambda_0=\mu(\lambda)$ (vector components $\mu(\lambda)$ are defined in a separate block).

Step 5 We calculate the values of the functional $I_0=I(\lambda_0)$.

Step 6 Compare $I_0<I(\lambda)$. If the inequality holds, then we assume that $\lambda=\lambda_0$, $I(\lambda)=I_0$ and go to step 4. Otherwise, the procedure ends and λ is the optimal value.

Step 7 Output parameters: λ, $I(\lambda)$.

The following operations are performed in the calculation unit $\lambda_0 = \mu(\lambda)$:

Step 1 The following parameters are set: λ, ℓ, N, a, b.

Step 2 Calculate the vector $\phi(\lambda) = a(b,\lambda) - b(a,\lambda)$.

Step 3 Set $\mu = (0,0,...,0), k = 1$.

Step 4 Set $\phi_{\max} = 10^{60}, i = 1$.

Step 5 Checking the i-components of the vector μ. If $\mu_i = 0$, then go to the next step, otherwise - to step 8.

Step 6 Checking the i-components of the vectors $\phi_{\max}, \phi$. If $\phi_{\max} > \phi$ go to the next step, otherwise go to step 8.

Step 7 Assign $\phi_{\max} = \phi_i, m = i$.

Step 8 $i = i + 1$.

Step 9 If $i \geq N$, then go to step 5. Otherwise, go to the last step.

Step 10 Assign $\mu_m = 1, k = k + 1$.

Step 11 If $k \geq \ell$, then go to step 4, otherwise $\lambda_0 = \mu$ and the procedure ends.

The results obtained apply to all criteria for the non-informativity of attributes defined by functionals, which in principle can be reduced to form (4).

Suppose that the components of the vector a are positive, and the vector b is strictly greater than zero, i.e.

$$a = \left(a^1, a^2, ..., a^N\right);\ a^j \geq 0,\ b = \left(b^1, b^2, ..., b^N\right);\ b^j > 0,\ j = \overline{1, N},$$

and

$$\frac{a^1}{b^1} \leq \frac{a^2}{b^2} \leq ... \leq \frac{a^N}{b^N}. \tag{8}$$

On the set Λ^ℓ we consider problem (5).

From the definition of the set ℓ – non-informative vectors Λ^ℓ it follows that the cardinality of the set-in question is C_N^ℓ. That means, with full enumeration, to find the minimum value of the functional $I(\lambda)$ for a given ℓ it is calculated C_N^ℓ times. It follows that for each given ℓ there exists a ℓ – informative vector $\lambda_\ell \in \Lambda^\ell$ that provides

$$I(\lambda_\ell) = \min_{\lambda \in \Lambda} I(\lambda), \tag{9}$$

where $\ell = \overline{1, N}$.

To solve problem (9), a partial enumeration method is proposed. This method consists in finding such a ℓ – informative vector λ, in which the given criterion of non-informativity $I(\lambda)$ reaches its minimum. In this case, the minimum of the function $I(\lambda)$ is found using the enumeration procedure for all $\lambda = \left(\lambda^1, \lambda^2, ..., \lambda^N\right)$, $\lambda^k \in \{0,1\}$ for which $\sum_{k=1}^{N} \lambda^k = \ell$.

To ensure the efficiency of enumeration, we strictly order all possible ℓ – non-informative vectors λ as follows. The first in this order is the vector having the first ℓ components equal to one, the rest zero, and the last vector has the last ℓ components equal to one, the rest zero.

Suppose given some ℓ – non-informative vector $\lambda_r = \left(\lambda_r^1, \lambda_r^2, ..., \lambda_r^N\right)$, located in a certain place of the considered order. We define a rule for choosing the vector $\lambda_{r+1} = \left(\lambda_{r+1}^1, \lambda_{r+1}^2, ..., \lambda_{r+1}^N\right)$ immediately following it.

If $\lambda_r^1 = *,\ \lambda_r^2 = *, ..., \lambda_r^k = *, \lambda_r^{k+1} = 1, \lambda_r^{k+2} = 0, \lambda_r^{k+3} = 0, ..., \lambda_r^N = 0$, then $\lambda_{r+1}^1 = \lambda_r^1,\ \lambda_{r+1}^2 = \lambda_r^2, ..., \lambda_{r+1}^k = \lambda_r^k, \lambda_{r+1}^{k+1} = 0, \lambda_{r+1}^{k+2} = 1, \lambda_{r+1}^{k+3} = \lambda_r^{k+3}, ..., \lambda_{r+1}^N = \lambda_r^N$.

If $\lambda_r^1=1,\ \lambda_r^2=1,...,\lambda_r^k=1,\lambda_r^{k+1}=0,\lambda_r^{k+2}=0,...,\lambda_r^j=0,\lambda_r^{j+1}=1,...,\lambda_r^N=1,$ then $\lambda_{r+1}^1=\lambda_r^1,\ \lambda_{r+1}^2=\lambda_r^2,...,\lambda_{r+1}^{k-1}=\lambda_r^{k-1},\lambda_{r+1}^{k+1}=1,\lambda_{r+1}^{k+2}=1,...,\lambda_{r+1}^{k+N-j+1}=1,\ \lambda_r^{k+N-j+2}=0,...,\lambda_{r+1}^N=0.$

Schematically, this rule is presented in the form

$$(*,*,...*,\underbrace{1,0},0,...,0),(*,*,...,*,\underbrace{1,0},\underbrace{0,...,0,1},\underbrace{1,...,1}).$$

It was proved in [8] that the above rule guarantees an exhaustive search of all possible options, starting from the first and ending with the last.

Given (8), the enumeration process ends when the vector is formed $\lambda=(0,1,1,...1,0,0,...,0)$.

The number of functional calculations $I(\lambda)$ in this method for a given $\ell\geq 2$ is determined as

$$C_N^\ell-C_{N-1}^\ell=\frac{N!}{(N-\ell)!\ell!}-\frac{(N-1)!}{(N-\ell-1)!\ell!}=\frac{\prod_{i=1}^{\ell}(N-\ell+i)-\prod_{i=1}^{\ell}(N-\ell-1+i)}{\ell!}=$$

$$=\frac{\ell\prod_{i=1}^{\ell-1}(N-\ell+i)}{\ell!}.$$

Thus, from the last expression it follows that the number of iterations in calculating the functional in this method is significantly less than with full enumeration.

From the results of the proposed scheme, it is easy to formulate a simple algorithm for minimizing the criterion of non-informativeness of feature subsystems by the method of partial enumeration. This algorithm is implemented as follows:

Step 1 Assuming $\lambda=\left(\underbrace{1,1,...1}_{\ell},0,0,...,0\right)$.

Step 2 Calculation $I(\lambda)$.

Step 3 Assuming $I_{\max}=I(\lambda),\ \lambda_{\max}=\lambda$.

Step 4 Determination $\lambda=Q(\lambda)$ (following the rule).

Step 5 Calculation $I(\lambda)$.

Step 6 Check: if $I_{\min}>I(\lambda)$, then assign $I_{\min}=I(\lambda),\ \lambda_{\min}=\lambda$. Otherwise, proceeds directly to the next step.

Step 7 Check: if $\lambda=\left(0,\underbrace{1,1,...1}_{\ell},0,0,...,0\right)$, then the procedure ends and $\lambda_{\max}$ defines the best subsystem of features. Otherwise, go to step 4.

The proposed algorithms were tested on examples of solving a number of practical problems, one of which is associated with the assessment of the operational state of a steeply inclined conveyor installed in the quarry of a mining and metallurgical plant.

This conveyor can be in one of three states: dangerous, pre-emergency, safe. The signals characterizing the dangerous state of the complex represented the first class of objects (K_1), the signals characterizing the pre-emergency state – the second class of objects (K_2), the signals characterizing the safe state – the third class of objects (K_3). The number of features characterizing each object was 9.

Each class contains the same number of objects, equal to 5056. Thus, each object can be represented in the form of a vector $x_{ij}=\left(x_{ij}^1,\ x_{ij}^2,\ ...\ ,\ x_{ij}^9\right)$, where x_{ij}^k is the $k-$th sign of the $j-$th object $i-$of the th class, where $k=\overline{1,9};\ j=\overline{1,5056};\ i=\overline{1,3}$.

The task of determining the main indicators characterizing the state of the conveyor was reduced to

solving the optimization problem

$$\begin{cases} I(\lambda) = \dfrac{(a,\lambda)}{(b,\lambda)} \to opt, \\ \lambda \in \Lambda^{\ell}. \end{cases}$$

After determining ℓ-non-informative feature sets ($\ell = \overline{1,9}$), n-l informative features were determined, based on these informative features, the recognition problem was solved using the "k-nearest neighbors" method, which allowed us to assess the degree of "usefulness" of each of these feature sets from the point of view their influence on the quality of recognition of objects of the control sample.

As a result of solving this problem, the most non-informative set of signs x_1, x_3, x_4, x_6, x_7 was determined, and an informative set of signs x_2, x_5, x_8, these informative signs represent the values of the signals coming from the vertical components of three-component sensors $D_i, i = \overline{1,3}$.

Based on these results, domain experts developed recommendations for preventing emergencies during the operation of the conveyor complex.

3. Conclusion

The task of determining non-informative features is reduced to the optimization problem, i.e. the transition from a N-dimensional initial system to a $N - \ell$-dimensional subsystem of features, in which ℓ provides the minimum value of this measure of the non-informativeness of features. The solution of the optimization problem will provide an informative description of objects with the least number of features.

For the selection of non-informative signs, the Fisher type criterion of non-information was used. A method has been developed for solving the problem of optimizing the choice of non-information features, which uses a vector function that indicates the direction of the most rapid decrease in the functional. Also, for this criterion of non-informativeness, a complete enumeration method for the selection of non-informative features is proposed.

Methods and algorithms for determining non-informative feature sets are proposed, the implementation of which is associated with minimizing a given heuristic criterion for information content. The principles for constructing these methods are, in fact, close to the principles on which the methods for determining informative sets of attributes are based. The difference between them is due only to the fact that instead of the task of maximizing the information content criterion (when determining informative sets of attributes), the minimization problem (when determining non-informative sets of signs) is solved.

References

[1] Ayvazyan S A, Buchstaber V M, Enukov I S and Meshalkin L 1989 *Applied Statistics: Classification and Dimension Reduction* (Moscow: Finance and Statistics) p 607
[2] Gorelik A L and Skripkin V A 2004 *Recognition methods* (Moscow: Sov.radio) p 262
[3] Zhuravlev Yu 1978 *Cybernetics problems* vol 33 (Moscow: Nauka) pp 5-68
[4] Kutin G I 1981 *Proc. International radioelectronics* vol 9 pp 54–70
[5] Zagoruyko N G 1972 *Recognition methods and their application* (Moscow: Sov. radio) p 208
[6] Cheponis K A 1988 *Methods, criteria and algorithms used in the conversion, selection and selection of features in data analysis* (Lithuania: Vilnius) p 150
[7] Fazilov Sh and Mamatov N 2007 *Reports of the II International Conference Problems of Management and Informatics* (Kirgizstan: Bishkek) pp 59-68
[8] Fazilov Sh and Mamatov N 2019 *Journal of Physics: Conf. Series* 1210 (2019) 012043 doi:10.1088/1742-6596/1210/1/012043