УДК 519.87

Selecting Informative Variables in the Identification Problem

Eugene D. Mihov*
Oleg V. Nepomnyashchiy†
Institute of Space and Information Technology
Siberian Federal University
Kirensky, 26, Krasnoyarsk, 660074
Russia

Received 23.06.2016, received in revised form 14.08.2016, accepted 14.09.2016

The problem of multidimensional object classification with small training sample is considered. The following algorithms of estimating variable informativeness are considered: Ad, Del, AdDel.

A new algorithm for selecting informative variables is proposed. It is based on the optimization of the coefficient vector of the kernel fuzziness. Some modification of this algorithm is also discussed.

The comparative analysis of existing methods for selecting informative variables is presented.

Keywords: classification, small training sample, informative variable, optimization of the coefficient vector of the kernel fuzziness.
DOI: 10.17516/1997-1397-2016-9-4-473-480.

Introduction

Nowadays, the classification problem is solved by many ways. One of the widespread classification methods uses neural networks. Classification methods based on parametric model and non-parametric algorithms are also used. First of all one must select object variables that are used for classification.

Modern classification objects can be characterized by many variables, but not all variables reflect the object membership to a particular class. The variables that do not reflect the object membership to a particular class are uninformative, and those variables that reflect the object membership to a particular class are informative. Selecting the most informative variables is an important issue.

The modern classification objects can be characterized by many variables. For example, to identify the class of the patient disease one can carry out the large number of tests, and obtain hundred or more variables characterizing the patient health, but not all variables are informative.

To classify an object, it is necessary to have a large training sample. The number of variables that characterize an object directly depends on the size of training sample to construct an accurate model. However, large training samples are not often available. We cite the Chairman of the Russian Foundation for Basic Research on Mathematics, Mechanics and Computer Science expert Council academician Evgeniy Moiseev: "Classification on the basis of small sample, in the sense of mathematical statistics, is a very important task. We dealt with the medical problem that has 20 variables, and the size of training sample is equal to 600. According to common rules, it is impossible to draw a conclusion on the basis of 600 sample elements if there are 20 variables."

One should note that for such tasks the estimation of variable informativeness has the particular importance. It is difficult to create the decision rule for large number of variables and a

*edmihov@mail.ru
†ONepomnuashy@sfu-kras.ru

large training sample is required. If the decision rule is created with the use of only the most informative variables, the classification results are more accurate. The reason is that the model has lower dimension.

Therefore, the estimation of variable informativeness is a topical problem.

1. Classification

Classification is a general process related to categorization, the process in which ideas and objects are recognized, differentiated, and understood.

Classification is categorization of objects and phenomena into groups, classes and ranks, according to their characteristic differences and similarities.

Let us show the standard classification task Fig. 1.

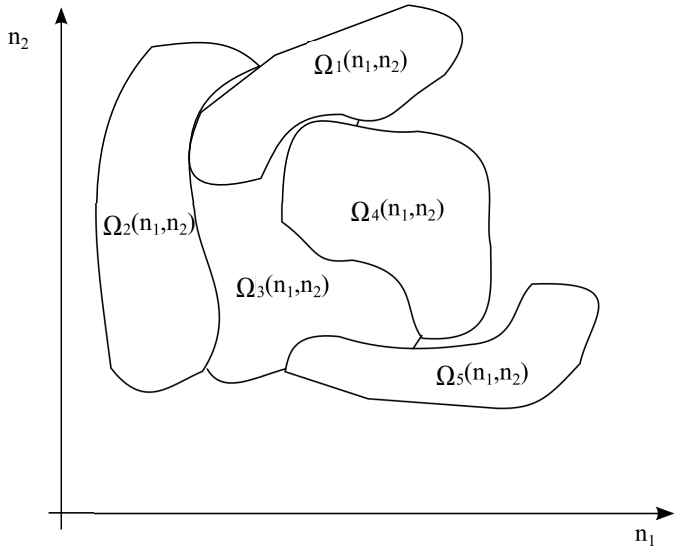

Fig. 1. The standard identification task

We use the following notation in Fig. 1: n_j is the variable j, $\Omega_i(n_1, n_2)$ is the domain where class i is defined.

The aim of classification is to find $\Omega_i(n_1, n_2), O_j(n_{1j}, n_{2j}) \in \Omega_i(n_1, n_2)$, where $O_j(n_{1j}, n_{2j})$ is the classification object with variables $n_1 = n_{1j}$, $n_2 = n_{2j}$.

We use the following method to determine the probability that object belongs to some class

$$P(n_{1o}, \ldots, n_{mo})_t = \prod_{i=1}^{m} \Phi(\frac{M(n_{it}) - n_{io}}{cs_i}), \tag{1}$$

m is the number of variables, P_t is the probability that object belongs to the class t, $O(n_{1o}, \ldots, n_{mo})$ is the object that is characterized by variables $n_{1o}, \ldots, n_{mo}, M(n_{it})$ is the expectation of variable$i$ for class t, $\Phi(*)$ is the kernel function and cs_i is the kernel fuzziness parameter for the variable i.

Kernel function $\Phi(*)$ satisfies the following conditions

$$\frac{1}{c_s} \int_{-\infty}^{+\infty} \Phi(\frac{t - t_i}{c_s}) dt = 1, \tag{2}$$

$$\lim_{c_s \to 0} \frac{1}{c_s} \int_{-\infty}^{+\infty} \varphi(t)\Phi(\frac{t-t_i}{c_s})dt = \varphi(t_i). \tag{3}$$

One should note that the choice of kernel fuzziness parameter vector has a direct impact on the model accuracy. These parameters determine the influence of sample elements on $P(O(n_{1o}, \ldots, n_{mo}))$ (Fig. 2).

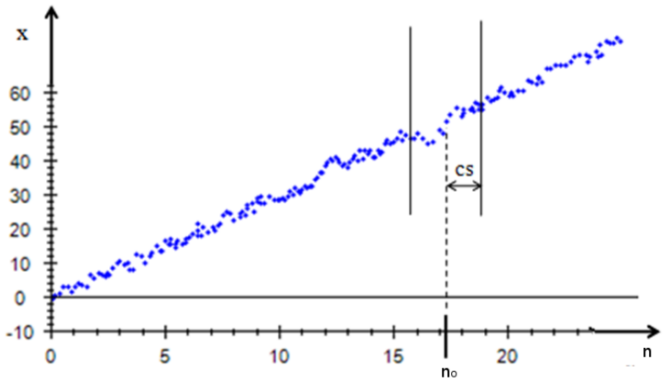

Fig. 2. Determination of the kernel fuzziness parameter

Fig. 2 shows that only sample elements for which $|n - n_o| < cs$ have the influence on n_o. Moreover, sample elements for which $|n - n_o| \to 0$ have a greatest impact on n_o.

2. Variable informativeness

The variables that do not reflect object membership to some class are uninformative, and those variables that reflect object membership to the class are informative.

Certainly the degree of informativeness is the relative term and it is rather difficult to define it.

There are two methods to estimate variable informativeness. They are direct and indirect methods.

In direct method one should find $\vec{cs} = (cs_1, cs_2, \ldots, cs_m)$ and

$$R(\vec{n}, \vec{cs}) = \frac{r}{s}, \tag{4}$$

$R(\vec{cs}) \to 0$, where $R(\vec{cs})$ is the average classification error (4), s is the total number of classified objects, and r is the number of correctly classified objects.

Indirect method supposes that the variable informativeness is estimated with the use of distribution properties of variables: $M(n_1), \ldots, M(n_m), D(n_1), \ldots, D(n_m)$.

The most popular indirect method is based on the Fisher criterion: $I(n_i) \to \infty, D(n_i) \to 0$ and $\max(M(n_i) - M(n_j)) \to \infty$, $j \neq i$, where $i(n_i)$ is the degree of informativeness of variable n_i.

In what follows we consider direct methods to estimate informativeness.

3. The estimation of the informativeness of variable

Let us assume that one need to choose the most informative n variables out of m variables $(n < m)$. There are several algorithms to perform this task.

Del algorithm.

It is necessary to exclude variable n_1, and calculate $R(n_2, \ldots, n_m)$, using only variables $n_2, \ldots, n_m$ here $R(n_2, \ldots, n_m)$ is the average classification error. Then the same calculations are performed with variables $n_2, \ldots, n_m$, and $R(n_1, \ldots, n_m), R(n_1, \ldots, n_{m-1})$ are obtained, respectively. The variable with the least degree of informativeness is found by the rule $\max R(n_1, \ldots, n_{i-1}, n_{i+1}, \ldots, n_m) \to \min I(n_i)$, where n_i is the variable with the least degree of informativeness. Variable with the least degree of informativeness is eliminated. The algorithm is repeated until the most informative n variables are found.

The number of iterations is calculated by the formula

$$L = m + (m-1) + (m-2) + \ldots + (n+1) = \sum_{i=1}^{m-n} (m-i) \tag{5}$$

Ad algorithm.

One should estimate $R(n_i)$ for each n_i and then choose the most informative variable by the rule $\min R(*) \to \max I(*)$. Thus, the first most informative variable is found. Then we add variables $n_1, \ldots, n_{i-1}, n_{i+1}, n_m$ to the found variable and obtain the set of pairs $(n_i, n_1), \ldots, (n_i, n_m)$. Then $R(n_i, n_1), \ldots, R(n_i, n_m)$ are calculated, and the most informative set of variables is selected using the rule given above. This operation is repeated until the most informative n variables are found. The number of iterations is equal to number of iterations of Del algorithm.

AdDel algorithm.

Methods given above are included in the class of so-called "greedy" algorithms.

These methods find only local optimum value. The global optimum value is $\vec{n}_{min}, n_1, \ldots, n_r$, $R(\vec{n}_{min}) \to min$.

One can use AdDel algorithm to increase the probability to find $\vec{n}_{min}$. First we find a_1 the most informative variables, using the Ad method. Then we delete a_2 ($a_2 < a_1$) variables, using the Del method. Calculations are repeated until the most informative n variables are found.

This algorithm has one special quality. The quality criterion (in this case, the average error of classification) varies as shown in Fig. 3.

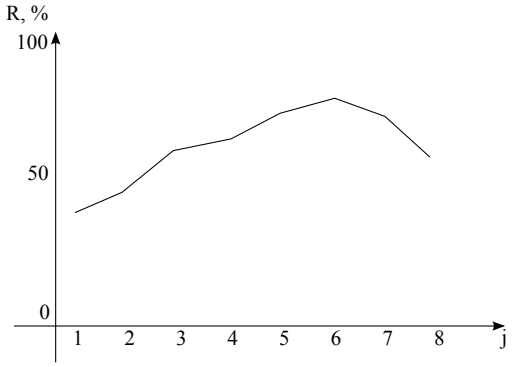

Fig. 3. Relationship between the error of classification and the number of variables

Fig. 3 shows that an increase in the number of variables leads at first to the increase in the accuracy of the model. Further increase in the number of variables reduces the accuracy of the model. This feature of function $R(\vec{n})$ can be used to obtain the optimal number of variables. The optimality criterion is the average error of classification.

The algorithm based on the optimization of the kernel fuzziness parameter.

Non-parametric method based on the kernel approximation is used as a method of classification. One of the important parameters of this method is the vector of kernel parameters $\vec{cs} = (cs_1, \ldots, cs_m)$. Parameter cs_i is used to set the weight of variable n_i.

One need to find such vector $\vec{cs}$ that $R(\vec{n}, \vec{cs}) \to 0$. In other words, we need to solve an optimization problem.

To find the least informative variable the rule $\max cs_i \to \min I(n_i)$, $i = 1, \ldots, m$ is employed. The algorithm uses both direct and indirect ways to estimate informativeness of variables.

4. Computational experiments

Between 2 and 40 variables are used to classify the object. Plots show a relationship between the number of calculations of $R(\vec{n}, \vec{cs})$ (C) and the number of variables (N).

Variables are separated into 3 groups with respect to the dispersion value (D).

To select the most informative variable we use the following methods: algorithm Ad, algorithm Del, algorithm AdDel and the algorithm based on the optimization of vector $\vec{cs}$.

First, we consider the algorithm Ad. The number of calculations for the algorithm Ad is shown in Fig. 4.

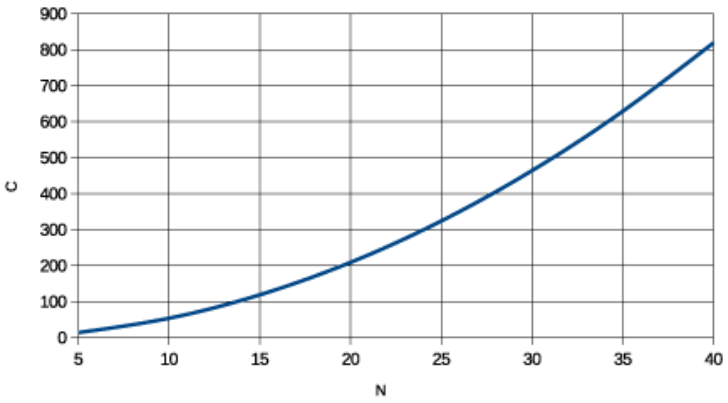

Fig. 4. Number of calculations versus the number of variables for the algorithm Ad

Fig. 4 demonstrates that the relationship is non-linear, and selection of the most informative variables demands considerable computational resources when the number of variables increases.

Now we consider the algorithm Del. The number of calculations for the algorithm Del is shown in Fig. 5.

Fig. 5 demonstrates that the relationship is also non-linear.

Algorithms Ad and Del are so-called "greedy" algorithms. This means that they find local optimum values but they do not always find the global optimum value. That is why the AdDel algorithm is more useful method to estimate the variable informativeness.

Let us demonstrate the selection of the most informative variables with the use of the AdDel algorithm. The AdDel algorithm is based on the Ad and Del algorithms. First we use the Ad algorithm to find the most informative $a_1$variables then we use the Del algorithm to reject a_2 $(a_2 < a_1)$ variables. In this computer experiment $a_1 = 3, a_2 = 1$.

The number of calculations for the algorithm AdDel is shown in Fig. 6

Fig. 6 demonstrates that the number of calculations increases exponentially with increasing number of variables. This means that the AdDel algorithm demands even more computational resources than algorithms presented in Figs. 4, 5.

Now we consider the selection of the most informative variables with the use of the algorithm based on the optimization of vector $\vec{cs}$.

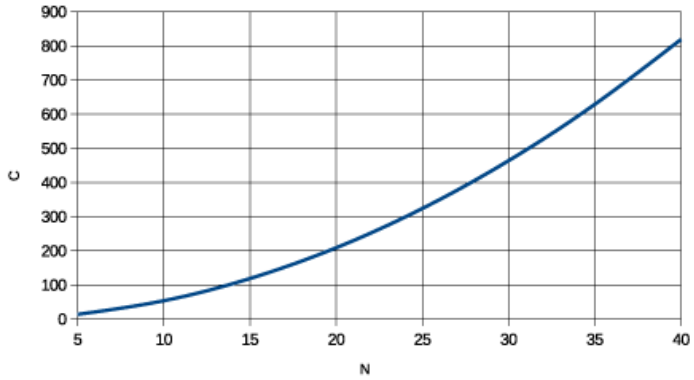

Fig. 5. Number of calculations versus the number of variables for the algorithm Del

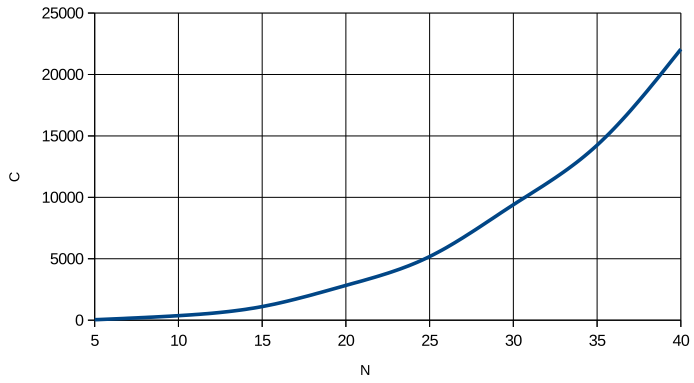

Fig. 6. Number of calculations versus the number of variables for the algorithm AdDel

Let us look more closely at this algorithm. We need to find a vector $\vec{cs}, R(\vec{n}, \vec{cs}) \to 0$, so it is necessary to solve the optimization problem. We use the Neddler-Midd method to solve this problem. The rate of convergence depends on the initial vector $(\vec{cs}_1, \ldots, \vec{cs}_n)$.

The number of calculations for the algorithm based on the optimization of vector $\vec{cs}$ is shown in Fig. 7.

Fig. 7 demonstrates that the number of calculations increases linearly with increasing number of variables for the algorithm based on the optimization of vector $\vec{cs}$. This method requires fewer calculations than methods shown in Figs. 4, 6 in the case when object is characterized by large number of variables.

It is often required to build mathematical model in case of small training sample. To estimate the variable informativeness in such cases is very difficult problem. That is why modified method for estimating the variable informativeness based on the optimization of vector $\vec{cs}$ is proposed.

Variables are divided into groups, according to their informativeness. The degree of informativeness is defined by an expert. Each group of variables has its own coefficient cs_i.

Then optimization problem is simplified because the number of parameters for optimization is reduced. The mathematical model becomes more accurate because it is easier to obtain the decision rule for a small set of parameters than to obtain the decision rule for a large set of parameters.

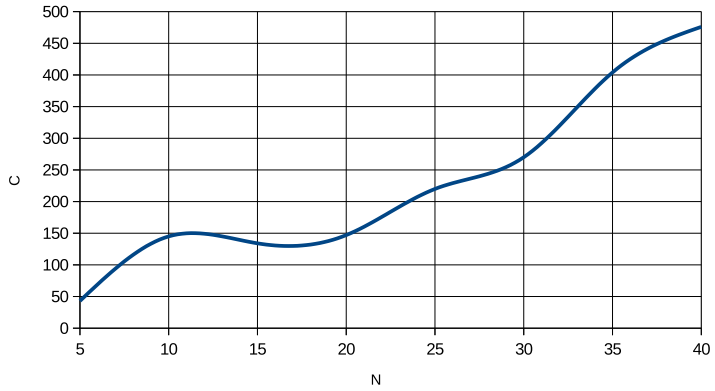

Fig. 7. Number of calculations versus the number of variables for the algorithm based on the optimization of vector $\vec{cs}$

A flaw of this modification is that uninformative variable can be identified as informative variable and vise versa due to the wrong expert decision.

The number of calculations for the modified algorithm based on the optimization of vector $\vec{cs}$ is shown in Fig. 8

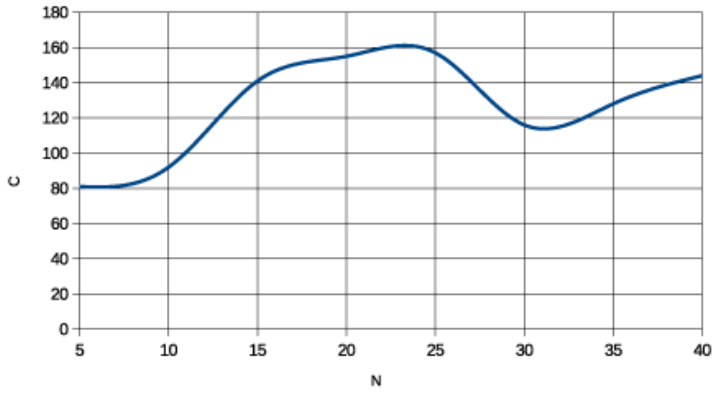

Fig. 8. Number of calculations versus the number of variables for the modified algorithm based on the optimization of vector $\vec{cs}$

Fig. 8 demonstrates that the number of calculations is almost independent of number of variables. This characteristic of the method is very important in the case when an object is characterized by a large number of variables. The accuracy of the forecast is increased for small samples.

There are situations in classification problems when $R(\vec{n}_1, \vec{cs}) \approx 0, R(\vec{n}_2, \vec{cs}) \approx 0, \vec{n}_1 \neq \vec{n}_2$ and $\vec{n}_1 \neq \vec{n}_2$. In other words, classification problem can be solved using different sets of variables.

Conclusion

Various algorithms for estimating the informativeness of variables were considered. They are Ad, Del and AdDel algorithms. Advantages and disadvantages of algorithms are discussed.

The algorithm for estimating the informativeness of variables based on the optimization of vector $\vec{cs}$ and modification of this algorithm were also considered.

The comparative analysis of these algorithms was carried out.

The study was performed by a grant from the Russian Science Foundation (project no. 16-19-10089).

References

[1] N.G.Zagoruyko, Cognitive data's analysis, Novosibirsk, Academic Press GEO, 2013 (in Russian).

[2] A.V.Medvedev, Some notes on H–models for non-inertia systems with a delay, *Vestnik SibGAU*, **54**(2014), no. 5, 24–34 (in Russian).

[3] E.D.Mihov, Parameter's of core's smooth optimization in the case of nonparametric modeling, *Vestnik SibGAU*, **16**(2015), no. 2, 338–342 (in Russian).

Выделение информативных признаков в задаче классиффикации

Евгений Д. Михов
Олег В. Непомнящий
Институт космических и информационных технологий
Сибирский федеральный университет
Киренского, 26, Красноярск, 660041
Россия

В статье представлена проблема классификации многомерных объектов при малой выобрке. Рассмотрены следующие алгоритмы оценки информативности признаков: Ad, Del, AdDel.

Предложен новый алгоритм информативности признаков, который основан на оптимизации вектора коэффициентов размытости ядра, а также его модификация.

Проведен сравнительный анализ существующих методов с предложенным.

Ключевые слова: классификация, малая обучающая выборка, информативность признаков, оптимизация вектора коэффициентов размытости ядра.