ISSN 0005-1179, Automation and Remote Control, 2016, Vol. 77, No. 12, pp. 2212–2226. © Pleiades Publishing, Ltd., 2016.
Original Russian Text © V.V. Tsurko, A.I. Michalski, 2016, published in Avtomatika i Telemekhanika, 2016, No. 12, pp. 136–154.

DATA ANALYSIS

The Contrast Features Selection with Empirical Data

V. V. Tsurko[a] **and A. I. Michalski**[b]

Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia
e-mail: [a] *v.tsurko@gmail.com,* [b] *ipuran@yandex.ru*

Received January 21, 2015

Abstract—The problem of selection the most informative features is reduced to an optimization problem for the average risk functional whose maximization is equivalent to maximization of informational distance between distributions of features in two classes. We consider a maximization procedure for the average risk functional via empirical risk, estimating the divergence between them, with Rademacher complexity. The proposed method has been applied efficiently to problems of selection parameters important to separate the states of technological processes. We show an experimental comparison of the developed approach with other widely known feature selection techniques.

DOI: 10.1134/S0005117916120109

1. INTRODUCTION

As the technologies for data collection and storage are developing, industrial systems become more complex, and the technological basis for biological experiments and mass inspections improves, there arise new problems that machine learning methods can be applied to solve. These problems are characterized by a sharp increase in the number of considered factors, i.e., problem dimension, while the number of experimental observations remains limited.

Technical systems used in research and industry are characterized, due to the ever-increasing complexity, by a large number of sensors and devices intended to control their state and ensure safety. For instance, only one wing edge of a Columbia shuttle hosts 132 inertial sensors that generate data with frequency 20 kHz [1]. A nuclear power plant is a system characterized by high structural complexity, high degree of interaction between elements, and the presence of a human operator in the control circuit.

Modern technological processes for industrial production require tracking a large number of parameters and even human participation in the control over the process state. For reliable monitoring of the state in such systems, real time automated control methods are being developed, and their efficiency to a large extent depends on the number of parameters by which one makes a decision that the system approaches a dangerous state [2]. Reducing the number of controlled parameters can not only improve performance and reliability of automated state diagnostic system but also take some load off the operators servicing various equipment, thus reducing the role of "human factor" by decreasing information flows.

There are three basic reasons why a large number of uninformative features have a negative effect on the quality of a learning algorithm [3, 4]. First, as the number of features increases statistical reliability of the machine learning algorithm deteriorates. As the number of features increases, average training error usually decreases, while the average error on test data unknown to the algorithm first decreases, then passes a minimum point and begins to increase back. Second, as the number of features in a problem grows the number of training set objects required for reliable classification increases as well. Finally, usually an increase in the number of features also greatly increases the running time of a learning algorithm.

The main advantages of selecting informative features include an improvement in the algorithm's accuracy and its generalization power, improving the stability of its operation, better data visualization and interpretation, reduction in dimensionality and costs for data storage, reducing the algorithm's training time and its operation time in real life conditions [4].

Feature selection algorithms are subdivided into three main classes [5]: algorithms embedded into the main training algorithm (embedded approach), methods that serve as a wrapper over the main algorithm, which run the training algorithm several times with different parameters and choose the best version (wrapper approach), and filters that are applied to problem features before the main algorithm runs (filter approach).

Training algorithms that use the embedded feature selection algorithm select important features sequentially during training based on the training set and an internal criterion for the algorithm's accuracy. A classical example of such algorithms is given by logical classification algorithms.

The second class of feature selection approaches is independent of the basic learning algorithm but uses the main algorithm as a subprocedure. These methods enumerate subsets of features and choose a subset on which the accuracy of the base learning algorithm is best.

The third class of feature selection methods contains filtering algorithms that choose important features even before the main learning algorithm begins [6]. The simplest example of a filtering algorithm is to order features in decreasing order of correlation with the target feature and choose k features with the largest resulting correlation. The best known filtering feature selection algorithms include RELIEF [7] and FOCUS [8]. Filtering feature selection algorithms may also be considered to include methods that transform the feature space, e.g., principal components analysis [9] or independent components analysis [10].

There exists a class of methods that select features for which the distance between distributions in the two classes is maximal [11]. One advantage of such methods is that here one selects features that are directly related to the reconstructed target without consideration an essentially intermediate model of feature interaction. To evaluate the distance between distributions in two classes one often uses an informational measure of divergence, e.g., Kullback–Leibler distance [12]. The work [13] uses symmetrized Kullback–Leibler distance between distributions approximated with a mixture of parameterized density functions. The work [14] lists several modifications of the Kullback–Leibler distance that lead to better results. Since actual distribution laws for features in classes are not known, in these methods they are estimated from empirical data with additional hypotheses on the distribution of features. The accuracy of the resulting estimates depends on the sample size and parametric complexity of the class of distributions.

In this work, we propose a feature selection technique based on consideration of the average risk functional whose maximization is equivalent to maximization of symmetrized Kullback–Leibler distance and on selection a subset of features that achieves maximum for an estimate of this functional over empirical data. Unlike previous work [11, 14], in the contrasting feature selection we do not use a parametric representation for the distributions in two classes. A significant novelty of the proposed approach is that we maximize not the empirical risk, i.e., an estimate of informational distance with empirical data, but an estimate of the average risk functional which does not depend on the random sample. We maximize the average risk functional by maximizing empirical risk with a special term that accounts for the value of uniform deviation of empirical risk from the average over the features subset, which lets us find the optimal set of informative features for a given training sample.

2. METHOD OF CONTRASTING DISTRIBUTIONS

The method of contrasting distributions arises in selecting features whose distributions are the most different in comparison of the two classes. Let us describe the problem setting for the choice of

contrasting features as an average risk maximization problem and consider methods for maximizing average risk via empirical risk with an estimate which is adapted to data and uses the Rademacher penalty function.

2.1. Average Risk

Let $X \in \mathbb{R}^n$ be the set of objects, $Y = \{0, 1\}$,—the set of classes, P, an unknown probability distribution on $X \times Y$, (x, y), a random pair from distribution P. Object $x = (x(1), \ldots, x(n))$ is an n-dimensional vector, vector coordinates mean different features, $p(x|y = 0)$ and $p(x|y = 1)$ are distribution densities in the two classes. Let $\varphi_0(x)$ and $\varphi_1(x)$ be functions that estimate the true unknown densities $p(x|y = 0)$ and $p(x|y = 1)$ respectively.

The mean risk functional can be introduced as a characteristic of the measure of distribution functions. For this purpose, we define the loss function

$$\mathrm{L}(x, y, \varphi_0, \varphi_1) = -y \ln \varphi_0(x) - (1 - y) \ln \varphi_1(x) \tag{1}$$

and consider its expectation, the average risk functional

$$\mathrm{M}(\varphi_0, \varphi_1) = -\mathbb{E}_{xy} \left[y \ln \varphi_0(x) + (1 - y) \ln \varphi_1(x) \right], \tag{2}$$

where $\mathbb{E}_{xy}$ denotes expectation of random values x and y.

To justify the form of the proposed mean risk functional, let us consider the Bayesian optimal classification rule that predicts class 0 for objects that satisfy

$$\ln \frac{p(x|y = 0)}{p(x|y = 1)} > t$$

and predicts class 1 otherwise. Here t equals the log ratio of prior probabilities of the two classes

$$t = \ln \frac{P(y = 1)}{P(y = 0)}.$$

The probability of wrong classification for an object from class 0, i.e., type I error, equals here

$$P_0(t) = \int\limits_{x\colon \ln \frac{p(x|y=1)}{p(x|y=0)} > t} p(x|y = 0)dx.$$

If in the construction of the classification rule we replace the distribution in the opposite class $p(x|y = 1)$ with some other distribution $\varphi_1(x)$, the probability of wrong classification for an object from class 0 becomes equal to

$$\hat{P}_0(t, \varphi_1) = \int\limits_{x\colon \ln \frac{\varphi_1(x)}{p(x|y=0)} > t} p(x|y = 0)dx.$$

The difference $\hat{P}_0(t, \varphi_1) - P_0(t)$ shows how much the probability of incorrect classification for an element from class 0 changes if elements from class 1 have distribution $\varphi_1(x)$, and the log ratio of prior probabilities between classes is t. Integrating this difference by parts, we get a characteristic for the change in the type I classification error in class 0 that does not depend on the ratio of prior probabilities for the classes,

$$\begin{aligned} D_0(\varphi_1) &= \int\limits_{-\infty}^{+\infty} \Big(\hat{P}_0(t, \varphi_1) - P_0(t) \Big) dt \\ &= \int\limits_{-\infty}^{+\infty} \left(\ln \frac{\varphi_1(x)}{p(x|y = 0)} - \ln \frac{p(x|y = 1)}{p(x|y = 0)} \right) p(x|y = 0)dx. \end{aligned}$$

With similar considerations, we can get a characteristic for the change in type I classification error in class 1, replacing distribution $p(x|y=0)$ with distribution $\varphi_0(x)$

$$D_1(\varphi_0) = \int\limits_{-\infty}^{+\infty} \left(\hat{P}_1(t,\varphi_0) - P_1(t)\right) dt$$
$$= \int\limits_{-\infty}^{+\infty} \left(\ln \frac{\varphi_0(x)}{p(x|y=1)} - \ln \frac{p(x|y=0)}{p(x|y=1)}\right) p(x|y=1)dx.$$

A similar interpretation of informational distance has been given in [15]. Let us transform each term in the average risk functional

$$\mathbb{E}_{xy}\left[y \ln \varphi_0(x)\right] = \int\limits_{-\infty}^{+\infty} \ln \varphi_0(x) p(x|y=1)dx$$
$$= D_1(\varphi_0) - \int\limits_{-\infty}^{+\infty} \ln p(x|y=0) p(x|y=1)dx,$$
$$\mathbb{E}_{xy}\left[(1-y) \ln \varphi_1(x)\right] = \int\limits_{-\infty}^{+\infty} \ln \varphi_1(x) p(x|y=0)dx$$
$$= D_0(\varphi_1) - \int\limits_{-\infty}^{+\infty} \ln p(x|y=1) p(x|y=0)dx.$$

As a result, we get the following expression:

$$\mathrm{M}(\varphi_0,\varphi_1) = -\mathbb{E}_{xy}\left[y \ln \varphi_0(x) + (1-y) \ln \varphi_1(x)\right]$$
$$= \mathbb{E}_{xy}\left[y \ln p(x|y=0) + (1-y) \ln p(x|y=1)\right]$$
$$- P(y=1) D_1(\varphi_0) - P(y=0) D_0(\varphi_1).$$

Thus, maximizing the average risk functional with respect to the distribution exactly corresponds to minimizing the sum of type I errors in two classes with weights equal to prior probabilities of the classes.

To select informative features, we use the average risk functional among distributions of sets of n features. We denote the class of all possible such distributions by Φ_n. The maximal value of average risk functional attainable in this class shows how well one can theoretically solve the classification problem with these features. To design a constructive algorithm for maximizing the average risk functional with empirical data, let us consider specific elements from the class Φ_n, namely Bayesian estimates of n-dimensional histograms.

Suppose that the values of coordinates $x(j)$ are divided into τ_j intervals; then $k = \prod_{j=1}^{n} \tau_j$ is the number of intervals in the n-dimensional histogram, $\sigma_1, \ldots, \sigma_k$ are the n-dimensional interval for subdividing the values of x. If the prior distribution of probabilities is uniform on the k-dimensional simplex, then $\varphi_y^b(x)$, a Bayesian estimate of the n-dimensional histogram of the distribution in the class y, is defined by the following formula from [3, p. 63]

$$\varphi_y^b(x) = \sum_{i=1}^{k} \mathbb{I}\{x \in \sigma_i\} \frac{n_i^y + 1}{\ell_y + k},$$

where $y = 0, 1$, $\mathbb{I}\{x \in \sigma_i\}$ is the indicator which equals one if x belongs to the interval σ_i and zero otherwise, ℓ_y is the size of an independent sample from class y, n_i^y is the number of sample elements from class y that fall into the σ_i bin of the histogram.

It is clear that

$$\sup_{\varphi_0,\varphi_1\in\Phi_n} \mathrm{M}(\varphi_0,\varphi_1) \geqslant \mathrm{M}(\varphi_0^b,\varphi_1^b),$$

and therefore, maximizing the average risk on Bayesian estimates of n-dimensional histograms, we choose the set of features where it is possible to construct a rule with a small probability of incorrect classification. Instead of Bayesian histogram estimates we could use other distributions, but for average risk maximization with empirical data Bayesian estimate histograms allow to efficiently find an estimate of the uniform deviation of empirical risk from the average, which is used in the informative feature selection algorithm.

2.2. Empirical Risk

Since the distributions of vector x in two classes are not known, we cannot immediately maximize the average risk functional. Instead, we can use its estimate based on experimental data, namely empirical risk [3].

Let $x_1^y,\ldots,x_{\ell_y}^y$ be a sample from class y, $y=0,1$. Then a Bayesian estimate for the probability to fall into the ith interval for each class has the form

$$\varphi_y^b(i)=\frac{n_i^y+1}{\ell_y+k},$$

where $n_i^y=\sum_{j=1}^{\ell_y}\mathbb{I}\{x_j^y\in\sigma_i\}$, $y=0,1$.

The following constraints hold:

$$0<c\leqslant\varphi_y^b(i),\quad i=1,\ldots,k,\quad c=\frac{1}{k+\max(\ell_0,\ell_1)}; \tag{3}$$

$$\sum_{i=1}^{k}\varphi_y^b(i)=1,\quad y=0,1. \tag{4}$$

Empirical risk is the average penalty function over the sample

$$\mathrm{M}_e(\varphi_0^b,\varphi_1^b)=-\frac{1}{\ell_0+\ell_1}\left(\sum_{i=1}^{k}n_i^1\ln\varphi_0^b(i)+\sum_{i=1}^{k}n_i^0\ln\varphi_1^b(i)\right). \tag{5}$$

It is easy to check that the expectation of the empirical risk for given functions $\varphi_0^b(x)$ and $\varphi_1^b(x)$ equals the average risk value for them, and as the number of observations ℓ_0 and ℓ_1 grows to infinity, the value of the empirical risk with probability one converges to the average risk value. In the considered case, functions $\varphi_0^b(x)$ and $\varphi_1^b(x)$ are not fixed but rather determined by the random sample. To account for this fact in average risk maximization, in order to get a correct estimate for the risk we have to correct the value of the empirical risk [3]. One approach to such a correction is to introduce Rademacher complexity.

2.3. Rademacher Complexity

Rademacher complexity is a measure of complexity for a class of real functions, introduced to statistical learning theory by V. Kolchinskii in 1999. It can be interpreted as the maximal covariance of functions from this class with random (Rademacher) noise [16]. The more complex a set of functions is, the higher are chances to find there a function which is similar to arbitrary random noise and get an empirical risk value which is significantly different from the average risk value for this function. We can account for this with an additional terms in the formulas, the Rademacher penalty. To get an expression for the Rademacher penalty in the problem at hand, we

represent the empirical risk functional as

$$\mathrm{M}_e(\varphi_0^b,\varphi_1^b)=-\frac{1}{\ell_0+\ell_1}\left(\sum_{i=1}^{\ell_1}\ln\varphi^b_{0,x_i^1}+\sum_{i=1}^{\ell_0}\ln\varphi^b_{1,x_i^0}\right),$$

where φ^b_{y,x_i^t} is the Bayesian estimate of the probability of a sample element of class y to fall into the same histogram interval where element x_i^t of sample t has fallen.

Let $\delta_1^0,\ldots,\delta_{\ell_0}^0,\delta_1^1,\ldots,\delta_{\ell_1}^1$ be a sequence of independent identically distributed random values that take values $+1$ and (-1) with probability $1/2$ and independently of the sample $(x_1^0,\ldots,x_{\ell_0}^0,$ $x_1^1,\ldots,x_{\ell_1}^1)$, called a Rademacher sequence, and let F be the class of Bayesian histogram estimates constructed for all possible subsets of the set of features.

The Rademacher penalty is defined, following [16], as

$$R(F)=\sup_{\varphi_0^b,\varphi_1^b\in F}\left|\frac{1}{\ell_0+\ell_1}\left(\sum_{i=1}^{\ell_1}\delta_i^1\ln\varphi^b_{0,x_i^1}+\sum_{i=1}^{\ell_0}\delta_i^0\ln\varphi^b_{1,x_i^0}\right)\right|.$$

Summing up variables δ_i^y corresponding to the same intervals, we get

$$R(F)=\sup_{\varphi_0^b,\varphi_1^b\in F}\left|\frac{1}{\ell_0+\ell_1}\left(\sum_{i=1}^{k}\Delta_i^1\ln\varphi_0^b(i)+\Delta_i^0\ln\varphi_1^b(i)\right)\right|, \tag{6}$$

where $\Delta_i^y=\sum_{j=1}^{\ell_y}\delta_j^y\mathbb{I}\{x_j^y\in\sigma_i\}$, σ_i is the ith interval of the n-dimensional histogram, $y=0,1$.

Note that by construction it holds that $\sum_{i=1}^k\Delta_i^y\leqslant\ell_y$. We denote $\Delta^y=(\Delta_1^y,\ldots,\Delta_k^y)$ and $\bar y=1-y$.

The following lemma and theorem let us find the value of Rademacher penalty in the class of Bayesian histogram estimates.

Lemma. *In the class of Bayesian histogram estimates F, the value*

$$Q(F,\Delta^y)=\max_{\varphi_{\bar y}^b\in F}\sum_{i=1}^{k}\Delta_i^y\ln\varphi_{\bar y}^b(i)$$

equals:

- *if $\exists t:\Delta_t^y<0$ and $\Delta_i^y\leqslant 0,\ i=\overline{1,k}$, then for $j=\arg\max_i\Delta_i^y$*

$$Q(F,\Delta^y)=\Delta_j^y\ln\left(1-c(k-1)\right)+\sum_{i=1,i\neq j}^{k}\Delta_i^y\ln c,\quad c=(k+\ell_y)^{-1},$$

- *if $\Delta_i^y\leqslant 0,\ i=\overline{1,s}$ and $\Delta_i^y>0,\ i=\overline{s+1,k}$, then*

$$Q(F,\Delta^y)=\sum_{i=1}^{s}\Delta_i^y\ln c+\sum_{i=s+1}^{k}\Delta_i^y\ln\frac{\Delta_i^y(1-cs)}{\sum_{j=s+1}^k\Delta_j^y},\quad c=(k+\ell_y)^{-1}.$$

Theorem. *In the class of Bayesian histogram estimates F, the Rademacher penalty is computed as*

$$R(F)=\frac{1}{\ell_0+\ell_1}\max\left\{Q(F,\Delta^0)+Q(F,\Delta^1);Q(F,-\Delta^0)+Q(F,-\Delta^1)\right\}. \tag{7}$$

Proofs of the lemma and the theorem are given in the Appendix.

2.4. An Average Risk Estimate Based on Rademacher Complexity

Values of the Rademacher penalty and empirical risk are used to get an uniform estimate for the average risk functional with an inequality obtained by Kolchinskii in [16]. For a class of loss

functions uniformly bounded by a constant U and for every $t > 0$ it holds that

$$P\left\{\sup_{f\in F}|\mathrm{M}(f)-\mathrm{M}_e(f)|\geqslant 2R(F)+\frac{3tU}{\sqrt{\ell_0+\ell_1}}\right\}\leqslant\exp\left(-\frac{t^2}{2}\right).$$

Applying this inequality in the class of Bayesian histogram estimates leads to the following statement.

Statement. *In the class of Bayesian histogram estimates F, with probability at least $1-\eta$ it holds that*

$$\mathrm{M}(\varphi_0^b,\varphi_1^b)>\mathrm{M}_e(\varphi_0^b,\varphi_1^b)-2R(F)-\frac{3\sqrt{-2\ln\eta}\ln\left(k+\max\left(\ell_0,\ell_1\right)\right)}{\sqrt{\ell_0+\ell_1}}. \tag{8}$$

Proof of the statement is given in the Appendix.

2.5. Algorithm for Contrasting Features Selection

The algorithm for selection of contrasting features (the contrasting distributions algorithm) solves the problem of finding such a subset of features that Bayesian estimate histograms $\varphi_0^b(x)$ and $\varphi_1^b(x)$ maximize the average risk functional estimate constructed with empirical data. This corresponds to looking for features whose distributions in the two classes are maximally different in terms of cross-entropy.

The contrasting method is adapted to the problem with a method similar to structural risk minimization proposed by Vapnik and Chervonenkis in [3].

Suppose that the object $x=(x(1),\ldots,x(n))$ consists of n features (coordinates), $C_n=\{1,\ldots,n\}$ is the set of all features, and we will construct a vector $\pi_{C_m}(x)$ consisting of a subset of $C_m\subseteq C_n$ coordinates of vector x.

The contrasting algorithm consists of two stages: on the first stage we construct an ordered sequence of subsets of features, and on the second stage choose a subset that maximizes the average risk functional estimate. The first stage consists of n steps.

On the first step, we enumerate all features one by one and choose feature i that maximizes empirical risk (5):

$$i=\arg\max_{j=1,\ldots,n}\mathrm{M}_e\left(\varphi_0^b\left(\pi_{\{j\}}(x)\right),\varphi_1^b\left(\pi_{\{j\}}(x)\right)\right).$$

The selected feature i is included into the first constructed subset in sequence $C_1=\{i\}$.

On the second step, we enumerate all possible pairs of features where one feature has been fixed on the previous step, and the other is different. We choose a pair of features $\{i,j\}$ that maximizes empirical risk:

$$j=\arg\max_{d=1,\ldots,n,\,d\neq i}\mathrm{M}_e\left(\varphi_0^b\left(\pi_{\{i,d\}}(x)\right),\varphi_1^b\left(\pi_{\{i,d\}}(x)\right)\right).$$

We construct the subset of features $C_2=\{i,j\}$ that satisfies $C_1\subset C_2$.

On the third step, we enumerate all possible triples of features, two of which have been fixed on previous steps, and the third is different. We construct a triple $\{i,j,d\}$ that maximizes empirical risk and the corresponding subset $C_3=\{i,j,d\}$, $C_1\subset C_2\subset C_3$.

We repeat the process on subsequent steps, stopping when all features have been considered and we have constructed a sequence of feature subsets $C_1\subset C_2\subset C_3\subset\ldots\subset C_n$. After constructing the sequence of subsets we proceed to the second stage of the algorithm, maximizing the average risk estimate.

On the second stage of the algorithm, for every subset of C_i out of the constructed sequence we compute empirical risk (5) and average risk estimate defined by the right-hand side of inequality (8).

The algorithm chooses a subset of features C_i that maximizes the average risk estimate

$$C_i\colon\, i = \arg\max_{j=1,\ldots,n} \mathrm{M}\left(\varphi_0^b\left(\pi_{C_j}(x)\right), \varphi_1^b\left(\pi_{C_j}(x)\right)\right).$$

3. APPLICATIONS OF THE CONTRASTING DISTRIBUTIONS ALGORITHM

3.1. Simulation Experiment

To test the operation of the proposed feature selection algorithm, we have applied it to simulated data. We considered a two-class problem with 500 objects in each class, each object characterized by 100 features. 98 features had identical normal distributions in the two classes; distribution parameters were chosen randomly for every feature, with the mean lying in the interval from 10 to 20 and variance from 1 to 100. The other two informative features also had a normal distribution, but with significantly different means in the two classes. In the experiments, means of distributions for informative features in two classes were made sequentially closer and closer to each other in order to study the separability of the two classes with the method of contrasting distributions for small differences between two classes.

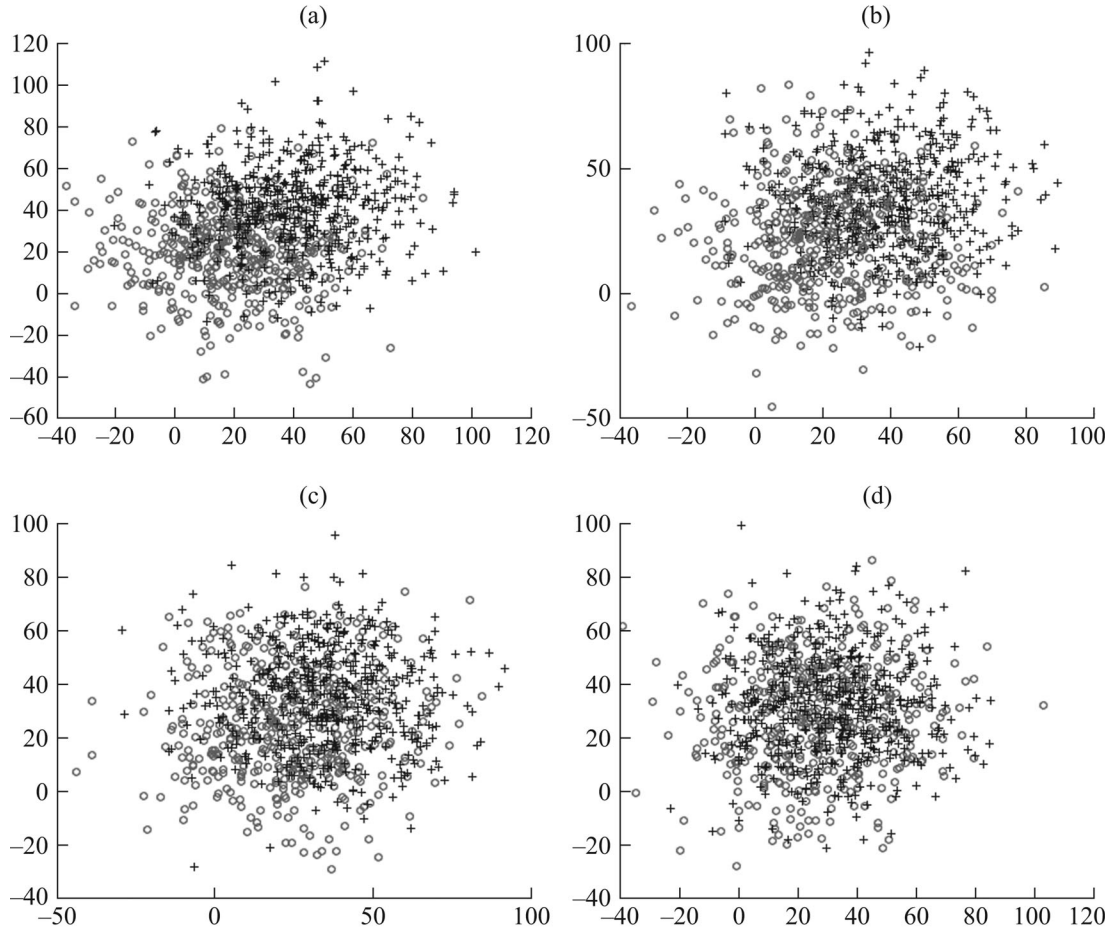

Fig. 1. Model experiment. Features with different distributions in classes: (a) centers of classes (20, 20) and (40, 40), (b) centers of classes (22.5; 22.5) and (37.5; 37.5), (c) centers of classes (25, 25) and (35, 35), (d) centers of classes (27.5; 27.5) and (32.5; 32.5).

Figure 1 shows an illustration of the four experiments we have conducted. The plots show objects of two classes, and the vertical and horizontal axes show the values of the first and second informative features in question.

On Fig. 1a objects of two classes are maximally separable: two features of the first class have a normal distribution centered at point (20, 20), and two features of the second class have a normal distribution centered at (40, 40), with variances equal to 20. In this experiment, the contrasting distributions algorithm successfully finds the two informative features out of 100, 98 of which have identical distribution laws.

Figures 1b and 1c illustrate experiments where two informative features had distributions centered at (22.5; 22.5) and (37.5; 37.5) in the second experiment and (25, 25) and (35, 35) in the third experiment. In these experiments, the contrasting distributions algorithm also successfully found the two features in question.

Figure 1d illustrates the fourth experiment, where centers of the classes are maximally close to each other and equal to the points (27.5; 27.5) and (32.5; 32.5), with variance of 20. In this case, the classes are already hard to separate visually, and the contrasting distributions algorithm also does not find the necessary pair of features out of 100 available ones.

3.2. Feature Selection for an Industrial Process Control

The contrasting distribution algorithm has been applied to analyze real life data on the state of an industrial process made available by an industrial software developer company under NDA. The data represent readings of 10 process parameters in time. Parameters have real values. Data have been divided by experts into two classes corresponding to the two states of system operation. The first class includes 562 measurements of ten parameters; the second class, 258 measurements of the same parameters.

The problem reduced to finding a subset of features that maximize the average risk functional, and it was solved with the method of contrasting distributions. Here the values of each parameter measured at different time moments were assumed to be independent, unlike parameters measured at the same time moment.

We applied the method of contrasting distributions to the data. Results obtained by the contrasting feature selection algorithm were compared with results obtained by other well-known feature selection algorithms. In particular, we have applied to the data the RELIEF algorithm [7], two feature selection techniques based on correlations, namely CFS [17] and IBM SPSS Modeler feature selection [18], principal components analysis [9], and independent components analysis [10]. To choose the best set of features and test the quality of the result, we used 10-fold cross-validation: the entire sample was divided into 10 disjoint blocks of nearly equal size (up to rounding). The partition was stratified, i.e., we preserved the same proportions of classes in the blocks as in the entire sample. Each block, in turn, became the test set, feature selection and transformation algorithms were applied to the training set composed of the other nine blocks, and after feature selection (or

Table 1. A comparison of feature selection results for different algorithms in Experiment 2

Feature selection algorithm	Subsets of features	Percent of correctly classified objects ± CI, %
Contrasting distributions	**10, 1**	**93.56 ± 3.55**
RELIEF	10, 2, 7, 3, 5, 1, 8, 6	91.1 ± 4.63
Feature selection based on correlation, CFS	4, 5, 6, 10	91.1 ± 3.14
Feature selection based on correlation, IBM	3, 7	73.05 ± 4.67
Principal components analysis		89.51 ± 2.9
Independent components analysis		87.68 ± 1.27

transformation) the data were classified by a naive Bayes classifier whose quality was evaluated on the test set. Table 1 shows a comparison of the classification quality results on sets of features obtained with different algorithms.

A comparison of classification accuracy shows that the best results are achieved on the set of features selected by the algorithm of contrasting distributions. An important advantage of the proposed algorithm is also the fact that it has selected a smaller number of features than other algorithms.

3.3. Feature Selection for Semiconductor Production Control

The algorithm of contrasting distributions has been applied to a problem from the open library of machine learning tasks UCI [19]. The objects of this problem are produces semiconductors, 1567 objects in total, the features are 590 signals detected by sensors in the production of semiconductors and used to control the quality of the resulting final product. All objects have been divided into two classes: faulty items and non-faulty items. The problem was to select features by which one can reliably predict whether an object belongs to one of the two classes.

To solve this problem and demonstrate the efficiency of the methods proposed in this work, we have implemented a scheme completely similar to the one described in Section 3.2: first various feature selection algorithms have been applied to the data, then a naive Bayes classifier was applied to the selected features, and its quality was evaluated by 10-fold cross-validation. Table 2 shows a comparison of the results of different feature selection algorithms on this problem.

Table 2. A comparison of feature selection algorithms in Experiment 3

Feature selection algorithm	Subsets of features	Percent of correctly classified objects ± CI, %
Contrasting distributions	**406, 56, 410, 177**	**93.17 ± 0.56**
Principal components analysis		91.77 ± 1.93
Independent components analysis		85.77 ± 2.36
No feature selection	1–590	79.45

4. CONCLUSION

In this work, we have proposed an algorithm for choosing a subset of features based on the informational distance between distributions of features in classes. We have constructed an average risk functional that formalizes this distance, but real distribution laws for objects in classes are not known, so average risk is estimated by empirical data. We have considered estimate that adapt to data and are based on Rademacher complexity. In this work, we have obtained the exact value of the Rademacher penalty term for a class of multidimensional histograms and showed a lower bound on the considered average risk functional with Rademacher penalty.

The process of selection a set of features that maximize an average risk estimate has been formalized in the algorithm of contrasting distributions. The algorithm has been applied to selecting informative features in a model problem and also to data from two real life control problems for industrial processes; for one of the problems, the data were taken from an open repository of machine learning datasets. In all problems, the algorithm of contrasting distributions has shown good results and high accuracy, exceeding other popular feature selection techniques.

When estimating the uniform deviation of empirical risk from the average in the class of Bayesian histogram estimates, as an alternative to Rademacher complexity we have also considered the VC dimension introduced by Vapnik and Chervonenkis [3], which is based on the combinatorial

dimension of a class of functions. When choosing parameters related to the state of the system described in the paper, the latter approach leads to overly high estimates, and as a result the right-hand side of inequality (8) turns out to be negative already for two parameters. This follows from a relatively small amount of experimental data and the universality of the method of Vapnik and Chervonenkis which does not require one to solve an additional maximization problem, as in the case of computing Rademacher complexity. For sufficiently large amounts of data, this method does lead to reliable results, and an example of its successful application for finding diseases that accompany cancer is given in [20].

By introducing Rademacher complexity, we have obtained an estimate for the uniform deviation of empirical risk from the average for a smaller number of observations, but this is achieved at the cost of losing universality of the approach since we now have to compute the supremum of the mean value of the loss function on a Rademacher sequence of random numbers. For the class of Bayesian histogram estimates, the value of this supremum is given by the theorem, but in the general case its computing may represent a nontrivial problem.

APPENDIX

Proof of Lemma. Consider the case when all coefficients are nonpositive:

$$\exists t\colon \Delta_t^y < 0, \quad \Delta_i^y \leqslant 0, \quad i = 1, \ldots, k.$$

Consider the function

$$r(\varphi_{\bar{y}}^b) = r(\varphi_{\bar{y}}^b(1), \ldots, \varphi_{\bar{y}}^b(k)) = \sum_{i=1}^{k} \Delta_i^y \ln \varphi_{\bar{y}}^b(i).$$

Since coordinates of the gradient $\nabla r(\varphi_{\bar{y}}^b) = \left(\frac{\Delta_1^y}{\varphi_{\bar{y}}^b(1)}, \ldots, \frac{\Delta_k^y}{\varphi_{\bar{y}}^b(k)}\right)$ are negative, function $r(\varphi_{\bar{y}}^b)$ is maximized at a vertex of the simplex

$$\Gamma = \left\{ \varphi_{\bar{y}}^b(i) : 0 < c \leqslant \varphi_y^b(i), i = 1, \ldots, k; \sum_{i=1}^{k} \varphi_{\bar{y}}^b(i) = 1 \right\}.$$

Vertices of simplex Γ are points with k coordinates, and values of $k-1$ coordinates equal c, and the value of one of the coordinates equals $1 - c(k-1)$. The simplex vertex number j has the form $v_j = (c, \ldots, 1 - c(k-1), \ldots, c)$, where the coordinate which is not equal to c occupies the jth position.

Therefore, function $r(\varphi_{\bar{y}}^b)$ at vertex v_j takes value

$$r(v_j) = \sum_{i=1, i \neq j}^{k} \Delta_i^y \ln c + \Delta_j^y \ln (1 - c(k-1)).$$

It is clear that

$$\max_{\varphi_{\bar{y}}^b \in F} r\left(\varphi_{\bar{y}}^b\right) = \max_{j=1,\ldots,k} r(v_j).$$

Then we choose as j the index corresponding to the maximal value of Δ_j^y and find that

$$Q(F, \Delta^y) = \sum_{i=1, i \neq j}^{k} \Delta_i^y \ln c + \Delta_j^y \ln (1 - c(k-1)),$$

where $j = \arg\max_i \Delta_i^y$.

Consider the case when coefficients Δ_i^y take values of arbitrary sign. Without loss of generality we will assume that coefficients are ordered in nondecreasing order:

$$\Delta_1^y \leqslant \ldots \leqslant \Delta_s^y \leqslant 0 < \Delta_{s+1}^y \leqslant \ldots \leqslant \Delta_k^y.$$

We construct the Lagrangian

$$L(\varphi_{\bar{y}}, \lambda, \mu) = \sum_{i=1}^{k} \Delta_i^y \ln \varphi_{\bar{y}}^b(i) - \lambda \left(\sum_{i=1}^{k} \varphi_{\bar{y}}^b(i) - 1 \right) + \mu^{\mathrm{T}} \left(\varphi_{\bar{y}}^b(i) - c \right),$$

$$\frac{\partial L}{\partial \varphi_{\bar{y}}^b(i)} = \frac{\Delta_i^y}{\varphi_{\bar{y}}^b(i)} - \lambda + \mu_i, \quad \mu_i \geqslant 0,$$

where $\mu = (\mu_1, \ldots, \mu_k)$ are Lagrange multipliers, and T denotes transposition.

Let us find the critical points of the Lagrange system:

$$\frac{\Delta_i^y}{\varphi_{\bar{y}}^b(i)} - \lambda + \mu_i = 0, \quad i = 1, \ldots, k,$$

$$\sum_{i=1}^{k} \varphi_{\bar{y}}^b(i) = 1,$$

$$\mu_i \left(\varphi_{\bar{y}}^b(i) - c \right) \geqslant 0, \quad i = 1, \ldots, k,$$

$$\mu_i \geqslant 0, \quad \varphi_{\bar{y}}^b \geqslant c, \quad i = 1, \ldots, k.$$

Let J_c be the set of indices for which function $\varphi_{\bar{y}}^b(i)$ equals constant c:

$$J_c = \left\{ j \colon \varphi_{\bar{y}}^b(j) = c \right\}.$$

Let $N = |J_c|$ be the power of the set J_c, I_c is the complement of the set J_c.

In this notation, we have

$$\varphi_{\bar{y}}^b(j) = c, \quad \mu_j = \lambda - \frac{\Delta_i^y}{c}, \quad \mu_j \geqslant 0, \quad j \in J_c,$$

$$\varphi_{\bar{y}}^b(i) = \frac{\Delta_i^y}{\lambda}, \quad \varphi_{\bar{y}}^b(i) \geqslant c, \quad \mu_i = 0, \quad i \in I_c.$$

We express the value of λ using the fact that the sum of $\varphi_{\bar{y}}^b(i)$ equals one:

$$cN + \frac{1}{\lambda} \sum_{i \in I_c} \Delta_i^y = 1,$$

$$\lambda = \frac{1}{1 - Nc} \sum_{i \in I_c} \Delta_i^y.$$

All $\varphi_{\bar{y}}^b(i)$ cannot be equal to c simultaneously since $ck = k/(k+l) < 1$; consequently, it holds that $N < k$ and $I_c \neq \emptyset$. The set I_c is nonempty, and the first equality together with $cN < 1$ implies that $\sum_{i \in I_c} \Delta_i^y \neq 0$, so parameter λ is defined in all possible stationary points.

To solve the problem, we have to find such sets J_c for which the Lagrange system will be feasible. Let us study the solution with regard to the imposed inequality constraints, $\mu_j \geqslant 0$, $j \in J_c$

and $\varphi^b_{\bar{y}}(i) \geqslant c, i \in I_c$; we get that

$$\mu_j = \frac{1}{1-cN}\sum_{i\in I_c}\Delta^y_i - \frac{\Delta^y_j}{c} \geqslant 0, \quad j \in J_c,$$

$$\varphi^b_{\bar{y}}(i) = \Delta^y_i(1-cN)\left(\sum_{i\in I_c}\Delta^y_i\right)^{-1} \geqslant c, \quad i \in I_c.$$

If $\sum_{i\in I_c}\Delta^y_i < 0$, we get that

$$\Delta^y_j \leqslant \frac{c}{1-cN}\sum_{i\in I_c}\Delta^y_i < 0, \quad j \in J_c,$$

$$\Delta^y_i \leqslant \frac{c}{1-cN}\sum_{i\in I_c}\Delta^y_i < 0, \quad i \in I_c.$$

Then Δ^y_i is negative, and this case has been considered in the beginning of the proof. If $\sum_{i\in I_c}\Delta^y_i>0$, we get that

$$\Delta^y_j \leqslant \frac{c}{1-cN}\sum_{i\in I_c}\Delta^y_i, \quad j \in J_c, \tag{A.1}$$

$$\Delta^y_i \geqslant \frac{c}{1-cN}\sum_{i\in I_c}\Delta^y_i > 0, \quad i \in I_c. \tag{A.2}$$

The set I_c does not contain indices of negative coefficients Δ^y_i. We let $J_c = \{1,\dots,s\}$ and show that no coefficient can either be excluded from J_c or added to J_c.

Let us try to add index $s+1$ to the set J_c. Let $J_c = \{1,\dots,s+1\}$ and $I_c = \{s+2,\dots,k\}$, and let us check inequalities (A.1), (A.2). By assumption $\Delta^y_{s+1} > 0$, consequently, by construction $\Delta^y_{s+1} \geqslant 1$. Also by construction, coefficients $\sum_{i\in I_c}\Delta^y_i \leqslant \sum_{i=1}^k \Delta^y_i - \Delta^y_{s+1} \leqslant \ell - 1 < \ell$ and condition $k > s$ implies the following chain of inequalities:

$$\frac{c}{1-cN}\sum_{i\in I_c}\Delta^y_i < \frac{c\ell}{1-cN} = \frac{\ell}{\ell+k-s-1} \leqslant 1 \leqslant \Delta^y_{s+1}.$$

Hence, we have obtained $(s+1) \in I_c$. Larger indices $s+2,\dots,k$ cannot occur in J_c since (A.1) and (A.2) imply that if $i \in I_c$, $j \in J_c$ then $\Delta^y_j \leqslant \Delta^y_i$.

Thus, we conclude that constraints (A.1), (A.2) hold if and only if $J_c = \{1,\dots,s\}$ and $I_c = \{s+1,\dots,k\}$.

The problem's solution will have the following form:

$$\varphi_{\bar{y}}(j) = c, \quad j = 1,\dots,s,$$

$$\varphi_{\bar{y}}(i) = \frac{\Delta^y_i(1-sc)}{\sum_{t=s+1}^k \Delta^y_t}, \quad i = s+1,\dots,k.$$

Then the maximal value of the function equals

$$Q(F,\Delta^y) = \sum_{i=1}^s \Delta^y_i \ln c + \sum_{i=s+1}^k \Delta^y_i \ln\frac{\Delta^y_i(1-cs)}{\sum_{j=s+1}^k \Delta^y_j},$$

as needed.

Proof of Theorem. Expression (7) results from formula (6) by taking the supremum of absolute value as $\sup|A| = \max(\sup A; \sup -A)$ and replacing the supremum with maximum that can be computed according to the lemma.

Proof of Assertion. Bayesian histogram estimates satisfy the following inequalities:

$$1/(\max(\ell_0,\ell_1)+k) \leqslant \varphi_y^b < 1;$$

this implies that the class of loss functions (1) is uniformly bounded:

$$\begin{aligned}\left|\mathrm{L}(x,y,\varphi_0^b,\varphi_1^b)\right| &= \left|y\ln\varphi_0^b(x)+(1-y)\ln\varphi_1^b(x)\right| \\ &\leqslant \max\left(|\ln\varphi_0^b|,|\ln\varphi_1^b|\right) \leqslant \ln(\max(\ell_0,\ell_1)+k) = U.\end{aligned}$$

Let us fix the probability of satisfying the inequality with value $\eta = \exp(-t^2/2)$, then it will hold that

$$P\left\{\sup_{\varphi_0^b,\varphi_1^b\in F}\left|\mathrm{M}(\varphi_0^b,\varphi_1^b)-\mathrm{M}_e(\varphi_0^b,\varphi_1^b)\right| < 2R(F)+\frac{3\sqrt{-2\ln\eta}\ln(\max(\ell_0,\ell_1)+k)}{\sqrt{\ell_0+\ell_1}}\right\} > 1-\eta.$$

Resolving this inequality, we get that with probability at least $1-\eta$ a lower bound on the mean risk functional can be represented as

$$\mathrm{M}(\varphi_0^b,\varphi_1^b) > \mathrm{M}_e(\varphi_0^b,\varphi_1^b) - 2R(F) - \frac{3\sqrt{-2\ln\eta}\ln(k+\max(\ell_0,\ell_1))}{\sqrt{\ell_0+\ell_1}},$$

as needed. This completes the proof of the statement.

REFERENCES

1. Iverson, D.L., Data Mining Applications for Space Mission Operations System Health Monitoring, *Proc. SpaceOps 2008 Conf.*, ESA, EUMETSAT, AIAA, Heidelberg, Germany, May 2008.
2. Kostyukov, V.N. and Naumenko, A.P., Analysis of Modern Methods and Means for Monitoring and Diagnostics of Pneumatic Pumps. Part 1. Online Monitoring Systems, *V Mire NK*, 2010, no. 1 (47), pp. 64–70.
3. Vapnik, V.N. and Chervonenkis, A.Ya., *Teoriya raspoznavaniya obrazov* (Image Recognition Theory), Moscow: Nauka, 1974.
4. Wolf, L. and Shashua, A., Features Selection for Unsupervised and Supervised Inference: The Emergence of Sparsity in a Weight-Based Approach, *J. Machine Learning Res.*, 2005, vol. 6, pp. 1855–1887.
5. Blum, A. and Langley, P., Selection of Relevant Features and Examples in Machine Learning, *AI*, 1997, vol. 97, pp. 245–271.
6. John, G.H., Kohavi, R., and Pfleger, K., Irrelevant Features and the Subset Selection Problem, *Proc. 11th Int. Conf. on Machine Learning*, Morgan Kaufmann Publishers, 1994, pp. 121–129.
7. Kira, K. and Rendell, L., The Feature Selection Problem: Traditional Methods and a New Algorithm, *10th National Conf. on Artificial Intelligence*, Cambridge: MIT Press, 1992, pp. 129–134.
8. Allmuallim, H. and Dietterich, T.G., Learning with Many Irrelevant Features, *Proc. 9th National Conf. on Artificial Intelligence*, San Jose: AAAI, 1991, pp. 547–552.
9. Jolliffe, I.T., *Principal Component Analysis*, New York: Springer-Verlag, 1986.
10. Comon, P., Independent Component Analysis. A New Concept, *Signal Process.*, 1994, vol. 36, pp. 287–314.
11. Koller, D. and Sahami, M., Toward Optimal Feature Selection, *Proc. 13th Int. Conf. on Machine Learning*, Morgan Kaufmann Publishers, 1996, pp. 284–292.
12. Kullback, S. and Leibler, R.A., On Information and Sufficiency, *Annals Math. Statist.*, 1951, vol. 22, no. 1, pp. 79–86.
13. Novovicova, J., Pudil, P., and Kittler, J., Divergence Based Feature Selection for Multimodal Class Densities, *IEEE Trans. Patt. Anal. Machine Intelligen.*, 1996, vol. 18 (2), pp. 218–223.

14. Coetzee, F.M., Correcting Kullback–Leibler Distance for Feature Selection, *Patt. Recognit. Lett.*, 2005, vol. 26, no. 11, pp. 1675–1683.

15. Eguchi, S. and Copas, J., Interpreting Kullback–Leibler Divergence with the Neyman–Pearson Lemma, *J. Multivariate Anal.*, 2006, vol. 97, pp. 2034–2040.

16. Koltchinskii, V. and Panchenko, D., Rademacher Process and Bounding the Risk of Function Learning, in *High Dimension. Probab. II*, Gine, D.E., Wellner, J., Eds., Basel: Birkhauser, 1999, pp. 443–457.

17. Hall, M.A., Correlation-based Feature Selection for Discrete and Numeric Machine Learning, *Proc. 17th Int. Conf. on Machine Learning* (ICML-00), Morgan Kaufmann Publish., 2000.

18. *IBM SPSS Modeler 14.2 Algorithms Guide*, available at ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/.

19. Asuncion, A. and Newman, D.J., *UCI Machine Learning Repository* (http://www.ics.uci.edu/∼mlearn/MLRepository.html), Irvine, CA: University of California, School of Information and Computer Science, 2007.

20. Tsurko, V.V. and Mikhal'skii, A.I., Statistical Analysis of the Relation between Cancer and Accompanying Diseases, *Usp. Gerontologii*, 2013, vol. 26, no. 4, pp. 766–774.

This paper was recommended for publication by L.A. Mironovskii, a member of the Editorial Board