Features Selection by Distributions Contrasting

Varvara V. Tsurko[1] and Anatoly I. Michalski[1,2]

[1] Institute of Control Sciences Russian Academy of Sciences,
Profsouznaya str. 65, 117997, Moscow, Russian Federation
v.tsurko@gmail.com
[2] National Research University Higher School of Economics,
Bolshoy Tryokhsvyatitelsky Pereulok 3, 109028, Moscow, Russian Federation
amikhalsky@hse.ru

Abstract. We consider the problem of selection the set of features that are the most significant for partitioning two given data sets. The criterion for selection which is to be maximized is the symmetric information distance between distributions of the features subset in the two classes. These distributions are estimated using Bayesian approach for uniform priors, the symmetric information distance is given by the lower estimate using Rademacher penalty and inequalities from the empirical processes theory. The approach was applied to a real example for selection a set of manufacture process parameters to predict one of two states of the process. It was found that only 2 parameters from 10 were enough to recognize the true state of the process with error level 8%. The set of parameters was found on the base of 550 independent observations in training sample. Performance of the approach was evaluated using 270 independent observations in test sample.

Keywords: classification, features selection, information distance between in-class distributions, Rademacher penalty, set of manufacture process parameters

1 Introduction

Machine learning algorithms are widely applied in problems from different areas. In applications general methods of data mining such as classification, regression, estimation of distributions often are modified to meet specific conditions of the problem which give a rise to new settings and to new methods of the problem solution. Features selection problem gives such example. In real world problems not all parameters have the same significance for classification or prediction of the dependent variable value. Because of limited amount of experimental data one needs to reduce the size of feature space to increase the statistical reliability of the result. Principal component analysis or Fourier decomposition are examples of such methods.

The universal way to construct a machine learning algorithm is to introduce a functional and to minimize it. When one builds a regression model the target is to reduce the mean error between a prediction and a real value of a process.

The classification algorithm aims to minimize the probability of an error in the class label assignment.

The problem of features selection can be set in the similar manner. One is to form a set of features for which distributions in two classes are the most different. In other words to minimize minus distance between the distributions. Such settings arise when one looks for the data description to have high classification result. This approach allows to find factors which are more relevant to one class in comparison with the other.

The difference between classes can be measured as the divergence between features distributions in the classes. It is important that distributions and the divergence between them are to be evaluated on empirical data. Similar task of features selection based on distributions divergence was considered in [1], where frequencies of features occurring in the data were calculated and features with the maximal difference of these frequencies in different classes were selected. In [7] the method of features selection was focused on the maximization of dependence between selected features and the target variable.

In the paper we consider the features selection problem in case of two classes which are described by the conditional distributions of the features. The goal of the method is to find the set of features for which conditional distributions in classes have the maximal difference. The difference between distributions is characterized by a functional of average risk. Maximization of the average risk is equivalent to the Kullback-Leibler divergence maximization between the two in class distributions. Because the real distributions are unknown, we estimate them on the empirical data. The value of the average risk is estimated using the Rademacher penalty term which takes into account the complexity of distribution functions in the subset of features. The proposed method selects a set of features that provides the maximum of the average risk with the guaranteed prescribed probability. The method was applied to experimental data of a manufacturing process.

2 Distributions Contrasting Algorithm

2.1 Average Risk

A lot of data analysis problems like classification, regression, probability density reconstruction could be formulated in terms of average risk minimization. Problem of distribution contrasting can be set in the same way. Let x be a random vector of continuous features, $\varphi_0(x)$ and $\varphi_1(x)$ be probability density functions (pdfs) which estimate the conditional distribution of x under hypotheses H_0 and H_1 respectively, y be a class label variable which takes values 0 or 1 and states the number of hypothesis. The loss function is defined in following form

$$f_{\varphi_0,\varphi_1}(x,y) = -y \ln \varphi_0(x) - (1-y) \ln \varphi_1(x).$$

The functional of average risk is defined as an expectation of the loss function:

$$M(\varphi_0, \varphi_1) = -E_{x,y}(y \ln \varphi_0(x) + (1-y) \ln \varphi_1(x)), \tag{1}$$

where expectation is taken by joint distribution of x and y.

The functional has the meaning of $\varphi_0(x)$ and $\varphi_1(x)$ pdfs crossentropy weighted by a priori probabilities of the hypotheses H_0 and H_1. Minimization of it by $\varphi_0(x)$ and $\varphi_1(x)$ is equivalent to probability density reconstruction under different hypotheses. If the class label y is unobserved in the data then minimization of (1) is equivalent to a clusterization problem [6].

In this paper we consider a different problem of average risk maximization. Rewriting the functional of average risk gives

$$M(\varphi_0, \varphi_1) = I(\varphi_0, \varphi_1) - E_{x,y}(y \ln p(x|H_1) + (1-y) \ln p(x|H_0)),$$

where

$$I(\varphi_0, \varphi_1) = -E_{x,y}\left(y \ln \frac{\varphi_0(x)}{p(x|H_1)} + (1-y) \ln \frac{\varphi_1(x)}{p(x|H_0)}\right).$$

Maximization of the average risk functional by $\varphi_0(x)$ and $\varphi_1(x)$ is equivalent to maximization of $I(\varphi_0, \varphi_1)$ which is close to the Kullback-Leibler divergence between two pairs of distributions $\varphi_0(x), p(x|H_1)$ and $\varphi_1(x), p(x|H_0)$ [4]. In other words, optimally selected functions $\hat{\varphi}_0, \hat{\varphi}_1$ should maximally differ from the true pdfs of x under the hypothesis H_1 and H_0 respectively.

For probability density functions $\varphi_0(x), \varphi_1(x)$ we use Bayesian estimates of conditional distributions $p(x|H_0)$ and $p(x|H_1)$ obtained from given data with formula [10]

$$\varphi_y^b(x) = \sum_{i=1}^{k} I(x \in \sigma_i) \frac{(1-y)n_i + ym_i + 1}{(1-y)l_0 + yl_1 + k}, \tag{2}$$

where k is the number of bins in the histogram used in estimations, $I(x \in \sigma_i)$ is indicator which equals to 1 if vector x belongs to bin σ_i, l_0 and l_1 denotes the sizes of independent samples obtained under hypotheses H_0 and H_1 respectively, n_i and m_i are the numbers of observations from independent samples obtained under hypotheses H_0 and H_1 respectively that put into the bin σ_i. Given formula represents Bayesian probability estimation in case of uniform prior distribution on a k-fold simplex.

With defined estimations the average risk for $\varphi_0^b(x)$ and $\varphi_1^b(x)$ is obtained by substituting them in (1). Maximization of it in the way explained further gives the set of features for which the Bayesian estimates of conditional distributions under different hypotheses in these classes maximally differ. It allows us to conclude that the constructed set includes features for which the hypothesis can be tested with less error than in case of using all features in the data sets.

2.2 Empirical Risk

Let $x_1^0, \ldots, x_{l_0}^0$ be a sample with the pdf $p(x|H_0)$ and $x_1^1, \ldots, x_{l_1}^1$ be a sample with the pdf $p(x|H_1)$. The formula for pdf (2) is based on Bayesian estimates

of histograms with k bins for first and the second classes. We denote these estimates [10] of probability density of the bin i as

$$\varphi_0^b(i) = \frac{n_i + 1}{\sum\limits_{j=1}^{k} n_j + k},\ \varphi_1^b(i) = \frac{m_i + 1}{\sum\limits_{j=1}^{k} m_j + k}. \tag{3}$$

It is clear, that

$$0 < c \leq \varphi_y^b(i),\ y = 0, 1,\ i = 1, \ldots, k,\ c = \frac{1}{k + l_0 + l_1}; \tag{4}$$

$$\sum_{i=1}^{k} \varphi_y^b(i) = 1,\ y = 0, 1, \tag{5}$$

and the vector $\varphi_y^b = (\varphi_y^b(1), \ldots, \varphi_y^b(k))$ which satisfies conditions (4), (5) defines a histogram. Finally, let F denote a set of histograms pairs $(\varphi_0^b, \varphi_1^b)$ calculated from empirical data using (3) for all subsets of features set.

The average risk $M(\varphi_0^b, \varphi_1^b)$ reflects the divergence between Bayesian estimates and real densities. In order to evaluate the average risk we calculate an empirical risk by using average of empirical data instead of expectation. Applying $E_y(y) = l_1/(l_0 + l_1)$ we get a formula for the functional of empirical risk

$$M_e(\varphi_0^b, \varphi_1^b) = -\frac{1}{l_0 + l_1} \left(\sum_{i=1}^{k} m_i \ln \varphi_0^b(i) + \sum_{i=1}^{k} n_i \ln \varphi_1^b(i) \right). \tag{6}$$

The relation between the average risk and the empirical risk was discussed in [9]. Consideration of this relation allows us to switch from the maximization of the average risk problem to the maximization of the empirical risk corrected by a penalty term. A form of the penalty term will be discussed below.

2.3 Rademacher Penalization

The functional of empirical risk (6) can be presented in the form

$$M_e(\varphi_0^b, \varphi_1^b) = -\frac{1}{l_0 + l_1} \left(\sum_{i=1}^{l_1} \ln \varphi_{0,x_i^1}^b + \sum_{i=1}^{l_0} \ln \varphi_{1,x_i^0}^b \right),$$

where $\varphi_{y,x_i^t}^b$ – the Bayesian estimate of conditional probability under the hypothesis H_y for the bin to which x_i^t belongs.

Let $\delta_1^0, \ldots, \delta_{l_0}^0, \delta_1^1, \ldots, \delta_{l_1}^1$ be a sequence of independent and identically distributed random variables which take values +1 and -1 with probability 1/2 each independently of $(x_1^0, \ldots, x_{l_0}^0, x_1^1, \ldots, x_{l_1}^1)$. The Rademacher penalty term [3, 5] is defined then as

$$R = \sup_{\varphi_0^b, \varphi_1^b \in F} \left| \frac{1}{l_0 + l_1} \left(\sum_{i=1}^{l_1} \delta_i^1 \ln \varphi^b_{0,x_i^1} + \sum_{i=1}^{l_0} \delta_i^0 \ln \varphi^b_{1,x_i^0} \right) \right|. \tag{7}$$

With Δ_i^y denoting the sum of δ_t^y that correspond to the same bin i it could be represented as

$$R = \sup_{\varphi_0^b, \varphi_1^b \in F} \left| \frac{1}{l_0 + l_1} \sum_{i=1}^{k} \left(\Delta_i^1 \ln \varphi_0^b(i) + \Delta_i^0 \ln \varphi_1^b(i) \right) \right|.$$

In order to solve the optimization problem we remove the modulus be representing the penalty term in form

$$R = \frac{1}{l_0 + l_1} \max \{A, -A\}, \tag{8}$$

where

$$A = \sup_{\varphi_0^b \in F} \sum_{i=1}^{k} \Delta_i^1 \ln \varphi_0^b(i) + \sup_{\varphi_1^b \in F} \sum_{i=1}^{k} \Delta_i^0 \ln \varphi_1^b(i).$$

Thus, in order to find the optimal solution of initial problem we consider the optimization subproblem, which is to find

$$R' = \sup_{\varphi^b \in F} \sum_{i=1}^{k} \Delta_i \ln \varphi^b(i). \tag{9}$$

The solution is given by the rules 1-3:

1. If $\Delta_i > 0,\ i = 1 \ldots, k$ then

$$R' = \sum_{i=1}^{k} \Delta_i \ln \frac{\Delta_i}{\sum\limits_{t=1}^{k} \Delta_t}$$

2. If $\Delta_i \leq 0,\ i = 1, \ldots, k$ then

$$R' = \sum_{i=1}^{k} \Delta_i \ln c + \Delta_m \ln \frac{1 - c(1-k)}{c},$$

   where $m = \arg\max_i \Delta_i$
3. If $\Delta_i > 0,\ i = 1, \ldots, s$ and $\Delta_i \leq 0,\ i = s+1, \ldots, k$ then

$$R' = \sum_{i=1}^{s} \Delta_i \ln \frac{\Delta_i (1 - c(k-s))}{\sum\limits_{t=1}^{s} \Delta_t} + \sum_{i=s+1}^{k} \Delta_i \ln c.$$

It's important to notice that particular relation between Δ_i, c and size of empirical data set is significant. The solution for more general case, when Δ_i can take any values, is quite similar, but the rules become more complex.

By substitution of extremal values of (9) into (8) we obtain the value for the Rademacher penalty term.

2.4 Average Risk Evaluation

Values of the penalty term and the empirical risk can be used for estimation of the average risk using the symmetrization inequality [2, 3]. For the class of functions uniformly bounded by a constant U and for all $t > 0$ the following holds

$$P\left\{\sup_{\varphi\in F}|M(\varphi)-M_e(\varphi)|\geq 2R+\frac{3tU}{\sqrt{l_0+l_1}}\right\}\leq\exp\left(-\frac{t^2}{2}\right). \tag{10}$$

For Bayesian estimates (3) it is valid that $\frac{1}{l_0+l_1+k}\leq\varphi_y^b<1$, from which we obtain $0<|\ln\varphi_y^b|\leq\ln(l_0+l_1+k)=U$.

Using (10) and fixing the probability $\eta=\exp\left(-\frac{t^2}{2}\right)$ we derive the following inequality

$$P\left\{\sup_{\varphi\in F}|M(\varphi)-M_e(\varphi)|<2R+\frac{3\sqrt{-2\ln\eta}\ln(l_0+l_1+k)}{\sqrt{l_0+l_1}}\right\}\geq 1-\eta.$$

Hence, with the probability not less than $1-\eta$ the lower bound of the functional of average risk is

$$M(\varphi)>M_e(\varphi)-2R-\frac{3\sqrt{-2\ln\eta}\ln(l_0+l_1+k)}{\sqrt{l_0+l_1}}. \tag{11}$$

2.5 Distributions Contrasting Algorithm

We consider a set of features $X=(f_1,f_2,\ldots,f_n)$ measured in two different classes. The goal is to find such subset X_j of X for which two classes maximally differ in terms of the conditional distributions divergence. There are two stages: first task is to form the sequence of features subsets, second is to define the subset satisfied the goal.

We start with building histograms of k bins for each feature in a class. Then the value of the empirical risk (6) is calculated and the feature with maximum value of the empirical risk forms the beginning of sequence. Without restricting the generality let the feature f_1 be the one with a maximum value of the empirical risk functional, so $X_1=f_1\in X$.

Then all possible pairs of features are constructed with one feature f_1 obtained in the first step and two-dimensional histograms for each pair in two classes are built. Selecting the pair with the maximum value of the empirical risk, e.g. the pair will be (f_1,f_2), leads to the second subset in the target sequence $X_2=(X_1,f_2)=(f_1,f_2)\in X$ which is a superset for X_1.

At the third step we create all possible triplets of features with two features fixed on the previous step obtaining X_3 in a similar way. The process continues until all features are put in the sequence. As the result of the first stage we obtain the sorted subsets of the features.

At the second stage of the algorithm we evaluate the average risk value (1) using the estimation (11) for the features in X_j for each $j\in\{1,\ldots,n\}$. The

result is a set of estimates of the average risk : $M_{X_1}, M_{X_2}, \ldots, M_X$. Finally, we select the average risk M_{X_j} with the maximum value and corresponding subset of features X_j.

Proposed algorithm of distributions contrasting has $O(n^2)$ complexity.

3 Experimental Results

The algorithm of distributions contrasting was evaluated using the empirical data from a real manufacturing process presented by time records of 10 parameters. The two states of the process were matter of interest and data were labeled by an expert with the class label for each time point: 562 points in the first class and 258 points in the second class. For evaluation purpose the data were divided into a test sample and a training sample in proportion 1 : 2.

Then the distributions contrasting algorithm was applied to the training sample. We sorted the sequence of ten parameters and evaluated lower bound of the average risk. The value of the empirical risk and the 90% lower bound of the average risk for the features sequences are shown in the Table 1.

Table 1. Results of the distributions contrasting

Number of parameters in the set	Parameters included in the set	Empirical risk	Lower bound of average risk
1	10	4.799	2.1275
2	10, 1	6.884	**2.9038**
3	10, 1, 4	10.1013	1.9058
4	10, 1, 4, 5	13.4695	0.6662
5	10, 1, 4, 5, 2	16.8365	-0.4554
6	10, 1, 4, 5, 2, 7	20.2038	-1.5773
7	10, 1, 4, 5, 2, 7, 3	23.5711	-2.6551
8	10, 1, 4, 5, 2, 7, 3, 6	26.9384	-3.738
9	10, 1, 4, 5, 2, 7, 3, 6, 8	30.3057	-4.8208
10	10, 1, 4, 5, 2, 7, 3, 6, 8, 9	33.673	-5.9037

The lower bound of average risk riches its maximum on the pair of features. With increasing number of features in the set the lower bound of the average risk goes down. So the optimal number of parameters to distinguish two given states of the system is two and it includes parameters #10 and #1.

To verify the results both training and test samples were classified using different sets of parameters described above. The Naive Bayes Classifier [6] was used for classification. Figure 1 illustrates errors of classification for ten classifiers. Each pair of columns reflects classification errors on the training sample and on the test sample. The vertical axis describes the absolute value of error and horizontal axis shows the number of features in the set.

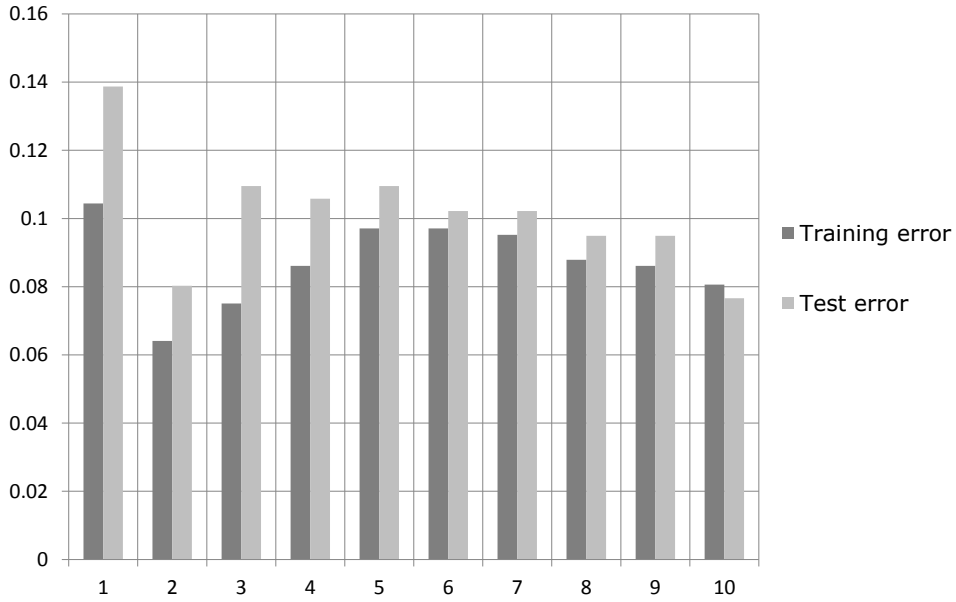

Fig. 1. Naive Bayes Classifier errors

The figure 1 shows that the error on the training sample, which was used for features selection, is minimal for the set composed by parameters #10 and #1. Exactly the same set was selected by the algorithm of distributions contrasting as optimal one to distinguish the two classes. The training error is about 6%, the test error is minimal on the same set of features and equals 8%. For sets with more features both errors are greater because of higher dimension of feature space. For the one parameter set both errors are greater because single parameter is not enough for good classification of the considering process states. Results of Naive Bayes classification show that the set of features obtained by the algorithm of distributions contrasting gives precise and stable results of classification.

4 Conclusion

The algorithm of features selection proposed in the paper is based on the divergence between distributions in the classes. Two classes considered in the presented example were related to the two different states of the manufacture process. The constructed features subset contained two parameters which were enough to predict the state of the process with the high precision. This was confirmed by control sample classification using Naive Bayes approach.

In other applications classes can be formed differently and features set can have more specific meaning. One of example is described in [8]. The features space was formed by ICD-10 codes of diseases that a person had at the end

of his life. The class label in that case was formed by the logical condition "did the person have a cancer?". By selection the features set for which lower bound estimate for distance between distributions of diseases in the two classes had maximal value we found a list of diseases related to the cancer. All those diseases act cancer stimulation role and should be cured at the initial state to lower the risk of cancer incidence.

References

1. Bay S.D., Pazzani M.J.: Detecting group differences: mining contrast sets. Data mining and knowledge discovery. 5, pp. 213–246 (2001)
2. Koltchinskii V.: Rademacher penalties and structural risk minimization. In: IEEE Transactions on Information Theory (1999)
3. Koltchinskii V.: Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems. LNM, vol. 2033. Springer, Heidelberg (2008)
4. Kullback S., Leibler R.A.: On information and sufficiency. The Annals of Mathematical Statistics. Vol. 22, No. 1, pp. 79–86 (1951)
5. Lozano F.: Model selection using Rademacher Penalization. In: the Second ICSC Symposia on Neural Computation (NC2000). ICSC Adademic (2000)
6. Manning C., Raghavan P., Schutze H.: An Introduction to Information Retrieval. Cambridge University Press, Cambridge (2009)
7. Song L., Smola A., Gretton A., Bedo J., Borgwardt K.: Feature selection via dependence maximization. Journal of Machine Learning Research. 13, pp. 1393–1434 (2012)
8. Tsurko V., Michalski A.: Statistical Analysis of Links between Cancer and Associated Diseases (in Russian). Adv. geront. 26, No. 4, pp. 766–774 (2013)
9. Vapnik V.: Statitical Learning Theory. Wiley Interscience (1998)
10. Vapnik V., Chervonenkis A.: Pattern Recognition Theory (in Russian). Nauka, Moscow (1974)