Feature Selection by Reodering According to their Weights

Jiřina, Marcel
2004

CENTRUM APLIKOVANÉ KYBERNETIKY

České vysoké učení technické v Praze - fakulta elektrotechnická

Feature Selection by Reordering according to their Weights

Technical report

Marcel Jiřina and Marcel Jiřina, jr.

www@c-a-k.cz

2004

Institute of Computer Science
Academy of Sciences of the Czech Republic

Feature Selection by Reordering according to their Weights

Marcel Jiřina and Marcel Jiřina, jr.

Technical Report No. V-919

October 2004

Abstract

Feature selection serves for both reduction of the total amount of available data (removing of valueless data) and improvement of the whole behavior of a given induction algorithm (removing data that cause deterioration of the results). This paper discusses this problem in more detail. A method of proper selection of features for an inductive algorithm is discussed. The main idea behind the method consists in proper descending ordering of features according to a measure of new information contributing to previous valuable set of features. The measure is based on comparing of statistical distributions of individual features including mutual correlation. A mathematical theory of the approach is described. Results of the method applied to real-life data are shown

Keywords:
Multivariate data, classification, feature selection, feature weight, correlation.

1 Introduction

Feature selection is a process of data preparation for their consequential processing. Simply said, the feature selection filters out unnecessary variables. There are two aspects for feature selection. The first one and much older aspect is the time requirement for processing of large amount of data during learning as well as recall phases of machine learning [7]. The other aspect is a finding that results of the induction algorithm (classification, recognition or also approximation and prediction) may be worse due to the presence of unnecessary features than with optimal feature selection [3].

There exist two essentially different views, so called filter model and wrapper model [6]. In filter model features are selected independent of the induction algorithm. Wrapper models (methods) are tightly bound to an induction algorithm. Another approach can be more quantitative stating that each feature has some "weight" for its use by induction algorithm. There are lots of approaches trying to define and evaluate feature weights, usually without any relation to induction algorithm, e.g. [3], [7], [8].

2 Problem Formulation

The suggested method is based on a selection of relevant (appropriate) feature set from a given set. This can be achieved without the need of a metric on the feature sets. In fact a proper ordering of features or feature sets is sufficient. There should be a measure for this ordering. The measure need not be necessarily a metrics in pure sense. It should give a tool for evaluating how much a particular feature brings new information to the set of features already selected.

Two problems then arise. First, to find a proper measure mentioned and second, to find a criterion or level of this measure below which the corresponding features can be omitted without loss of information.

3 The Method

The suggested method consideres features with relation to classification into one of two classes. It means that to each data sample corresponds a feature more, the class to which it belongs. We denote classes by 0 and 1 here. For each sample of so-called learning set the class is known. The method for stating the measure of feature weight utilizes comparisons of statistical distributions of individual features and for each feature separately for each class. Comparison of distributions is derived from testing hypothesis whether two probability distributions are from the same source or not. The higher the probability that these distributions are different the higher is the influence of particular feature (variable) to proper classification. In fact, we do not evaluate correlation probability between a pair of features, but between subsets corresponding to the same class only.

After these probabilities are computed, the ordering of features is possible. The first feature should bring maximal information for good classification, the second one a little less including ("subtracting") also correlation with the first, the third again a little less including correlations with two preceding features etc.

3.1 Outline - Feature Weights

The standard hypothesis testing is based on the following considerations: Given some hypothesis, e.g. two distributions are the same, or two variables are correlated. To this hypothesis some variable V is defined, e.g. the maximal difference between probability distribution functions or correlation coefficient. To this variable a probability p is assigned; its value is computed from value of V and often using some other information or assumptions. Then some level (threshold) P is chosen. If $p \geq P$ the hypothesis is assured, otherwise rejected. Sometimes instead of p the $1 - p$ is used and thus P and the test must be modified properly.

The logic used in this paper is based on somethig "dual" to the considerations above: Let $q = 1 - p$ be some probability (we call it the probability levels of rejection of hypothesis), $Q = 1 - P$ be some level. If $q < Q$ the hypothesis is assured, otherwise rejected. The larger q, the more likely the hypothesis is rejected (for the same level Q or P). It is just what we need. In fact, the weights assigned to individual features are probability levels q related to rejection of hypotheses that distributions are the same and variables are correlated.

Let F_i be feature. For the first ordering of individual features (variables) F_1 , F_2 , ... as to their influence on proper classification we use the probability levels of rejection p_{ii} of the hypothesis that the probability distributions of the feature F_i for the class 0 and for the class 1 are the same. This first ordering does not respect any correlation of variables.

To include influence of correlations let us denote by p_{ij0} and p_{ij1} probability levels of rejection that distributions of variables for class 0 are correlated and that distributions of variables for class 1 are correlated, respectively. Moreover, let p_{ii} be probability level of rejection that distributions of the feature F_i for the class 0 and for the class 1 are the same. How to get these numbers is discussed in the next section. Taking all probability levels together, we have two triangular matrices, one for p_{ij0} and another for p_{ij1}, i,j = 1, 2, ..., n. All results of pairwise distribution comparisons or correlations can be written in square matrix $n \times n$ as follows

$$M = \begin{bmatrix} p_{11} & p_{120} & \Lambda & p_{1n0} \\ p_{211} & p_{22} & \Lambda & p_{2n0} \\ \mathrm{M} & \mathrm{M} & \mathrm{O} & \mathrm{M} \\ p_{n11} & p_{n21} & \Lambda & p_{nn} \end{bmatrix} .$$

In this matrix in diagonal entries are probability levels of rejection of hypothesis that that for feature of a given index and for class 0, and class 1 the distributions are the same. In the upper triangular part there are probability levels p_{ij0} for class 0, and in the bottom triangular part the probability levels p_{ij1} for class 1.

In the beginning the ordering of features is arbitrary. We now sort rows and columns in descending order according to diagonal elements $p_{ii}$of the matrix M. After it, first, we reassign indexes according to this ordering (and store information about original ordering of features, i.e. original indexes). The first feature now is a feature having the largest difference in distributions for both classes. The second feature has lesser difference in distributions for both classes and can be possibly somehow correlated to the preceding feature, etc.. Then, first, we state correlation coefficient for class 0 of variables 1 and 2, second, correlation coefficient for class 1 of variables 1 and 2 getting then probability levels p_{120} and p_{211} of rejection that distributions are correlated. The lesser these probabilities, the stronger correlation between features F_1 and F_2 exists. Mutual relations for first two features are shown in Table 1.

Table 1. Mutual relations for first two features

feature	class *0*		class *1*
F_1	distribution of F_1 for class 0	$\leftarrow p_{11} \rightarrow$	distribution of F_1 for class 1
	β p_{120}	β $\sqrt{(p_{120}\, p_{211})}$	β p_{211}
F_2	distribution of F_2 for class 0	$\leftarrow p_{22} \rightarrow$	distribution of F_2 for class 1

In this table it is shown that the mutual dependence between features F_1 and F_2 for different classes is expressed as the geometric mean of probability levels p_{120} and p_{211}.

Let us define independence level of feature F_i on preceding features F_k, $k < i$ by formula:

$$\pi_i = p_{ii} \prod_{k=1}^{i-1} \sqrt{p_{ik0}\, p_{ki1}} \quad . \qquad (1)$$

According to this formula the probability level of rejection that distributions for class 0 and 1 are the same is modified by measure of dependence on preceding variables. In fact, we define independence level as probability levels of rejection that classes are the samemultiplied by the geometric mean of probability levels p_{ik0} and p_{ki1} . of rejection that one or the other class is correlated to preceding features.

We associate these probability levels to corresponding rows and columns and again we sort rows and columns according to π_i in descending order. After it we again compute π_i according to (1) using new ordering and new indexing of variables. This step is repeated until no change in ordering occurs. It was found that this process converges fast but we have no convergence proof up to now.

By this procedure features are reordered from original arbitrary ordering to new ordering such that the first feature has the largest π_i , and the last the smallest π_i .

In this context the variable π_i is a measure of how much new information we get using variable i or how much information we loose when deleting it. It has been seen that this measure includes probability of correlation between variables as well as noise which causes lessening of p_{ii} similar way as smaller difference between distributions for class 0 and class 1. Differentiation between classes is essential here. We cannot use simply p_{ij} between features because information on classes would be lost.

4 Theory

4.1 Measure of new information

Unlike methods published we do not group features into different feature sets but we would like to order them according to some measure. The measure should express how much the next feature brings new information to the preceding collection of features. We speak about measure of information but it need not be just entropy; we propose to use some probability.

Definition

Let an ordered feature set of F_1, F_2, ... and two classes $c \in \{0,1\}$ be given. The measure of new information from the feature F_i is given by

$$\pi_i = p_{ii} \prod_{k=1}^{i-1} \sqrt{p_{ik0} p_{ki1}} \ ,$$

where p_{ii} is probability level of rejection that distributions of F_i for class 0 and for class 1 are the same, and p_{ik0} and p_{ki1}, k = 1, 2, ..., i-1 are probability levels of rejection that features F_i and F_k for class 0 and class 1 are correlated. p_{ii} is probability level of rejection resulting from a corresponding test, e.g. . Kolmogorov - Smirnov test [9] or Cramér – von Mises test [1]. Determination of the probability levels p_{ik0} and p_{ki1} is described in the next subsection.

The π_i is, in fact, our measure that distributions of the feature F_i are different and, at the same time, the feature for one and for the other class are not correlated with corresponding parts of all preceding features.

Correlation probability

For calculation of the probability levels p_{ik0} and p_{ki1} we use a standard approach [4]. First we transform a correlation coefficient of two features into probability that corresponding features are correlated. Let the distribution of these two features be a two-dimensional normal distribution with parameters u_1, u_2, s_1, s_2 and ρ, where u_1, u_2 are means, s_1, s_2 dispersions of the two features, and ρ is (a priori known) correlation coefficient between these two features. The pairs of features in individual samples are, in fact, selection from this two-dimensional distribution and the statistical distributions of features F_1 and F_2 are marginal distributions of the two-dimensional distribution. Let there be n random selections of pairs F_1 and F_2, and let empirical correlation coefficient be r.

We need probability levels p_{ij0} and p_{ij1} of rejection that features i and j for class 0 and for class 1 are correlated. Each of these probability levels is calculated as $1 - p$, where p is the probability that corresponding pair of features F_i and F_j for classes 0 and 1 are correlated. The procedure for quantifying p is described below.

Let $\rho = 0$, then the statistics $t = r\sqrt{n-2}\big/\sqrt{1-r^2}$ has the Student's distribution $t(n-2)$ with $n-2$ degrees of freedom and does not depend on parameters u_1, u_2, s_1, s_2, [4].

More general approach according to Fischer [4] takes into account nonzero value of ρ. Let us use a transformation

$$Z = \frac{1}{2} \ln \frac{1+r}{1-r} \ , \tag{2}$$

This random variable has approximately normal distribution with mean value

$$E\{Z\} = \frac{1}{2} \ln \frac{1+\rho}{1-\rho} \tag{3}$$

and dispersion

$$D\{Z\} = \frac{1}{n-3}$$

for $n > 10$ and $|\rho|$ not too close to 1. The random variable

$$U = \sqrt{n-3}\,(Z - E\{Z\}) \tag{4}$$

has approximately normal distribution $N(0,1)$ for $n > 10$. Note that Z and U have these approximately normal distributions for any distribution of original random variables with finite means and dispersions.

There is one problem. Even for really uncorrelated distributions for limited number of samples a nonzero correlation coefficient ρ_0 is found. Consecutively, from the distributions above a nonzero probability p_0 that features are correlated follows. We solve this by introducing two steps. In the first step we compute the mean half-with of the correlation coefficient ρ_0 under the assumption of uncorrelated distributions. This is done simply stating probability $p = 0.75$. In fact, we consider the mean of the right hand half of distribution here. From this probability and number n of samples at hand it follows

$$\bar{\rho}_0 = \frac{e^{2z} - 1}{e^{2z} + 1} \;, \tag{5}$$

where $z = \Phi^{-1}(p_0)/\sqrt{n-3}$ and $\Phi^{-1}(p_0) \approx 0.674490$ is the inverse value of standard normal distribution for probability $p_0 = 0.75$. The interval $<-\bar{\rho}_0,\ \bar{\rho}_0>$ contains ρ_0 with probability 0.5 .

From it follows that variables with absolute value of empirical correlation coefficient r equal to or close to $\bar{\rho}_0$, are, in fact, probably uncorrelated. So we set value of $\bar{\rho}_0$ as a priori correlation coefficient ρ of the two-dimensional normal distribution above. Using this value as ρ in (3) and as value of empirical correlation coefficient r to (2), we get probability $p = 0.5$ from (4). This value follows intuitive consideration that in this case we have truly no information whether features are correlated or not, both cases are equally possible. The procedure for computation is simple, $p = \Phi(U)$ is the probability for standard normal distribution, where U is given by (4) using (3), (2) and ρ is computed as ρ_0 using (5).

4.2 Local Geometry of Feature Space and Dimensionality-induced Degradation

We will analyze behavior of the simplest nearest neighbor method (k-NN method) with respect to the presence of some irrelevant features. Let us have n dimensional space and j relevant features, n-j irrelevant features, and k nearest neighbors to point x. The most distant k-th neighbor let be X and its distance from point x be d_X.

The distance of the i-th neighbor $n_i=(n_{i1},n_{i2}, \ldots, n_{in})$ from point x is

$$d_i = \sqrt{\sum_{i=1}^{j}(n_i - x_i)^2 + \sum_{i=j+1}^{n}(n_i - x_i)^2} \le d_X \;. \tag{6}$$

The second sum is a random variable. Its distribution approaches to normal distribution with the number of irrelevant features going to infinity and is independent of particular neighbor. When using relevant features only, it holds

$$d_{ir} = \sqrt{\sum_{i=1}^{j}(n_i - x_i)^2} \le d_{Xr} \;, \tag{7}$$

where d_{Xr} is the distance of the farthest of all k neighbors from point x in this case.

The problem is that d_{ir}'s are different from d_i's and the difference is caused by random part in (6). This has two consequences. First, sets of k neighbors are different. For some neighbors both (6) and (7) are valid, for some points it holds either (6) or (7) only. When k neighbors according to (7) are considered relevant, then not all k neighbors according to (6) can be relevant. Second, it has been shown [5] that for limited number of samples even for small dimensions the position of neighbors can be influenced by boundary effect. Irrelevant features make dimension larger and also strengthen the boundary effect. In the final effect it again influences which k points are selected as nearest neighbors.

Considering relevant features only, we expect that ratio of number of points of one and of the other class gives good estimation about ratio of corresponding probability densities. Irrelevant features cause that this is valid for part, say k_1 neighbors of all k neighbors selected according to (6). For other (k-k_1) neighbors it need not hold because selection of these points is influenced by some random number without connection to true probability density distribution. We can estimate that the estimation of the probability distribution function or classification gets the worse the larger difference between d_X and d_{Xr}, i.e. the larger number of irrelevant features.

Even for relevant features with smaller influence the boundary effect may cause that such features may behave as irrelevant. The error caused by boundary effect may be larger than error when such features are not used. So, the monotonicity assumption [2] is not valid even if weighting assigns nonzero influence to all features.

For quantitative estimation let us consider slightly different nearest-neighbor method. Let all features be standardized (normalized) to zero mean and unit dispersion. Let us consider that we have unlimited number of data samples. Let us consider most simple approach, the (most) naive nearest neighbor approach using L_2 metrics, i.e. Euclidean space. Let there be a point x of unknown class. Let us build two balls around this point, one of radius r_0 equal to the distance to the nearest point of one class 0, and the other of radius r_1 equal to the distance to the nearest point of the other class 1. Let us consider r_0 and r_1 as average values of these radii. The probability that point x belongs to the class 0 is equal to $p_0 = V_0 / (V_0 + V_1)$, where V_0 and V_1 are volumes of the corresponding balls. After simple arrangement we obtain

$$p_0 = \frac{r_0^n}{r_0^n + r_1^n} \quad . \tag{8}$$

We could also use so-called Bayes ratio $p_{BAYES\,0} = V_0 / V_1 = r_0^n / r_1^n$. The radii r_i, $i \in \{0,1\}$, are given by formula

$$r_i = \sqrt{\sum_{k=1}^{n} (x_k - f_{ki})^2} \quad , \tag{9}$$

where x_k is the k-th coordinate of point x, and f_{ki} is k-th feature of the nearest point of class i.

Let features with indexes 1, 2, ..., j be relevant and features with indexes j+1, ..., n be irrelevant. Then (9) can be rewritten as follows

$$r_i = \sqrt{\sum_{k=1}^{j} (x_k - f_{ki})^2 + \sum_{k=j+1}^{n} (x_k - f_{ki})^2} \quad . \tag{10}$$

The second sum corresponds to irrelevant features. If for each irrelevant feature distributions for class 0 and class 1 are the same, then the second terms for both classes are the same. Omitting the second term, we can get radius of the ball in j-dimensional space of relevant features and using (8) we get estimation of probability needed. By the use of (10) additional terms cause that some constants are added and value of $p_0(x)$ is thus distorted. The larger number of irrelevant features, the larger is the second term in (10) in comparison to the first term, and the larger is the distortion of $p_0(x)$.

Induction – classification algorithms need not use or assume just Euclidean geometry in the feature space. Let us assume metrics L_s, $s \in (0; \infty)$. Some well-known metrics are L_1 – absolute, L_2 – Euclidean, and L_∞ - max. For all cases, except the last the (10) has now form

$$r_i^s = \sum_{k=1}^{j} |x_k - f_{ki}|^s + \sum_{k=j+1}^{n} |x_k - f_{ki}|^s \quad .$$

It is seen again that the term corresponding to irrelevant features distorts the estimation. In the case of L_∞ - max metrics, the maximal difference can arise from irrelevant features the more often the larger number of irrelevant features occurs. The final consequence is the same.

5 Results

The suggested method is demonstrated on a task of feature ordering of UCI MLR real-life databases [10]. Table 2 gives reordering of features and corresponding probabilities that particular feature brings new information to the preceding set of features. It is given for Heart, Vote, Splice, Spam, Shuttle, Ionosphere, German, and Adult databases. In all cases the set of data normally used for learning was used for features reordering. In table numbers of features in the original data set are given but ordered in diminishing influence, i.e. descending Pcor($S_k \neq S_l$). Pcor($S_k \neq S_l$) denotes our measure that distributions of the feature F_i are different and, at the same time, the feature for one and for the other class are not correlated with corresponding parts of all preceding features (in the table all features above the feature considered), see (1). In Fig. 1 Pcor($S_k \neq S_l$) is given for features in the same order.

Table 2. Reordering of features for eight data bases from UCI MLR

Database	Heart	Vote	Splice	Spam	Shuttle	Ionosphere	German	Adult
1	13	3	29	52	9	6	1	6
2	12	7	30	53	1	12	3	8
3	3	6	34	7	7	4	6	5
4	9	8	32	16	3	32	21	10
5	8	11	31	57	8	14	20	1
6	10	5	28	5	5	20	9	4
7	11	4	36	19	2	33	14	13
8	1	12	25	21	6	22	2	11
9	2	14	26	55	4	16	10	9
10	5	15	23	25		18	17	2
11	4	13	40	2		28	4	7
12	7	1	41	56		10	12	12
13	6	10	24	3		24	23	3
14		9	22	12		8	5	14
15		2	48	11		2	...	
16			43	9		26		
17			20	27		7		
18			21	17		21		
19			15	10		3		
20			19	1		9		
			...	...		...		

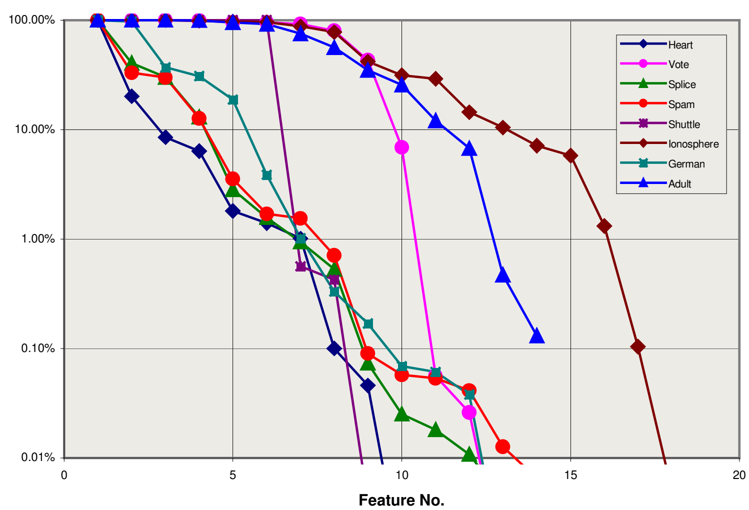

Fig. 1. Dependence of the π_i on feature number after reordering for eight databases from UCI MLR

6 Conclusion

We have presented a procedure for evaluating feature weights based on the idea that we need not evaluate subsets of features or build some metrics in the space of feature subsets. It was shown that instead of metrics some ordering would suffice. This is much weaker condition then metric. In fact, we need ordering of features from the point of view of the ability of feature possibly bring something new to the set of features already selected. If features are properly ordered we need not measure any distance. Knowledge that one feature is more important than the other should be sufficient. Having features already ordered, the question on proper feature set selection is reduced from combinatorial complexity to linear or at worst polynomial complexity – depending on the induction algorithm.

References

[1] Csörgö, S., Faraway, J.J.: The Exact and Asymptotic Distributions of Cramér-von Mises Statistics. J.R. Statist. Soc. B vol. 58, No. 1 (1996) 221-234

[2] Dash, M., Liu, H.: Consistency-based search in feature selection. Artificial Intelligence 151 (2003) 155–176

[3] Dong, M., Kothari, R.: Feature subset selection using a new definition of classifiability. Pattern Recognition Letters 24 (2003) 1215–1225

[4] Hátle, J, Likeš, J.: Basics in probability and mathematical statistics (in Czech), SNTL/ALFA Praha (1974)

[5] Jiřina, M., Jiřina, M., jr.: Features of Nearest Neighbors Distances in High-Dimensional Space. Technical Report No. 913, Inst. of Computer Science, Prague, Czech Republic, (2004), 15 pp.

[6] John, J.K., Kohavi, R., Pfleger, K.: Irrelevant features and the Subset Selection problem. In: Machine Learning: Proc. of the Eleventh Int. Conf. ed. Cohen,W., Hirsh,H., Morgan Kaufmann Publishers, San Francisco, Ca., USA (1994) 121-129

[7] Koller, D., Sahami, M.: Toward Optimal Feature Selection. Proc. of the Thirteenth Int. Conf. on Machine Learning. Morgan Kaufmann Publishers, San Francisco, Ca., USA, Morgan-Kaufman (1996) 284-292

[8] Last, M., Kandel, A., Maimon, O.: Information-theoretic algorithm for feature selection. Pattern Recognition Letters 22 (2001) 799-811

[9] Smirnov, N.: Table for estimating the goodness of fit of empirical distributions, Annals of Math. Statist. 19 (1948) 279-281

[10] UCI Repository of Machine Learning Databases. http//www.ics.uci.edu/ ~mlearn/MLRepository.html