A Divergence Criterion for Classifier-Independent Feature Selection

Naoto Abe, Mineichi Kudo, Jun Toyama, and Masaru Shimbo

Division of Systems and Information Engineering,
Graduate School of Engineering,
Hokkaido University, Sapporo 060-8628, Japan.
{chokujin,mine,jun,shimbo}@main.eng.hokudai.ac.jp

Abstract. Feature selection aims to find the most important feature subset from a given feature set without degradation of discriminative information. In general, we wish to select a feature subset that is effective for any kind of classifier. Such studies are called *Classifier-Independent Feature Selection*, and Novovičová *et al.*'s method is one of them. Their method estimates the densities of classes with Gaussian mixture models, and selects a feature subset using Kullback-Leibler divergence between the estimated densities, but there is no indication how to choose the number of features to be selected. Kudo and Sklansky (1997) suggested the selection of a minimal feature subset such that the degree of degradation of performance is guaranteed. In this study, based on their suggestion, we try to find a feature subset that is minimal while maintainig a given Kullback-Leibler divergence.

1 Introduction

The goal of feature selection is often said to be to find a subset of a given size from a given feature set such that the subset has the most discriminative information. However, the goal of feature selection has recently changed to finding that is most effective for classifiers without the size of the subset given. In large-scale problems (over 50 features), there seem to be many *garbage features*, which have a bad influence on the construction of classifiers. It is, therefore, expected that the performance of classifiers can be improved by removing such garbage features.

Many methods have been proposed for feature selection [1,2], such as Sequential Forward/Backword Floating Search (SFFS/SBFS) method [3]. These methods select a feature subset that maximizes a criterion function based on the recognition rate of a classifier chosen beforehand. Such an approach is useful when we know what kind of classifiers will be used. It is, however, more desirable to select a feature subset that is universally effective for any classifier. Such an approach is called *Classifier-Independent Feature Selection* [4,5] and Novovičová *et al.*'s method [6] is one such method.

In Novovičová *et al.*'s method, class-conditinal densities are estimated from given data with Gaussian mixture models, and a feature subset of a given size that maximizes Kullback-Leibler (K-L) divergence [7] between the densities is

F.J. Ferri et al. (Eds.): SSPR&SPR 2000, LNCS 1876, pp. 668–676, 2000.

selected. However, from the viewpoint of classifier-independent feature selection, it is desirable to find a feature subset that is as small as possible but includes all information necessary for classification. Therefore, we think it is more important to select a feature subset on the basis of performance. Such a trial has been carried out by Kudo and Sklansky [8,9,10].

In this study, we use K-L divergence to evaluate the performance of a feature subset in Novovičová *et al.*'s method. That is, we use K-L divergence in double roles to rank features and to evaluate how useful a chosen subset is. Some experiments were carried out to confirm the effectiveness of the proposed method.

2 Novovičová *et al.*'s Method

In Novovičová *et al.*'s method [6], a class-conditional density is estimated from a training sample set of the class using the following Gaussian mixture model:

$$p(\mathrm{x}|\omega) = \sum_{m=1}^{M} \alpha_m^{\omega} \prod_{i=1}^{D} \left\{ f_0(x_i|\mathbf{b}_{0i})^{1-\phi_i} f(x_i|\mathbf{b}_{mi}^{\omega})^{\phi_i} \right\}, \tag{1}$$
$$\Phi = (\phi_1, \phi_2, \ldots, \phi_D) \in \{0,1\}^D,$$

where M is the number of components and α_m^{ω} are weights of components satisfying $\sum_{m=1}^{M} \alpha_m^{\omega} = 1$. Also, $\mathbf{b}_{mi}^{\omega}$ and $\mathbf{b}_{0i}$ are the parameters specifing the components, Φ is the parameter to indicate a feature subset, and D is the number of given features. The function f is a Gaussian specified by parameter $\mathbf{b}_{mi}^{\omega}$, and f_0 is a background Gaussian distribution specified by $\mathbf{b}_{0i}$. The parameters α_m^{ω}, $\mathbf{b}_{mi}^{\omega}$ and $\mathbf{b}_{0i}$ are estimated by the EM algorithm [11] as maximum likelihood estimators. A vector Φ indicates which features are used and which features are ignored: $\phi_i = 1$ for feature i to be used and $\phi_i = 0$ for feature i to be the background. The key is that the density form of each component (1) is independent with respect to features so that we can evaluate individual features independently. The dependency among feautres is absorbed in the mixture. To measure how far two densities are from each other, we use K-L divergence:

$$J(\Phi) = \sum_{\omega \in \Omega} P(\omega) \mathrm{E}_{\omega} \left\{ \log \frac{p(\mathbf{x}|\omega)}{p(\mathbf{x}|\Omega - \omega)} \right\}.$$

Here, Ω denotes the set of classes and $P(\omega)$ denotes a priori probability of class ω. E_{ω} is the expectation over $\mathbf{x}$ from class ω, and $p(\mathbf{x}|\omega)$ is the class-conditional probability density functions defined by (1). K-L divergence measures the separability between two different distributions. If K-L divergence in a feature subset is 0, the feature subset has no discriminative information. In Novovičová *et al.*'s method, K-L divergence is given as the sum of that of each feature. They rank each feature in order of its K-L divergence, and select the feature subset of a given size by removing some features with a small K-L divergence.

3 Proposed Method

3.1 How to Select the Number of Features

K-L divergence is monotone increasing with respect to the number of features. When all features are used, K-L divergence takes the maximum value. As seen in [10] , the most important characteristic of evaluation functions for classifier-independent feature selection is the monotonicity of the functions. Therefore, we use K-L divergence as our evaluation function of a feature subset. Features that contribute only a little to increasing K-L divergence is thought to be garbage features. Thus, a feature subset is chosen by the following two steps. First, features are sorted in order of the values of K-L divergence. Second, the smallest number of features D_α such that attains α-degradation of the K-L divergence of the full feature set is selected (Fig. 1)[9].

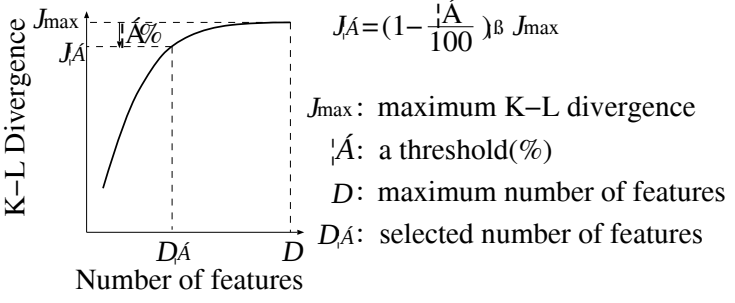


Fig. 1. Method for determining the number of features

3.2 Fake K-L Divergence

If the number of training samples is sufficiently large, the estimated distribution is expected to be close to the true distribution. Then, the estimated K-L divergence is also reliable. However, in practice, because of a limited number of training samples, the calculated K-L divergence can increase by adding a feature without discriminative information. We call it *fake K-L divergence*. Indeed, in an example using a uniform distribution, fake K-L divergence is observed (Fig. 2). In this example, two classes share the same uniform distribution in $[0, 1]^d$; thus, the true K-L divergence is zero for all d. A Gaussian is used for the estimation of the uniform distribution. From Fig. 2, we see larger fake K-L divergence for a smaller number of samples and also it is proportional to the dimensionality d. The fake K-L divergence is approximated by $3.07d/N$, where d is the number of features and N is the number of traning samples. It is possible to use this score $(3.07/N)$ as the amount of fake K-L divergence due to a feature without discriminative information when N samples are given. However, since this sometimes

overestimates the true fake divergence, we use the following estimation. When D is the number of original features, we use the difference between the K-L divergence of Φ_{D-1} and that of Φ_D. This is because, after sorting of features in order of K-L divergence, the last Dth feature is expected to have no discriminative information. Thus, we can regard this difference as the increase due to the fake K-L divergence when one garbage feature is added. Then the influence of fake K-L divergence of d garbage features is estimated by

$$\bar{J}(d) = \{J(\Phi_D) - J(\Phi_{D-1})\} \times d \qquad (2)$$
$$d = 1, 2, \ldots, D.$$

Accordingly, we have a more accurate estimation of $J(\Phi_d)$ by substructing the fake K-L divergence $\bar{J}(\Phi_d)$ from $J(\Phi_d)$, where $\bar{J}(\Phi_d)$ depends on size N.

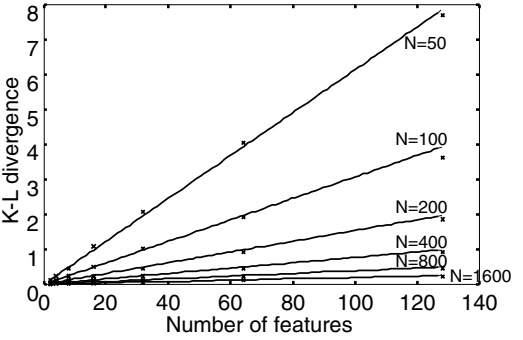


Fig. 2. Fake K-L divergence using a uniform distribution. N is the number of samples.

3.3 Number of Components

In Novovičová *et al.*'s method, there is also no indication how to decide the number of components M in (1). We determine the number of components by the Minimum Description Length (MDL) principle [12]. Varying the number of components in the proper range, we adopt the number of components M for each class that minimizes the following MDL value:

$$\begin{aligned} \mathrm{MDL} &= -L + \frac{1}{2} m \log N \\ m &= M(1 + 2D) \\ L &= \sum_{\mathbf{x} \in \mathbf{X}_\omega} \log p(\mathbf{x}|\omega) \end{aligned} \qquad (3)$$

Here, M is the number of components, D is the number of features, N is the number of samples, and $\mathbf{X}_\omega$ denotes the set of training samples of class ω. The

number of free parameters m is derived from equation (1). To avoid the dependency for initialized parameters, we carried out this experiments 20 times with different initialized parameters and selected the number of components that is minimized (3) the most times.

4 Experiments

We dealt with three real data sets. The threshold α is taken as 1.0% or 0.5%, and the number of components M is determined by the MDL criterion. If all the garbage features are removed properly, the recognition rate is expected to be improved for any kind of classifier. Four classifiers that were used to evaluate the goodness of the selected feature subset are the Bayes linear classifier, the Bayes quadratic classifier, the C4.5 decision tree [13] classifier, and the one-nearest neighbor (1-NN) classifier.

1. *Mammogram*: A mammogram database [9].

   The database is a collection of 86 mammograms from 74 cases. The 65 features are of 18 features characterizing calcification (number, shape, size, etc.) and 47 texture features (histogram statistics, Gabor wavelet response, edge intensity, etc.). There are two classes, one of benign and one of malignant tumors (57 and 29 samples, respectively).

2. *Sonar*: A sonar database [14].

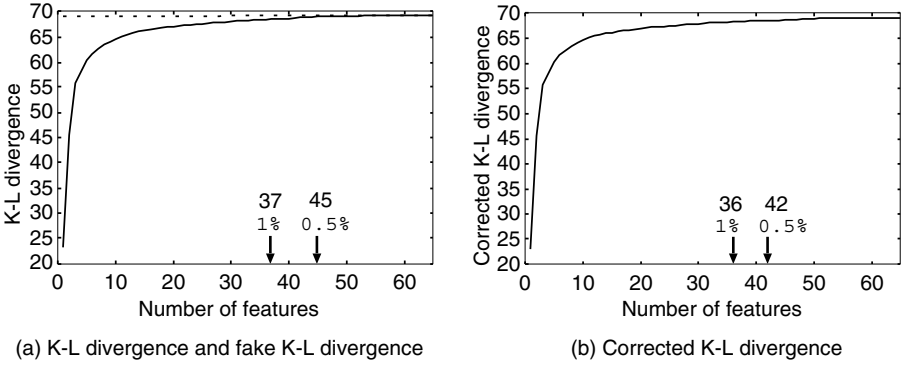
   The task is to discriminate between sonar signals bounced off a metal cylinder and those bounced off a roughly cylindrical rock using 60 features, each of which describes the energy within a particular frequency band, integrated over a certain period of time. The database consists of 111 patterns obtained by bouncing sonar signals off a metal cylinder at various angles and under various conditions and 97 patterns obtained from rocks under similar conditions.

3. *Wdbc*: A Wisconsin breast cancer database [14].

   Thirty features (radius, texture, perimeter, area, etc.) were computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. There are two classes, one of benign and one of malignant tumors (357 and 212 samples, respectively).

The selected number of components were all one for every dataset. For *mammogram* data, the K-L divergence and the estimated fake K-L divergence are shown in Fig 3(a), and the corrected K-L divergence (K-L divergence minus fake K-L divergence) is shown in Fig 3(b). By the correction of K-L divergence, the selected number of features became less than that in the case before correction. In the following, the corrected K-L divergence are used in all experiments. The values of K-L divergence and recognition rates for every number $d (d = 1, 2, \ldots, D)$

are shown in Fig 4. Recoginition rates as well as results by another classifier-independent feature selector, SUB [5] are shown in Tables 1. Here, the recognition rates are calculated by applying the leave-one-out technique to the data.



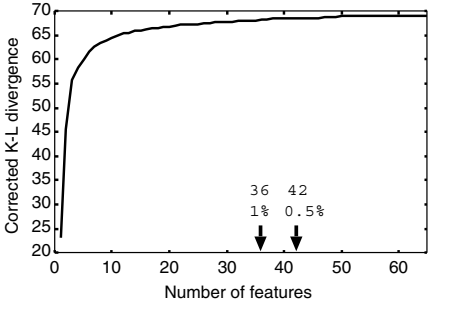
(a) K-L divergence and fake K-L divergence (b) Corrected K-L divergence

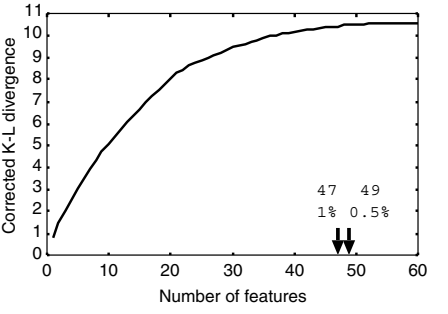
Fig. 3. K-L divergence (—) and fake K-L divergence (- -) for *mammogram* data. In Figure (a), the K-L divergence and the estimated fake K-L divergence are arranged to share the same point at the right end.

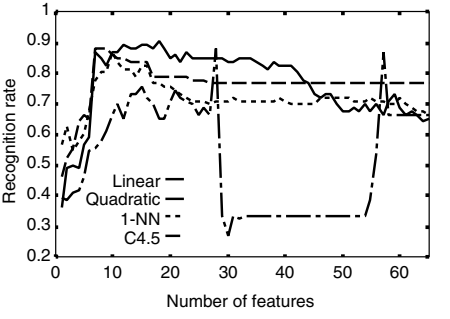
In *mammogram* data, the selected feature subset succeeded in improving or at least maintaining the performance of classifiers except for the quadratic classifier. This exception is because the covariance matrix used in the quadratic classifier became singular owing to the smaller number samples than the dimensionality plus one. In *sonar* data, the recognition rates of the classifiers except for the linear classifier were improved. The densities of the two classes in the *sonar* data share almost the same mean vector, so the linear classifier did not work well. In *wdbc* data, the classification rates of all classifiers were improved, or at least maintained, compared with the case where all features were used. In total, the fundamental effectiveness of the proposed method as a classifier-independent algorithmn was confirmed. In these experimemts, the value of threshold α was set to a small constant of 1% or 0.5%. The larger is the value of α, the smaller is the number of features selected. The felexibility is left to the user.

5 Discussion

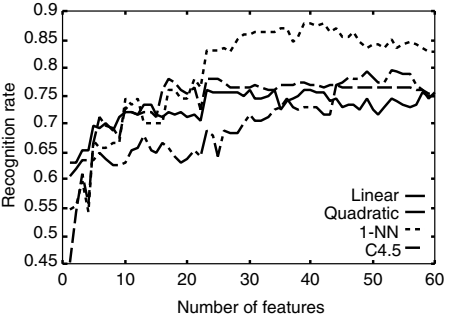
The effectiveness of our approach depends on two factors. One factor is how well the mixture model approximates the densities, and another factor is how well K-L divergence expresses the true performance of a feature subset. In our experiments, the divergence curve is sufficiently smooth and becomes flat as becoming larger number of features. This allows us to choose a fairly small value of α. Ideally, the value of α should be almost zero. If the divergence reflects



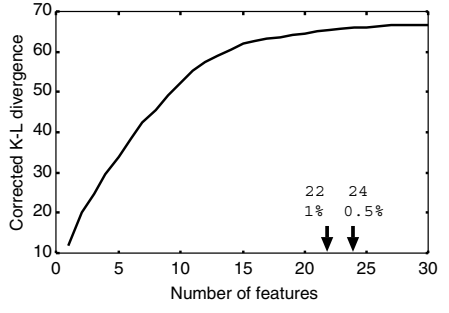


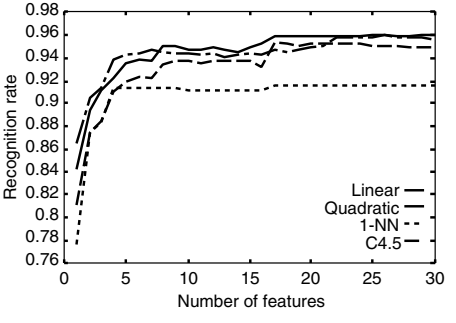


(a) mammogram data



(b) sonar data





(c) wdbc data

Fig. 4. Results of selection of the number of features and the recognition rate.

exactly the true performance of the feature subset, the flatness shows that there are actually some garbage features.

It is very difficult to determine which feature subset is best in the sense of classifier-independent feature selection. Of course, if we have the Bayes classifier, this task can be performed easily. However, in that case every feature contributes to a certain degree to the classification, even if the degree is small. Then, the problem is reduced to finding a compromise between the performance and the cost for measurement of the features.

One method for determining whether or not the selected subset is really effective is to confirm whether the performance has been improved in as many classifiers as possible. Another method is to focus on the performance of the best classifier. For example, the quadratic classifier has a high recognition rate at size 28 in the *mammogram* data. This suggests that the feature subset for classifier-independent feature selection must be a subset larger than 28. In this regard, our feature subset is preferable to that of SUB. For practical use, the user also can select a larger value of α in balance of the priorities of the measuremental cost and performance.

Table 1. Recognition rate obtained by the leave-one-out technique. Values in parentheses are the number of selected features. An up-arrow means on that recognition rate was improved compared with the case that all features were used and a down-arrow means on that recognition rate was degraded.

(a) *mammogram* data

Classifier	Recognition rate[%]			
	Proposed method		SUB	ALL
	α=1.0% (36)	α=0.5% (42)	(10)	(65)
Linear	82.6↑	80.2↑	89.5	65.1
Quadratic	33.7↓	33.7↓	88.4	66.3
1-NN	70.9↑	69.8↑	77.9	66.3
C4.5	76.7	76.7	76.7	76.7

(b) *sonar* data

Classifier	Recognition rate[%]			
	Proposed method		SUB	ALL
	α=1.0% (47)	α=0.5% (49)	(35)	(60)
Linear	73.6↓	74.5↓	76.0	75.0
Quadratic	77.9↑	79.3↑	82.2	75.5
1-NN	86.5↑	83.7↑	84.6	82.7
C4.5	76.4↑	76.4↑	68.8	75.5

(c) *wdbc* data

Classifier	Recognition rate[%]			
	Proposed method		SUB	ALL
	α=1.0% (22)	α=0.5% (24)	(24)	(30)
Linear	96.0↓	96.0↓	97.4	96.1
Quadratic	95.8↑	95.8↑	96.0	95.6
1-NN	91.6	91.6	90.5	91.6
C4.5	95.3↑	95.3↑	94.6	94.9

6 Conclusion

We proposed a method for finding a feature subset from the viewpoint of classifier-independent feature selection. The fundamental effectiveness of the proposed method was comfirmed by the results of experiments conducted on three real data sets. Further examination of the effectiveness of the proposed method using more classifiers is needed.

Acknowledgment

We are most grateful to Proffessor Sklansky at University California, Irvine for providing his database.

References

1. Yu, B., and Yuan, B.: A More Effecient Branch and Bound Algorithm for Feature Selection. Pattern Recognition **26**(1993) 883–889
2. Siedlecki, W., and Sklansky, S.: A Note on Genetic Algorithms for Large-Scale Feature Selection. Pattern Recognition Letters **10**(1989) 335–347
3. Pudil, P., Novovičová, J., and Kittler, J.: Floating Search Methods in Feature Selection. Pattern Recognition Letters **15**(1994) 1119–1125
4. Holz, H.J., and Loew, M.H.: Relative Feature Importance: A Classifier-Independent Approach to Feature Selection. In: Gelsema E.S. and Kanal L.N. (eds.) Pattern Recognition in Practice IV Amsterdam:Elsevier (1994) 473–487
5. Kudo, M., and Shimbo, M.: Feature Selection Based on the Structural Indices of Categories. Pattern Recognition **26**(1993) 891–901
6. Novovičová, J., Pudil, P., and Kittler, J.: Divergence Based Feature Selection for Multimodal Class Densities. IEEE Transactions on Pattern Analysis and Machine Intelligence **18**(1996) 218–223
7. Boekee, D.E., and Van der Lubbe, J.C.A.: Some Aspects of Error Bounds in Feature Selection. Pattern Recognition **11**(1979) 353–360
8. Kudo, M., and Sklansky, J.: A Comparative Evaluation of Medium- and Large-Scale Feature Selectors for Pattern Classifiers. In: 1st International Workshop on Statistical Techniques in Pattern Recognition Prague Czech Republic (1997) 91–96
9. Kudo, M., and Sklansky, J.: Comparison of Algorithms that Select Features for Pattern Classifiers. Pattern Recogntion **33**-1(2000) 25–41
10. Kudo, M., and Sklansky, J.: Classifier-Independent Feature Selection for Two-Stage Feature Selection. Advances in Pattern Recognition, Lecture Notes in Computer Science **1451**(1998) 548–554
11. Dempster, A.P., Laird, N.M., and Rubin, D.B.: Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of Royal Statistical Society **39**(1977) 1–38
12. Ichimura, N.: Robust Clustering Based on a Maximum Likelihood Method for Estimation of the Suitable Number of Clusters. The Transactions of the Institute of Electronics Information and Communication Engineers **8**(1995) 1184–1195 (in Japanese)
13. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann San Mateo CA (1993)
14. Murphy, P.M., and Aha, D.W.: UCI Repository of machine learning databases [Machine-readable data repository]. University of California Irnive, Department of Information and Computation Science (1996)