Classifier-Independent Feature Selection for Two-Stage Feature Selection

Mineichi Kudo[1] and Jack Sklansky[2]

[1] Division of Systems and Information Engineering
Graduate School of Engineering
Hokkaido University, Sapporo 060-8628, Japan.
E-mail:mine@main.eng.hokudai.ac.jp
[2] Department of Electrical Engineering
University of California, Irvine, California 92697, USA.
E-mail:sklansky@uci.edu

Abstract. The effectiveness of classifier-independent feature selection is described. The aim is to remove garbage features and to improve the classification accuracy of all the practical classifiers compared with the situation where all the given features are used. Two algorithms of classifier-independent feature selection and two other conventional classifier-specific algorithms are compared on three sets of real data. In addition, two-stage feature selection is proposed.

1 Introduction

Feature selection is one of the most important issues in pattern recognition. The process is very useful for (1) reducing of the cost of extracting features, (2) improving of the classification accuracy of a practical classifier, and (3) improving the reliability of the estimation of the performance. A large number of algorithms have been proposed for feature selection, and some comparative studies [1, 2, 3] have also been carried out. Most algorithms for feature selection use a criterion based on a specific classifier ("classifier-specific feature selection"). If the classifier to be used is known, these algorithms are useful. However, in many cases, what we want is a "universally effective feature subset." A trial to find such a feature subset is called "classifier-independent feature selection" [4]. If a given problem is large, many garbage features may exist in the measured features, and these garbage features can reduce the performance of classifiers constructed from a limited number of training samples. Removing such garbage features is effective independently of classifiers. Even if we choose a classifier, when the cost of constructing the classifier is very high, e.g., in neural networks, or the estimation of the performance requires so much time, e.g, the leave-one-out estimator, or both, classifier-specific feature selection becomes infeasible.

In this paper, we discuss how well classifier-independent feature selection works in large problems with respect to the number of features (large-scale feature selection).

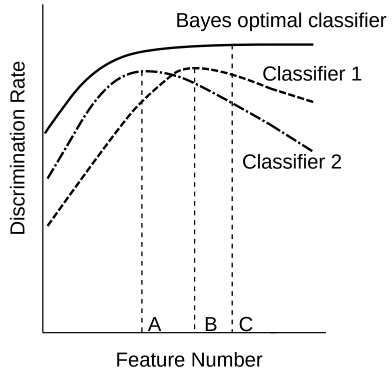

Fig. 1. A conceptual relationship between the number of features and discrimination rate.

2 Two-stage feature selection

We also propose a two-stage approach in which classifier-independent feature selection is first performed, followed by classifier-specific feature selection.

In the first stage, we hope to remove garbage features. A conceptual graph is shown in Fig. 1. Our aim is to reduce the size of a given feature set without a large degradation of the potential discrimination performance, ideally measured by the correct discrimination rate Bayes optimal classifier.

If we determine a classifier to be used, we can use an algorithm of classifier-specific feature selection and maximize the classification accuracy as in point A or B in Fig.1. However, a feature subset that is optimal for one classifier is not necessarily optimal for another classifier. What we really want is a feature subset that is universally effective for any kind of classifier. Therefore, we find a feature subset for which any classifier is improved in its classification accuracy (point C in Fig.1). This can be done by removing garbage features, because garbage features have no discriminative information. Unfortunately, in realistic situations, almost all features have some amount of discriminative information. All we can therefore hope for is to reduce the number of features at the expense of slight degradation of the potential discrimination performance. Such a feature subset is point C in Fig.1. To do this, we need a criterion function that is monotonic with respect to the number of features, as Bayes correct recognition rate is, because such a function has a high possibility of reflecting the potential discrimination performance. For this purpose, we cannot use a specific classifier because such a classifier shows a peaking phenomenon with respect to the number of features and we cannot know when we should terminate the search procedure.

Next, we proceed to the second stage in which a classifier is chosen and then some conventional techniques are used with a criterion function on the basis of the classifier. In this two-stage feature selection, we can expect that: 1) the computation time to find a feature subset would be less than that of the case where a classifier-specific algorithm is applied directly to the whole feature set, and 2) a feature subset that is better than one found through examination of the whole feature set may be found. The latter can happen when the number of original features is very large, for example, more than one hundred. By the limited ability of algorithms to search candidate solutions, algorithms may fail to find better solutions in such large-scale problems.

3 Algorithms

Some algorithms suitable for classifier-independent feature selection have been proposed [4, 5, 6, 7]. These algorithms do not assume any classifier in their criteria, and many of them are based on an approximation of class-conditional density. In addition, the criteria in two of them are monotonic with respect to the number of features. In fact, SUB [5] uses the maximum size of hyper-rectangles enclosing training samples of one class only, and DIV [7] uses divergence.

In our experiment, the latest versions of SUB and DIV were used. Some parts of SUB have been updated: 1) the hyper-rectangles are obtained by a randomized algorithm [8]; and 2) the set of hyper-rectangles including individual training samples maximally is used to obtain the criterion value instead of the single largest hyper-rectangle, where the size is measured by the number of training samples included in the hyper-rectangle; and 3) all features with a score that is less than 0.1 times the average score of features are removed in each step. The value of the parameter for termination is 0.98. DIV is also partly improved by the use of a “dogs and rabbits strategy [9]” for the initial populations of the EM algorithm. The number of component densities in the mixture model is 5 for all classes. In DIV, the desired number of features must be specified, and we adopted the same numbers as those obtained by the other algorithms.

4 Experiments

Used datasets are the following.

1. **Mammogram:** A mammogram database.
   The database is a collection of 86 mammograms from 74 cases which are gathered from the University of California, San Francisco (UCSF), the Mammographic Image Analysis Society (MIAS), and the University of California, Los Angels (UCLA). The 65 features selected include 18 features characterizing calcification (number, shape, size, etc.) and 47 texture features (histogram statistics, Gabor wavelet response, edge intensity, etc.). There are two classes of benign and malignant (57 and 29 samples, respectively).
2. **Sonar:** A sonar database [10].
   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. **Mushroom:** A mushroom database [10].
   The task is to assess the edibility of a large mushroom samples. There are two classes of edible and poisonous and with 4208 and 3916 samples, respectively. The 22 categorical features are converted to 125 numeral features. We chose 1000 samples (500 of each class) randomly for training and the remaining 7124 samples for test.

For a comparison, the sequential forward (backward) floating algorithm, SFFS(SBFS), and the genetic algorithm, GA, were used in a fashion that they try to find the feature subset with the best criterion value, where the nearest neighbor correct recognition rate with the leave-one-out technique was adopted as the criterion value. GA was carried out 16 times in total with four sets of parameters of (p_c=0.6, p_m=0.4, T=50, $P1$), (0.6, 0.4, 50, $P2$), (0.8, 0.1, 50, $P1$), and (0.8, 0.1, 50, $P2$), where p_c is the probability of crossover and p_m is the probability of mutation, T is the maximum number of generations, and $P1$ and $P2$ are two different initializations of population. For details, see [3].

Five classifiers were used for evaluating the selected feature subset. Among these, C4.5 [11] is a decision tree, Subclass [8] is a classifier using hyper-rectangles, and 1-NN is the nearest neighbor rule.

The results are shown in Tables 1- 3.

Table 1. Results of mammogram data. The figures in parentheses are the numbers of selected features.

Classifier	Leave-one-out Recognition Rate (%)							
	FULL	SUB	DIV	DIV	DIV	SFFS	SBFS	GA
	(65)	(10)	(13)	(10)	(7)	(13)	(7)	(29)
Linear	65.1	**89.5**	87.2	82.6	87.2	65.2	88.4	86.0
Quadratic	66.2	**88.4**	76.7	79.1	86.0	61.6	32.6	65.1
C4.5	72.1	76.7	83.7	84.9	**88.4**	69.8	86.0	79.1
Subclass	68.6	80.2	79.1	79.1	76.7	69.8	**86.0**	75.6
1-NN	66.3	77.9	72.1	76.7	79.1	90.7	89.5	**91.9**

In mammogram data, DIV outperformed SFFS and SBFS in almost all cases when the number of features is 7 or 13. In sonar data, DIV is better than GA in the same number of features. In mushroom data, SUB dramatically reduced the number of the original features while maintaining almost the same classification accuracy. In addition, SUB gave better solutions in almost all cases than those of

Table 2. Results of sonar data. The figures in parentheses are the numbers of selected features.

Classifier	Leave-one-out Recognition Rate (%)							
	FULL	SUB	DIV	DIV	DIV	SFFS	SBFS	GA
	(60)	(35)	(40)	(35)	(30)	(18)	(19)	(40)
Linear	75.0	76.0	73.1	76.9	75.5	**79.3**	74.0	73.1
Quadratic	75.5	**82.2**	78.4	79.3	78.8	81.3	77.4	74.0
C4.5	64.9	68.8	69.7	70.2	72.6	70.2	**73.1**	64.9
1-NN	82.7	84.6	83.2	83.2	83.2	95.2	94.7	**95.7**

Table 3. Results of mushroom data. The figures in parentheses are the numbers of selected features. SFFS, SBFS and GA cannot be carried out due to their long computation time.

Classifier	Recognition Rate for Test Samples (%)				
	FULL	SUB	DIV	DIV	DIV
	(125)	(20)	(25)	(20)	(15)
Linear	52.0	**99.8**	98.0	98.0	92.4
Quadratic	52.0	52.0	52.0	52.0	52.0
C4.5	**99.9**	99.4	98.0	98.8	92.4
Subclass	**99.7**	99.4	82.5	98.0	80.1
1-NN	**99.9**	99.8	98.0	98.0	92.4

DIV in these three data and succeeded in improving the classification accuracy of all four classifiers compared with the situation where all given features are used. Although GA found the best solution for 1-NN, 1-NN is just one used for the evaluation.

From these results, we can say: 1) classifier-specific feature selection algorithms are not satisfactory for classifier-independent feature selection as they tend to overfit the classifier used; and 2) classifier-independent feature selection algorithms can improve the classification accuracy of almost all classifiers compared with the situation where all the features are used.

We proceeded to the second stage on mammogram data. Here, the feature subset of size 10 selected by SUB was chosen as the initial feature set. The leave-one-out correct recognition rate with a linear classifier was used as the criterion function for classifier-specific feature selection in the second stage. SFS (the sequential forward search algorithm), SBS(the sequential backward search algorithm), SFFS, and SBFS were carried out. The results are shown in Table 4.

Table 4. Comparison between two-stage feature selection and single feature selection on mammogram data. In the two-stage feature selection, SUB chose 10 features first and then classifier-specific algorithms were carried out. In the second stage, a linear classifier correct recognition rate estimated by the leave-one-out technique was used as the criterion value. For the single feature selection, a 1-NN correct recognition rate estimated by the leave-one-out technique was used as the criterion value. The figures in parentheses are the numbers of selected features.

	Recognition Rate (%), Time (s) and Evaluation Number								
		Two-stage Feature Selection				Single Feature Selection			
	FULL	SUB+	SUB+	SUB+	SUB+	SFFS	SBFS	GA	(std. dev.)
		SFS	SBS	SFFS	SBFS				(16 trails)
	(65)	(6)	(6)	(6)	(6)	(7)	(13)	(22.0)	(± 9.6)
Rate	65.1	93.0	93.0	93.0	93.0	90.7	89.5	90.0	(± 1.4)
Time	0	2104*	2217*	5833*	3637*	261	298	968	(± 5)
Ev. Num.	1	55	55	155	96	11414	13047	6588	(± 1)

∗ includes time consumed by SUB (169 seconds).

In an acceptable time (1,935 seconds for SFS, 2,048 seconds for SBS, 5,664 seconds for SFFS, 3,468 seconds for SBFS on SUN Ultrasparc Station I), these algorithms found a feature subset of six features with the criterion value of 93.0%. This is the best feature subset found on mammogram data. It should be noted that, in our naive calculation, the time consumed in one estimation of the correct recognition rate of a linear classifier is about one-thousand-times longer than the case when 1-NN is used. Therefore, classifier-specific algorithms with a linear classifier could not be applied to the whole feature set in an acceptable time. The two-stage feature selection needed much smaller numbers of evaluations than those of single feature selection (see Table 4).

5 Conclusion

We compared some algorithms suitable for classifier-independent feature selection. Through experiments, we confirmed that algorithms not requiring a specific classifier are necessary for this purpose. Also, two-stage feature selection was confirmed to be effective in the following two points: 1) the computation time to find a feature subset is reduced compared with the situation where a classifier-specific algorithm is applied directly to the whole feature set, and 2) a feature subset found by this two-stage feature selection can be better than one found through examination of the whole feature set, because we can use more time-consuming evaluations and classifiers in the second stage.

Acknowledgment

The authors are grateful to Dr. Pudil for providing the program and to the UCI repository of machine learning databases for providing many kinds of databases. This work was carried out within the Japan-U.S. Cooperative Science Program of the U.S. National Science Foundation (NSF) and the Japan Society for the Promotion of Science (JSPS). The authors thank NSF and JSPS for their financial support.

References

1. Ferri, F. J., Pudil P., Hatef M., Kittler J.: Comparative study of techniques for large-scale feature selection. In Gelsema E. S. and Kanal L. N. eds. Pattern Recognition in Practice IV Elsevier Science B. V. 1994 403–413
2. Jain A., Zongker D.: Feature selection: Evaluation, application, and small sample performance. IEEE Trans. Pattern Anal. Machine Intell. **19**(1997) 153–157
3. Kudo M., 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
4. Holz H. J., 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., Shimbo M.: Feature selection based on the structural indices of categories. Pattern Recognition **26**(1993) 891–901
6. Pudil P., Novovičová J., Kittler J.: Feature selection based on the approximation of class densities by finite mixtures of special type. Pattern Recognition **28**(1995) 1389–1397
7. Novovičová J., Pudil P., Kitler J.: Divergence based feature selection for mulimodal class densities. IEEE Trans. Pattern Anal. and Machine Intell. **18**(1996) 218–223
8. Kudo M., Yanagi S., Shimbo M.: Construction of class regions by a randomized algorithm: A randomized subclass method. Pattern Recognition **29**(1996) 581–588
9. McKenzie P., Alder M.: Initializing the em algorithm for use in gaussian mixture modelling. In Gelsema E. S. and Kanal L. N. eds. Pattern Recognition in Practice IV Amsterdam:Elsevier 1994 91–105
10. Murphy P. M., AhaD. W.: UCI Repository of machine learning databases [Machine-readable data repository]. University of California, Irivne, Department of Information and Computation Science 1996
11. Quinlan J. R.: C4.5: Programs for Machine Learning. Morgan Kaufmann San Mateo CA 1993

This article was processed using the LaTeX macro package with LLNCS style