ESANN'2005 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 27-29 April 2005, d-side publi., ISBN 2-930307-05-6.

Modified Backward Feature Selection by Cross Validation

Shigeo Abe
Graduate School of Science and Technology, Kobe University

Rokkodai, Nada, Kobe, Japan

Abstract. Since training of a classifier takes time, usually some criterion other than the recognition rate is used for feature selection. This may, however, leads to deteriorating the generalization ability by feature selection. To overcome this problem, in this paper, we propose modified backward feature selection by cross validation. Initially, we determine the candidate set which consists of the features that do not deteriorate the generalization ability, if each is deleted from the initial set of features. If the generalization ability is not deteriorated even if all the candidate features are deleted, we terminate the algorithm. Otherwise, we delete by backward deletion the candidate feature that improves the generalization ability the most, and determine the candidate set that is a subset of the current candidate set. We iterate the above procedure until the candidate set is empty. We evaluate our method using support vector machines for some benchmark data sets and show that many features are deleted without deteriorating the generalization ability.

1 Introduction

Feature selection is one of the important steps to construct a pattern classification system with high generalization ability [1, pp. 205–247]. Although support vector machines are robust for a large number of input features, by feature selection we can improve their generalization ability.

The feature selection methods for support vector machines are classified into two: backward or forward feature selection based on some selection criterion [2, 3, 4]; and SVM-based feature selection, in which a feature selection criterion is added to the objective function [5, 6, 7, 8, 9] or forward feature selection is done by changing the margin parameter [10].

The selection criterion used in the literature is, except for some cases [2, 3], the margin [4, 11, 12, 8]. In addition, in most cases, linear support vector machines are used. But, since we want to reduce features without deteriorating the generalization ability, we need to check it after feature selection.

In this paper we discuss reducing features by modified backward feature

selection combined with cross validation. To accurately delete the unnecessary features we use as a selection criterion the generalization ability estimated by cross validation. In contrast to the conventional backward feature selection method, which deletes one feature that has the highest selection criterion, if deleted, at a time, we simultaneously delete the features with the generalization ability, if deleted, higher than or equal to that with all the features. If too many features are deleted, we backtrack and delete one feature that has the highest generalization ability, if deleted, and repeat the above procedure until there is no feature to be deleted.

In Section 2, we discuss the modified backward feature selection method and in Section 3, using some benchmark data sets we demonstrate that the features are deleted without deteriorating the generalization ability.

2 Modified Backward Feature Selection

Usually we select a feature selection criterion other than the recognition rate, because it is time consuming to evaluate the recognition rate. But if the number of training data is small, the recognition rate can be used as a feature selection criterion.[1]

In the following we discuss modified backward feature selection using the generalization ability estimated by cross validation as a selection criterion.

In the backward feature selection, starting from the initial set of features we temporally delete each feature and calculate the value of selection criterion and delete the feature with the highest value of the selection criterion from the set. In the modified backward selection, we delete all the features that improve the generalization ability, if deleted. In addition, if the generalization ability is decreased if a feature is deleted, we consider that the feature is indispensable for classification and we exclude it from the candidate of deletion.

Let the initial set of selected features be F^m, where m is the number of input variables, and the recognition rate of the validation set by cross validation be R^m.

We delete the ith ($i = 1, \ldots, m$) feature temporally from F^m and estimate the generalization ability by cross validation. Let the recognition rate of the validation set be R_i^m. We iterate this procedure for all i ($i = 1, \ldots, m$). Then we rank features according to R_i^m. We call this process *backward feature ranking*.

To speed up backward feature ranking, we consider that the features that satisfy

$$R_i^m < R^m \tag{1}$$

are indispensable for classification and thus they cannot be deleted.

Thus, we set the set of features that are candidates for deletion and also for further feature ranking

$$S^m = \{i \,|\, R_i^m \geq R^m,\, i \in \{1, \ldots, m\}\}. \tag{2}$$

[1] Professor N. Kasabov's lecture at Kobe University on June 1, 2004 showed usefulness of this criterion.

We call S^m candidate set. If S^m is empty, we consider that there is no feature to delete and stop deleting the feature. If only one feature is included, we consider that this feature can be deleted and stop deleting the features.

Assume that S^m includes more than one feature. Then, we set

$$F^k = F^m - S^m \tag{3}$$

where $k = m - |S^m|$ and $|S^m|$ denotes the number of elements in S^m. If

$$R^k \geq R^m, \tag{4}$$

we consider that F^k is the reduced set of features that realize the generalization ability the same with or higher than that with the initial set of features and stop deleting the features.

If (4) is not satisfied, we restore F^k to F^m and delete $\arg\max_{i \in S^m} R_i^m$ from F^m:

$$F^{m-1} = F^m - \{\arg\max_{i \in S^m} R_i^m\}. \tag{5}$$

We call this process *backward feature deletion*. According to the assumption, further feature deletion is possible only for the features in $S^m - \{\arg\max_{i \in S^m} R_i^m\}$. Thus we set the candidate set S^{m-1}:

$$S^{m-1} = \{i \mid R_i^{m-1} \geq R^m, i \in S^m - \{\arg\max_{i \in S^m} R_i^m\}\}. \tag{6}$$

We iterate the above procedure for the feature set F^{m-1} and the candidate set S^{m-1} until the candidate set is empty.

The advantages of our method are as follows:

1. Features can be deleted without deteriorating the generalization ability.
2. If all the features in the candidate set are deleted, the computation time is reduced considerably compared to the conventional backward feature selection method.
3. By determining the candidate set using the current candidate set, the number of steps for backward feature deletion is reduced.

3 Performance Evaluation

We evaluated the proposed method using the data sets listed in Table 1. The first four data sets were used in [1, pp. 205–237]. The diagnosis data set is for classifying skin image data into one of four classes.

Since the generalization abilities of L1 and L2 support vector machines do not differ very much, we used L1 support vector machines for the first four data sets and the L2 support vector machines for the diagnosis data set. Since optimal features change as the kernels are changed. We first determine the optimal kernels for each classification problem by 5-fold cross validation and for

Table 1: Benchmark data specification

Data	Inputs	Classes	Training data	Test data
Iris	4	3	75	75
Numeral	12	10	810	820
Thyroid	21	3	3772	3428
Blood cell	13	12	3097	3100
Diagnosis	45	4	498	1200

the determined kernels we performed feature selection. Namely, we estimated the generalization ability by 5-fold cross validation for a given kernel changing the value of C.

Table 2 shows the results. The "Deleted" column lists the features deleted according to the algorithm, and "Validation" and "Test" columns show the recognition rates of the validation sets and test data sets, respectively. If the recognition rate of the training data is not 100%, it is shown in the bracket. The column "C" lists the value of C determined by cross validation for the training data set. For the iris and numeral data sets we used a polynomial kernels with degree 2, for the blood cell and thyroid data sets, polynomial kernels with degree 4, and for the diagnosis data set, linear kernels.

For the numeral, thyroid, and diagnosis data sets, by deleting all the features in the candidate set from the initial set, the recognition rate for the validation data was higher than or equal to that of the initial set. Thus, the features in the candidate set were all deletable. The recognition rates of the test data sets with the deleted features were higher than those with the initial sets of features for thyroid and numeral data sets and comparable for the diagnosis data set. For the thyroid data set, we could delete 18 features from 21 features and for the diagnosis data set, 25 features out of 45.

For the iris data set, the candidate set S^4 included all four features. Thus, we deleted the two features with higher recognition rates, i.e., the second and third features. But since the recognition rate of the validation set was lower than that with all the features, we backtracked to backward feature selection; since the second and third features had the same recognition rate for the validation set, we deleted the second and third features separately. The resulting recognition rates were higher than that with all the features.

For the blood cell data, deletion of the candidate set S^{13} from the initial set of features resulted in degradation of the generalization ability. Thus, the first feature, which has the maximum generalization ability, if deleted, was deleted by backward feature deletion. The candidate set S^{12} was then calculated as $\{6, 8, 9, 10, 13\}$. Since deletion of all the features in the candidate set did not deteriorate the generalization ability, deletion was terminated. The

Table 2: Feature selection by cross validation. The numerals in the parentheses in the "Deleted" column list the remaining features

Data	Deleted	C	Validation (%)	Test (%)
Iris	None	5000	94.67	93.33
	2, 3	50	93.33 (98.00)	97.33 (98.67)
	2	10	96.00 (97.33)	97.33 (98.67)
	3	500	96.00 (99.00)	96.00 (98.67)
Numeral	None	1	99.51 (99.97)	99.63
	3, 4, 10, 12	1	99.51 (99.94)	99.76
Blood cell	None	1	93.77 (96.23)	93.23 (96.51)
	(4, 7, 12)	10	90.73 (94.54)	89.87 (93.83)
	1, 6, 8, 9, 10, 13	1	94.25 (96.00)	92.45 (95.93)
Thyroid	None	10^5	97.96	97.93
	(3, 8, 17)	10^4	98.52 (99.76)	98.48 (99.81)
Diagnosis	None	1	71.69 (79.62)	73.17 (78.71)
	25 features	1	71.89 (78.66)	73.08 (78.38)

resulting recognition rate of the test data was slightly lower than that with all the features.

In [1, pp. 205–237], exception ratios that are defined by overlapping regions of fuzzy regions are used for selection criteria. Some of the features selected by the exception ratios are included in the features selected by the proposed method. But there is no data set whose selected features are the same. The recognition rates of the iris, numeral, and blood cell test data sets were comparable with or better than those by multilayer neural networks and fuzzy systems. But, that of the thyroid test data set was inferior to that by the fuzzy min-max classifiers, which showed the best recognition rate of 99.42% with all 21 features [1, pp. 177–184].

4 Conclusions

In this paper, we proposed the modified backward feature selection method by cross validation. First, we determine the candidate set that does not deteriorate the generalization ability, if one feature in the set is deleted. Then if the generalization ability is not deteriorated if all the features in the candidate set of features are deleted, we stop deleting the features. Otherwise, we delete one

feature in the candidate set from the set of features and repeat the above procedure. We evaluated our method using support vector machines and showed that many features can be deleted without deteriorating the generalization ability.

References

[1] S. Abe. *Pattern Classification: Neuro-fuzzy Methods and Their Comparison*. Springer-Verlag, London, UK, 2001.

[2] S. Mukherjee, et al. Support vector machine classification of microarray data. Technical Report AI Memo 1677, MIT, 1999.

[3] T. Evgeniou et al. Image representations for object detection using kernel classifiers. In *Proc. Asian Conference on Computer Vision (ACCV 2000)*, pages 687–692, 2000.

[4] I. Guyon et al. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1–3):389–422, 2002.

[5] P. S. Bradley and O. L. Mangasarian. Feature selection via concave minimization and support vector machines. In *Proc. Fifteenth International Conference on Machine Learning (ICML '98)*, pages 82–90, 1998.

[6] J. Weston et al. Feature selection for SVMs. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 668–674. MIT Press, 2001.

[7] J. Weston et al. Use of the zero-norm with linear models and kernel methods. *Journal of Machine Learning Research*, 3:1439–1461, 2003.

[8] S. Perkins, K. Lacker, and J. Theiler. Grafting: Fast, incremental feature selection by gradient descent in function space. *Journal of Machine Learning Research*, 3:1333–1356, 2003.

[9] Y. Grandvalet and S. Canu. Adaptive scaling for feature selection in SVMs. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 569–576. MIT Press, 2003.

[10] M. Brown. Exploring the set of sparse, optimal classifiers. In *Proceedings of Artificial Neural Networks in Pattern Recognition (ANNPR 2003)*, pages 178–184, 2003.

[11] J. Bi et al. Dimensionality reduction via sparse support vector machines. *Journal of Machine Learning Research*, 3:1229–1243, 2003.

[12] A. Rakotomamonjy. Variable selection using SVM-based criteria. *Journal of Machine Learning Research*, 3:1357–1370, 2003.