ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 206, 2007

*Eugeniusz Gatnar**

FEATURE SELECTION AND MULTIPLE MODEL APPROACH IN DISCRIMINANT ANALYSIS

Abstract. Significant improvement of model stability and prediction accuracy in classification and regression can be obtained by using the multiple model approach. In classification multiple models are built on the basis of training subsets (selected from the training set) and combined into an ensemble or a committee. Then the component models (classification trees) determine the predicted class by voting.

In this paper some problems of feature selection for ensembles will be discussed. We propose a new correlation-based feature selection method combined with the wrapper approach.

Key words: tree-based models, aggregation, feature selection, random subspaces.

1. INTRODUCTION

Tree-based models are popular and widely used because they are simple, flexible and powerful tools for classification and regression. Unfortunately they are not stable, i.e. a small change in a predictor value could lead to a quite different model. To solve this problem, single models $C_1(\mathbf{x}), ..., C_K(\mathbf{x})$ are combined into one global model $C^*(\mathbf{x})$.

In classification the component models vote for the predicted class

$$C^*(\mathbf{x}) = \arg\max_y \left\{ \sum_{m=1}^{M} I(\hat{C}_m(\mathbf{x}) = y) \right\} \tag{1}$$

Several variants of aggregation methods have been proposed so far. They manipulate training cases (random sampling) or predictors (random selection) or values of the y (system of weights) or involve randomness directly (Gatnar 2001).

* Professor, Institute of Statistics, Katowice University of Economics, Katowice, Poland.

The method developed by T. K. Ho (1998) has been called "Random subspaces" (RSM). Each component model $C_m(\mathbf{x})$ in the ensemble is fitted to the training subsample U_m containing all cases from the training set but with randomly selected features. Varying the feature subsets used to fit the component classifiers results in their necessary diversity.

This method is very useful, especially when data are highly dimensional, or some features are redundant, or the training set is small compared to the data dimensionality. Similarly, when the base classifiers suffer from the "curse of dimensionality".

The RSM uses a parallel classification algorithm, in contrast to boosting or adaptive bagging that are sequential. It does not require specialised software or any modification of the source code of the existing ones.

A disadvantage of the RSM is the problem of finding the optimal number of dimensions for random subspaces. T. K. Ho (1998) proposed to choose half of the available features while L. Breiman (2001) – the square root of the number of features, or twice the root.

In order to obtain the appropriate number of variables we need to apply a feature selection procedure to the initial number of variables chosen at random.

2. FEATURE SELECTION FOR ENSEMBLES

The aim of feature selection is to find the best subset of variables. In general there are three approaches to feature selection for ensembles:

- filter methods, that filter undesirable features out of the data before classification,
- "wrapper methods", that use the classification algorithm itself to evaluate the usefulness of feature subsets,
- "ranking methods" that score individual features.

Filter methods are the most common used methods for feature selection in statistics. In particular they are the correlation-based methods and we can divide them into three groups: simple correlation-based selection methods, advanced correlation-based selection methods, and contextual merit-based methods.

For example, the method proposed by N. C. Oza and K. Tumar (1999) belongs to the first group. It ranks the features by their correlations with the class. This approach is not effective if there is a strong feature interaction.

The correlation feature selection (CFS) method developed by M. Hall (2000) is advanced because it also takes into account correlations between pairs of features. The CFS value of a set of features F_m is calculated as

$$\mathrm{CFS}(F_m) = \frac{L_m \cdot \bar{r}_j}{\sqrt{L_m + L_m(L_m - 1) \cdot \bar{r}_{ij}}} \tag{2}$$

where:

$\bar{r}_j$ – the average feature-class correlation,
$\bar{r}_{ij}$ – the average feature-feature correlation,
L_m – the number of features in the set F_m.

The wrapper methods generate sets of features. Then they run the classification algorithm using features in each set and evaluate resulting models using 10-fold cross-validation. R. Kohavi and G. H. John (1997) proposed a stepwise wrapper algorithm that starts with an empty set of features and adds single features that improve the accuracy of the resulted classifier. Unfortunately, this method is only useful for data sets with relatively small number of features and very fast classification algorithms (e.g. trees). In general, the wrapper methods are computationally expensive and very slow.

The RELIEF algorithm (Kira, Rendell 1992) is an interesting example of ranking methods for feature selection. It draws instances at random, finds their nearest neighbors, and gives higher weights to features that discriminate the instance from neighbors of different classes. Then those features with weights that exceed a user-specified threshold are selected.

3. PROPOSED METHOD

We propose to reduce the dimensionality of random subspaces using a filter method based on Hellwig heuristic (CFSH). The method is a correlation-based feature selection and consists of two steps.

1. Iterate $m = 1$ to M:

- choose at random half of the data set features ($L/2$) to the training subset U_m,
- determine the best subset F_m of features in U_m according to the Hellwig method,
- grow and prune the tree using the subset F_m.

2. Finally combine the component trees using majority voting.

The heuristic proposed by Z. Hellwig (1969) takes into account both class-feature correlation and correlation between pairs of variables. The best

subset of features is selected from among all possible subsets $F_1, F_2, ..., F_M$ that maximises the so-called "integral capacity of information":

$$H(F_m) = \sum_{j=1}^{L_m} h_{mj} \tag{3}$$

where L_m is the number of features in the subset F_m and h_{mj} is the capacity of information of a single feature x_j in the subset F_m

$$h_{mj} = \frac{r_c^2}{1 + \sum_{i=1, i \neq j}^{L_m} r_{ij}} \tag{4}$$

In the equation (4) r_c is a class-feature correlation, and r_{ij} is a feature-feature correlation.

The correlations r_{ij} are computed using the formula of symmetrical uncertainty coefficient (Press *et al.* 1989) based on the entropy function

$$r_{ij} = \frac{2[\mathrm{E}(x_i) + \mathrm{E}(x_j) - \mathrm{E}(x_i, x_j)]}{\mathrm{E}(x_i) + \mathrm{E}(x_j)} \tag{5}$$

because the variable y representing class is nominal. The measure (5) lies between 0 and 1. If the two variables are independent, then it equals 0, and if they are dependent, it equals 1.

Continuous features have been discretised using the contextual technique of U. M. Fayyad and K. M. Irani (1993).

Unfortunately, maximising the formula (3) in some cases does not lead to the most accurate model from among all models generated by the CFSH method, so the further improvement of the aggregated model accuracy can be achieved combining the CFSH methods with the wrapper approach.

In order to obtain the best subset of features we propose to choose the best model (in terms of classification error) from among the top 5 models containing sets of features generated by the CFSH method. The algorithm contains 3 steps:

1. Choose the top 5 feature subsets $F_1, F_2, ..., F_5$ that maximize the value of (3).

2. Build the models $C(F_1), ..., C(F_5)$ and calculate the classification error for each of them using the appropriate test set: $e[C(F_1)], ..., e[CF_5)]$.

3. Choose the subset F^* that gives model with the lowest classification error:

$$F^* = \arg\min_k \{e[C(F_k)]\} \tag{6}$$

4. EXAMPLE

In order to compare prediction accuracy of ensembles for different feature selection methods we used 9 benchmark data sets from the Machine Learning Repository at the UCI (Blake *et al.* 1998). Results of the comparisons are presented in Tab. 1. For each data set an aggregated model has been built containing $M = 100$ component trees[1].

Table 1

Benchmark data sets

Data set	Number of examples in the learning set	Number of examples in the test set	Number of features	Number of classes
DNA	2 000	1 186	180	3
Letter	15 000	5 000	16	26
Satellite	4 435	2 000	36	6
Soybean	600	83	35	19
German credit	900	100	24	2
Segmentation	2 000	310	19	8
Sick	3 400	372	29	2
Anneal	800	98	38	5
Australian credit	600	90	15	2

Classification errors have been estimated for the appropriate test sets and presented in Tab. 2.

Table 2

Classification errors for the data sets (in %)

Data set	Single model (tree)	CFS	CFSH	Wrapper approach
DNA	6.40	5.20	4.51	4.78
Letter	14.00	10.83	5.84	5.34
Satellite	13.80	14.87	10.32	10.28
Soybean	8.00	9.34	6.98	7.05
German credit	29.60	27.33	26.92	26.72
Segmentation	3.70	3.37	2.27	2.14
Sick	1.30	2.51	2.14	2.12
Anneal	1.40	1.22	1.20	1.20
Australian credit	14.90	14.53	14.10	14.04

[1] In order to grow trees we have used the Rpart procedure written by T. M. Therneau and E. J. Atkinson (1997) for the *S*-PLUS and *R* environment.

In this paper we have proposed a modification of the correlation-based feature selection method for classifier ensembles based on the Hellwig heuristic. The wrapper approach gives more accurate aggregated models than those built with the CFSH correlation-based feature selection method.

REFERENCES

Blake C., Keogh E., Merz C. J. (1998), *UCI Repository of Machine Learning Databases*, Department of Information and Computer Science, University of California, Irvine (CA).

Breiman L. (2001), *Random forests*, "Machine Learning", **45**, 5–32.

Fayyad U. M., Irani K. B. (1993), *Multi-interval discretisation of continuous-valued attributes*, [in:] *Proceedings of the XIII International Joint Conference on Artificial Intelligence*, Morgan Kaufmann, San Francisco, 1022–1027.

Gatnar E. (2001), *Nieparametryczna metoda estymacji i regresji*, Wydawnictwo Naukowe PWN, Warszawa.

Hall M. (2000), *Correlation-based feature selection for discrete and numeric class machine learning*, [in:] *Proceedings of the 17th International Conference on Machine Learning*, Morgan Kaufmann, San Francisco.

Hellwig Z. (1969), *On the problem of the optimal selection of predictors*, "Statistical Revue", **3–4** (in Polish).

Ho T. K. (1998), *The random subspace method for constructing decision forests*, IEEE Trans. on Pattern Analysis and Machine Learning, **20**, 832–844.

Kira A., Rendell L. (1992), *A practical approach to feature selection*, [in:] *Proceedings of the 9th International Conference on Machine Learning*, D. Sleeman, P. Edwards (eds.), Morgan Kaufmann, San Francisco, 249–256.

Kohavi R., John G. H. (1997), *Wrappers for feature subset selection*, "Artificial Intelligence", **97**, 273–324.

Oza N. C., Tumar K. (1999), *Dimensionality reduction through classifier ensembles. Technical Report, NASA-ARC-IC-1999-126*, Computational Sciences Division, NASA Ames Research Center.

Press W. H., Flannery B. P., Teukolsky S. A., Vetterling W. T. (1989), *Numerical recipes in Pascal*, Cambridge University Press, Cambridge.

Therneau T. M., Atkinson E. J. (1997), *An introduction to recursive partitioning using the RPART routines*, Mayo Foundation, Rochester.

Eugeniusz Gatnar

DOBÓR ZMIENNYCH A PODEJŚCIE WIELOMODELOWE W ANALIZIE DYSKRYMINACYJNEJ

W pracy przedstawiono podstawowe zagadnienia związane z doborem zmiennych w podejściu wielomodelowym (*multiple-model approach*) w analizie dyskryminacyjnej.

Podejście wielomodelowe polega na budowie K modeli, prostych (składowych) $C_1(x), ..., C_K(x)$, które są następnie łączone w jeden model zagregowany $C^*(x)$, np. w oparciu o zasadę majoryzacji

$$\hat{C}^*(\mathbf{x}) = \arg\max_y \left\{ \sum_{m=1}^{M} I(\hat{C}_m(\mathbf{x}) = y) \right\}.$$

Znane z literatury metody agregacji modeli różnią się przede wszystkim sposobem tworzenia prób uczących $U_1, \ldots, U_K$, w oparciu o które powstają modele składowe. Jedną z najprostszych jest metoda losowego doboru zmiennych do modeli składowych.

Aby jednak zmienne te wpływały na jakość budowanych modeli zaproponowano wykorzystanie metody doboru zmiennych spośród tych, które zostały wylosowane. W tym celu zmodyfikowano metodę korelacyjną Hellwiga.