ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 216, 2008

*Eugeniusz Gatnar**

COMBINING DIFFERENT TYPES OF CLASSIFIERS

ABSTRACT. Model fusion has proved to be a very successful strategy for obtaining accurate models in classification and regression. The key issue, however, is the diversity of the component classifiers because classification error of an ensemble depends on the correlation between its members.

The majority of existing ensemble methods combine the same type of models, e.g. trees. In order to promote the diversity of the ensemble members, we propose to aggregate classifiers of different types, because they can partition the same classification space in very different ways (e.g. trees, neural networks and SVMs).

Key words: multiple-model approach, model fusion, classifier ensemble, diversity measures.

I. INTRODUCTION

Fusion of classification models is commonly used in classification in order to improve classification accuracy. In this approach K component (base) models $C_1(\mathbf{x}),..., C_K(\mathbf{x})$ are combined into one global model (ensemble) $C^*(\mathbf{x})$, for example using majority voting:

$$C^*(\mathbf{x}) = \arg\max_y \left\{ \sum_{k=1}^{K} I(C_k(\mathbf{x}) = y) \right\}. \quad (1)$$

Tumer i Ghosh (1996) proved that the classification error of the ensemble $C^*(\mathbf{x})$ depends on the diversity of the ensemble members. In other words, the higher diversity of component models, the lower classification error of the combined model.

* Ph. D., Chair of Statistics, Katowice University of Economics, Katowice, Poland.

The high accuracy of the classifier ensemble $C^*(\mathbf{x})$ is achieved if the members of the ensemble are "weak" and diverse. The term "weak" refers to classifiers that have high variance, e.g. classification trees, nearest neighbors, and neural nets.

Diversity among classifiers means that they are different from each other, i.e. they misclassify different examples. This is obtained by using different training subsets, assigning different weights to instances or selecting different subsets of features (subspaces).

Several variants of aggregation methods have been developed so far. They differ in two aspects: the way the subsets to train component classifiers are formed and the method the base classifiers are combined. Generally, there are three approaches have been developed to obtain diversity among component models:

- Manipulating training examples, e.g. *Bagging* (Breiman, 1996); *Boosting* (Freund and Shapire, 1997) and *Arcing* (Breiman, 1998).
- Manipulating input features: *Random subspaces* (Ho, 1998); *Random split selection* (Amit and Geman, 1997), *Random forests* (Breiman, 2001).
- Manipulating output values: *Adaptive bagging* (Breiman, 1999); *Error-correcting output coding* (Dietterich and Bakiri, 1995).

II. BASE CLASSIFIERS

Existing ensemble methods combine the same type of models built for different subsets of observations, e.g. RandomForest developed by Breiman (2001), or different subsets of features, e.g. Feature Subspaces developed by Ho (1998). In order to improve the diversity of the ensemble members, we proposed to fuse classifiers of different types.

We used 6 types of classifiers.

- k -Nearest Neighbors,
- Linear Discriminants:

$$f_j(\mathbf{x}) = \mathbf{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_j - \frac{1}{2} \boldsymbol{\mu}_j^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_j \tag{2}$$

and Quadratic Discriminants:

$$f_j(\mathbf{x}) = -\frac{1}{2} \log \left| \boldsymbol{\Sigma}_j \right| - \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1} (\mathbf{x} - \boldsymbol{\mu}_j), \tag{3}$$

for the j -th class.

- **Classification Trees:**

$$f(\mathbf{x}) = \sum_{k=1}^{K} \alpha_k I(\mathbf{x} \in R_k), \tag{4}$$

where R_k are disjoint regions in the feature space $I(\mathbf{x} \in R_k) = \prod_{l=1}^{L} I\left(v_{kl}^{(d)} \le x_l \le v_{kl}^{(g)}\right)$.

- **Neural Networks with one hidden layer:**

$$\mathbf{y}_j = g_j\left(v_{0j} + \sum_{m=1}^{M} v_{mj}\mathbf{z}_m\right), \tag{5}$$

where $\mathbf{z} = [z_1, z_2, z_3, ..., z_M]$ is a set of variables in the hidden layer:

$$\mathbf{z}_m = h\left(w_{0m} + \sum_{n=1}^{N} w_{nm}\mathbf{x}_n\right) \tag{6}$$

and h is an activation function. We chosen the sigmoid function $h(u) = \dfrac{1}{1+\exp(-u)}$ to activate the neurons.

- **Support Vector Machines (SVM):**

$$f(\mathbf{x}) = \sum_{(\mathbf{x}_i, y_i) \in V} l_i\, y_i K(\mathbf{x}_i, \mathbf{x}) + \hat{\alpha}_0, \tag{7}$$

where $V = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_S, y_S)\}$ is the set of support vectors and $\mathrm{K}(\mathbf{u}, \mathbf{v}) = \exp(-\gamma\|\mathbf{u} - \mathbf{v}\|^2)$ is the Gaussian kernel function.

III. CLASSIFICATION OF EUROPEAN COUNTRIES

In order to compare the accuracy of the two types of ensembles we performed the classification of European countries based on the World Bank data and the AMECO database.

The World Bank classifies economies based on the gross national income (GNI) per capita[1] to one of four classes:

- H – high income ($10,726 and more),
- UM – upper middle income ($3,466–$10,725),
- LM – low middle income ($876–$3,465),
- L – low income ($875 or less).

This was the true class for each country in the training set.

The AMECO database is the annual macro-economic database of the European Commission's Directorate General for Economic and Financial Affairs. It contains data for EU-25, the euro area, EU Member States, candidate countries and other OECD countries (United States, Japan, Canada, Switzerland, Norway, Iceland, Mexico, Korea, Australia and New Zealand).

The database contains a selection of about 700 variables, e.g. Gross Savings, Final Consumption Expenditure of General Government, Exports of Goods and Services, Imports of Goods and Services, Unemployment Rate, etc.

We have collected 780 observations in the training set:

- 15 countries (old EU-15 members) observed in the years 1970–2005.
- 14 countries (10 new EU-25 members and 4 candidate countries: Bulgaria, Romania, Turkey, Croatia) observed in the years 1991–2005.

In order to classify the European countries we started with single classification models, and Figure 1 shows how they divide the classification space. We used the 10-fold cross-validation to assess their performance and the estimated classification errors are presented in the table 1.

Then we combined classifiers of the same type using bagging, boosting and random subspace method. Their errors are presented in Table 1.

Table 1

Classification errors for different combining methods

Method	3-NN	LDA	QDA	Tree	Nnet	SVM
Single model (CV)	12.79%	25.13%	22.67%	17.44%	15.38%	19.21%
Bagging	12.88%	23.46%	21.92%	12.31%	13.08%	18.46%
Boosting	12.59%	21.56%	20.05%	13.15%	12.78%	17.63%
Random Subspace	12.21%	20.97%	19.34%	12.23%	11.83%	17.45%

[1] www.worldbank.org/Home/Data/CountryClassification.

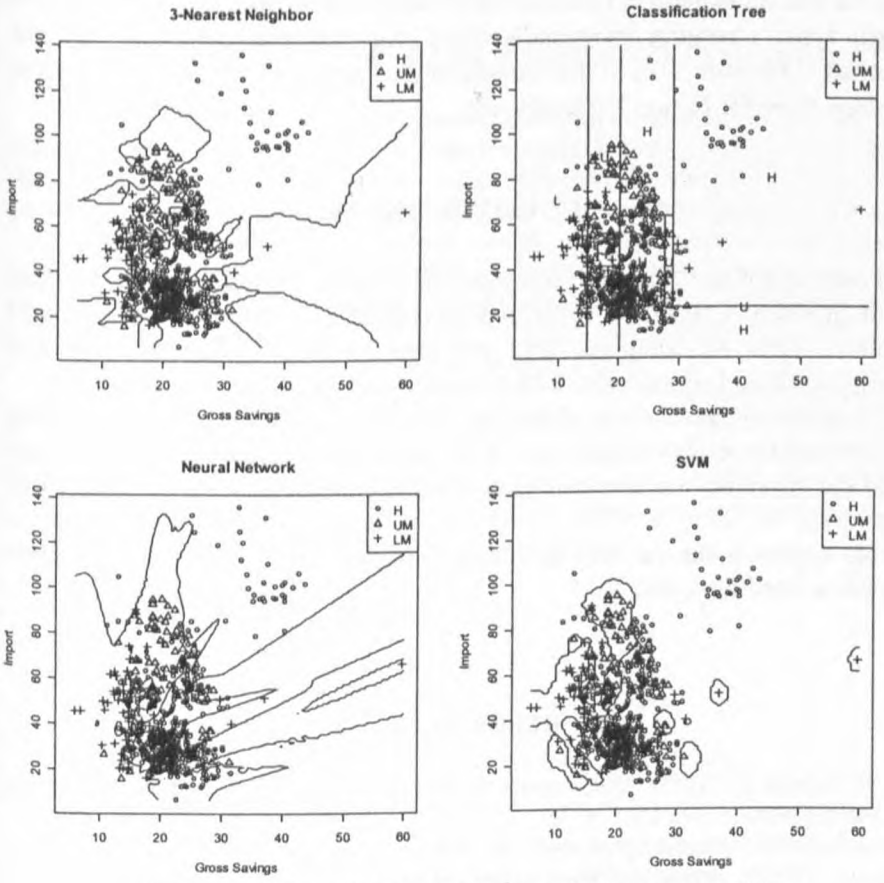

Figure 1. Partition of the classification space by 3-nearest neighbors, trees, neural nets and SVMs

Then we combined classification models of different types using majority voting (1) and we observed significant improvement of the classification accuracy. The results are shown in table 2. We have pruned trees, while NNets and SVMs have been tuned over supplied parameter ranges with the „tune" function from the e1071 library in the package R.

Table 2

Classification errors for combining different models

Ensemble	Error
6 models	11.54%
60 models	10.28%

In the second experiment we created 60 ensemble members as 10 classifiers of each type, changing their parameters, e.g. the parameter „k" for the k -Nearest Neighbors, size of the Classification Trees, the number of neurons in the hidden layer for Neural Networks, etc.

IV. CONCLUSIONS

In our experiments we have combined classifiers of different types, i.e. Linear and Quadratic Classifiers, Trees, Neural Networks, SVM models and Nearest Neighbors. Then we compared their performance with the ensembles formed using the standard fusion methods like *bagging* or *boosting*.

The obtained results showed that ensembles of classifiers of different types outperformed those of the same type. They are more accurate because the members of the ensemble are diverse and divide the classification space in very different way.

This approach can be used in difficult domains, e.g. in economics, pattern recognition, medicine, etc.

REFERENCES

Amit Y., Geman D. (1997): Shape quantization and recognition with randomized trees, *Neural Computation*, 1545–1588.

Breiman L. (1996): Bagging predictors. *Machine Learning*, 24, 123–140.

Breiman L. (1998): Arcing classifiers. *Annals of Statistics*, 26, 801–849.

Breiman L. (1999): Using adaptive bagging to debias regressions. Technical Report 547, Department of Statistics, University of California, Berkeley.

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

Dietterich T., Bakiri G. (1995): Solving multiclass learning problem via error-correcting output codes. *Journal of Artificial Intelligence Research*, 2, 263–286.

Freund Y., Schapire R.E. (1997): A decision-theoretic generalization of on-line learning and an application to boosting, *Journal of Computer and System Sciences* 55, 119–139.

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

Tumer K., Ghosh J. (1996): Analysis of decision boundaries in linearly combined neural classifiers. *Pattern Recognition* 29, 341–348.

Eugeniusz Gatnar

ŁĄCZENIE RÓŻNYCH RODZAJÓW MODELI DYSKRYMINACYJNYCH

Łączenie modeli okazało się być bardzo efektywną strategią poprawy jakości predykcji modeli dyskryminacyjnych. Kluczowym zagadnieniem, jak wynika z twierdzenia Tumera i Ghosha (1996), jest jednak stopień różnorodności agregowanych modeli, tzn. im większa korelacja między wynikami klasyfikacji tych modeli, tym większy błąd.

Większość znanych metod łączenia modeli, np. RandomForest zaproponowany przez Breimana (2001), agreguje modele tego samego typu w różnych przestrzeniach cech. Aby zwiększyć różnice między pojedynczymi modelami, w referacie zaproponowano łączenie modeli różnych typów, które zostały zbudowane w tej samej przestrzeni zmiennych (np. drzewa klasyfikacyjne i modele SVM).

W eksperymentach wykorzystano 5 klas modeli: liniowe i kwadratowe modele dyskryminacyjne, drzewa klasyfikacyjne, sieci neuronowe, oraz modele zbudowane za pomocą metody *k*-najbliższych sąsiadów (k-NN) i metody wektorów nośnych (SVM).

Uzyskane rezultaty pokazują, że modele zagregowane powstałe w wyniku łączenia różnych modeli są bardziej dokładne niż gdy modele składowe są tego samego typu.