ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 225, 2009

*Eugeniusz Gatnar**

MEASURES OF DIVERSITY AND THE CLASSIFICATION ERROR IN THE MULTIPLE-MODEL APPROACH

Abstract

Multiple-model approach (model aggregation, model fusion) is most commonly used in classification and regression. In this approach K component (single) models $C_1(\mathbf{x})$, $C_1(\mathbf{x})$, ..., $C_K(\mathbf{x})$ are combined into one global model (ensemble) $C^*(\mathbf{x})$, for example using majority voting:

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

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.

Since several diversity measures for classifier ensembles have been proposed so far in this paper we present a comparison of the ability of selected diversity measures to predict the accuracy of classifier ensembles.

Key words: Multiple-model approach, Model fusion, Classifier ensemble, Diversity measures.

1. Introduction

Several variants of aggregation methods have been developed in the past decade. 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 to obtain the training subsets:

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

- 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).

Having a set of classifiers they can be combined using one of the following methods:

- averaging methods, e.g. average vote and weighted vote;
- non-linear methods, e.g. majority vote (the component classifiers vote for the most frequent class as the predicted class), maximum vote, Borda Count method, etc.;
- stacked generalisation, where the classifiers are fitted to training subsamples obtained by leave-one-out cross-validation (Wolpert, 1992).

2. Diversity

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 neighbours, and neural nets.

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

Tumer and Ghosh (1996) proved that the classification error of the ensemble $C^*(\mathbf{x})$ depends on the diversity of the ensemble members:

$$e(C^*) = e^B(C^*) + \frac{1 + r(K-1)}{K} e(C_i), \tag{2}$$

where $e^B(C^*)$ is the Bayes error, r is average correlation coefficient between errors of component classifiers, K is the number of base classifiers, and $e(C_i)$ is an error of individual classifier.

As we can see in formula (2) the ensemble error decreases with decrease of the correlation between the component classifiers.

In general, Sharkey and Sharkey (1997) introduced four levels of diversity:

- Level 1 – no more than one classifier is wrong on each example.
- Level 2 – up to half of classifiers could be wrong for each example (majority vote is always correct).
- Level 3 – at least one classifier is correct for each example.
- Level 4 – none of the classifiers is correct for some examples.

The level of diversity among candidate classifiers determines the method they should be combined. For example, the majority vote is good for the classifiers that exhibit the level 1 an 2 diversity. Otherwise some more complex methods, e.g. stacked generalization, are more appropriate.

Several combining methods that take into account the diversity of classifiers have been proposed in the literature. For example, Rosen (1996) presented a combination method that incorporates an error-decorrelation penalty term. It allows component classifiers to make errors which are uncorrelated. Hashem (1999) proposed the use of relative accuracy of component classifiers as weights in linear combinations of the members of an ensemble. Oza and Tumer (1999) developed a method named *input decimation* that eliminates features low correlated with the class. Zenobi and Cunningham (2001) have used the hill-climbing search for feature subsets that is guiding by a diversity measure.

Recently, Melville and Mooney (2004) developed a method called DECORATE that reduces the classification error of the ensemble by increasing diversity. It adds artificial examples to the original training set, i.e. examples oppositely labelled.

3. Measures of diversity

We can simply measure the agreement between two classifiers $C_i(\mathbf{x})$ and $C_j(\mathbf{x})$ as:

$$Agreement(C_i, C_j) = \frac{1}{N}\sum_{n=1}^{N} I(C_i(\mathbf{x}_n) = C_j(\mathbf{x}_n)) \tag{3}$$

but this takes into account both correct and incorrect classifications of the component models.

In order to overcome this drawback we define the „oracle" output (O) of the classifier $C_k(\mathbf{x})$ as:

$$O_k(\mathbf{x}_i)=\begin{cases}1 & C_k(\mathbf{x}_i)=y_i\\ 0 & C_k(\mathbf{x}_i)\neq y_i\end{cases} \tag{4}$$

In other words, the value of $O_k(\mathbf{x})=1$ means that the classifier $C_k(\mathbf{x})$ is correct, i.e. it recognizes the true class (y) of the example $\mathbf{x}$, and $O_k(\mathbf{x})=0$ means that the classifier is wrong. The relationship between a pair of classifiers is presented in two-way contingency table (Table 1).

Table 1

Oracle labels for two classifiers

	$O_i(x)=1$	$O_i(x)=0$
$O_j(x)=1$	a	b
$O_j(x)=0$	c	d

Source: own study.

The most simple measure of diversity (or rather similarity) of two classifiers is the binary version of the Pearson's correlation coefficient:

$$r(i,j)=\frac{ad-bc}{(a+b)(c+d)(a+c)(b+d)} \tag{5}$$

Partridge and Yates (1996), and Margineantu and Diettrich (1997) have used a measure named **within-set generalization diversity** (kappa statistics):

$$K(i,j)=\frac{2(ac-bd)}{(a+b)(c+d)+(a+c)(b+d)} \tag{6}$$

It measures the level of agreement between two classifiers with the correction for chance.

Skalak (1996) has proposed the disagreement measure.

$$DM(i,j)=\frac{b+c}{a+b+c+d} \tag{7}$$

This is the ratio between the number of examples on which one classifier is correct and the other is wrong to the total number of examples.

Giacinto and Roli (2001) have introduced a measure named **compound diversity**:

$$CD(i,j) = \frac{d}{a+b+c+d} \tag{8}$$

This measure is also named "double-fault measure" because it is the proportion of the examples that have been misclassified by both classifiers.

Kuncheva *et al.* (2000) recommended the **Yule's Q statistics** as diversity measure:

$$Q(i,j) = \frac{ad-bc}{ad+bc} \tag{9}$$

The Yule's Q statistics is the original measure of dichotomous agreement, designed to be analogous to the correlation. This measure is pairwise and symmetric and varies between –1 and 1. A value of "0" indicates statistical independence of classifiers, positive values mean that the classifiers have recognized the same examples correctly and negative values – that the classifiers commit errors on different examples.

Gatnar (2005) proposed the **Hamann's coefficient**:

$$H(i,j) = \frac{(a+d)-(b+c)}{a+b+c+d} \tag{10}$$

This binary similarity coefficient is simply the difference between the matches and mismatches as a proportion of the total number of entries. It ranges from –1 to 1. A value of "0" indicates an equal number of matches to mismatches, "–1" represents perfect disagreement and "1" – perfect agreement.

In the case of pairwise measures, the overall value of diversity for the ensemble $C^*(\mathbf{x})$ is computed as the mean:

$$Diversity(C^*) = \frac{2}{K(K-1)} \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} Diversity(C_i, C_j) \tag{11}$$

Several non-pairwise measures have been also developed to estimate the diversity between classifiers $C_1(\mathbf{x})$, $C_1(\mathbf{x})$, ..., $C_K(\mathbf{x})$.

Hansen and Salamon (1990) proposed the measure of **difficulty**:

$$DI = \frac{1}{K}\sum_{k=1}^{K}\left(p_V(k) - \overline{p}_V\right)^2 \tag{12}$$

It is in fact the variance of variable $p_v(k)$ representing the proportion of classifiers that correctly classify an example **x** chosen at random.

Partridge and Krzanowski (1997) have introduced the **generalized diversity** measure:

$$GD = 1 - \frac{p(2)}{p(1)} \tag{13}$$

where:

$$p(1) = \frac{1}{K}\sum_{k=1}^{K} k \cdot p_k \quad \text{and} \quad p(2) = \frac{1}{K(k-1)}\sum_{k=1}^{K} k(k-1)p_k \tag{14}$$

and p_k is thef probability that k classifiers mischassify an eample **x chosen at random.**

Cunnigham and Carney (2000) used the entropy function:

$$IEN = \frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{I} -\frac{N_i^n}{K}\log_I\left(\frac{N_i^n}{K}\right) \tag{15}$$

where N_i^n is number of base classifiers that misclassified $\mathbf{x}_n$ to class C_i.

4. Experiments

In order to compare the ability of the diversity measures to detect the accuracy of combined classifier we followed the synthetic experiment presented in (Kuncheva *et al.*, 2000). We have generated two artificial sets of classifier ensembles of known classification accuracy.

In the first experiment we have used a test set of 10 examples ($N = 10$) and 3 classifiers: $C_1(\mathbf{x}), C_2(\mathbf{x}), C_3(\mathbf{x})$ (i.e. $K = 3$) and each has the same classification error $e(C_i) = 0.4$ (6 out of 10 examples are recognized correctly). That gave the total number of 28 different combinations of classification results for the test set.

We have used all the measures of diversity mentioned above and majority vote combining. The Pearson's correlation coefficients between the ensemble error and the diversity are presented in Table 2.

Table 2

Pearson's correlation with the ensemble error in experiment 1

Diversity measure	Correlation with the error
K	0.209
DM	0.387
CD	0.408
Q	0.421
H	0.532
DI	0.324
GD	0.543
IEN	0.412

Source: own study.

In the second experiment we have generated a test set of 100 examples ($N = 100$) and also 3 classifiers: $C_1(\mathbf{x}), C_2(\mathbf{x}), C_3(\mathbf{x})$ (i.e. $K = 3$) each has the same classification error $e(C_i) = 0.4$ (60 out of 100 examples are recognized correctly). That gives the total number of 36 151 different combinations of classification results.

We have used all the measures of diversity and majority vote combining. Pearson's correlation coefficients between the ensemble error and the diversity are presented in Table 3.

Table 3

Pearson's correlation with the ensemble error in experiment 2

Diversity measure	Correlation with the error
K	0.387
DM	0.402
CD	0.478
Q	0.471
H	0.598
DI	0.539
GD	0.678
IEN	0.578

Source: own study.

5. Conclusions

In this paper we compared the ability of several diversity measures to detect the accuracy of a classifier ensemble.

As the result of two experiments we conclude that the Hamann's coefficient is the best diversity measure among the pairwise diversity measures, while in the group of non-pairwise measures the Partridge and Krzanowski's measure is the recommended.

In general, we observed that non-pairwise measures better identify the diversity among component models (which is the main reason of the prediction error) than the pairwise ones.

References

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.

Cunnigham P., Carney J. (2000), *Diversity versus quality in classification ensembles based on feature selection*, [in:] *Proceedings of European Conference on Machine Learning*, LNCS, vol. 1810, Springer, Berlin, 109–116.

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

Fleiss J. L. (1981), *Statistical methods for rates and proportions*, John Wiley and Sons, New York.

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.

Gatnar E. (2001), *Nonparametric method for classification and regression*, PWN, Warszawa (in Polish).

Gatnar E. (2005), *A diversity measure for tree-based classifier ensembles*, [in:] *Data analysis and decision support*, eds D. Baier, R. Decker, L. Schmidt-Thieme, Springer-Verlag, Heidelberg–Berlin, 30–38.

Giacinto G., Roli F. (2001), *Design of effective neural network ensembles for image classification processes*, "Image Vision and Computing Journal", **19**, 699–707.

Hansen L. K., Salamon P. (1990), *Neural network ensembles*, "IEEE Transactions on Pattern Analysis and Machine Intelligence", **12**, 993–1001.

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

Kuncheva L., Whitaker C., Shipp D., Duin R. (2000), *Is independence good for combining classifiers*, [in:] *Proceedings of the 15th International Conference on Pattern Recognition*, Barcelona, Spain, 168–171.

Kuncheva L., Whitaker C. (2003): *Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy*, "Machine Learning", **51**, 181–207.

Margineantu M. M., Dietterich T. G. (1997), *Pruning adaptive boosting*, [in:] *Proceedings of the 14th International Conference on Machine Learning*, Morgan Kaufmann, San Mateo, 211–218.

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.

Partridge D., Krzanowski W. J. (1997), *Software diversity: practical statistics for its measurement and exploitation*, "Information and software Technology", **39**, 707–717.

Partridge D., Yates W. B. (1996), *Engineering multiversion neural-net systems*, "Neural Computation", **8**, 869–893.

Sharkey A., Sharkey N. (1997), *Diversity, selection, and ensembles of artificial neural nets*, [in:] *Neural Networks and their applications*, NEURAP-97, 205–212.

Skalak D. B. (1996), *The sources of increased accuracy for two proposed boosting algorithms*, [in:] *Proceedings of the American Association for Artificial Intelligence AAAI-96*, Morgan Kaufmann, San Mateo.

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

Wolpert D. (1992), *Stacked generalization*, "Neural Networks", **5**, 241–259.

Eugeniusz Gatnar

Miary zróżnicowania modeli a błąd klasyfikacji w podejściu wielomodelowym

Podejście wielomodelowe (agregacja modeli), stosowane najczęściej w analizie dyskryminacyjnej i regresyjnej, polega na połączeniu M modeli składowych $C_1(\mathbf{x}), \ldots, C_M(\mathbf{x})$ jeden model globalny $C^*(\mathbf{x})$:

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

Tumer i Ghosh (1996) udowodnili, że błąd klasyfikacji dla modelu zagregowanego $C^*(\mathbf{x})$ zależy od stopnia podobieństwa (zróżnicowania) modeli składowych. Inaczej mówiąc, najbardziej dokładny model $C^*(\mathbf{x})$ składa się z modeli najbardziej do siebie niepodobnych, tj. zupełnie inaczej klasyfikujących te same obiekty.

W literaturze zaproponowano kilka miar pozwalających ocenić podobieństwo (zróżnicowanie) modeli składowych w podejściu wielomodelowym.

W artykule omówiono związek znanych miar zróżnicowania z oceną wielkości błędu klasyfikacji modelu zagregowanego.