Heterogenous Committees with Competence Analysis

Norbert Jankowski & Krzysztof Grąbczewski
Department of Informatics, Nicolaus Copernicus University
ul. Grudziądzka 5, 87–100 Toruń, Poland
[norbert,kgrabcze]@phys.uni.torun.pl
http://www.phys.uni.torun.pl/kis

Abstract

We explore some new types of committees in search of hybrid models successful in many different classification benchmarks. To provide a reliable comparison of the ensembles we restrict the task to some constant configuration of committee members for each benchmark. We were looking for new types of committees which, in such configuration, would be as much accurate and stable as possible. The paper focuses on some ideas of heterogenous committees with different ways of their members competence estimation. Heterogenous committee members adapt in different ways are able to solve different problems. Measuring the competence of committee members helps in making competent and accurate decisions.

1. Introduction

In general, adaptive models are combined into ensembles (committees) to overcome some disadvantages of the base algorithms. For example we can not solve a multiclass problem with a single model of binary linear discrimination (or other binary classifiers). In such cases committees of linear models are usually constructed to solve the problem (for example, such committee may contain one linear model per class or one linear model per class pair—see [8, 10] for more). Research presented in this paper aims at building such committees for supervised learning problems that maximize classification accuracy and stability.

Let the training data set for a classification task be defined as

$$\mathcal{S} = \{\langle \mathbf{x}_i, c(\mathbf{x}_i)\rangle : i = 1, \ldots, n\}. \tag{1}$$

Each pair $\langle \mathbf{x}_i, c(\mathbf{x}_i)\rangle$ represents a single data vector (instance) $\mathbf{x}_i$ and its corresponding class label $c(\mathbf{x}_i)$. Without loss of generality we may assume that the class labels are integers from 1 to m.

The simplest committee used for classification is the *voting committee*. The decision module of a voting committee simply counts the votes for each class according to the rule that each member has a single unit vote. The voting committee winner class for given vector $\mathbf{x}$ is computed as

$$V(\mathbf{x}) = \arg \max_{c=1,\ldots,m} \sum_{j=1}^{N} 1_{F_j(\mathbf{x})=c}, \tag{2}$$

where N is the committee members count, $F_j(\mathbf{x})$ is the class label predicted for instance $\mathbf{x}$ by j-th member, and 1_t is the truth value of expression t (1 for true and 0 for false).

The most commonly used alternative to plain voting is the *weighting of probabilities* scheme, where the probability that given vector $\mathbf{x}$ belongs to the i-th class

$$p^w(i|\mathbf{x}) = \frac{1}{N} \sum_{j=1}^{N} p(i|\mathbf{x}, F_j). \tag{3}$$

Here $p(i|\mathbf{x}, F_j)$ defines the probability that vector $\mathbf{x}$ is classified to i-th class by j-th committee member. If for given model type of the committee member the probability $p(i|\mathbf{x}, F_j)$ is not defined directly, it may be approximated in several ways (for example by softmax [1] of submodel outputs or in the worst case by binary decision—if all members approximate probabilities by binary decisions the weighting scheme is equivalent to the voting scheme). The winner class is simply defined as the most probable class:

$$W(\mathbf{x}) = \arg \max_{i=1,\ldots,m} p^w(i|\mathbf{x}). \tag{4}$$

The committees consisting of models of different types are called *heterogenous committees*. Committee members may also (or instead) differ in instances distribution used for learning. Some examples of such committees are bootstrap aggregation [4] or boosting algorithms [9, 20]. Typically such committees were constructed from popular decision trees like C 4.5 [19] or CART [5]. Training data presented to the committee members may be (re-)constructed by the committee in several other ways. For example committee members may learn on (different) subsets of attributes (using several feature selection algorithms).

Other types of committees—stacking and grading—were presented in [24, 22, 3, 25] and [21]. In the place of voting or weighting schemes these committees combine models via stacking. After submodels' learning the decision module uses a meta-level learning to determine where the submodels perform well and where they make mistakes using validation parts of data. Such knowledge is used to make the final decision of the committee. To approximate model certainty a linear regression [24, 22] or meta-decision trees [25] may be used. Other combining schemes were proposed in [14, 16].

Some preliminary results using another type of committees with competence (computed in different way than those presented in this paper) have already been investigated in our group [7].

2. Heterogenous committees with competence

Homo- or Heterogenous committees? Some kinds of committees are dedicated to homogenous environments, for example boosting committees. Many others use heterogenous base models (the majority of voting committees and weighting committees). The advantage of heterogenous models is that such committees may be more resistant to some data-derived traps, while using homogenous system it may be very difficult to obtain an interesting result even using very many submodels. This is caused by better or worse ability of the chosen classification algorithm to adapt to the particular problem (see section 3). If a committee is able to check which models are adequate (competent) to given dataset and to use this information in the decision module, then such committee may take more suitable decisions. In the case of homogenous committees, when the base model is not adequate for given dataset, it may be very hard or even impossible to obtain results similar to a single model but more eligible for the task. Selection of the types of base models for heterogenous committees should not be accidental. Such models should cover possibly huge spectrum of different kinds of models, and as a consequence a huge spectrum of datasets (section 3 confirms the advantages of such diversity).

Decisions reflecting competence. If we are able to estimate the competence $C(F, \mathbf{x})$ of a decision taken by given model F for given data vector $\mathbf{x}$ as a real number (usually in the $[0, 1]$ interval) then we may reflect it in the final decision of the committee in a number of ways.

One of the possibilities is the *winner take all* strategy, which given a vector $\mathbf{x}$ makes the classification decision

$$WTA_C(\mathbf{x}) = F_i(\mathbf{x}), \qquad i = \arg \max_{j=1,\dots,N} C(F_j, \mathbf{x}). \quad (5)$$

Another way is a generalization of the weighting com-

mittee (3):

$$p_C^w(i|\mathbf{x}) = \frac{\sum_{j=1}^{N} C(F_j, \mathbf{x})p(i|\mathbf{x}, F_j)}{\sum_{c=1}^{m} \sum_{j=1}^{N} C(F_j, \mathbf{x})p(c|\mathbf{x}, F_j)}. \quad (6)$$

Assuming equal competence of each model regardless the value of $\mathbf{x}$, we get the special case of (3). The committee decision for data vector $\mathbf{x}$ is

$$W_C(\mathbf{x}) = \arg \max_{i=1,\dots,m} p_C^w(i|\mathbf{x}). \quad (7)$$

Similarly, we may introduce a competence factor to the voting committees (2):

$$V_C(\mathbf{x}) = \arg \max_{c=1,\dots,m} \sum_{j=1}^{N} C(F_j, \mathbf{x}) 1_{F_j(\mathbf{x})=c}. \quad (8)$$

For most algorithms, it is difficult (or even impossible) to define their competence on the basis of the information from the dataset and the learning process (because of the bias-variance dilemma [1, 8]). Even if it is possible for some algorithms then it may be difficult to compare different measures obtained in different ways.

CV-committees and competence estimation. It seems that a reasonable solution to this problem is the cross-validation committee approach. The CV-committees were successfully used by us in [13].

The idea is to estimate the competence of given learning algorithm by running a CV test (computing the average performance on the validation parts of the data). At the same time the models trained in the CV process may be used as a voting or weighting committee composing a new classification model. Another advantage of using CV-committee occurs when a given model parameters values may lead to different results when training on datasets of different size—in such case it is much safer to used the CV models trained on a subset of the whole data, but already validated than to train a model on the whole dataset (larger sample) but with no validation.

Please notice that in our approach to the committees with competence analysis of the members, the single adaptive models are replaced by CV-committees yielding committees of CV-committees.

Global competence. The validation obtained directly from a CV-committee is a good measure of the overall competence of such models. It means that we get a competence factor independent of the data vector being classified. The competence measures which reflect general suitability of the submodels for given task and are called *global*.

In the case of using CV-committees as a committee members, denoting by $V(F_j)$ the average classification accuracy of the j-th CV-committee F_j on the validation parts

of the data, we get a simple definition of the global competence:

$$C_G(F_j, \mathbf{x}) = V(F_j). \qquad (9)$$

Local competence. An alternative approach to measuring the competence of committee members is to pay much attention to the reliability of base models' decisions in a vicinity of the classified object. This way we get some *local* measures of model competence.

The simplest way to define a local competence is to check the classification accuracy in the neighborhood of vector $\mathbf{x}$. Because usually we have a finite set of training vectors, we need to estimate such accuracy on the basis of the training data—we can just check how precisely the predicted class labels correspond to the original class labels $c(\mathbf{v})$ $(F_j(\mathbf{v}) = c(\mathbf{v}))$ for a set of nearest neighbors of $\mathbf{x}$. So, the local competence of j-th model in a committee may be defined by:

$$C_L(F_j, \mathbf{x}) = \frac{1}{|N_\mathbf{x}|} \sum_{\mathbf{v} \in N_\mathbf{x}} 1_{F_j(\mathbf{v})=c(\mathbf{v})}, \qquad (10)$$

where $N_\mathbf{x}$ is a set of neighbors of $\mathbf{x}$ and $|N_\mathbf{x}|$ is the number of its elements. $N_\mathbf{x}$ may be determined as the set of k vectors nearest to the instance $\mathbf{x}$ or may contain all the vectors inside a small hypersphere around $\mathbf{x}$: $N_\mathbf{x} = \{\mathbf{w} : ||\mathbf{x} - \mathbf{w}|| < r\}$.

Local weighted competence. A modification of the local competence (10) to adjust it to the distances between the analyzed vector $\mathbf{x}$ and its neighbors yields a measure of *local weighted competence*:

$$C_{LW}(F_j, \mathbf{x}) = \frac{1}{|N_\mathbf{x}|} \sum_{\mathbf{v} \in N_\mathbf{x}} \frac{1_{F_j(\mathbf{v})=c(\mathbf{v})}}{1 + ||\mathbf{x} - \mathbf{v}||}. \qquad (11)$$

It is important that using the two indices of local competence in committees does not require any learning in addition to that of the committee members (just like in the case of voting or weighting committees). However, some additional computations are required to determine the set of neighbors of given vector when it is to be classified by the committee.

Local competence with CV-committee. The local competence measures (10) and (11) may be overoptimistic, because they rely on the decisions of the trained models on the data they used for training. As a consequence, the models which overfit the data (like 1 Nearest Neighbor, which is always maximally accurate on the training data, but usually not too good in tests on unseen data) are regarded as most competent.

When CV-committees are used as a committee neighbors, one can think of another way to calculate local competence too avoid the overestimation.

The classification accuracy of the j-th base model of the committee in the neighborhood of a point $\mathbf{x}$ may be calculated by checking the answers of the appropriate members of the CV-committee. For a given neighboring vector $\mathbf{v}$, the appropriate member is meant as the one which did not see $\mathbf{v}$ in its learning process. This leads to the definition of

$$C_L^{CV}(F_j, \mathbf{x}) = \frac{1}{|N_\mathbf{x}|} \sum_{\mathbf{v} \in N_\mathbf{x}} 1_{F_j^k(\mathbf{v})=c(\mathbf{v})}. \qquad (12)$$

where k is the index of the sumodel of CV-committee F_j, which did not train on $\mathbf{v}$ (the submodel is denoted by F_j^k).

In this way the competence is validated and should be more trustful, although the fact that for a given neighbor we estimate the accuracy on the basis of just one submodel of the CV-committee while classification is performed by voting of all the members, may bring a suspicion that this estimation is overpessimistic. Taking to the account, that the competence should correspond to the accuracy on unseen data, rejects the suspicion.

Note that as the previous measures, $C_L^{CV}(F_j, \mathbf{x})$ can be computed on the basis of the correctness of classification of a few points from the original training data (dataset used to train j-th CV-committee). This means that computationally is not more expensive than learning of CV-committee because the validated accuracies for all the training vectors may be easily determined just after the learning phase of the CV-committee. The only additional effort is, again, the search for nearest neighbors of the vector to be classified.

Local weighted competence with CV-committee. In a similar way to the above local competence which uses CV-committee the distance-weighted version of the competence can be defined:

$$C_{LW}^{CV}(F_j, \mathbf{x}) = \frac{1}{|N_\mathbf{x}|} \sum_{\mathbf{v} \in N_\mathbf{x}} \frac{1_{F_j^k(\mathbf{v})=c(\mathbf{v})}}{1 + ||\mathbf{x} - \mathbf{v}||}. \qquad (13)$$

Global and local competence in one system. Global and local competence are not mutually exclusive—they may be used together in a single committee. It is enough to see, that any product of different competence measures may be used as a new competence measure in all three voting schemes, we analyze here: the winner takes all (5), weighting with competence (7) and voting with competence (8).
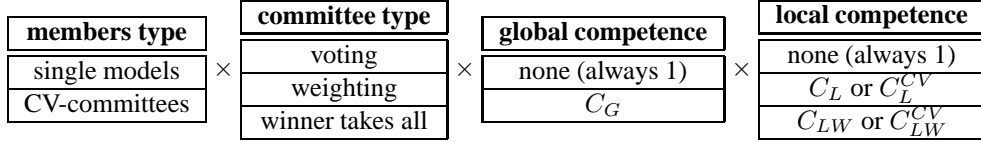
This way we may create for example the following competence combinations:

- local and global competence:

$$C_{G+L}(F, \mathbf{x}) = C_G(F, \mathbf{x})\, C_L(F, \mathbf{x}), \qquad (14)$$

- local weighted and global competence:

$$C_{G+LW}(F, \mathbf{x}) = C_G(F, \mathbf{x})\, C_{LW}(F, \mathbf{x}), \qquad (15)$$

members type		committee type		global competence		local competence
single models		**voting**		**none (always 1)**		**none (always 1)**
CV-committees	$\times$	weighting	$\times$	C_G	$\times$	C_L or C_L^{CV}
		winner takes all				C_{LW} or C_{LW}^{CV}

Figure 1. Scheme of possible committee configurations.

- local and global competence for CV-committees:

$$C_{G+L}^{CV}(F,\mathbf{x}) = C_G(F,\mathbf{x})\, C_L^{CV}(F,\mathbf{x}), \qquad (16)$$

- local weighted and global competence for CV-committees:

$$C_{G+LW}^{CV}(F,\mathbf{x}) = C_G(F,\mathbf{x})\, C_{LW}^{CV}(F,\mathbf{x}). \qquad (17)$$

3. Results

Using the concepts presented above we could test quite large number of different committee configurations. The scheme of all the combinations is presented in figure 1. Please notice that using single models as committee members excludes using C_L^{CV} and C_{LW}^{CV} local competence measures, however using CV-committees allows each of the 5 possibilities regarding local competence. Thus we could build 48 different committees according to this scheme, however some of them would not make much sense (although technically feasible), for example the winner takes all technique with no competence measure.

To make this discussion readable (also because of the space limit) we analyze 15 different committees configurations (most interesting from the point of view of this work). Each committee was composed of heterogenous members. As the base models we have chosen algorithms characterized by different inner structure, different learning strategies and as a consequence different properties of usability. The parameters of each algorithm were the same for all the benchmarks. The committee members were: the k nearest neighbors (kNN) algorithm [6] with $k = 5$, Separability of Split Value decision tree [11, 12], Naive Bayes classifier [18] and two kinds of Support Vector Machines [23, 2, 15]: with Gaussian kernels (dispersion = 0.1 and C = 10) and with linear kernels (C = 1). Each committee was composed of the five base models mentioned above or of five CV-committees (each using one of the five base algorithms as the type of its members). The CV-committees used 10-fold cross-validation (had 10 members). The neighborhood of a vector (for the purpose of competence evaluation) was defined as 5 nearest neighbors.

The 15 committees are organized in three groups (see table 1). The first group of committees consists of five committees of single base models—the results of the single models are presented in the first five-column block of the table. The other two groups use CV-committees of the base models instead of single models. The rightmost group contains five committees which take advantage of the global competence measure (9).

Benchmark tests were performed on datasets from the UCI machine learning repository [17]. Tests were computed on 17 datasets (australian credit, balance scale, german numeric, glass, cleveland «heart» disease, image, ionosphere, chess «kr-vs-kp», Ljubljana breast cancer, liver disorders, pima indian diabetes, sonar, tic-tac-toe, voting records, vowel, waveform and wine) which differ significantly in number of features and instances. Each model was tested using 10 repetitions of 10-fold cross-validation, and averaged accuracies are presented in table 1. For each benchmark, the best result is labelled with an asterisk, and other results which do not significantly differ from the best one are typeset in **bold** font. Significantly worse results are not typed in bold. To compute statistical significance the paired t-test was used with confidence level of 95%.

The row labeled "#best" in table 1 shows how many times given model was the best or statistically insignificantly different. The row "#g-best" presents the sums of "#best" for the groups of models.

It can be seen that the inner diversity heterogeneous committees results in the ability to deal with different benchmarks—nearly always at least one of base models perform as good as the best model, composing a good base for the committees which take competence into account.

An important observation is that introducing local and global competence increases the numbers of the best or insignificantly different models—see the rows "#g-best" and "#best". The "#best" counter of the voting committees is 9 while the best committees were 14 times at the top. Whereas the best base model were at the top only 5 times, and Naive Bayes was the winner just one time. Committees using C_{G+L}^{CV} and C_{G+LW}^{CV} are nearly never significantly worse than the best model for given test dataset. The most unsuccessful committees were defined by WTA scheme, so we suggest using weighting instead.

As opposed to stacking models the committees with competence do not use additional (meta level) learning while the performance is not decreased—the best committees with competence are nearly always the best (up to

Model	Base models					Committees of base models					Committees of CV-committees of base algorithms									
Dataset	kNN	SSV Tree	Naive Bayes	SVM	SVM lin	V	W	WTA_{C_L}	W_{C_L}	$W_{C_{LW}}$	V	W	$WTA_{C_L^{CV}}$	$W_{C_L^{CV}}$	$W_{C_{LW}^{CV}}$	V_{C_G}	W_{C_G}	$WTA_{C_{G+L}^{CV}}$	$W_{C_{G+L}^{CV}}$	$W_{C_{G+LW}^{CV}}$
Australiam	83.9	85.3	80.0	83.2	85.2	85.2	85.2	83.1	85.2	85.0	85.3	85.4	84.3	85.6	85.5	85.4	85.6	85.3	85.6*	85.6
Balance	82.2	78.6	90.6	88.6	84.5	89.4	90.6	82.9	90.5	90.4	89.7	90.7	86.9	90.4	90.5	94.4*	90.7	88.9	90.6	90.6
German	72.1	72.3	72.8	72.7	76.5	76.0	76.7	72.3	75.7	75.7	75.9	76.6	73.2	76.6	76.5	75.8	76.7*	74.0	76.6	76.6
Glass	66.5	69.2	47.5	62.5	37.1	62.7	69.8	70.6	71.4	71.5	64.7	69.9	71.3	72.5*	72.4	68.5	70.1	70.3	72.1	72.1
Heart	81.7	77.4	83.6*	78.8	83.1	82.6	82.7	77.7	82.2	82.2	82.5	82.7	80.7	82.4	82.3	82.6	82.6	80.9	82.5	82.8
Image	94.6	95.8	79.8	94.9	85.6	95.3	96.9	96.3	97.3	97.3*	95.0	96.8	96.3	96.9	96.9	96.3	96.8	95.8	97.0	97.1
Ionosphere	84.0	87.3	82.5	94.7*	86.9	92.9	93.0	91.6	93.3	93.3	92.8	93.0	90.5	93.2	93.6	92.5	93.1	94.3	93.5	93.4
Kr-vs-kp	94.0	98.5	87.9	99.5*	94.9	98.8	99.2	96.0	99.4	99.4	98.8	99.2	97.1	99.5	99.4	98.8	99.3	99.4	99.4	99.4
L.breast	70.3	72.0	71.0	71.7	71.0	74.2*	73.2	72.8	72.4	72.7	74.1	73.3	71.5	72.6	72.8	73.9	72.8	71.7	72.8	72.9
Liver	61.1	66.8	56.1	71.0	68.7	70.6	71.1	64.5	71.0	71.1	70.6	71.0	67.0	70.4	71.3*	70.8	71.2	68.6	71.0	71.0
Pima	73.9	73.4	75.5	75.1	76.9	76.6	76.7	73.6	75.8	75.7	76.7	76.8	74.3	76.5	76.7	76.9	77.0*	75.2	76.5	76.5
Sonar	82.0	74.3	67.8	78.3	73.8	81.2	81.3	82.3	82.6	82.9	81.7	82.8	81.5	83.5*	83.1	81.8	82.1	82.4	82.6	82.5
TicTacToe	84.3	95.6	69.5	98.4*	65.3	87.9	89.3	84.3	90.3	90.3	87.9	89.8	84.0	91.0	90.9	87.9	96.1	98.3	96.5	96.4
Vote	94.0	95.0	90.4	95.8	96.1*	95.8	95.7	94.3	96.0	96.0	95.8	95.7	94.7	95.7	95.7	95.7	95.8	95.4	95.9	96.0
Vowel	87.4*	85.8	79.2	79.0	40.2	81.6	85.1	87.0	86.4	86.3	83.7	85.0	86.5	86.1	86.2	85.8	85.6	86.9	86.6	86.5
Waveform	81.5	77.3	80.9	82.0	78.8	83.9	84.9	81.3	84.6	84.6	84.1	85.2*	81.7	85.0	85.0	84.3	85.0	82.1	84.8	84.8
Wine	97.1	89.7	97.2	97.7	96.2	98.0	98.5	96.9	98.5	98.5	98.7	99.0*	96.9	98.9	98.8	98.6	99.0	97.0	98.8	98.9
#best	2	3	1	5	5	9	10	4	11	12	9	13	5	13	13	9	13	9	14	14
#g-best			16						46					53					59		

Table 1. Results comparison of heterogenous committees with local and global competence

V — voting committee (2), W — weighting committee (4), W_C — weighting committee with competence (7), WTA_C — winner takes all with competence (5)

C_L — local competence (10), C_{LW} — local weighting competence (11)

C_L^{CV} — local competence for CV-commitees (12), C_{LW}^{CV} — local weighting competence for CV-commitees (13)

C_G — global competence (9), C_{G+L}^{CV} — global and local competence for CV-commitees (16), C_{G+LW}^{CV} — global and local weighting competence for CV-commitees (17)

statistical significance), similar behavior was observed by Zenko et. al. [25].

The results of [25] show that if the accuracy of C4.5 is significantly worse than for example kNN or Naive Bayes then boosting or bagging with C4.5 does not help to increase the accuracy sufficiently and such models do not win with some single models like kNN or Naive Bayes. This can be seen as a trap of homogenous committees while heterogenous ensembles may be successful in a huge spectrum of benchmarks, thanks to the different and self–complementary nature of the base models.

4. Conclusions

The heterogenous committees augmented by competence analysis in local and global versions become more accurate and stable. Such ensembles perform successfully for a wide range of different problems.

It was shown that single models, even the best ones, were not so accurate as presented committees and that the best committees significantly outperform the plain voting or weighting committees.

Comparing to stacking or grading committees, the competence based committees presented here do not need additional meta-level learning and are characterized by similar ability to classify unseen data.

References

[1] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[2] B. E. Boser, I. M. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. In D. Haussler, editor, *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, Pittsburgh, PA, 1992. ACM Press.

[3] L. Breiman. Stacked regressions. *Machine Learning*, 24:49–64, 1996.

[4] L. Breiman. Bias-variance, regularization, instability and stabilization. In C. M. Bishop, editor, *Neural Networks and Machine Learning*, pages 27–56. Springer-Verlag, 1998.

[5] L. Breiman, J. H. Friedman, A. Olshen, and C. J. Stone. *Classification and regression trees*. Wadsworth, Belmont, CA, 1984.

[6] T. M. Cover and P. E. Hart. Nearest neighbor pattern classification. *Institute of Electrical and Electronics Engineers Transactions on Information Theory*, 13(1):21–27, Jan. 1967.

[7] W. Duch, L. Itert, and K. Grudziński. Competent undemocratic committees. In L. Rutkowski and J. Kacprzyk, editors, *Neural Networks and Soft Computing. Proceedings of the 6th International Conference on Neural Networks and Soft Computing (ICNNSC)*, Advances in Soft Computing, pages 412–417, Zakopane, Poland, June 2002. Springer-Verlag.

[8] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley, 2 edition, 1997.

[9] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, Aug. 1997.

[10] J. Friedman, T. Hastie, and R. Tibshirani. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, 2001.

[11] K. Grąbczewski and W. Duch. A general purpose separability criterion for classification systems. In *Proceedings of the 4th Conference on Neural Networks and Their Applications*, pages 203–208, Zakopane, Poland, June 1999.

[12] K. Grąbczewski and W. Duch. The Separability of Split Value criterion. In *Proceedings of the 5th Conference on Neural Networks and Their Applications*, pages 201–208, Zakopane, Poland, June 2000.

[13] K. Grąbczewski and N. Jankowski. Mining for complex models comprising feature selection and classification. In I. Guyon, S. Gunn, M. Nikravesh, and L. Zadeh, editors, *Feature extraction, foundations and Applications*. Springer, 2004.

[14] R. A. Jacobs, M. L. Jordan, S. J. Nowlan, and J. E. Hinton. Adaptive mixtures of local exports. *Neural Computation*, 79(3), 1991.

[15] S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, and K. R. K. Murthy. Improvements to Platt's SMO algorithm for SVM classifier design. *Neural Computation*, 13:637–649, 2001.

[16] R. Maclin. Boosting classifiers regionally. In *Proceeding of AAAI*, 1998.

[17] C. J. Merz and P. M. Murphy. UCI repository of machine learning databases, 1998. http://www.ics.uci.edu/∼mlearn/MLRepository.html.

[18] T. Mitchell. *Machine learning*. McGraw Hill, 1997.

[19] J. R. Quinlan. Programs for machine learning. San Mateo, CA: Morgan Kaufmann, 1993.

[20] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26(5):1651–1686, 1998.

[21] A. K. Seewald and J. Fürnkranz. Evaluation of grading classifiers. In *Advances in Intelligent Data Analysis: Proceedings of the Fourth International Symposium (IDA-01)*, Berlin, 2001. Springer-Verlag.

[22] K. M. Ting and I. H. Witten. Stacked generalization; when does it work? In *Proceedings of the 15th International Join Conference on Artificial Intelligence*, 1997.

[23] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, 1995.

[24] D. H. Wolpert. Stacked generalization. *Neural Networks*, 5:241–259, 1992.

[25] B. Zenko, L. Todorovski, and S. Dzeroski. A comparison of stacking with meta decision trees to bagging, boosting, and stacking with other methods. In *Proceedings of the 2001 IEEE International Conference on Data Mining*, pages 669–670. IEEE Computer Society, 2001.