Generalization Capability of Homogeneous Voting Classifier Based on Partially Replicated Data

Jacek Jelonek

Institute of Computing Science
Poznań University of Technology
Piotrowo 3A
60-965 Poznań, Poland
jacek.jelonek@cs.put.poznan.pl

Abstract

The generalization error is one of the most important features taken into account in performance evaluation and verification of any classifier. We propose a voting system based on homogenous base classifiers (HVC) which ensures a better generalization capability than any of its components. The principle of this idea consists in the differentiation of a learning data set for each base classifier and does not relate to their architecture, learning algorithm or any parameters. Therefore, this concept could be applied to decision trees, neural networks, statistical classifiers and many others. In order to show the superiority of HVC we performed several comparative tests on some data sets using cross-validation.

Introduction

Recently, the concept of combining multiple classifiers into one classification system has become very popular. The problem has been approached in many ways depending on the aims of the researcher. The main purpose of creating a complex multi-classifier system is to obtain better classification performance than the performance offered by its components - individual base classifiers. The classification capability, however, is not always as important as, for example, the possibility of handling a very large, distributed database. If that is the case, the whole learning data set is partitioned into a number of smaller disjoint training subsets and each of them is applied to one distinct classifier (Chan & Stolfo 1995a). Another situation where multi-classifier systems could be successfully applied is, when solving a complex pattern recognition task. For example, in image recognition (understanding) numerous types of features, such as continuous and/or binary values, discrete labels, histograms and structural primitives could usually be extracted to represent and classify objects. In such a situation, it is reasonable to apply different dedicated classifiers that cooperate.

In multi-classifier systems, particular classifiers produce their own classification prediction which is then aggregated with others to produce one final prediction. The aggregation can be made in many different ways. The simplest aggregation consists in using the majority rule in a *simple voting* system. The final prediction is chosen as classification with plurality of votes. Variation of a simple voting rule is a *weight voting* rule (Littlestone & Warmuth 1994). In such a case, each base classifier is associated with a weight that is updated during the learning phase of classification. The values of weights reflect how accurately the classifier has performed on the learning data set. The final prediction is based on the predictions of base classifiers and their weights. More elaborate systems contain sophisticated architecture, for example the *arbiter*, *combiner trees* (Chan & Stolfo 1995b) or *hierarchical trees* (Jordan & Jacobs 1994).

Thus, in designing a complex multi-classifier system it is crucial to decide what kind of architecture should be applied and what type of differentiation of classifiers should be used. There are three major sources of differentiation: differentiation of algorithms (e.g. neural network classifier - NN, decision trees, Bayes classifier, instance-based classifier - IBL), differentiation of the input data space (Jordan & Jacobs 1994) and differentiation of the learning data set (Chan & Stolfo 1995c).

Using the above remarks, we can describe the Homogeneous Voting Classifier proposed in this paper. It is a simple voting system consisting of a number of identical base classifiers which are differentiated by a learning data set. The differentiation means, in this case, selection of a learning sample for each base classifier from the learning data set. The selected samples may have a non-empty intersection, i.e. may be replicated, but their sum covers the entire learning data set. The key idea of differentiating a learning data set seems to be similar to (Chan & Stolfo 1995c) but the purpose is different. Their aim was to deal with a large database by splitting the learning data set into classifiers running, for example, in a distributed environment. Their results show that the performance of the multi-classification system based on replicated data sets (replication factor Δ = 5...30%) is not better than the performance of the same system based only on disjoint partitioned data (or single base classifier). The limit of replication to 30% is justified if the data set is very large. However, if the data set is smaller, this constraint could be rejected. In the HVC, we allow the replication factor to be over 75%. The results are very encouraging since this classifier gives a higher accuracy than any of its base components. In a word, the present paper could be considered as complementary to (Chan & Stolfo 1995c).

In the next sections, we consider the overlearning problem encountered in many classification systems and we introduce our idea of differentiating the learning data set as a way of avoiding that problem. We describe the HVC system in which the proposed differentiation is used and present the results of a comparative study. Conclusions are drawn in the final section.

Overlearning Problem

Approximation and generalization errors are considered in the performance evaluation and verification of any classifier. The approximation error reflects a number of misclassifications on a learning data set and the generalization error could be defined in a simplified way as a number of misclassifications on a testing data set. It is usually assumed that learning and testing data sets are disjoint and constitute the whole available data set. The main danger of the learning phase of many classifiers is overlearning. It consists in overfitting to learning examples and, in consequence, in decreasing the generalization capability of a classifier. That problem is especially common in neural network classifiers. Several solutions to that problem have been found. The basic one is especially useful if the available data set is small. It is based on a heuristic argument, which states that the simplest model will in general achieve the best generalization or approximation. Following this argument, one may try to choose as simple a net structure as possible, or state additional constraints on the weights to limit network complexity. A more sophisticated solution concerns a dynamic modification of an architecture - the *cascade correlation model*, or applying algorithms based on weights pruning. Another solution that helps prevent overlearning consists in using a "smart" stop condition of the learning phase, defined usually as the minimum value of the generalization error. A very interesting concept of the control of neural network generalization has been proposed by (Röbel 1994). The key idea consists in a dynamic selection of learning samples.

Not only neural network classifiers have to deal with the overlearning problem. The tendency to increase generalization capability is seen also in the instance-base learning algorithms. The proposal of *nearest mode* algorithms instead of *nearest neighbor*, or the elaboration of next versions of IBL classifiers proposed by (Aha, Kibler, & Albert 1991) could be examples of this tendency. In the case of decision trees, new versions of pruning methods also contribute to the reduction of the overlearning effect of the classification systems.

Our general idea is to increase the generalization capability of a classifier by applying a reduced learning data set to it. This approach could, however, lead to the decreasing of the approximation capability of the classifier because a part of the learning data set would not be used at all and the ratio of cardinality of a learning set to a testing set would grow worse. A natural solution to that problem is to combine a number of base well generalizing classifiers, defined on different (but very similar - we suggest that the replication factor should be over 80%) learning data sets, into one voting based classification system - HVC.

Homogeneous Voting Classifier

In designing a voting-base classifier based on partially replicated data, there are two free parameters: a number of base classifiers and a replication factor Δ that have to be set. To simplify the computational tests, we decided to combine these parameters. In our approach, the number of base classifiers determines the replication factor Δ. The procedure to obtain an HVC system based on N base-classifiers is the following:

1. Divide the whole learning data set into N disjoint subsets. Their cardinality should be the same (difference - plus/minus one record).
2. Create N base classifiers (for example C4.5, Neural Network or IBL) and apply, for each of them, the N-1 learning data subsets obtained in step number 1. The replication factor Δ equals $100*(N-1)/N$ [%] (for $N > 2$).
3. Verify prediction of the whole HVC by applying the testing data set for each base classifier and use the simple voting rule to obtain final classification.

The above procedure allows creation of a Homogeneous Voting Classifier based on partially replicated data, where a replication factor Δ equals: 0.0; 66.7; 75.0; 80.0; 83.3; 85.7; 87.5 or 88.9%. In our opinion, this range of value of a replication factor (especially from 80 to 90%) is the most promising in consideration of the generalization performance of HVC. The above algorithm is illustrated on Figure 1.

Results

Six data sets have been used in the computational verification of the HVC concept. Four of them: BUPA Liver Disorders, GLASS Identification Database, IRIS Plants Database and PIMA Indians Diabetes Database have been obtained from UCI Repository of machine learning database (Murphy & Aha 1994). We decided to choose them because of characteristic of theirs, that is required in our computational experiment, i.e., there are only numerical values of the attributes (no missing ones) and number of cases is in range from 100 to 1000. The COOC Brain Tumours Database and HIST Brain Tumours Database are our own medical data sets. Table 1 describes the base features of the data sets.

The generalization capability of the HVC based on C4.5, Neural Network and IBL3 base classifier has been verified by applying, for each data set, a ten times 10-fold cross validation test - a well known estimator of generalization error. So, an average and standard deviation of accuracy of classification could have been calculated. This computational test has been repeated nine times for a varying number of base classifiers.

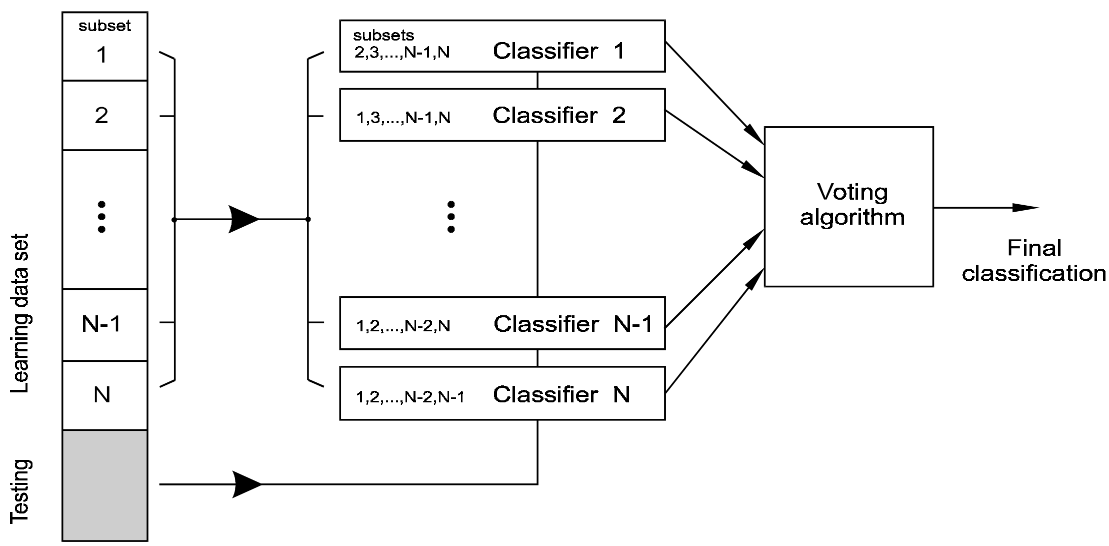

Figure 1: Splitting and replication of the learning data set for HVC.

Data set	Instances	Attributes	Classes	Distribution
BUPA Liver Disorders	345	6	2	145, 200
COOC Brain Tumours Database	700	22	14	14 x 50
GLASS Identification Database	214	9	6	70,17,76,13,9,29
HIST Brain Tumours Database	700	17	14	14 x 50
IRIS Plants Database	150	4	3	50, 50, 50
PIMA Indians Diabetes Database	768	8	2	500, 268

Table 1: Characteristics of experimental data sets.

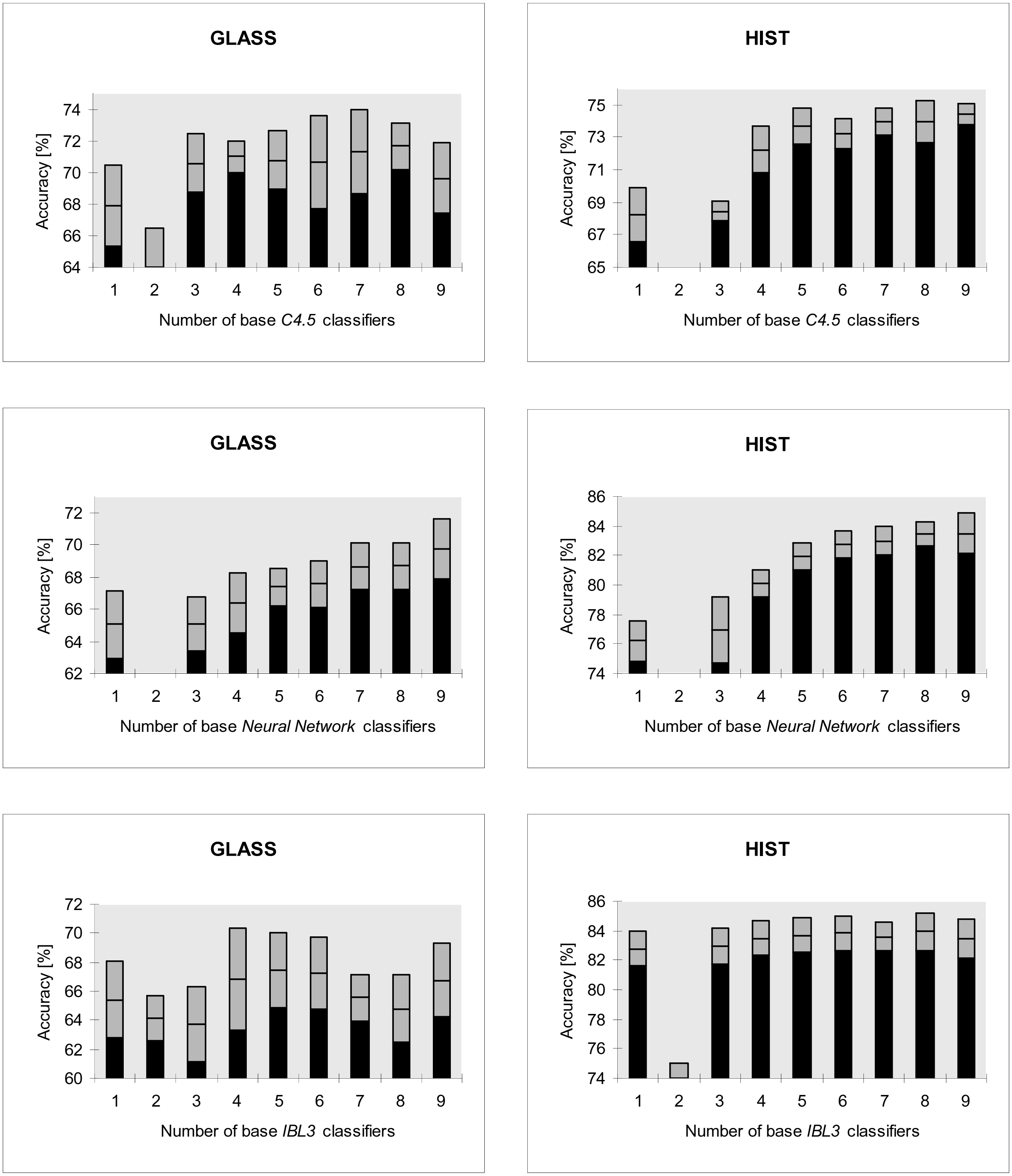

Figure 2: Accuracy of HVC based on C4.5, Neural Network and IBL3 classifiers applied for GLASS and HIST data sets.

Data set	Accuracy of single C4.5 classifier [%]	Number of base C4.5 classifiers - N	Data replication factor Δ [%]	Accuracy of HVC for N - base classifier [%]	Average improvement [%]
BUPA	64.5 ± 1.8	9	88.9	69.6 ± 2.0	5.1
COOC	55.4 ± 2.2	9	88.9	62.2 ± 1.3	6.8
GLASS	67.9 ± 2.6	8	87.5	71.6 ± 1.5	3.7
HIST	68.2 ± 1.7	9	88.9	74.4 ± 0.7	6.2
IRIS	95.5 ± 0.5	6	83.3	95.9 ± 1.1	0.4
PIMA	70.8 ± 0.8	7	85.7	72.4 ± 0.4	1.6

Table 2: Classification results of HVC (based on C4.5) and single C4.5 classifier.

Data set	Accuracy of single NN classifier [%]	Number of base NN classifiers - N	Data replication factor Δ [%]	Accuracy of HVC for N - base classifier [%]	Average improvement [%]
BUPA	67.6 ± 1.8	9	88.9	69.9 ± 1.2	2.3
COOC	62.4 ± 1.3	9	88.9	68.4 ± 1.2	6.0
GLASS	65.0 ± 2.1	9	88.9	69.7 ± 1.9	4.7
HIST	76.2 ± 1.4	8	87.5	83.5 ± 0.8	7.3
IRIS	95.3 ± 0.8	9	88.9	95.9 ± 1.1	0.6
PIMA	76.4 ± 0.8	9	88.9	76.8 ± 0.6	0.4

Table 3: Classification results of HVC (based on Neural Network) and single Neural Network classifier.

Data set	Accuracy of single IBL3 classifier [%]	Number of base IBL3 classifiers - N	Data replication factor Δ [%]	Accuracy of HVC for N - base classifier [%]	Average improvement [%]
BUPA	57.0 ± 3.5	3	75.0	60.4 ± 3.2	3.4
COOC	62.7 ± 1.8	3	75.0	63.4 ± 1.6	0.7
GLASS	65.4 ± 2.6	5	80.0	67.4 ± 2.6	2.0
HIST	82.8 ± 1.2	8	87.5	84.0 ± 1.3	1.2
IRIS	93.3 ± 1.9	5	80.0	95.0 ± 0.9	1.7
PIMA	68.2 ± 1.4	2	0.0	71.2 ± 1.3	3.0

Table 4: Classification results of HVC (based on IBL3) and single IBL3 classifier.

Conclusions

We have verified the usefulness of applying the HVC system for six medium-sized data sets. The improvement of classification accuracy depends on the data set and on the base classifier used. The Homogeneous Voting Classifier based on C4.5 and Neural Network performed relatively well. Depending on a data set, the average improvement varies up to 7.3%. In the case of IBL3 base classifier, the improvement is insignificant (especially in consideration of the standard deviation of classification accuracy). It is not a surprising result - it is just confirmation of the fact that the generalization capability of IBL3 classifier is much better than C4.5 or Neural Network ones and therefore it is difficult to expect any additional improvement by using methods such as HVC.

When the number of the base classifier is two (N=2) the results are usually much worse than for other values. There are two reasons. First, in that case the replication factor Δ equals 0, e.g., the learning sets are disjoint. Second, the simple voting systems perform poorly when a number of votes is only two (no majority of votes).

It is difficult to predict an optimal number of base classifiers in the HVC system (which is equivalent to replication factor Δ in our approach). We especially focused on a replication factor Δ varying from 75 to 90%. This range of value of Δ seems to be the most promising. In future research, the number of base classifiers and the replication factor should be treated independently.

Acknowledgments

We thank Roman Słowiński for comments on the first draft of this paper.

We wish to thank Poznań Supercomputing and Networking Center affiliated to the Institute of Bioorganic Chemistry at the Polish Academy of Sciences in Poznań for making the computational part of our work possible.

This work has been supported by the grant KBN 8 T11F 010 08p02.

References

Books with Multiple Authors

Breiman, L.; Friedman, J.H.; Olshen, R.A.; and Stone, C.J. 1986. *Classification and Regression Trees*. Belmont, CA: Wadsworth.

Quinlan, J.R. 1993. *C4.5: programs for machine learning*. Morgan Kaufmann Publishers, CA: San Mateo.

Journal Articles

Jordan, M.I., and Jacobs, R.A. 1994. Hierarchical Mixture of Experts and the EM algorithm. *Neural Computation* 6:181-214.

Littlestone, N., and Warmuth, M.K. 1994. The weighted majority algorithm. *Information and Computation* 108(2):212-261.

Turney, P. 1994. Theoretical Analyses of Cross-Validation Error and Voting in Instance-Based Learning. *Journal of Experimental and Theoretical Artificial Intelligence* 6:331-360.

Turney, P. 1994. A Theory of Cross-Validation Error. *Journal of Experimental and Theoretical Artificial Intelligence*. 6:361-391.

Quinlan, J.R. 1986. Induction of decision trees. *Machine Learning.* 1:81-106.

Proceedings Papers

Aha, D.W.; Kibler, D.; and Albert, M.K. 1991. Instance-Based Learning Algorithms. Machine Learning 6: 37-66.

Chan, P., and Stolfo, S. 1995a. A Comparative Evaluation of Voting and Meta-learning on Partitioned Data. In Proceedings of Twelfth International Conference on Machine Learning: 90-98.

Chan, P., and Stolfo, S. 1995b. Learning Arbiter and Combiner Trees from Partitioned Data for Scaling Machine Learning. International Conference on Knowledge Discovery and Data Mining: 39-44.

Kwok, S.W., and Carter, C. 1990. Muliple decision trees. In *Uncertainty in Artificial Intelligence* 4:327-335. Elsevier Science Publishers B.V., 1990.

University Technical Report

Röbel, A. 1994. The Dynamic Pattern Selection Algorithm: Effective Training and Controlled Generalization of Backpropagation Neural Networks, Technical Report, Institut für Angewandte Informatik, FG Informatik in Natur- und Ingenieurwissenschaften, Technische Univer-sität Berlin.

Forthcoming Publication

Chan, P., and Stolfo, S. 1995c. Scaling Learning by Meta-Learning Over Disjoint and Partially Replicated Data. (Extended Abstract), via [http://www.cs.columbia.edu/~sal/hpapers/metalrep.ps].

Other

Murphy, P.M., and Aha, D.W. 1994. UCI Repository of machine learning database [http://www.ics.uci.edu/~mlearn/MLRepository.html], Irvine, CA: University of California, Department of Information and Computer Science.