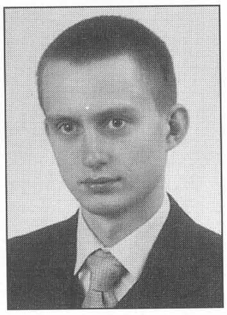
Tomasz ROGALA, Andrzej BRYKALSKI
SZCZECIN UNIVERSITY OF TECHNOLOGY, FACULTY OF ELECTRICAL ENGINEERING, INSTITUTE OF ELECTRONICS TELECOMMUNICATIONS AND COMPUTER SCIENCE

Practical aspects of combining multiple classifiers into the committee machines

mgr inż. Tomasz ROGALA



absolwent Wydziału Elektrycznego Politechniki Szczecińskiej (2002, kierunek: elektronika i telekomunikacja). Obecnie uczestnik III roku studiów doktoranckich na tym wydziale. Jego zainteresowania naukowe dotyczą analizy i przetwarzania sygnałów oraz szeroko pojętych metod sztucznej inteligencji. Przygotowuje pracę doktorską poświęconą przetwarzaniu oraz rozpoznawaniu sygnałów bioelektrycznych powstających w układzie wzrokowym.

e-mail: Tomasz.Rogala@ps.pl

dr hab. inż. Andrzej BRYKALSKI, prof. nadzw. PS

absolwent Wydziału Elektrycznego Politechniki Szczecińskiej (1979, specjalność: automatyka i metrologia elektryczna). Od 1979 zatrudniony w Politechnice Szczecińskiej, gdzie na Wydziale Elektrycznym uzyskał stopień doktora nauk technicznych (1983). Stopień doktora habilitowanego nauk technicznych uzyskał na Wydziale Elektrotechniki i Technik Informacyjnych, Uniwersytetu Technicznego Ilmenau, Niemcy (1991). Obecnie profesor nadzwyczajny Politechniki Szczecińskiej w Instytucie Elektroniki, Telekomunikacji i Informatyki, kierownik Zakładu Podstaw Informatyki. Od początku pracy zawodowej zajmował się badaniami dotyczącymi komputerowej analizy zagadnień wiroprądowych, ekranowania elektromagnetycznego, procesów dyfuzyjnych i modelowania dynamiki tych procesów, wybranych problemów akustycznych, a obecnie przetwarzaniem sygnałów biomedycznych. Jest autorem i współautorem blisko 50 artykułów w czasopismach naukowych, poświęconych głównie tej tematyce, a także kilkudziesięciu prac prezentowanych na konferencjach naukowych.

e-mail: Andrzej.Brykalski@ps.pl

Abstract

Committee machines are ensembles of relatively simple classifiers, able to achieve high accuracy and overcome computational problems using the *divide and conquer* principle. This paper discusses the main ideas underlying the design of committee machines. At the beginning architectures of such structures are introduced briefly. Their interesting properties are demonstrated using artificial dataset, called "two spiral problem", which is a popular benchmark for evaluating classification algorithms. Then, recent applications of committee machines in signal identification are presented. Finally, a real-life problem of the automatic identification of electroretinograms, electrical signals used in ophthalmic diagnosis, is discussed. The results suggest that the high efficiency of committee machines, compared to the single multilayer perceptron networks, may be significantly decreased by an important constraint, which is the limited number of cases in dataset.

Streszczenie

Struktury noszące nazwę komitetów (ang. *committee machines*) to zgrupowania względnie prostych klasyfikatorów, pozwalające na skuteczne rozwiązywanie skomplikowanych problemów klasyfikacyjnych, dzięki zastosowaniu zasady „dziel i zwyciężaj". Niniejsza praca opisuje idee leżące u podstaw takich konstrukcji i przedstawia ich interesujące własności w oparciu o popularny zestaw danych testowych zwany problemem dwóch spiral. Następnie przedstawione zostają aktualne przykłady zastosowań komitetów w identyfikacji skomplikowanych sygnałów. W dalszej części pracy opisano wyniki zastosowań komitetów w badaniach autorów nad identyfikacją elektroretinogramów – jednowymiarowych sygnałów elektrycznych wykorzystywanych w diagnostyce okulistycznej. Rezultaty obliczeń sugerują, że teoretycznie bardzo wysoka skuteczność komitetów, w stosunku do osiąganej przy użyciu pojedynczych sieci neuronowych, może zostać obniżona przez ograniczoną ilość danych uczących oraz szczególnie nietypowy rozkład przypadków w przestrzeni cech.

1. Introduction

Since development of the multilayer perceptron (MLP) architecture [1,2,5] the neural networks have been massively applied in many fields of engineering. However there exist certain problems, that despite their low dimensionality are difficult to solve by the MLPs, because of the sophisticated nature of the feature space. Additionally, every recognition task can become too difficult if the number of training samples is too small compared to the length of single input vector. This phenomenon is explained in [5] using the statistical measure called Vapnik-Chevroninkis dimension.

Nonetheless, these problems can be solved using simple structures, if individual predictors work together according to the *divide and conquer* principle. Such a group of classifiers is called the committee machine. These structures are recently increasing in popularity. Interesting examples of their application from past and current year are the papers by Fernandes et. al. [3,4] and Statmatatos and Widmer [8]. The first two works concern automated forest fire detection by analysis of visual light spectra. Due to the unequal distribution of categories in learning set, the initial algorithm resulted in unacceptable false alarm rate. Application of the cascade of multiple classifiers enabled lowering this ratio, without the loss of overall sensitivity. Such approach is called *boosting by filtering* and is explained later in the paper. The third paper deals with automatic identification of pianists basing on advanced set of features describing individual properties of the performed piece of music. It proves the usefulness of committee machines when the feature set is very sophisticated compared to the number of learning examples. Dividing the input feature space into regions controlled by dedicated individual experts allowed significant increase of correct classifications rate. Partitioning the feature space into areas controlled by different classifiers is called *mixture of experts* (MoE), and is presented in the next section as well.

2. Architectures of committee machines

According to Haykin [8], committee machines may be classified into two major categories. The first one contains all the structures, which combine the responses of several predictors, without involving the input signal into decision mechanism. The simplest form of a committee machine is ensemble averaging. The outputs of multiple predictors, which are usually of the same type, are averaged (or generally - combined linearly) to produce the final answer of the system. Such architecture is presented in figure 1. It has been proved [5] that it performs not worse, which in practice means usually better, than a single classifier of the same type.

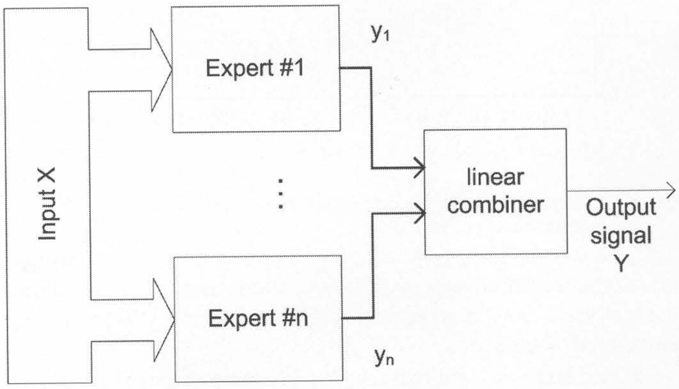


Fig. 1. Static committee machine based on ensemble averaging

The second solution belonging to the first group of committee machines is called boosting. The term *boosting* signifies improving the efficiency of a *weak learner* algorithm (which means an algorithm which performs only slightly better than chance) to an arbitrarily high value, by training each classifier in an ensemble using a different subset of training examples (see fig. 2). There are few ways of obtaining this effect, but in this paper we focus on *AdaBoost* algorithm [1,2,5]. Its main advantage is that in an ensemble of total n experts, the predictors with high index i tend to *specialize* in classifying the difficult examples (misclassified by the predictions with lower i). Detailed description of the algorithm, along with pseudo-code implementation, can be found in [2,10]. It is worth noting that when a classifier used as an individual expert performs better than the *weak learner*, the benefits of boosting decrease.

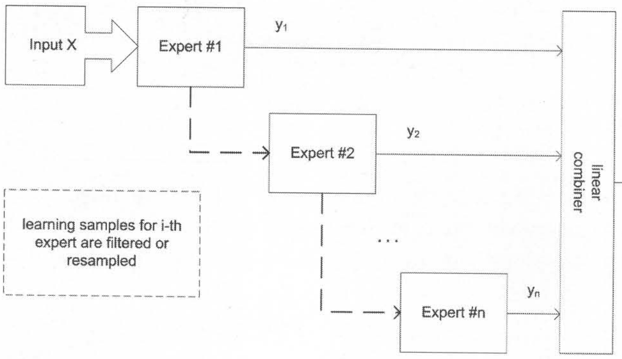


Fig. 2. General idea of the boosting procedure

Second category, according to Haykin's taxonomy [5], describes structures involving the input signal into the mechanism of combining individual responses. Two kinds of such structures are most commonly mentioned in the literature: *mixture of experts*, in which the responses of experts are nonlinearly combined by single gating network, and *hierarchical mixture of experts*, where the individual responses are combined by several gating networks arranged in hierarchical fashion. The simpler version is presented in figure 3.

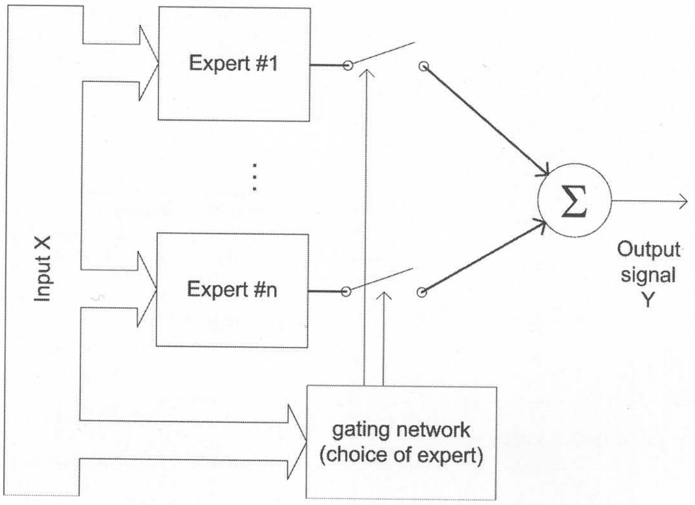


Fig. 3. Sample mixture of experts - input signal is involved in the choice of predictor

3. Sample problem – separation of two spirals

The properties of different committee machine solutions were presented using artificial dataset called "two spiral problem" [5]. Its scatter plot has been shown in figure 4. The task consists of two categories of examples in 2-D feature space, both having form of a spiral and beginning in the same point. Distinction between classes is a difficult problem due to highly nonlinear nature of the task.

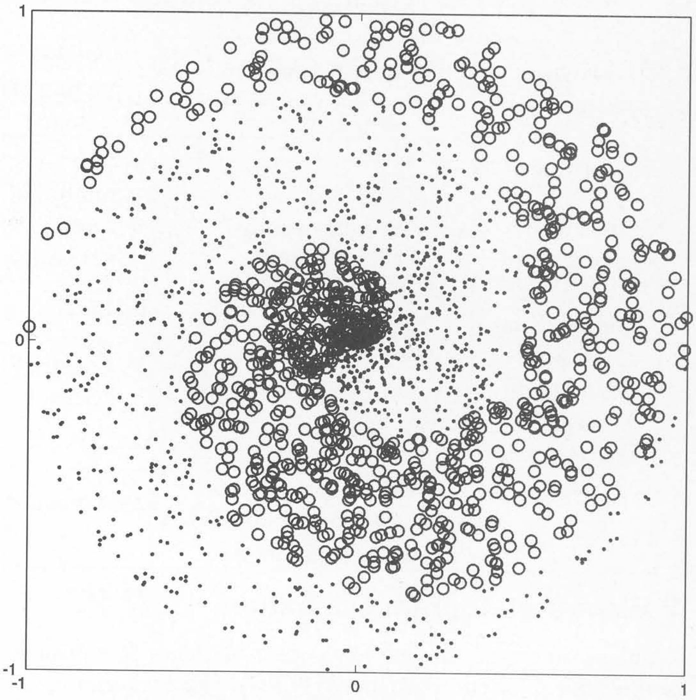


Fig. 4. Scatter plot of the "two spiral problem" dataset in 2-D space

The computations were performed in Matlab environment, using numerical procedures provided in [10] and own code, on a 1,8 GHz Pentium M machine. The efficiency of each classifier was evaluated using 10-fold cross-validation test [1,2,5], in order to prevent the dependency of the results from choice of test subset.

Both artificial "two spiral" problem, and electroretingram signal classification task (described further ahead) were solved using: single neural network classifier, ensemble of classifiers (simple voting), boosting algorithm and mixture of experts. The latter solution was implemented in very simplified form. Usually the gating network is trained in similar manner like the experts in the committee. In described situation the gating network was replaced by a simple function detecting to which quarter of the coordinate system does a particular example belong. Thus the training and test subsets were divided into four approximately equal parts, and used to train four experts. Each of the experts was responsible for a quarter of the X-Y plane (see fig. 4). Such an operation caused the distribution of the categories in each of the four training set to be significantly easier to learn.

The efficiency and computation time of all the committee approaches were compared. In both cases, the basic expert was multilayer perceptron with five hidden neurons, trained using gradient descent algorithm with variable learning rate. Such an architecture dos not provide satisfying classification rate. For example, in case of two spiral problem, increasing the number of hidden neurons results in increase of classification rate up to even 95 % (for several hundred hidden neurons). However, designing the optimal MLP network was not the goal of the experiment. It was rather to prove that even a relatively simple classifier can provide high recognition accuracy, when joined into the structure of committee machine.

The results summarized in table 1 suggest, that the same simple classifier that provide unsatisfactory recognition rate, after incorporation into the committee machine can seriously increase its efficiency. The difference between single network result and the *AdaBoost* algorithm with 100 iterations is very significant. Since two proportions (recognition rate) are compared, the difference can be quantified using the χ^2 test [6]. The conclusion is correct with significance level p=0,0197 (χ^2 (1) = 5,44).

Tab. 1. Misclassification rate for different algorithms used for solving the two spiral problem

network	Type of committee	committee parameter	parameter value	comp. time [s]	classification rate (test set) [%]
multilayer perceptron (2 – 5 – 1) trained using gradient descent with variable learning rate	single network	none	none	17	60
	simple voting	number of voters (networks)	2	29	60
			5	60	74
			10	150	76
			20	306	84
	AdaBoost	boosting iterations	10	232	86
			25	456	87
			50	933	90
			100	1857	94
	mixture of experts (feature space divided into four quarters)	number of experts	4 with simplified gating network	29	83

IV Electroretinogram evaluation

Authors attempted to apply committee machines in automatic identification of electroretinograms (PERG). These waveforms are 1-D electrical signals representing electrical activity of the retina. The signals, their role and significance, measurement procedure and rules of evaluation have been described in detail in [7,8]. The research was conducted in cooperation with Chair and Clinic of Ophthalmology at the Pomeranian Medical University. Currently various classification algorithms and signal analysis methods are being applied and discussed.

The medical examination, called electroretinography, consists of stimulation of the retina by a specific light stimulus and acquisition of the response by two measurement electrodes. Analysis of the local extrema of the time-domain plot allows diagnosing vision disorders. However, due to significant amount of noise and artifacts, such an analysis can be a difficult task. The procedures of evaluation are standardized by the ISCEV (International Society for Clinical Electrophysiology of Vision) organization.

The most important fact about the data is that an individual signal is described using six parameters corresponding to the local extrema. Therefore, such a signal can be viewed as a point in R^6 space. Previous results [7,8] suggest, that data in such a space is difficult to partition. This is caused by noisy character of the analyzed plots and thus inaccurate designation of the features. Therefore full time-domain information (256 samples / waveform) was used.

All the waveforms were mapped into 2-D space using Principal Components Analysis (PCA) [1,2,5,8]. The first two principal components retained 85% of the whole dataset variance. Their values were transformed, to ensure that the mean value of both principal components in the dataset was equal to zero, and fall into

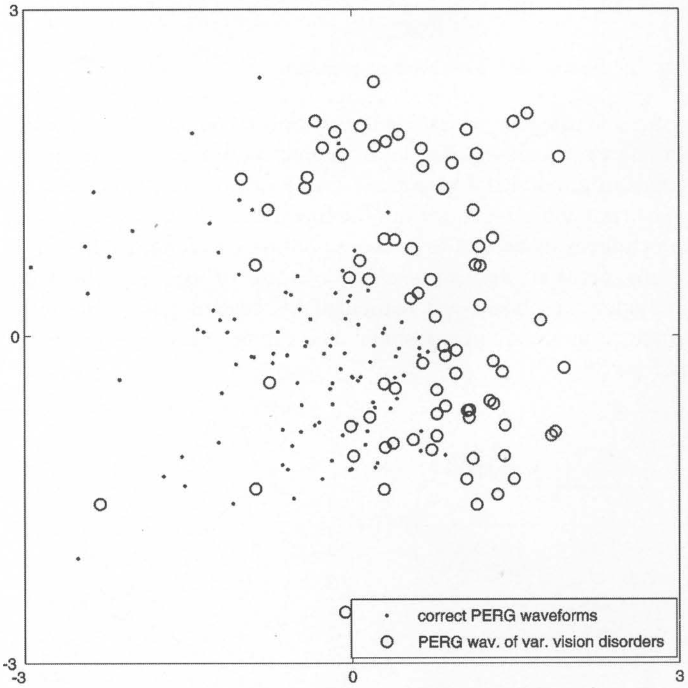


Fig. 5. Scatter plot of the PERG data mapped into 2-D space

Tab. 2. Misclassification rate for different algorithms used for PERG identification

network	type of committee	committee parameter	parameter value	comp. time [s]	classification rate (test set) [%]
multilayer perceptron (2 – 5 – 1) trained using gradient descent with variable learning rate	single network	none	none	2,51	83
	simple voting	number of voters (networks)	2	4,11	85
			5	9,21	85
			10	17,78	85
	AdaBoost	boosting iterations	10	21,47	84
			25	51,01	85
			50	107,15	82
			100	211,23	84
	mixture of experts	number of experts	could not be performed due to insufficient number of learning examples (feature space too sparse for further division)		

the <-3,3> range. Such a transformation is a lossy distortion of the PERG parameters values, but preserves the information about the separability of the dataset (layout of the points in the scatter plot remains unchanged). The scatter plot of the dataset, after mapping into first two principal components space, is presented in figure 5.

It is expected, that splitting the feature among multiple classifiers and combining them into committee machine, would alleviate the difficult distribution of categories. The solutions obtained using different algorithms are summarized in table 2. In each case the 5-fold cross-validation procedure was performed.

The data in table 2 signify the failure of *divide and conquer* principle in case of PERG waveform. The differences are statistically insignificant ($p \approx 0,9$). On the other hand, the results obtained by various committee machines are not worse than those obtained by single network. That may imply, that the main reason of the failure was insufficient number of cases in the dataset.

It can be also predicted intuitively. Methods like *AdaBoost* and *MoE* base on dividing the learning set into smaller parts, either by resampling the data or by dividing the feature space into disjoint regions. When the set is too small (data is too sparse in feature space), its further division is pointless. It results in obtaining very small subsets of data in relatively high dimensional space. Then, the "curse of dimensionality" phenomenon [1,2,5] starts to decrease the classification accuracy. Thus it would be advisable to repeat the experiment when more data become available.

5. Conclusions

The solution of two linearly inseparable problems using committee machines were presented. At the beginning both tasks were solved by multilayer perceptron network. In each case the network with low number of hidden neurons performed unsatisfactorily. When the network with identical architecture was used to form a committee, the performance of such classifier increased significantly. However, the benefits of using the committees in case of PERG recognition problem were reduced significantly due to limited number of training examples. In certain cases the results of the committee were equal to obtained using single network. On the other hand, the conclusions concerning two spiral problem, suggest that this approach should be tried again as soon as the sufficient amount of data become available.

Acknowledgements

This paper was supported by a grant from the State Committee for Scientific Research (KBN) number 3T11 023 28.

6. References

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

[2] Duda R.O., Hart. P.G., Stork D.G..: *Pattern Classification.* Wiley-Interscience (2001)

[3] Fernandes et. al.: Development of neural network committee machines for automatic forest fire detection using lidar. *Pattern Recognition* 37 (2004) pp. 2039-2047

[4] Fernandes et. al.: Design of committee machines for classification of single-wavelength lidar signals applied to early forest fire detection. *Pattern Recognition Letters* 26 (2005) pp. 625-632

[5] Haykin S.: *Neural networks: a comprehensive foundation.* Prentice Hall (1999)

[6] Montgomery D.C, Runger G..C.: *Apllied statistics and probability for engineers.* Wiley (2003)

[7] Rogala T., Brykalski A., Penkala K.: Certain aspects of bioelectrical signal smoothing. *Pomiary Automatyka Kontrola* 9/2004, pp. 21-24

[8] Rogala T., Brykalski A. .: Redukcja wymiarowości danych pomiarowych z wykorzystaniem liniowej i nieliniowej analizy składników głównych (PCA). *Pomiary Automatyka Kontrola* 2/2005, pp. 41-43

[9] Stamatatos E., Widmer G.: Automatic identification of music performers with learning ensembles. *Artificial Intelligence* 165 (2005) pp. 37-56

[10] Stork D.G., Yom-Tov E. : *Computer manual in Matlab to accompany pattern classification.* Wiley-Interscience (2004).

Tytuł: Praktyczne aspekty stosowania „komitetów" - układów połączonych klasyfikatorów.

Artykuł recenzowany

cd. ze str. 4

Trudno jest natomiast zaakceptować traktowanie *przewodnika* przez część środowiska metrologów jako swojego rodzaju biblię, co skutkuje pojawianiem się sporej liczby prac o kruchej podstawie naukowej. Dlatego trzeba przyznać sporo racji prof. Jaworskiemu, który w pracy [4] poddaje krytycznej analizie „grzechy twórców i zwolenników GUM-u teorii niepewności". Publikowanie tego rodzaju uwag jest bardzo istotne dla postawienia diagnozy stanu współczesnej metrologii lecz niezbędne jest zarazem podjęcie działań konstruktywnych mających na celu poprawę sytuacji. Niezbędna jest przede wszystkim dyskusja na różnych gremiach metrologicznych połączona z próbami publikowania jej efektów. Jest po temu wiele okazji, wśród których można wymienić Sympozjum nt. Niepewności Pomiarów, odbywające corocznie w Międzyzdrojach pod opieką prof. Kubisy, oraz konferencję „Podstawowe Problemy Metrologii" organizowaną przez ośrodek gliwicki. Dyskusje te niestety utrudnia duża różnorodność poglądów na temat problematyki niedokładności pomiaru widoczna w środowisku metrologów oraz wybujały rozrost terminologii nie zawsze interpretowanej jednoznacznie. Stąd - parafrazując słowa znanej piosenki - proponuję rozpocząć dyskusję pod hasłem „powróćmy do elementarza", które uznaję za na tyle istotne, że umieściłem je w tytule niniejszych uwag. Rozumiem to hasło w ten sposób, że w zaistniałej sytuacji niezbędne jest przede wszystkim uzgodnienie kanonu podstawowych pojęć (niezbędne będzie zapewne zrobienie użytku z tzw. „brzytwy Ockhama", czyli redukcja bytów do niezbędnego minimum), a następnie prowadzenie dyskursu wyłącznie językiem matematyki. Mam nadzieję, że Redakcja miesięcznika PAK udzieli swoich łamów wszystkim chętnym do zabrania głosu. Osoby, które z różnych względów nie chcą zabierać głosu publicznie, zachęcam do wysyłania uwag na podany poniżej mój adres poczty elektronicznej. Ja ze swej strony deklaruję, że postaram się zebrać i opublikować te uwagi.

Literatura

[1] Zięba A.: Natura zjawiska błędu pomiaru a konwencja GUM. Podstawowe Problemy Metrologii. Prace Komisji Oddziału PAN w Katowicach, Seria: Konferencje, Nr 8. Ustroń, 8-11 maj 2005, ss. 17-24.

[2] Metrologia w skrócie. Projekt EUROMET-u nr 595. Główny Urząd Miar, 2004.

[3] Guide to Expression of Uncertainty in Measurement. ISO 1993, 1995. Tłumaczenie polskie: Wyrażanie niepewności pomiaru. Przewodnik. Główny Urząd Miar, 1999.

[4] Jaworski J. M.: Lista grzechów twórców i zwolenników GUM-u teorii niepewności. Podstawowe Problemy Metrologii. Prace Komisji Oddziału PAN w Katowicach, Seria: Konferencje, Nr 8. Ustroń, 8-11 maj 2005, ss. 25-36.