AN APPLICATION OF OWA OPERATORS TO THE AGGREGATION OF MULTIPLE CLASSIFICATION DECISIONS

Ludmila I. Kuncheva[1]
Department of Electrical and Electronic Engineering
Imperial College, Exhibition Road, London SW7 2BT, UK
e-mail: L.Kuncheva@ic.ac.uk

Abstract

The paper considers a classification scheme made up by pooling together multiple classifiers and aggregating their decisions. The individual decisions are treated as degrees of membership assigned by the classifier to the object to be classified. We are interested in how the OWA operators compare to simple voting, linear and logarithmic techniques. In general, all the aggregation schemes appear to be of the same quality, superior to the single classifiers. It was found that OWA operators tend to generalize better than their competitors when the individual classifiers are overtrained. The idea is illustrated on a real and on an artificial data set.

Keywords: Pattern recognition, aggregation of multiple classifiers, committees of networks.

1 Introduction

Aggregating of multiple classification decisions borrows from human decision making the idea of working up a decision by collecting groups, teams, committees of individual experts. We consider a pool of classifiers each of which is supposed to make a classification decision, i.e., to assign the input object to a class from a predefined finite set of mutually exclusive classes. Provided all misclassifications are equally costly, the final goal is to minimize the number of misclassifications over the space containing all possible objects.

We will refer to the classifiers as "experts" and consider a simple scheme of expertise: the individual decisions are produced in one session without any

[1]On leave from CLBME, Bulgarian Academy of Sciences

communication between the experts. (For an excellent description of this correspondence the reader may refer to the paper by Ng and Abramson, 1992)[1].

Let $x \in \Re^n$ be the $n-$dimensional feature vector describing the object to be classified, and let $\Omega = \{\omega_1, \ldots, \omega_M\}$ be the set of classes. Three types of classifiers are detailed in [2]:

- Type 1 classifiers, the output of a classifier is a single class label. Therefore classifiers implement the mapping:

$$\psi_1 : \Re^n \rightarrow \Omega.$$

- Type 2 classifiers, the classifiers' outputs are some rank orderings on Ω, i.e.,

$$\psi_2 : \Re^n \rightarrow \mathcal{P}(\Omega).$$

  where $\mathcal{P}(.)$ stands for the class of all permutations of its argument set.

- Type 3 classifiers, each classifier yields some quantitative degrees of membership to each of the classes. Without losing generality we can confine the range of these degrees to the unit interval, so that each classifier implements the mapping:

$$\psi_3 : \Re^n \rightarrow [0,1]^M. \tag{1}$$

Note that instead of $\Re^n$ we can consider any feature space: continuous, discrete, or mixed, because the matter of interest here is the output of the classifier and not the classification mapping itself.

Depending on what output is adopted, a respective aggregation operator can be picked up. The simple voting scheme is applicable to any of the above type of classifiers ([3, 4, 2]). The main result is that given a set of *independent* classifiers whose classification accuracy exceeds 50 %, the more voters we have, the higher the classification accuracy is, the limit being the optimal (Bayesian) accuracy. Voting has been found to perform reasonably well in broad range of settings which, along with its simplicity, makes it an appealing choice.

Comparatively fewer studies address the case of rank orderings (type 2) [5, 6]. The vast majority of studies in aggregation are devoted to the third type of classifiers listed above. The most popular choice has been the **linear** aggregation pool both from heuristic and rigorous statistical point of view [7, 1, 8, 2, 9, 10].

There are many ways to implement the aggregation if we have classifiers of type 3. A consequence of the flexibility of fuzzy set theory is the abundance of aggregation operators that resemble some human decision-making rationale [11,

12, 13]. Choosing one of them for a particular practical problem might appear difficult. Since their introduction by Yager in 1988 [14], the Ordered Weighted Averaging (OWA) operators have been studied thoroughly with respect to their connections with the large family of fuzzy aggregation operators[15, 13]. Those operators incorporate most naturally the simplest aggregation perspectives: pessimistic, optimistic, indifferent, competition jury, etc. Their application however has not been a focus of attention so far.

It is worth mentioning that many catchy names have been used to label the combination of multiple classification decisions: committees of networks [16], adaptive mixture of experts [17], pandemonium system of reflective agents [18], divide-and-conquer methodology [19], change-glasses approach [20], etc. An interesting direction of research stems from the idea that the parameters of the aggregation operator may not be constant over the feature space but vary with the input, expressing some input-dependent characteristics of the experts like "competence", "variance", "confidence" etc. In fact, this leads to a hard or soft partitioning of the feature space, thereby allowing for applying different aggregation schemes on different parts of the feature space [17, 21, 18, 22, 23, 20].

In this paper we are interested in the application of the OWA operators to the aggregation of multiple classification decisions. We will assume that the parameters of the operators are constant over the feature space. Admittedly, an architecture based on variable parameters would outperform the constant-parameters one. Therefore we compare the results with those from the simple voting and the linear aggregation rule. The formal description of the problem is presented in Section 2. Section 3 contains the experimental results on the two-spirals data set. In section 5 some conclusions are drawn.

2 OWA operators for combining multiple classification decisions

We consider a scheme consisting of L individual classifiers. Let $y_i(x) \in [0,1]^M$ denote the output of the ith classifier for the input vector x, $i = 1, \ldots, L$. The aggregated value for the jth class is

$$y^{(j)}(x) = \mathcal{F}\left(y_1^{(j)}(x), \ldots, y_L^{(j)}(x); \theta\right),$$

where the superscript (j) denotes the jth component of the vector, and θ is the set of parameters. We assume that both $\mathcal{F}$ and θ do not vary over the feature space.

Definition. *An L-place OWA operator* $\mathcal{F}(a_1, a_2, \ldots, a_L; \theta)$ *is defined by the equation:*

$$\mathcal{F}(a_1, a_2, \ldots, a_L; \theta) = \theta_1\, b_1 + \theta_2\, b_2 + \ldots + \theta_L\, b_L$$

where b_i *is the ith largest element in the collection* $a_1, \ldots, a_L$, *and* $\theta = [\theta_1 \ldots \theta_L]^T$ *is a parameter vector associated with* $\mathcal{F}$, *such that*

1. θ_i *are nonnegative;*
2. $\sum_{i=1}^{L} \theta_i = 1$. □

It has been pointed out that OWA operators easily represent some of the widely used aggregation operators by the following parameter vectors

- minimum
$$\theta_{\min} = [\,0\; 0\; \ldots\; 0\; 1\,]^T$$
- maximum
$$\theta_{\max} = [\,1\; 0\; \ldots\; 0\; 0\,]^T$$
- average
$$\theta_{average} = \left[\, \frac{1}{L}\; \frac{1}{L}\; \cdots\; \frac{1}{L} \,\right]^T$$
- competition jury
$$\theta_{jury} = \left[\, 0\; \frac{1}{(L-2)}\; \cdots\; \frac{1}{(L-2)}\; 0 \,\right]^T \tag{2}$$

We will also use an OWA vector that can be viewed as a model of the linguistic quantifier "most". The interpretation is intended for the case where a linguistic expression is assigned to the aggregation operation, e.g. "Most experts agree on class ω_2".

- "most"
$$\theta_{\text{“}most\text{”}}(k) = \frac{k^2}{\sum_{i=1}^{L} i^2}, \quad k = 1, \ldots, L. \tag{3}$$

OWA operators possess the necessary features to be considered as an appropriate option for aggregation of classification decisions: they are monotonic on their arguments and idempotent. Note that we can assign an individual OWA vector for each class decision. By choosing the desirable linguistic expression we can either favor or neglect the class to a certain degree, and therefore move the classification boundary in the respective direction. This can be used as an alternative way to express the cost of the decision than by using a lost matrix.

If the classifiers' outputs are interpreted as posterior probabilities, aggregation by operators with monotonically decreasing coefficients $\theta_1 > \theta_2 > \ldots > \theta_L$ (e.g.,"most", maximum, etc.) favors more "confident" classifiers and punishes "skeptics". A scheme like "jury" (2) prevents rating high "confident ignorants" among the members of the committee. It is not clear in advance which of the OWA schemes will fit best the problem. This leads to the idea of estimating the OWA coefficients from the training data [24] instead of fixing them in advance. For this purpose we can use the nonnegative least square method on the sorted outputs and further rescaling so that the coefficients sum up to one. Indeed, this estimating procedure involves some imprecision (due to rescaling) but it is interesting to see whether some interpretable OWA shape can be picked up.

The competing aggregation techniques considered here are the simple voting, the plain logarithmic pool, and the weighted linear combination of classifiers' outputs with the least squares estimate of the coefficients on the training data.

3 Experiments

3.1 The two-spirals data set

The two-spirals data set is a challenging benchmark artificial problem for testing out classification techniques. It consists of two classes of 2-D vectors disposed as intertwined spirals (Fig. 1). The training and the test set, as provided initially are nearly the same which does not provide any space for testing the generalization performance of the classifier. Since we are interested in the generalization of the OWA aggregation of the individual decisions, we added random Gaussian noise to the training data with mean 0 and standard deviation 0.5 keeping the test set as originally designed.

The first-level classifiers ("experts", members of the committee) used in this study were radial-basis function (RBF) networks build on random subsets of the training set (contaminated with noise). For all experiments the **neural network toolbox** of `matlab` was used. In fact, the concrete implementation of the first-level classifiers is of no interest, provided they perform the mapping (1) and are trained independently, in the same experimental setting.

A series of experiments has been carried out showing similar results. An example is picked up to illustrate the better generalization performance of OWA classifier over its competitors. Admittedly, the most expressive example has been chosen. The tendency however was clearly presented in the rest of experiments.

Seven classifiers have been used, each one based on 29 nodes. The average of the classification accuracy on the training set of those classifiers was 62.87 %, and on the test set 58.04 %. We have deliberately selected the parameters of

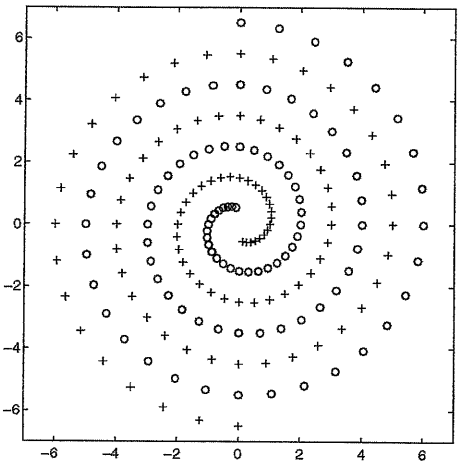

Figure 1:

the first-level classifiers so that the training would not lead to good results on the training set, and besides, we let for some overtraining. This has been done in order to study both the ability of the committee of classifiers to get better results than the single classifiers, and to check its generalization.

Figures 2 and 3 show the results from the example. The output of the combined classifier is plotted on each figure with the intensity of the gray level corresponding to the value of the function, 0 being presented by white, and 1 by black. The points from the original training set (before adding noise) which ideally should be assigned value 1 by the classifier is also depicted for reference. The classification accuracy on the test set using the respective aggregation technique shown in the brackets.

By "OWA" we denote the OWA aggregation with coefficients estimated by the nonnegative least squares algorithm (function `nnls` from `matlab`) on the *sorted* classifiers' outputs, and by "Linear" – weighted aggregation with coefficients estimated by the same function on the outputs.

For comparison, the same classification technique (RBF network) was run until the error goal was reached. This caused huge overtraining with training accuracy 100 %, and test one 49.48 %. The final result from this experiment is shown in Fig. 3. Clearly, the best choice in this example is the OWA aggregation scheme.

3.2 The "heart" data set

The second data set is taken from the database PROBEN1, ftp address:

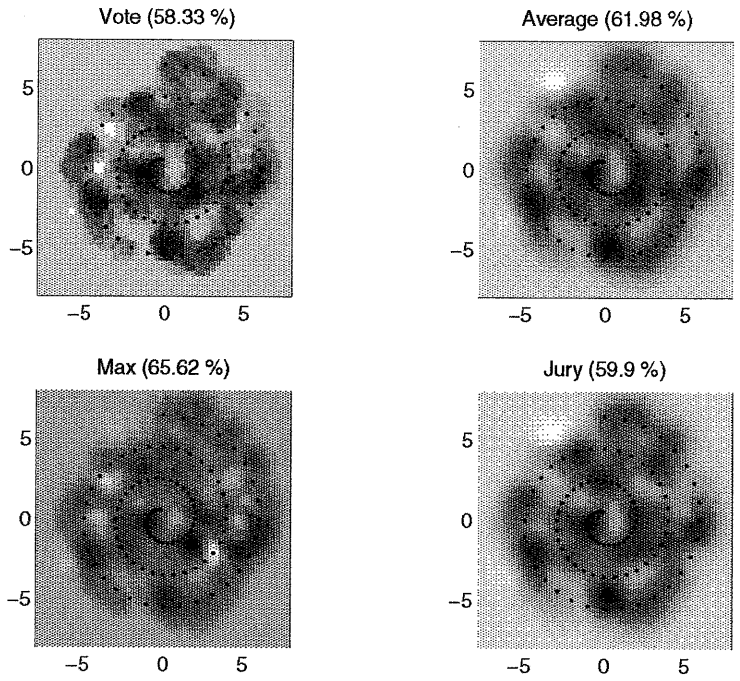

Figure 2:

`ftp://ftp.ira.uka.de/pub/neuron/proben1.tar.gz`.

A detailed description can be found in the Technical Report by Prechelt[25]. The data set has been supplied by Dr. Robert Detrano, V.A. Medical Center, Long Beach and Cleveland Clinic Foundation.

The data set consists of 303 patient records containing mixed variables (both continuous-valued and binary) taken from patient's history, clinical examinations, laboratory tests, etc. Two classes are considered depending on whether or not at least one of four major coronary blood vessels of a patient is reduced in diameter by more than 50 %.

The data set is called here **heart**, and the three partitions (heart1, heart2, and heart3) of the set into training and test parts are the same suggested by Prechelt[25]. Each partition comprises 228 training samples and 75 test ones.

The main task in this study has not been to achieve the best possible accuracy on the data set but to investigate the generalization abilities of some classification paradigms. Therefore no effort has been made to select optimal parameters with respect to the classification performance and the results are not supposed to be competing with those reported elsewhere.

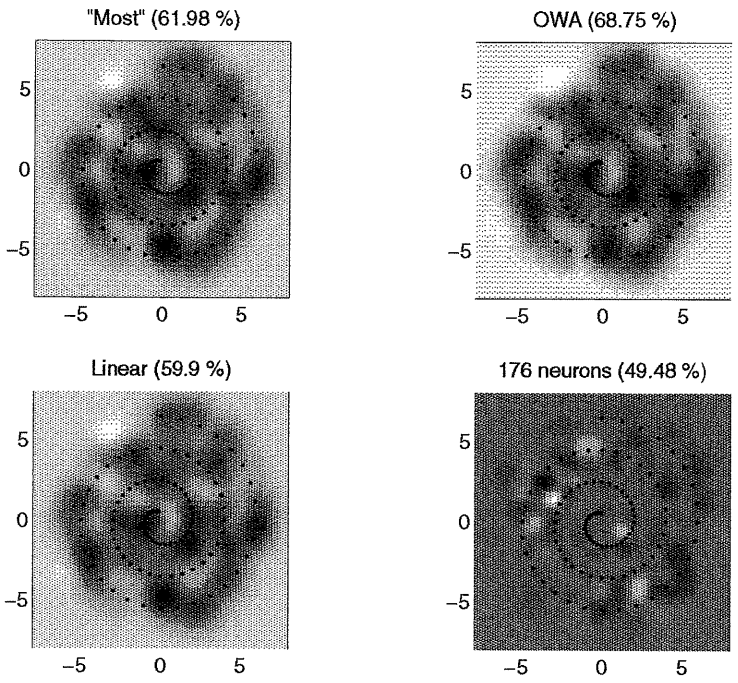

Figure 3:

After discarding some features with zero variance over the data set, the resultant feature set was reduced to 18 features. The fast back-propagation training algorithm from **neural network toolbox** was employed for different configurations of a multi-layer perceptron with one hidden layer. In each experiment, the first level classifiers used the same number of hidden nodes.

Tables 1 to 4 show the test results for the classification paradigms under consideration.

Five multi-layer perceptron networks (one hidden layer) have played the role of the "experts" in the committee. Four series of experiments have been carried out:

1 Every classifier has 2 hidden nodes. Randomly chosen 50 % of the *training* data set is used for training the classifier;

2 Classifiers have 2 hidden nodes each, the whole training data set is used;

3 Classifiers have 20 hidden nodes each, 50 % of the training set is used;

4 Classifiers have 20 hidden nodes each, the whole training set is used.

Classification paradigm	heart 1	heart 2	heart 3
Individual (average)	0.7013 (0.0604)	0.8285 (0.0488)	0.7483 (0.0274)
Individual (maximum)	0.7373 (0.0267)	0.8573 (0.0548)	0.7693 (0.0388)
Voting	0.7573 (0.0392)	0.8893 (0.0367)	0.8053 (0.0143)
Average	0.7733 (0.0178)	0.8853 (0.0328)	0.8080 (0.0275)
Linear	0.7653 (0.0180)	0.8880 (0.0275)	0.8067 (0.0261)
Logarithmic	0.7747 (0.0183)	0.8853 (0.0309)	0.8053 (0.0322)
Maximum	0.7813 (0.0220)	0.8920 (0.0231)	0.7987 (0.0317)
Jury	0.7707 (0.0207)	0.8853 (0.0394)	0.8053 (0.0303)
"Most"	0.7747 (0.0160)	0.8853 (0.0328)	0.8040 (0.0345)
OWA	0.7773 (0.0126)	0.8947 (0.0231)	0.8160 (0.0325)

Table 1: Test results from experiment #1: classification accuracy and standard deviations

Classification paradigm	heart 1	heart 2	heart 3
Individual (average)	0.7477 (0.0310)	0.8824 (0.0518)	0.8077 (0.0255)
Individual (maximum)	0.7547 (0.0220)	0.9053 (0.0193)	0.7947 (0.0351)
Voting	0.7707 (0.0084)	0.9093 (0.0105)	0.8267 (0.0166)
Average	0.7667 (0.0094)	0.9093 (0.0138)	0.8307 (0.0227)
Linear	0.7680 (0.0143)	0.9093 (0.0151)	0.8040 (0.0295)
Logarithmic	0.7667 (0.0094)	0.9067 (0.0126)	0.8320 (0.0220)
Maximum	0.7680 (0.0093)	0.9120 (0.0093)	0.8267 (0.0208)
Jury	0.7667 (0.0094)	0.9120 (0.0129)	0.8280 (0.0147)
"Most"	0.7667 (0.0094)	0.9067 (0.0126)	0.8320 (0.0220)
OWA	0.7693 (0.0090)	0.9093 (0.0123)	0.8227 (0.0141)

Table 2: Test results from experiment #2: classification accuracy and standard deviations

Classification paradigm	heart 1	heart 2	heart 3
Individual (average)	0.7104 (0.0483)	0.7509 (0.0636)	0.7288 (0.0399)
Individual (maximum)	0.7560 (0.0295)	0.8440 (0.0412)	0.7480 (0.0414)
Voting	0.7520 (0.0322)	0.8293 (0.1189)	0.7947 (0.0357)
Average	0.7613 (0.0231)	0.8773 (0.0439)	0.7920 (0.0383)
Linear	0.7600 (0.0259)	0.8760 (0.0333)	0.7787 (0.0261)
Logarithmic	0.7587 (0.0231)	0.8693 (0.0401)	0.7840 (0.0360)
Maximum	0.7573 (0.0287)	0.8773 (0.0406)	0.7813 (0.0351)
Jury	0.7493 (0.0325)	0.8813 (0.0317)	0.7947 (0.0328)
"Most"	0.7587 (0.0231)	0.8760 (0.0427)	0.7933 (0.0384)
OWA	0.7560 (0.0244)	0.8827 (0.0349)	0.7960 (0.0367)

Table 3: Test results from experiment #3: classification accuracy and standard deviations

Classification paradigm	heart 1	heart 2	heart 3
Individual (average)	0.7595 (0.0272)	0.8859 (0.0100)	0.7763 (0.0103)
Individual (maximum)	0.7733 (0.0208)	0.8747 (0.0157)	0.7773 (0.0227)
Voting	0.7787 (0.0129)	0.9053 (0.0284)	0.7880 (0.0247)
Average	0.7787 (0.0143)	0.9093 (0.0197)	0.7813 (0.0180)
Linear	0.7680 (0.0143)	0.8960 (0.0250)	0.7747 (0.0160)
Logarithmic	0.7800 (0.0130)	0.9093 (0.0197)	0.7800 (0.0181)
Maximum	0.7733 (0.0154)	0.9027 (0.0178)	0.7773 (0.0178)
Jury	0.7787 (0.0129)	0.9053 (0.0255)	0.7893 (0.0197)
"Most"	0.7800 (0.0130)	0.9107 (0.0178)	0.7773 (0.0178)
OWA	0.7733 (0.0126)	0.9040 (0.0186)	0.7760 (0.0138)

Table 4: Test results from experiment #4: classification accuracy and standard deviations

All training sessions have been stopped (without reaching the error goal) after the 200th epoch. Ten experiments have been carried out with each setting and the results presented in the tables are the average and the standard deviation.

The results are divided into three groups:

- One level classification results:
  - average of the constituents of the combined scheme;
  - the test accuracy of the classifier with the highest training accuracy.
- Standard aggregation techniques:
  - voting;
  - average;
  - linear – this corresponds to weighted aggregation where the coefficients are estimated via nonnegative least square function and rescaled afterwards;
  - logarithmic – the aggregation is performed by multiplying the classifiers' outputs.
- OWA group of aggregation techniques:
  - maximum;
  - jury;
  - "most"
  - OWA

  All of the notations are as described in the previous section.

It is interesting to see whether the OWA coefficients estimated from data can be matched to some interpretable profile. Figure 4 shows the averaged OWA profiles with partitions heart1, heart2, and heart3, for experiments # 1 to 4.

The OWA profiles are not expected to be identical for the two classes. It can be seen that there is no clear resemblance between the coefficients' profiles. It should be mentioned that the standard deviations of the coefficients were very large. This shows that for this particular task, the "optimal" OWA coefficients with respect to the least squares have been specific for each experiment.

4 Analysis and conclusions

The results show that OWA operators are of comparable quality to the most widely used aggregation operators of the same group (nonchanging parameters

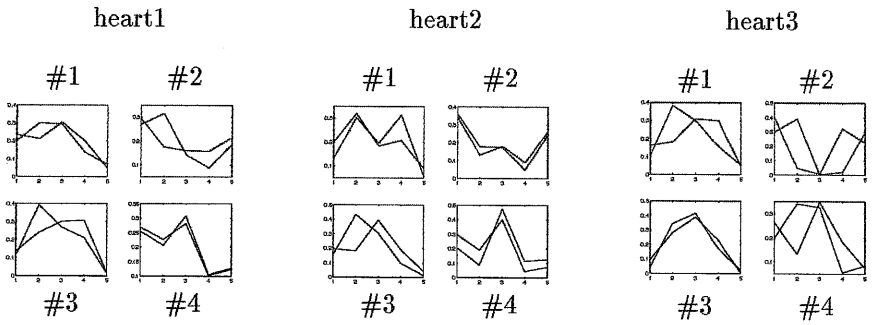

Figure 4: OWA coefficients

and aggregation type over the feature space). A better generalization capability can be expected due to the fact that they treat the classifiers as indistinguishable experts, maintaining only a profile with coefficients on the ordered decisions, rather than attaching coefficients to every expert. The risk in choosing the latter is that an "expert" whose decision fits the training set very well, and in fact overfits it, will be given most credit throughout the whole feature space. On the contrary, using OWA aggregation we do not put our "trust" in one classifier only but let the pool share it.

Indeed, OWA showed better generalization with the two-spiral data set. With the second data set, the difference in favor of OWA group of aggregation techniques is negligible. We may expect that OWA operators will perform better in case of overtrained first-level classifiers.

It is unlikely that the vast majority of practical classification problems will need sophisticated aggregation schemes. Therefore, following Occam's razor principle, the simplest aggregation techniques should be tried first [1]. OWA operators can be labeled as such, having at the same time some intellectually pleasing correspondence with human decision making.

5 Acknowledgements

This work has been supported by a fellowship kindly provided by the Royal Society and the Foreign & Commonwealth Office.

References

[1] Ng, K.-C., Abramson, B.: Consensus diagnosis: a simulation study, *IEEE Transactions on Systems, Man, and Cybernetics* **22** (1992) 916-928.

[2] Xu, L., Krzyżak, A., Suen, C.Y.: Methods of combining multiple classifiers and their applications to handwriting recognition, *IEEE Transactions on Systems, Man, and Cybernetics* **22** (1992) 418-435.

[3] Battiti, R., Colla, A.M.: Democracy in neural nets: voting schemes for classification, *Neural Networks* **7** (1994) 691-707.

[4] Lam, L., Suen, C.Y.: A theoretical analysis of the application of majority voting to pattern recognition, *Proc. 12th International Conference on Pattern Recognition, Jerusalem, Israel* (1994) 418-420.

[5] Ho, T.K., Hull, J.J., Srihari, S.N.: Decision combination in multiple classifier systems, *IEEE Transactions on Systems, Man, and Cybernetics* **16** 66-75.

[6] Tubbs, J.D., Alltop, W.O. Measures of confidence associated with combining classification results, *IEEE Transactions on Systems, Man, and Cybernetics* **21** (1991) 690-692.

[7] Jacobs, R.A.: Methods for combining experts' probability assessments, *Neural Computation* **7** (1995) 867-888.

[8] Benediktsson, J.A., Swain, P.H.: Consensus theoretic classification methods, *IEEE Transactions on Systems, Man, and Cybernetics* **22** (1992) 688-704.

[9] Hashem, S., Schmeiser, B., Yih, Y.: Optimal linear combinations of neural networks: An overview, *Proc. of the IEEE International Conference on Neural Networks, Orlando, Florida* (1994) 1507-1512.

[10] Trsep, V., Tanaguchi, M.: Combining estimators using nonconstant weighting functions, *in: Tesauro G., Touretzky, D.S., Leen, T.K., eds.: "Advances in Neural Information Processing Systems 7", MIT Press, Cambridge MA* (1995).

[11] Dubois, D., H. Prade. A review of fuzzy aggregation connectives, *Information Sciences*, **36**, 1985, 85-121.

[12] Bloch, I. Information combination operators for data fusion: A comparative review with classification, *IEEE Transactions on Systems, Man, and Cybernetics*, **26**, 1996, 52-67.

[13] Grabisch, M. On equivalence classes of fuzzy connectives – the case of fuzzy integral, *IEEE Transactions on Fuzzy Systems*, **3**, 1995, 96-109.

[14] Yager, R.R. On ordered weighted averaging aggregation operators in multicriteria decisionmaking, *IEEE Transactions on Systems, Man, and Cybernetics*, **18**, 1988, 183-190.

[15] Fodor, J. J.-L. Marichal, M. Roubens. Characterization of the ordered weighted averaging operators, *IEEE Transactions on Fuzzy Systems*, **3**, 1995, 236-240.

[16] Bishop, C. Neural Networks for Pattern Recognition, *Clarendon Press, Oxford*, 1995.

[17] Jacobs, R.A., Jordan, M.I. Adaptive mixture of local experts, *Neural Computation* **3** (1991) 79-87.

[18] Śmieja, F. The pandemonium system of reflective agents, *IEEE Transactions on Neural Networks*, **7**, 1996, 97-106.

[19] Chiang C.-C., Fu, H.-C.: A divide-and-conquer methodology for modular supervised neural network, *Proc. IEEE International Conference on Neural Networks, Orlando, Florida* (1994) 119-124.

[20] Kuncheva L.I., Change-glasses approach in pattern recognition, *Pattern Recognition Letters*, **14**, 1993, 619-623.

[21] Alpaydin, E. Combining global vs local perceptrons for classification, *Proc. International Conference on Soft Computing, SOCO'96, Reading, UK*, 1996, B291-297.

[22] Rastrigin, L.A., Erenshtein, R.H.: *Method of Gourd Recognition, Moscow, "Energoizdat"* (1981).(In Russian)

[23] Dasarathy, B.V., Sheela, B.V.: A composite classifier system design:concepts and methodology, *Proceedings of the IEEE* **67** (1979) 708-713.

[24] Filev, D., R.R. Yager. Learning OWA operator weights from data, *Proc. IIId IEEE Conference on Fuzzy Systems, Orlando, FL*, 1994, 468-473.

[25] L. Prechelt, PROBEN1 - A set of neural network benchmark problems and benchmarking rules, *Technical Report* # 21/94 (1994).