Control and Cybernetics

vol. **25** (1996) No. 2

Fuzzy aggregation of multiple classification decisions

by

Ludmila Kuncheva

Central Laboratory of Biomedical Engineering,
Bulgarian Academy of Sciences,
Acad. G. Bonchev Street, Block 105
1113 Sofia, Bulgaria

Abstract: A two-level classification scheme is considered where the first-level classifiers yield fuzzy decisions (degrees of support) for a certain class. These values are further aggregated in order to arrive at a final degree of support and to infer the class label. The aggregation operation proposed here uses degrees of consensus between first-level decisions for strengthening or alleviating the final decision support. An experimental example is presented in order to compare the proposed rule with some other aggregation connectives with respect to the classification accuracy.

1. Introduction

Two-level classification (combination of multiple classifiers) is a discipline gaining speed in the recent years although the idea emerged earlier, Barabash (1983), Dasarathy, Sheela (1979), Jozefczyk (1986), Rastrigin, Erenstein (1981). It stems from a decision making related experience that a collective of classifiers can exhibit a better overall classification performance than the best single classifier. The increasing interest in this direction has come to emphasize that, among the design criteria of a pattern classifier, the computational complexity is being subsequently shifted to a secondary plan, leaving the primary one to the classification accuracy. In other words, more complex classification structures are justifiable if this improves the classification performance.

Two main groups of two-level classification strategies may be distinguished depending on the type of the decisions of the first-level classifiers: *Complementary* and *Competitive.* Figure 1 shows a possible grouping of methods in two-level classification.

The studies in the *Complementary* stream assume that one of the first-level classifiers should be given the right to make decision for each single object. The

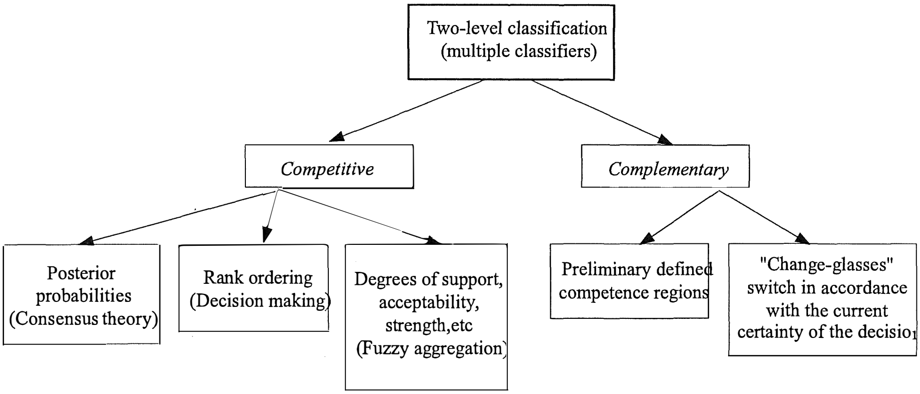

Figure 1.

main question is how to guess which of the classifiers is the best one for the object being classified. Two approaches have been suggested. The first one is the so called "Change-glasses", see Kuncheva (1993), Mitra, Kuncheva (1994), Ho, Hull, Srihari (1994) (with another catchy name "divide-and-conquer", Chiang, Fu 1994) in which the classifier is supposed to "realize" its degree of competence and if it is below a certain threshold, the paradigm switches to another (more precise) classification rule. The regions of uncertain decision may also be preliminarily learned, as proposed in Chiang, Fu (1994), Kuncheva, Mitra (1994), Mitra, Kuncheva (1994). The second approach is to partition the feature space into predefined regions of competence, one for each classifier. The object is first analyzed in order to find out in which competence region it falls and then the respective classifier makes the decision, Dasarathy, Sheela (1979), Rastrigin, Erenstein (1981). It should be mentioned that this strategy has given rise to many heuristic pattern classification techniques. Sometimes they are not even designated as two-level classifiers although the idea is implied.

The *Competitive* strategy seems to be the most commonly employed one. All the classifiers are supposed to yield classification decisions trying to guess the right class. Depending on what the classification outputs are (rank orders of the classes, Ho, Hull, Srihari 1994, Tubbs, Alltop 1991, posterior probabilities, binary values denoting crisp class labels, Lam, Suen 1994, Wernecke 1992, etc.) different aggregation techniques should be applied, Xu, Krzyzak and Suen (1992).

In the following it will be assumed that every first-level classifier yields a numerical estimate in the interval $[0, 1]$ for each class.

In a probabilistic framework, one could apply to this classification problem the so called consensus theory, Benediktsson, Sveinsson and Ersoy (1993), Benediktsson, Swain (1992), Bordley 1982, Ng and Abramson (1992). Having originated from multisource data analysis, these methods have recently been considered in the context of two-level pattern recognition where the classifiers are supposed to yield estimates of the posterior probabilities for the classes. The consensus theory provides guidelines to combine the individual decisions into a final estimate under different hypothesis of their interdependence.

When we treat the classifiers' outputs as degrees of acceptance, support, compliance, typicality, severity, strength of confirmation, etc., then the fuzzy aggregation rules appear to be more appropriate.

The present paper considers a fuzzy statement of a competitive strategy for two-level classification. A fuzzy operation is proposed for aggregating the decisions of several classifiers expressed as degrees of support for a given class. A refusal option is embedded at the second stage, which means that the decision concerning a given class includes the following three alternatives: {Yes, No, Refuse to decide}. The fuzzy aggregation rule itself is based on a degree of consensus between the first-level classifiers. The higher the degree, the greater the support to either positive or negative decision. The rule uses explicitly an aggregation connective that could be selected among the existing ones. As an appealing option we propose to use the OWA operators Yager (1987;1988) for such an aggregation. Sections 2 and 3 contain a formal statement of the problem along with some comments on the existing aggregation rules and their applicability to the two-level classification. The aggregation rule is described in Section 4. Section 5 presents an experimental illustration.

2. Two-level classification

The design process of a two-level classifier consists of two main steps: (1) Choice and training of the first-level classifiers; (2) Choice of the aggregation rule. The first stage, although tightly connected with the final result, will not be considered here. Only the surprising fact is worth mentioning that a vast number of authors consider neural networks as the first-level classifiers Alpaydin (1993), Battiti, Colla (1994), Benediktsson, Sveinsson and Ersoy (1993), Chiang, Fu (1994), Filippy, Costa, Pasero (1994), Hashem, Schmeiser, Yih (1994), Huang, Suen (1994), Jordan, Jacobs (1992), Nadal, Legault, Suen (1990), Poddar and Rao (1993), Sung-Bae, Kim (1992). It can be argued that much simpler and equally powerful conventional pattern classification techniques might lead to similar results. Perhaps the neural networks classifiers are preferred due to their ability to learn complex classification boundaries, neglecting sometimes the eventual generalization problems.

Formally, the second stage of the two-level classification paradigm can be directly mapped onto the decision making task (this analogy has been implied in almost all papers concerning the topic). Let $\Omega = \{\omega_1, \ldots, \omega_M\}$ be the set of classes (alternatives) and let x be an object generated by one of them (in general we can regard x as a d-dimensional feature vector). The problem to be solved is to infer the correct class knowing the classification decisions (individual preferences) of L classifiers (decision makers). Let $R = \{R_1, \ldots, R_L\}$ be the set of classifiers and let r_{ij} denote the assessment of the i-th classifier with respect to the j-th class. The term "*decision profile*" will be used to denote the $L \times M$ matrix $[r_{ij}]$. Values r_{ij} may denote degrees of support, strength of evidence, confidence, belief, acceptance rate, etc. Once they are to be used in a competitive type of a two-level classifier, we assume that they assess the same value (e.g. a degree of support) in the same scale.

The two-level classification environment considered here is characterized by the following additions to the classical pattern recognition setting:

- All classifiers operate simultaneously without interacting during the first-level classification process.
- All data of the object to be classified are supplied to the input of each classifier. Nevertheless, a classifier may be restricted to using only a subset of features "masking" the redundant ones.

The problem of inferring the correct class corresponds straightforwardly to the one of choosing the best alternative. It can be solved by using an appropriate aggregation rule. The objective, from a classification point of view, is to surpass the single-level classifiers with respetc to classification accuracy.

The "refuse" option is usually introduced when the cost of making a wrong decision is too high. The classifier is supposed to be able to measure a degree of confidence and to refuse to specify a class label in case of a low confidence. The two-level classification is especially convenient for such type of decision making because the classifiers can be regarded as experts and the confidence can be measured by the degree of consensus between them. In the ideal case the refuse rate $\mathbf{R}$ should be kept as low as possible while the accuracy $\mathbf{A}$ should be as high as possible. The result can be visualized as a curve $\mathbf{A}(\mathbf{R})$ on an *Accuracy-Refuse* plane. It can be formally proven that refusing to decide on the objects with lowest values of the maximal posterior probabilities leads to an increasing function $\mathbf{A}(\mathbf{R})$, Battiti, Colla (1994).

3. Aggregation rules

The choice of an optimal aggregation rule in the sense of the highest attainable classification accuracy requires the complete knowledge of probabilistic characteristics of the first-level classifiers which in most practical instances is not feasible. It can be shown that the approximation of the posterior probabilities as precisely as possible is a sufficient but not a necessary condition for obtaining an optimal two-level classifier. That is, we can elude a precise estimation (implied in the consensus theory) still being able to arrive at a good solution.

Furthermore, if we consider the outputs of the first-level classifiers in their fuzzy context, some constraints drop (e.g. the orthogonality property stating that the degree of support that each classifier yields for the classes should sum up to 1). Therefore, what seems to be particularly attractive is that this alleviation allows to use a rich palette of aggregation rules offered by fuzzy decision making. The rules range from the simplest max-min ones to the more sophisticated preference analyses, outranking theory, hybrid aggregators, etc. Krishnapuram, Lee (1992), Ramakrishnan, Rao (1992), Yager (1987;1988), Zimmermann (1987), Zimmermann, Zysno (1983).

Let us first consider the decision profile containing posterior probabilities, i.e. $r_{ij} = P_i(\omega_j/x)$. If all classifiers use the same features and yield the Bayesian posterior probabilities, then $P_1(\omega_j/x) = \ldots = P_L(\omega_j/x) = P(\omega_j/x)$. There is no need to build a two-level scheme since the optimal decision can be obtained using any of the classifiers. Let us suppose, however, that the classifiers use conditionally independent subsets of features $X^{(1)}, X^{(2)}, \ldots, X^{(L)}$ which form a partition on the set of features $\mathbf{X}$, i.e.

$$\bigcup_{i=1,\ldots,L} X^{(i)} = \mathbf{X};\ X^{(i)} \cap X^{(j)} = \emptyset,\ \forall i,j = 1,\ldots,L,\ i \neq j \tag{1}$$

Then the problem becomes to find a proper aggregation function in order to obtain the Bayesian posterior probability at the second level. Let $p(x_i/\omega_j)$ be the conditional probability density functions (p.d.f.s) for x_i, x_i taking values in the space defined by the subset of features $X^{(i)}$, $i = 1,\ldots,L$. Let $P(\omega_j)$ be the a priori probabilities for the classes, and $p(x_i)$ the unconditional p.d.f.s. The following formula expresses the posterior probability of correct classification at the second level

$$\begin{aligned} P(\omega_j/x) &= \frac{\prod_{i=1}^{L} p(x_i)}{p(x)} \cdot \frac{1}{P(\omega_j)^{L-1}} \cdot \prod_{i=1}^{L} P(\omega_j/x_i) = \\ &= M(x)\frac{1}{P(\omega_j)^{L-1}} \cdot \prod_{i=1}^{L} P(\omega_j/x_i), \end{aligned} \tag{2}$$

where $M(x)$ denotes a coefficient that does not depend on the class labels and, therefore, does not affect the classification decision, given x. In terms of the decision profile, the optimal aggregation in the Bayesian sense leads to the following formula:

$$r_j(x) = m_j M(x) \prod_{i=1}^{L} r_{ji}(x). \tag{3}$$

where m_j is a parameter independent of x. The above formula answers the question for the optimal aggregation rule in case of the Bayesian first-level classifiers based on conditionally independent subsets of features. It has been observed in a simulation experimental study of Ng and Abramson (1992) that in real tasks the above formula is weaker than other aggregation rules. This can be explained by its rather restrictive underlying assumptions. In real problems, neither the classifiers are Bayesian, nor the subsets of features are completely independent.

Then, other aggregation rules may be more appropriate. This is even more true when the values in the decision profile do not correspond to probabilities but to other characteristics that may not be clear cut.

The most popular choices are the weighted average although the use of a lot of different rules is equally justifiable (or better: equally not justifiable) with respect to the classification accuracy. Among these the following ones can be mentioned (see Krishnapuram, Lee, 1992, for more details):

The Generalized Mean Operator ($\rho \in [0, 1]$)

$$r_j(x) = \alpha_j \left(\sum_{i=1}^{L} w_{ji} r_{ji}(x)^{\rho} \right)^{1/\rho}$$

The Multiplicative γ-Operator ($\gamma \in [0, 1]$)

$$r_j(x) = \alpha_j \left(\prod_{i=1}^{L} r_{ji}(x)^{w_i} \right)^{(1-\gamma)} \left(1 - \prod_{i=1}^{L} (1 - r_{ji}(x)^{w_i} \right)^{\gamma}$$

The Additive γ-Operator

$$r_j(x) = \alpha_j \left[(1-\gamma) \prod_{i=1}^{L} r_{ji}(x)^{w_i} + \gamma \left(1 - \prod_{i=1}^{L} (1 - r_{ji}(x))^{w_i} \right) \right]$$

A special family of aggregation rules are the so called *Ordered Weighted Averaging* (OWA) operators Yager (1987;1988). This definition provides an abstraction which can conveniently handle heuristic views on the combination of the estimates (e.g. in forms of linguistic quantifiers). An OWA operator is presented by a weight vector W containing L values in the interval $[0, 1]$ which are associated with the ordered estimates. That is, the first weight is to be multiplied by the maximal one among the estimates, and the last weight, by the minimal value.

Formally, an OWA operator can be expressed as

$$r_j(x) = \sum_{i=1}^{L} w_i b_i(x)$$

where $b_i(x)$ is the i-th largest element in the collection $r_{j1}(x), \ldots, r_{jL}(x)$. In this formalism, the most popular aggregation rules can be expressed by the respective weight vector W:

$$[1/L, 1/L, \ldots, 1/L]^T : \textit{ Averaged Classifier} \tag{4}$$

$$[1, 0, \ldots, 0]^T : \textit{ Optimistic Aggregation Rule} \tag{5}$$

$$[0, 0, \ldots, 1]^T : \textit{ Pessimistic Aggregation Rule} \tag{6}$$

Following this representation one can define a lot of combination rules to reflect one's hypothesis about the relationships between the estimates, e.g.

$$[0,1,0,\ldots,0]: \textit{ The Rule of the Second Highest Maximum} \tag{7}$$

$$[0,0,\ldots,1,0]: \textit{ The Rule of the Second Lowest Minimum} \tag{8}$$

$$[0,1/(L-2),1/(L-2),\ldots,1/(L-2),0]: \text{ \textit{The Rule of the Competition Jury}} \tag{9}$$

These latter rules can be thought of as expressing the belief that the maximal and minimal estimates might be outliers.

Many other rules can be formulated stating appropriate profiles on the weight vector.

4. The proposed fuzzy aggregation rule

A wide group of aggregation methods are based on the concept of a consensus, see Berenstein, Kanal, Lavine (1986), Bezdek, Spillman, Spillman (1979), Day (1988), Fedrizzi, Kacprzyk, Zadrozny (1988), Fedrizzi, Kacprzyk (1989), Ho, Hull, Srihari (1994), Kacprzyk (1986), Kacprzyk, Fedrizzi, Nurmi (1992), Kuncheva (1994;1995), Zimmermann (1987). Although consensus is traditionally meant as a full and unanimous agreement we will adopt that it is a measurable parameter whose ultimate value corresponds to unanimity. A lot of excellent arguments in favor of continuous-valued measuring of the consensus have been raised in the literature (see, e.g., Kacprzyk, Fedrizzi, Nurmi 1992). Usually, the degree of consensus is used to monitor the evolution of group preferences or to form a set of alternatives that obey certain consensus requirements. Here we propose an aggregation rule that uses explicitly the degree of consensus between the decision makers. The main idea is to strengthen the acceptance or rejection rate of an alternative if the decision makers agree in their assessments on either of these. More formally this can be expressed by:

- if the decisions agree on an aggregated value above a certain threshold $\mathbf{T}$ from the interval $(0,1)$ we could increase the strength of support (the acceptance rate);
- if the decisions agree on an aggregated value below $\mathbf{T}$ we could even more "depress" the support;
- if the decisions disagree (regardless of the fact whether the resultant value is greater or less than $\mathbf{T}$) there are no reasons to change the aggregated value in either direction.

A model of neuron whose transition function implements the above rationales is proposed in Kuncheva (1994).

Let only one class be considered at a time, and let $y = [y_1,\ldots,y_L]^T$ denote the individual degrees of support of this class. We denote by $K(A(y), C(y), \mathbf{T}) \in$

$[0,1]$ the proposed aggregation rule where $A(y)$ stands for a conventional aggregation rule, $C(y)$ is a degree of consensus between the individual decisions y_i $i = 1,\ldots,L$, and $\mathbf{T}$ is a preliminarily fixed threshold. The following formula for K is proposed:

$$K(A(y),C(y),\mathbf{T}) = \frac{1}{1+\frac{1-\mathbf{T}}{\mathbf{T}}\exp(-\alpha C(y)(A(y)-\mathbf{T}))} \tag{10}$$

where α is a constant determining the strength of the influence of the degree of consensus on the final value. Note that a classical aggregation rule participates explicitly in the formula. Its value is only modified by the degree of consensus. If $A(y)$ and $C(y)$ are symmetric with respect to their arguments, K is also symmetric. Since the derivative of K on C is positive for $A(y) > \mathbf{T}$, and negative otherwise, the formula guarantees that the greater the degree of consensus, the stronger the acceptance support (if the aggregated value is above $\mathbf{T}$) and the greater the degree of consensus, the stronger the rejection support (if the aggregated value is below $\mathbf{T}$).

The way of measuring the degree of consensus $C(y)$ is another theme that has been partially discussed in Kuncheva (1992) (without being exhaustively studied).

In order to visualize the difference between a conventional aggregation rule and the proposed formula, the following problem is considered:

1. Let $y = [y_1, y_2]^T \in [0,1]^2$ be the vector containing the degree of support for an alternative provided by two first-level decision makers. Let $y_i = 0$ denote total rejection and $y_i = 1$ – total acceptance, $i = 1, 2$.
2. Since there are a lot of ways to introduce different competence of the experts, we will confine the discussion to the case of equal competence
3. After calculating the degree of support, we have to make a crisp decision either accepting or rejecting the alternative. We wish also to supply the rule with the ability to reply "refuse to decide". This can simply be done by choosing two thresholds: $\mathbf{T}_{accept}$ and $\mathbf{T}_{reject}$ and implementing the rule:
$$\text{Final Decision} = \begin{cases} \text{accept,} & \text{if } FS > \mathbf{T}_{accept} \\ \text{reject,} & \text{if } FS < \mathbf{T}_{reject} \\ \text{refuse to decide,} & \text{otherwise.} \end{cases}$$
The notation FS is used here for the value of the final support. In the case of classical aggregation rules it will be $A(y)$. In our case FS stands for $K(A(y),C(y),\mathbf{T})$.
4. As a measure of consensus we will use the *Highest Discrepancy*, Kuncheva (1992) defined by
$$C(y) = 1 - \max_{i,j=1,\ldots,L} |y_i - y_j|. \tag{11}$$
which for the current example is:
$$C(y) = 1 - |y_1 - y_2|.$$

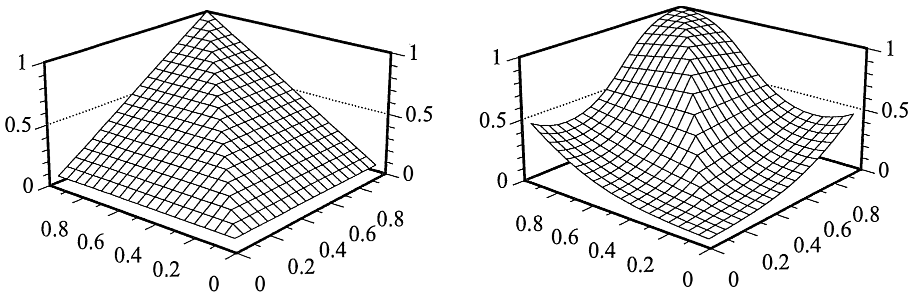

Figure 2.

Fig. 2 shows an example of decision surfaces for $A(y)$ and K versus y_1 and y_2 with $A(y) = \min\{y_1, y_2\}$ (pessimistic aggregation).

5. An illustrative example

The early detection of a hyaline membrane disease of a newborn is a problem of vital importance because the decision determines the necessary further treatment. Normally a clinical manifestation of the disease is highly obscured and a correct diagnosis is difficult to be stated. (It should be noted that our experimental study has only an illustrative meaning and is not supposed to give meaningful medical results). There is a high risk a premature newborn to be affected by this particular disease due to the immaturity of lungs. The problem is to predict whether the infant will suffer from a hyaline membrane disease or not, using some parameters measured before or immediately after delivery and others taken from the mother's history. A set of 6 parameters has been used including: birth weight, gestation age and maturity of the baby, Apgar index at two subsequent time moments after delivery, and blood pH. Due to a preliminary analysis the first and the last one have been excluded from further considerations. We will refer to the respective parameters as X_1, X_2, X_3, and X_4. The sample consists of 99 cases, 51 healthy and 48 affected children.

In order to form the first-level decisions, the linear discriminant analysis (LDA) was used on each variable. The degree of support of the hypothesis of being affected was measured by the posterior probability as estimated by the LDA program. These values are regarded as y_i-s: the individual classification decisions subject to aggregation.

The main goal of the experiment was to compare the behavior of the proposed rule with that of the respective aggregation connective embedded in the formula and, eventually, to demonstrate its advantages. For clarity it has been decided to use an equal competence of the experts since any adjustment of the weights could insert an artifact. Therefore, because of the nonoptimality of the aggregation, the absolute value of the classification accuracy provided by the aggregated decision may not be better than the best single one. This fact may hold true both for the simple aggregation rule and for its consensus modification through K and will not be further commented. Two measures of consensus have been used, see Kuncheva (1992): the *Highest Discrepancy* (introduced earlier (11)), and the *Integral Highest Discrepancy* measuring the highest deviation from the mean value:

$$C(y) = 1 - \max_{i=1,\dots,L} |y_i - \underline{y}| \tag{12}$$

where $\underline{y}$ is the mean value of $y_i - s$. For convenience the two measures will be called C_1, (12), and C_2, (11), respectively.

The following aggregation rules have been tried: (**P**) the Pessimistic Aggregation Rule (6), (**PP**) the Rule of Second Minimum (8), (**O**) the Optimistic Aggregation Rule (5), (**OO**) the Rule of Second Maximum (7), (**AV**) the Averaged Classifier (4), (**J**) the Rule of the Competition Jury (9), (**G**) and The Geometric Mean: a logarithmic aggregation rule akin to that presented by (3). (The bold notation correspond to these in the table.)

Each rule has been applied individually, and then through the proposed fuzzy consensus aggregation rule K.

The experiment included a smooth changing of the acceptance and rejection thresholds in a conjugated manner, i.e. $\mathbf{T}_{accept} = 1 - \mathbf{T}_{reject}$, so that $\mathbf{T}_{accept}$ increases from 0.5 to 1, and $\mathbf{T}_{reject}$ decreases respectively. The points for which FS (K or $A(y)$) exceeded $\mathbf{T}_{accept}$ have been assigned to the class "affected" while those below $\mathbf{T}_{reject}$, to the class "healthy". The points in between have been designated for "refuse". Usually, the classification results when the classifier has a refuse option are visualized on the plane **A** vs. **R** where **A** is the probability of correct classification of the accepted objects, and **R** is the overall probability of refuse (see Battiti, Colla, 1994). The ideal point in the plane $(\mathbf{R}, \mathbf{A})$ is that defined by coordinates $(0, 1)$. A criterion that can simultaneously account for **A** and **R** can be defined as follows:

$$U = \lambda A + (1 - \lambda)(1 - R)$$

with λ being a parameter in the interval $[0, 1]$. By changing monotonically $\mathbf{T}_{accept}$ (respectively $\mathbf{T}_{reject}$) the $\mathbf{A}(\mathbf{R})$ curve is obtained for each aggregation rule. In order to assess the curve we measured the average values of U along the curve. These values for the considered aggregation rules with several values of the parameter λ are presented in Table 1 along with those for the individual classifiers (the features). We have restricted the refuse rate **R** to vary up to 0.8 since for its higher values the assessment of the accuracy is based on too few objects and the result may be spurious.

λ	0.6	0.7	0.8	0.9
X_1	0.595	0.605	0.614	0.624
X_2	0.680	0.696	0.711	0.726
X_3	0.687	0.695	0.703	0.710
X_4	0.728	0.753	0.779	0.804
$\mathbf{P}$	0.674	0.700	0.726	0.751
$K(\mathbf{P}, C_1)$	0.715	0.712	0.708	0.705
$K(\mathbf{P}, C_2)$	0.709	0.709	0.709	0.709
$\mathbf{PP}$	0.712	0.729	0.745	0.762
$K(\mathbf{PP}, C_1)$	0.749	0.752	0.755	0.758
$K(\mathbf{PP}, C_2)$	0.738	0.745	0.752	0.759
$\mathbf{O}$	0.631	0.650	0.669	0.688
$K(\mathbf{O}, C_1)$	0.704	0.702	0.700	0.698
$K(\mathbf{O}, C_2)$	0.701	0.703	0.705	0.707
$\mathbf{OO}$	0.669	0.693	0.718	0.742
$K(\mathbf{OO}, C_1)$	0.732	0.738	0.744	0.750
$K(\mathbf{OO}, C_2)$	0.726	0.738	0.749	0.760
$\mathbf{AV}$	0.735	0.754	0.774	0.794
$K(\mathbf{AV}, C_1)$	0.763	0.771	0.780	0.788
$K(\mathbf{AV}, C_2)$	0.754	0.765	0.777	0.788
$\mathbf{J}$	0.659	0.684	0.709	0.734
$K(\mathbf{J}, C_1)$	0.730	0.724	0.718	0.713
$K(\mathbf{J}, C_2)$	0.722	0.721	0.720	0.719
$\mathbf{G}$	0.730	0.754	0.777	0.801
$K(\mathbf{G}, C_1)$	0.767	0.775	0.784	0.793
$K(\mathbf{G}, C_2)$	0.756	0.767	0.779	0.790

Table 1. Average values of U for different λ

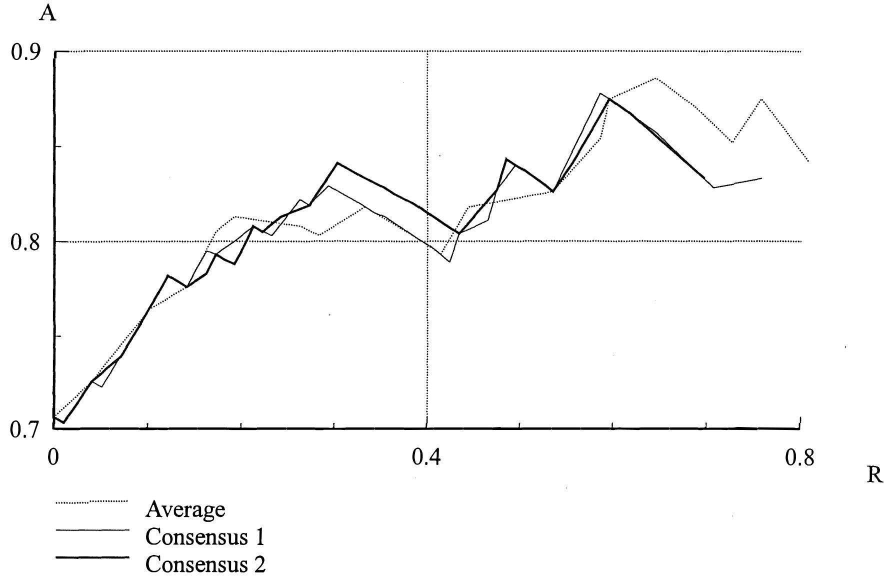

Figure 3.

It can be seen from the table that in most cases the proposed fuzzy consensus aggregation rule surpasses the embedded classical one. The two extreme cases are visualized in the plane **A** vs. **R**, in Figs. 3 and 4 respectively. The Averaged Classifier (4) appears to be the worst case according to the criterion U. It can be seen, however, that the curves defined by the rule without consensus and with consensus are quite comparable (Fig.3), even for some **R** the proposed rule is better. The best case is the Optimistic Aggregation Rule (5) where both curves obtained with consensus dominate the one without consensus (Fig.4). Note that in the above considerations, the final classification result is obtained regardless of what the type of the first-level classifiers is, provided they assess the same value in the same scale.

6. Discussion and conclusions

From the above considerations it can be concluded that the proposed aggregation formula including explicitly the degrees of consensus between the decision makers provides a versatile tool for decision support. It seems heuristically plausible since in the regions where the decision makers highly disagree, the crisp decision (either acceptance or rejection) becomes less likely to be made. On the other hand, the formula is more "generous" in cases where the decision makers

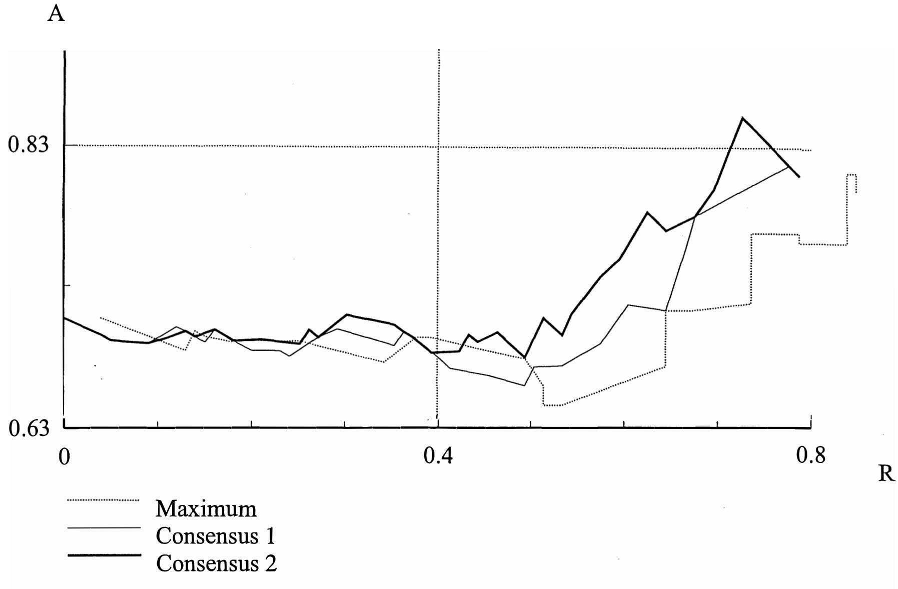

Figure 4.

agree, even though without much confidence as to the crisp decision.

The experimental illustration shows equal or better performance of the proposed formula compared to that of the respective conventional aggregation rule.

Acknowledgments

The author wishes to thank Dr. Nelly Jekova from the Institute of Obstetrics and Gynecology in Sofia for letting her use the data set, and for the helpful comments on it. The work is connected and partially supported by the contract No TH-468/94 with the National Scientific Fund, Bulgaria.

References

ALPAYDIN, E. (1993) Multiple networks for function learning. *Proc. IEEE International Conference on Neural Networks*, San Diego Calif, 9-14.

BARABASH, Y.L. (1983) *Collective Statistical Decisions in Recognition.* Moscow: Radio i Sviaz (in Russian).

BATTITI, R., COLLA, A.M. (1994) Democracy in neural nets: Voting schemes for classification. *Neural Networks*, **7**, 691-707.

BENEDIKTSSON, J.A., SVEINSSON, J.R. and ERSOY, O.K. (1993) Parallel consensual neural networks. *Proc. IEEE International Conference on Neural*

Networks, San Diego, Calif., 27-32.

BENEDIKTSSON, J.A., SWAIN, P.H. (1992) Consensus theoretic classification methods. *IEEE Transactions on Systems, Man, and Cybernetics*, **22**, 688-704.

BERENSTEIN, C., KANAL, L.N., LAVINE, D. (1986) Consensus rules. in: Kanal L.N. and J.F. Lemmer (eds.) *Uncertainty in Artificial Intelligence*, Elsevier Science Publishers, 27-32.

BEZDEK, J.C., SPILLMAN, B., SPILLMAN, R. (1979) Fuzzy relation spaces for group decision theory. *Fuzzy Sets and Systems*, **2**, 5-14.

BORDLEY, R.F. (1982) A multiplicative formula for aggregating probability assessments. *Management Science*, **28**, 1137-1148.

CHIANG, C.-C., FU, H.-C. (1994) A divide-and-conquer methodology for modular supervised neural network design. *Proc. IEEE International Conference on Neural Networks, Orlando, Florida*, 119-124.

DASARATHY, B., SHEELA, B.V. (1979) A composite classifier system design: Concepts and methodology. *Proc. of the IEEE*, **67**, 108-113.

DAY, W.H.E. (1988) Consensus methods as tools for data analysis, in: Bock H.H. (Ed.) *Classification and Related Methods for Data Analysis*, Elsevier Science Publishers, 317-324.

FEDRIZZI, M., KACPRZYK, J., ZADROZNY, S. (1988) An interactive multi--user decision support system for consensus reaching process using fuzzy logic with linguistic quantifiers. *Decision Support Systems*, **4**, 313-327.

FEDRIZZI, M., KACPRZYK, J. (1989) On the concept of fuzzy majority in consensus formation. *Technical report No ZPZC 220-22/89*, Systems Research Institute, Polish Academy of Sciences.

FILIPPY, E., COSTA, M., PASERO, E. (1994) Multi-layer perceptron ensembles for increased performance and fault-tolerance in pattern recognition tasks. *Proc. IEEE International Conference on Neural Networks, Orlando, Florida*, 2901-2906.

HASHEM, S., SCHMEISER, B., YIH, Y. (1994) Optimal linear combinations of neural networks: An overview. *Proc. IEEE Int. Conf. on Neural Networks, Orlando, Florida*, 1507-1512.

HO, T.K., HULL, J.J., SRIHARI, S.N. (1994) Decision combination in multiple classifier systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **16**, 66-75.

HUANG, Y.S., SUEN, C.Y. (1994) A method of combining multiple classifiers – a neural network approach. *Proc. 12th International Conference on Pattern Recognition*, Jerusalem, Israel, 473-475.

JORDAN, M.I., JACOBS, R.A. (1992) Hierarchies of adaptive experts. *Proc. Int. Conf. on Neural Networks, San Diego, Calif.*, 192-196.

JOZEFCZYK, J. (1986) Determination of optimal recognition algorithms in the two-level system. *Pattern Recognition Letters*, **4**, 413-420.

KACPRZYK, J., FEDRIZZI, M. (1988) A 'soft' measure of consensus in the setting of partial fuzzy preferences. *European Journal of Operational Re-*

search, **34**, 316-325.

KACPRZYK, J. (1986) Group decision making with a fuzzy linguistic majority. *Fuzzy Sets and Systems*, **18**, 105-118.

KACPRZYK, J., FEDRIZZI, M., NURMI, H. (1992) Group decision making and consensus under fuzzy preferences and fuzzy majority. *Fuzzy Sets and Systems*, **49**, 21-31.

KRISHNAPURAM, R., LEE, J. (1992) Fuzzy-connective-based hierarchical aggregation networks for decision making. *Fuzzy Sets and Systems*, **46**, 11-27.

KUNCHEVA, L.I. (1992) Additional criteria in fuzzy two-level pattern recognition. *Fuzzy Sets and Systems*, **47**, 281-289.

KUNCHEVA, L.I. (1993) "Change-glasses" approach in pattern recognition. *Pattern Recognition Letters*, **14**, 619-623.

KUNCHEVA, L.I. (1994) Pattern recognition with a model of fuzzy neuron using degree of consensus. *Fuzzy Sets and Systems*, 66, 241-250.

KUNCHEVA, L.I. (1995) Using degree of consensus in two-level fuzzy pattern recognition. *European Journal of Operational Research*, **80**, 365-370.

KUNCHEVA, L.I., MITRA, S. (1992) A two-level classification scheme trained by a fuzzy neural network. *12 International Conference on Pattern Recognition*, Jerusalem, Israel, 467-469.

LAM, L., SUEN, C.Y. (1994) A theoretical analysis of the application of majority voting to pattern recognition *Proc. 12th Int. Conf. on Pattern Recognition*, Jerusalem, Israel, 418-420.

MITRA, S., KUNCHEVA, L. (1994) Change-glasses pattern classification with a fuzzy neural network. *Proc. I Workshop on Fuzzy-Based Expert Systems, FUBEST'94, Sofia, Bulgaria, 28-30 September*, 29-32.

NADAL, C., LEGAULT, R., SUEN, C.Y. (1990) Complementary algorithms for the recognition of totally unconstrained handwritten numerals. *Proc. 10th International Conference on Pattern Recognition, Atlantic City, New Jersey*, 443-449.

NG, K.-C. and ABRAMSON, B. (1992) Consensus diagnosis: A simulation study. *IEEE Transactions on Systems, Man, and Cybernetics*, **22**, 916-928.

PODDAR, P. and RAO, P.V.S. (1993) Hierarchical ensemblance of neural networks. *Proc. Int. Conf. on Neural Networks, San Diego, Calif.*, 287-292.

RAMAKRISHNAN, R., RAO, C.J.M. (1992) The fuzzy weighted additive rule. *Fuzzy Sets and Systems*, **46**, 177-187.

RASTRIGIN, L.A., ERENSTEIN, R.H. (1981) *Method of Collective Recognition.* Moscow: Energoizdat (in Russian).

SUNG-BAE, C., KIM, J.H. (1992) Two design strategies of neural network for complex classification problems. *Proc. II International Conference on Fuzzy Logic and Neural Networks*, Iizuka, Japan, 759-762.

TUBBS, J.D., ALLTOP, W.O. (1991) Measures of confidence associated with combining classification rules. *IEEE Transactions on Systems, Man, and Cybernetics*, **21**, 690-692.

WERNECKE, K.-D. (1992) A coupling procedure for the discrimination of mixed data. *Biometrics*, **48**, 497-506.

XU, L., KRZYZAK, A. and SUEN, C.Y. (1992) Methods of combining multiple classifiers and their application to handwriting recognition. *IEEE Transactions on Systems, Man, and Cybernetics*, **22**, 418-435.

YAGER, R.R. (1987) On the aggregation of processing units in neural networks. *Proc. IEEE 1-st International Conference on Neural Nets, San Diego, Calif.*, II-327-II-333.

YAGER, R.R. (1988) On Ordered Weighted Averaging aggregation operators in multicriteria decision making. *IEEE Transactions on Systems, Man, and Cybernetics*, **18**, 183-193.

ZIMMERMANN, H.-J. (1987) *Fuzzy Sets, Decision Making and Expert Systems*, Kluwer Academic Publishers, Boston.

ZIMMERMANN, H.-J., ZYSNO, P. (1983) Decisions and evaluations by hierarchical aggregation of information. *Fuzzy Sets and Systems*, **10**, 243-260.