Published in: *Fuzzy Sets and Systems*,
122, (3), 2001, 401-407.

Using measures of similarity and inclusion for multiple classifier fusion by decision templates

Ludmila I. Kuncheva
School of Mathematics, University of Wales, Bangor
Bangor, Gwynedd, LL57 1UT, United Kingdom
e-mail: l.i.kuncheva@bangor.ac.uk

Abstract

Decision templates (DT) are a technique for classifier fusion for continuous-valued individual classifier outputs. The individual outputs considered here sum up to the same value (e.g., statistical classifiers, yielding some estimates of the posterior probabilities for the classes). First, the DT fusion algorithm is explained. Second, we show that two similarity measures (S_1 and S_2) and two inclusion indices (I_1 and I_2) between fuzzy sets (see Dubois and Prade, 1980) produce the same DT classifier. The equivalence is proven by showing that for every object submitted for classification, all 4 measures induce the same ordering on the set of class labels (through DT fusion), thereby assigning the object to the same class.

Keywords: Pattern recognition, multiple classifier fusion, aggregation, decision templates, measures of similarity and inclusion

1 Introduction

The objective of combination of a set of classifiers is to achieve a higher accuracy than the accuracy of the best individual in the set. There are generally two types of combination: classifier selection and classifier fusion (see [16]). The presumption in classifier *selection* is that each classifier is "an expert" in some local area of the feature space. For a feature vector $\mathbf{x} \in \Re^p$ submitted for classification, the classifier responsible for the vicinity of $\mathbf{x}$ should be given the highest credit for assigning the class label to $\mathbf{x}$. We can nominate either only one classifier to make the decision, as in [13], or more than one "local experts", as in [1, 7].

Classifier *fusion* assumes that *all* classifiers are trained over the whole feature space, and are thereby considered as competitive rather than complementary [12, 17]. Several fuzzy techniques have been proposed for classifier fusion, the most popular being the fuzzy integral [3, 5, 8, 15]. *Decision templates* (DT) [10, 11] is an intuitive and simple classifier fusion scheme. A series of experiments has been carried out comparing various DTs and 15 other classifier fusion methods [2, 9]. The results were generally in favor of DTs, and the best of all DTs were those based on the four measures studied here. DTs did not outperform all the rival methods in all experiments but in the long run rated better than methods such as Behavior-Knowledge Space BKS [6], which tended to overtrain; fusion by statistical classifiers, which was not always feasible; and assumption-based schemes such as Naive Bayes [17] whose assumptions do not always hold. Dempster Shafer aggregation implemented as suggested by Rogova [14] showed similar performance to the DTs but with much higher computational complexity. Table 1 is an excerpt of the results presented in [9]. Two data sets were used: the Satimage data (36 features (4 used in the experiments), 6 classes, 6435 objects) and the Phoneme data (5 features, 2 classes, 5404 objects) from ELENA database (anonymous ftp at ftp.dice.ucl.ac.be, directory pub/neural-nets/ELENA/databases). The table entries are the estimated testing classification accuracy (in %), averaged over 10 experiments with each data set. For each experiment, 2000 objects were taken at random

Table 1: Test classification accuracy with three classifier fusion methods (the average from 10 experiments, 2000 training objects taken at random as the training set in each experiment)

	Satimage data	Phoneme data
Best individual	80.62	75.17
Majority Vote	82.23	76.08
Average class support	83.88	75.91
DTs	84.88	77.45

as the training set, and the remaining part of the data was used for testing. Six individual classifiers (quadratic discriminant classifiers based on each combination of 2 features) were used with the Satimage data and 10 individual classifiers were used with the Phoneme data. Three fusion methods are displayed in the table: Majority vote, Average of the class support, and DTs (based on S_1). Each of the fusion methods improved on the best individual classifier (the accuracy is also shown in Table 1). In this example the DTs provide the highest classification accuracy. A result that appeared during the experiments was that the DT fusion using the four measures S_1, S_2, I_1 and I_2 produced the same classification accuracy. The connection was not straightforward and this motivated the current study.

Decision templates work by comparing a fuzzy set obtained from the individual classifier outputs for a given $\mathbf{x}$ with a template for each class using a measure of *similarity* (in a broad sense) between the two fuzzy sets. This paper proves the equivalence between four such measures: S_1, S_2, I_1, and I_2 (see [4]), for individual classifiers whose outputs sum up to the same (fixed) value. This type of classifiers is most often used in practice. For example, widely used statistical classifiers often produce some estimate of the posterior probabilities of the classes.

Section 2 explains the fuzzy template classifier fusion, Section 3 contains the proof of the equivalence of the four measures, and Section 4, the conclusions.

2 Decision templates for classifier fusion

Let $\mathbf{x} \in \Re^p$ be a feature vector and $\{1, 2, \ldots, c\}$ be the label set of c classes. Every mapping

$$D : \Re^p \to \{1, 2, \ldots, c\} \tag{1}$$

is called a *classifier*. A *fuzzy classifier* is the mapping

$$\tilde{D} : \Re^p \to [0, 1]^c. \tag{2}$$

with output $\mu_{\tilde{D}}(\mathbf{x}) = [\mu^1_{\tilde{D}}(\mathbf{x}), \ldots, \mu^c_{\tilde{D}}(\mathbf{x})]^T$. The components $\mu^i_{\tilde{D}}(\mathbf{x})$ can be regarded as "support" given by classifier $\tilde{D}$ for the hypothesis that $\mathbf{x}$ comes from class i. If $\tilde{D}$ is a statistical classifier, the degree of membership $\mu^i_{\tilde{D}}(\mathbf{x})$ is most often some estimate of the posterior probability $P(i|\mathbf{x})$. Besides the probabilistic interpretation, this degree can be viewed as typicalness, belief, certainty, possibility, etc., not necessarily coming from a statistical classifier. The decision of $\tilde{D}$ can be "hardened" so that a crisp class label $D(\mathbf{x}) \in \{1, 2, \ldots, c\}$ is assigned to $\mathbf{x}$, usually by the *maximum membership rule*:

$$D(\mathbf{x}) = k \quad \Longleftrightarrow \quad \mu^k_{\tilde{D}}(\mathbf{x}) = \max_{i=1,\ldots,c} \mu^i_{\tilde{D}}(\mathbf{x}) \tag{3}$$

Let $C = \{C_1, \ldots, C_L\}$ be the set of L individual classifiers. We denote by $C_i(\mathbf{x})$ the output of the ith classifier: $C_i(\mathbf{x}) = [d_{i,1}(\mathbf{x}), \ldots, d_{i,c}(\mathbf{x})]^T$. Typically, $d_{i,j}(\mathbf{x}) \in [0, 1]$. The individual classifier decisions are then aggregated (classifier fusion) to design a fuzzy classifier $\tilde{D}$

$$\tilde{D}(\mathbf{x}) = \mathcal{F}(C_1(\mathbf{x}), \ldots, C_L(\mathbf{x})), \tag{4}$$

where $\mathcal{F}$ is called *aggregation rule*. The class label for $\mathbf{x}$ is found by the maximum membership rule (3).

Definition 1. *The decision profile of a combination of classifiers, given* $\mathbf{x} \in \Re^p$*, is the matrix*

$$DP(\mathbf{x}) = \begin{bmatrix} d_{1,1}(\mathbf{x}) & \ldots & d_{1,j}(\mathbf{x}) & \ldots & d_{1,c}(\mathbf{x}) \\ \ldots & & & & \\ d_{i,1}(\mathbf{x}) & \ldots & d_{i,j}(\mathbf{x}) & \ldots & d_{i,c}(\mathbf{x}) \\ \ldots & & & & \\ d_{L,1}(\mathbf{x}) & \ldots & d_{L,j}(\mathbf{x}) & \ldots & d_{L,c}(\mathbf{x}) \end{bmatrix}. \tag{5}$$

Output of classifier $C_i(\mathbf{x})$

Let $Z = \{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$, $\mathbf{z}_j \in \Re^p$ be the (labeled) training data set.

Definition 2. *The decision templates of class* i *is the* $L \times c$ *matrix* $F_i = \{f_i(k,s)\}$ *whose elements are obtained by*

$$f_i(k,s) = \frac{\sum_{j=1}^{N} \; Ind(\mathbf{z}_j, i) \; d_{k,s}(\mathbf{z}_j)}{\sum_{j=1}^{N} \; Ind(\mathbf{z}_j, i)}, \tag{6}$$

where $Ind(\mathbf{z}_j, i)$ is an indicator function with value 1 if $\mathbf{z}_j$ has label i, and 0, otherwise [10, 11].

Thus, the fuzzy template for class i is the *average of the decision profiles* of the elements of the training set Z labeled in class i. When $\mathbf{x} \in \Re^p$ is submitted for classification, the DT scheme produces the soft class label:

$$\mu^i_{\tilde{D}}(\mathbf{x}) = \mathcal{S}(DP(\mathbf{x}), F_i)), \quad i = 1, \ldots, c, \tag{7}$$

where $\mathcal{S}$ is interpreted as *similarity*. The higher the similarity between the decision profile of the current $\mathbf{x}$ ($DP(\mathbf{x})$) and the fuzzy template for class i (F_i) is, the higher the support for that class ($\mu^i_{\tilde{D}}(\mathbf{x})$). Figure 1 illustrates how the DT scheme operates. The decision templates are calculated in advance using Z as in equation (6).

Regarding the two arguments of $\mathcal{S}$ as fuzzy sets on some universal set with $L \cdot c$ elements, various fuzzy measures of similarity can be used. Let A and B be fuzzy sets on $U = \{u_1, \ldots, u_n\}$. Here we consider the following two measures of similarity [4]

$$\mathcal{S}(A,B) \equiv S_1(A,B) = \frac{\parallel A \cap B \parallel}{\parallel A \cup B \parallel}, \tag{8}$$

where $\parallel \zeta \parallel$ is the relative cardinality of the fuzzy set ζ on U

$$\parallel \zeta \parallel = \frac{1}{n} \sum_{i=1}^{n} \mu_\zeta(u_i), \tag{9}$$

and

$$\mathcal{S}(A,B) \equiv S_2(A,B) = 1 - \parallel A \nabla B \parallel . \tag{10}$$

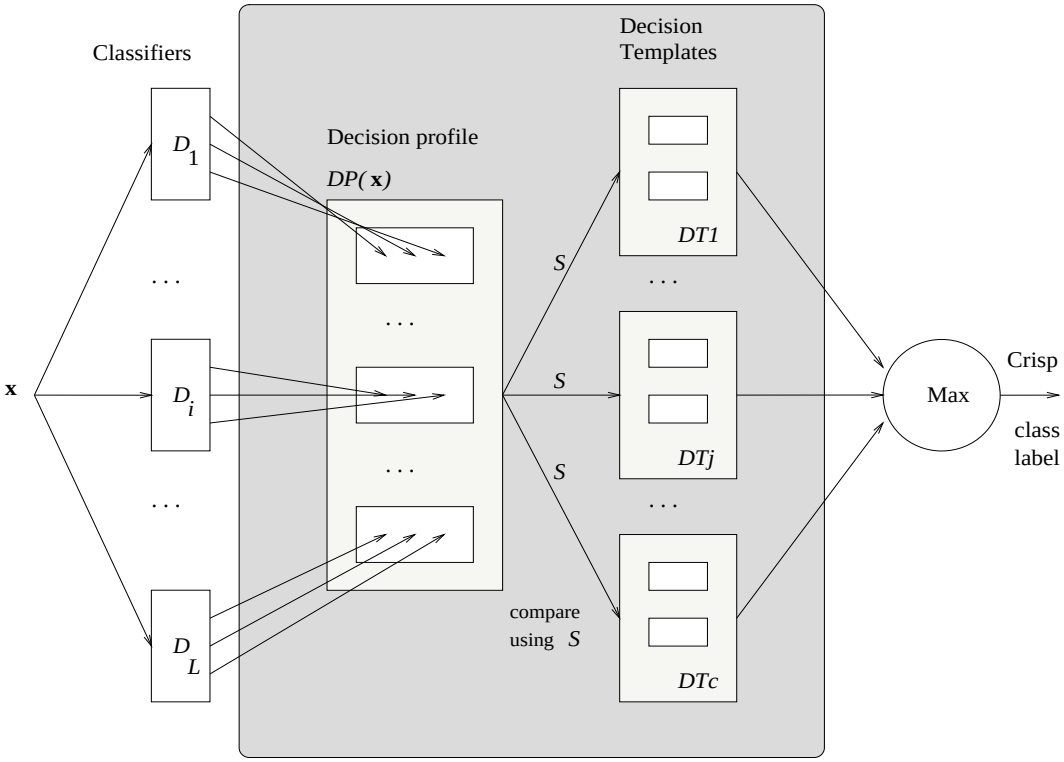

Figure 1: Operation of the decision templates (DT) classifier fusion scheme

where $A \nabla B$ is the symmetric difference defined by the Hamming distance:

$$\mu_{A \nabla B}(u) = |\mu_A(u) - \mu_B(u)|. \tag{11}$$

We also consider the following two indices of inclusion of A in B

$$\mathcal{S}(A, B) \equiv I_1(A, B) = \frac{\parallel A \cap B \parallel}{\parallel A \parallel}, \tag{12}$$

and

$$\mathcal{S}(A, B) \equiv I_2(A, B) = 1 - \parallel A| - |B \parallel \tag{13}$$

where $|-|$ is the bounded difference

$$\mu_{A|-|B}(u) = \max\{0, \mu_A(u) - \mu_B(u)\} \tag{14}$$

For intersection and union we use minimum and maximum, respectively. These four measures were of a special interest because they led to the most accurate (hardened) $\tilde{D}$ compared to DT using S_3, S_4, S_5, I_3, I_4, and the consistency measure C (see [4]), and also compared to another 15 fusion techniques in our preliminary experimental study [9]. These 4 measures were superior to the other 6, probably because they are *integral* rather than *point-wise*.

3 The equivalence of S_1, S_2, I_1, and I_2

We consider the (most common) case where the output of each classifier $C_i, i = 1 \ldots, L$ satisfy

$$\sum_{j=1}^{c} d_{i,j}(\mathbf{x}) = T, \;\; \forall \mathbf{x} \in \Re^p, \;\; T > 0. \tag{15}$$

When C_is are statistical classifiers, $d_{i,j}(\mathbf{x})$ are interpreted as posterior probabilities and therefore $T = 1$. We can regard the decision profile $DP(\mathbf{x})$ and the decision templates $F_j, j = 1, \ldots, c$ as fuzzy sets on a certain universal set M with $L \cdot c$ elements.

Two similarity measures S_q and S_r will be considered equivalent if the DT classifier fusion using any of them produces the same *crisp* class labeling. This means that for every $\mathbf{x} \in \Re^p$ submitted for classification, the maximal component of $\tilde{D}(\mathbf{x})$ using S_q has the same index (class label) as the maximal component of $\tilde{D}(\mathbf{x})$ using S_r. A sufficient (but not necessary) condition for this is that the two measures induce the same ordering of the class labels for $\mathbf{x}$, i.e.,

$$S_q(DP(\mathbf{x}), F_{j_1}) > S_q(DP(\mathbf{x}), F_{j_2}) \iff S_r(DP(\mathbf{x}), F_{j_1}) > S_r(DP(\mathbf{x}), F_{j_2}) \tag{16}$$

To simplify the notations we consider A, B and C as fuzzy sets on $U = \{u_1, \ldots u_n\}$ to denote respectively $DP(\mathbf{x}), F_{j1}$ and F_{j2} as fuzzy sets on the set of individual classifier outputs.

Proposition 1. *Let A, B and C be fuzzy sets on U such that $\| A \| = \| B \| = \| C \| = \frac{t}{n}, t > 0$. Then*

$$S_1(A, B) > S_1(A, C) \iff S_2(A, B) > S_2(A, C) \iff I_1(A, B) > I_1(A, C) \iff I_2(A, B) > I_2(A, C).$$

Proof 1. The proof will show that all 4 inequalities reduce to the same inequality.

Proof 1a. (S_1)

$$\| A \cap B \| = \frac{1}{n} \sum_{i=1}^{n} \min(\mu_A(u_i), \mu_B(u_i)) \tag{17}$$

Let I_{AB} be a set of indices (a subset of $\{1, 2, \ldots, n\}$) such that

$$I_{AB} = \{k \mid \mu_A(u_k) > \mu_B(u_k)\} \tag{18}$$

Then

$$\| A \cap B \| = \frac{1}{n} \left(\sum_{k \in I_{AB}} \mu_B(u_k) + \sum_{k \notin I_{AB}} \mu_A(u_k) \right) =$$

$$\frac{1}{n} \left(\sum_{k \in I_{AB}} \mu_B(u_k) + t - \sum_{k \in I_{AB}} \mu_A(u_k) \right) = \frac{1}{n} \left(t - \sum_{k \in I_{AB}} (\mu_A(u_k) - \mu_B(u_k)) \right). \tag{19}$$

For the denominator of S_1

$$\| A \cup B \| = \frac{1}{n} \sum_{i=1}^{n} \max(\mu_A(u_i), \mu_B(u_i)) =$$

$$\frac{1}{n} \left(\sum_{k \in I_{AB}} \mu_A(u_k) + \sum_{k \notin I_{AB}} \mu_B(u_k) \right) =$$

$$\frac{1}{n}\left(\sum_{k\in I_{AB}}\mu_A(u_k)+t-\sum_{k\in I_{AB}}\mu_B(u_k)\right)=\frac{1}{n}\left(t+\sum_{k\in I_{AB}}(\mu_A(u_k)-\mu_B(u_k))\right). \tag{20}$$

Let I_{AC} denote the same index set as in equation (18) but with respect to sets A and C. Similarly to equations (19) for $\| A\cap B\|$ and (20) for $\| A\cup B\|$, we get for $S_1(A,C)$

$$\| A\cap C\|=\frac{1}{n}\left(t-\sum_{k\in I_{AC}}(\mu_A(u_k)-\mu_C(u_k))\right). \tag{21}$$

and

$$\| A\cup C\|=\frac{1}{n}\left(t+\sum_{k\in I_{AC}}(\mu_A(u_k)-\mu_C(u_k))\right). \tag{22}$$

We introduce the following notations

$$\alpha_{AB}=\sum_{k\in I_{AB}}(\mu_A(u_k)-\mu_B(u_k)) \tag{23}$$

and

$$\alpha_{AC}=\sum_{k\in I_{AC}}(\mu_A(u_k)-\mu_C(u_k)) \tag{24}$$

Then from the definition of S_1 (8) and the inequality in the proposition it follows that

$$\frac{\| A\cap B\|}{\| A\cup B\|}>\frac{\| A\cap C\|}{\| A\cup C\|} \tag{25}$$

which gives

$$\| A\cap B\|\| A\cup C\|>\| A\cap C\|\| A\cup B\| \tag{26}$$

and

$$(t-\alpha_{AB})(t+\alpha_{AC})>(t+\alpha_{AB})(t-\alpha_{AC}) \tag{27}$$

which reduces to

$$\alpha_{AC}>\alpha_{AB}. \tag{28}$$

Proof 1b. (S_2)

First we consider the symmetric difference $A\nabla B$ (11)

$$\| A\nabla B\|=\frac{1}{n}\sum_{i=1}^{n}|\mu_A(u_i)-\mu_B(u_i)|=$$

$$=\frac{1}{n}\left(\sum_{k\in I_{AB}}(\mu_A(u_k)-\mu_B(u_k))+\sum_{k\notin I_{AB}}(\mu_B(u_k)-\mu_A(u_k))\right)=$$

$$=\frac{1}{n}\left(\sum_{k\in I_{AB}}(\mu_A(u_k)-\mu_B(u_k))+t-\sum_{k\in I_{AB}}(\mu_B(u_k)-\mu_A(u_k))\right)=$$

$$=\frac{1}{n}\left(t+2\sum_{k\in I_{AB}}(\mu_A(u_k)-\mu_B(u_k))\right)=\frac{1}{n}(t+2\alpha_{AB}). \tag{29}$$

The inequality for S_2 in the proposition can be rewritten as

$$1- \parallel A \nabla B \parallel > 1- \parallel A \nabla C \parallel, \tag{30}$$

or equivalently

$$1 - (t + 2\alpha_{AB}) > 1 - (t + 2\alpha_{AC}) \tag{31}$$

and

$$\alpha_{AC} > \alpha_{AB}. \tag{32}$$

Proof 1c. (I_1)
From the definition of I_1 (12) and the inequality in the proposition

$$\frac{\parallel A \cap B \parallel}{\parallel A \parallel} > \frac{\parallel A \cap C \parallel}{\parallel A \parallel} \tag{33}$$

and using the notations α_{AB} and α_{AC}

$$(t - \alpha_{AB}) > (t - \alpha_{AC}) \tag{34}$$

which reduces to

$$\alpha_{AC} > \alpha_{AB}. \tag{35}$$

Proof 1d. (I_2)
For the bounded difference $A| - |B$ (14)

$$\parallel A| - |B \parallel = \frac{1}{n} \sum_{i=1}^{n} \max\{0, \mu_A(u) - \mu_B(u)\} = \sum_{k \in I_{AB}} (\mu_A(u) - \mu_B(u)) = \alpha_{AB}. \tag{36}$$

Then from the inequality in the proposition

$$1- \parallel A| - |B \parallel > 1- \parallel A| - |C \parallel \tag{37}$$

it follows that

$$1 - \alpha_{AB} > 1 - \alpha_{AC}, \tag{38}$$

and

$$\alpha_{AC} > \alpha_{AB}, \tag{39}$$

which completes the proof. ∎

4 Conclusions

This paper considers classifier fusion using decision templates (DT). In the first part the DT fusion technique is explained. Two similarity measures (S_1 and S_2) and two inclusion indices (I_1 and I_2) were shown to be equivalent for the fuzzy template fusion if the individual classifier decisions sum up to the same (fixed) value. The proposition is based on the sufficient condition that if two similarity measures induce the same ordering on the set of class labels, the fusion will point to the same class label for both measures. In the proof, each of the four inequalities is taken separately and shown to reduce to the same inequality for all 4 measures.

Considering computational complexity, all four measures take linear time with respect to the number of elements of the universal set. In classifier fusion it is unlikely to have large number of individual classifiers.

Usually 5-10 classifiers appear to be accurate and different enough to form a useful group. The number of classes is typically not large either: from 2 to, e.g., 26 (in character recognition). This makes a universal set with up to 260 elements, and the computational complexity of the four measures is practically the same. The message is that when using DTs with statistical classifiers, only one of the 4 measures (any!) need be calculated.

References

[1] E. Alpaydin and M. I. Jordan. Local linear perceptrons for classification. *IEEE Transactions on Neural Networks*, 7(3):788–792, 1996.

[2] J.C. Bezdek, J.M Keller, R. Krishnapuram, and N.R. Pal. *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing*. Kluwer Academic Publishers, 1999.

[3] S.-B. Cho and J.H. Kim. Combining multiple neural networks by fuzzy integral and robust classification. *IEEE Transactions on Systems, Man, and Cybernetics*, 25:380–384, 1995.

[4] D. Dubois and H. Prade. *Fuzzy Sets and Systems: Theory and Applications*. Academic Press, NY, 1980.

[5] M. Grabisch and J.-M. Nicolas. Classification by fuzzy integral. *Fuzzy Sets and Systems*, 65:255–271, 1994.

[6] Y.S. Huang and C.Y. Suen. A method of combining multiple experts for the recognition of unconstrained handwritten numerals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17:90–93, 1995.

[7] R.A. Jacobs, M.I. Jordan, S.J. Nowlan, and G.E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3:79–87, 1991.

[8] J.M. Keller, P. Gader, H. Tahani, J.-H. Chiang, and M. Mohamed. Advances in fuzzy integration for pattern recognition. *Fuzzy Sets and Systems*, 65:273–283, 1994.

[9] L.I. Kuncheva, J.C. Bezdek, and R.P.W. Duin. Decision templates for multiple classifier fusion: an experimental comparison. *Pattern Recognition*, 34(2):299–314, 2001.

[10] L.I. Kuncheva, J.C. Bezdek, and M.A. Sutton. On combining multiple classifiers by fuzzy templates. In *Proc. NAFIPS'98*, pages 193–197, Pensacola, FL, 1998.

[11] L.I. Kuncheva, R.K. Kounchev, and R.Z. Zlatev. Aggregation of multiple classification decisions by fuzzy templates. In *Third European Congress on Intelligent Technologies and Soft Computing EUFIT'95*, pages 1470–1474, Aachen, Germany, August 1995.

[12] K.-C. Ng and B. Abramson. Consensus diagnosis: A simulation study. *IEEE Transactions on Systems, Man, and Cybernetics*, 22:916–928, 1992.

[13] L.A. Rastrigin and R.H. Erenstein. *Method of Collective Recognition*. Energoizdat, Moscow, 1981. (In Russian).

[14] G. Rogova. Combining the results of several neural network classifiers. *Neural Networks*, 7:777–781, 1994.

[15] D. Wang, J. M. Keller, C.A. Carson, K.K. McAdoo-Edwards, and C.W. Bailey. Use of fuzzy-logic-inspired features to improve bacterial recognition through classifier fusion. *IEEE Transactions on Systems, Man, and Cybernetics*, 28B(4):583–591, 1998.

[16] K. Woods, W.P. Kegelmeyer, and K. Bowyer. Combination of multiple classifiers using local accuracy estimates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:405–410, 1997.

[17] L. Xu, A. Krzyzak, and C.Y. Suen. Methods of combining multiple classifiers and their application to handwriting recognition. *IEEE Transactions on Systems, Man, and Cybernetics*, 22:418–435, 1992.