Minsk, Belarus, October 2-4, 2001

CLASSIFIERS FUSION WITH DATA DEPENDENT AGGREGATION SCHEMES

A. Lipnickas

Dept. of Applied Electronics, Kaunas University of Technology, Studentų 50, 3031, Kaunas, LITHUANIA, lipnick@soften.ktu.lt

ABSTRACT

In this paper we studied two different classifiers fusion algorithms exploiting the combination weights expressed over the entire data space and the combination with data dependent weights. The following aggregation schemes are employed in the study: the majority vote, the averaging, the combination via Choquet integral with the λ fuzzy measure, the combination via space partitioning and classifier selection approach, and the combination via Choquet integral with the data dependent λ - fuzzy measure.

1. INTRODUCTION

It is well known that a combination of many different neural networks can improve classification accuracy. The concept of combining proposed as early as 1965 [1] and has been studied by many authors. A variety of schemes have been proposed for combining multiple classifiers. The approaches used most often include the classifiers fusion algorithms with the combination weights expressed over the entire data space and data dependent combination weights. In classifiers fusion schemes, classifiers outputs are combined to achieve a "group decision". The most often used classifiers fusion schemes with the combination weights expressed over the entire data space are: the majority vote [2-4]; the probability schemes [5]; the weighted averaging [4, 6-9]; the Borda count [10]; the Bayes approach [2, 3], fuzzy connectives [11], combination through order statistics [12], and combination by a neural network [13]. and the fuzzy integral [14-16]. Combination with data dependent weights attempts to predict which group of classifiers is most likely to be correct for a given sample [4, 7, 17- 20]. The use of data dependent weights, when properly estimated, provides higher classification accuracy [18].

Numerous previous works on neural networks committees have shown that an efficient committee should consist of networks that are not only very accurate but also diverse in the sense that the networks make their independent errors in different regions of the input space [21, 22]. For a instance, the combination of two neural networks that agree everywhere cannot lead to any accuracy improvement, no matter how ingenious a combination method is employed.

It has been recently shown that the half&half bagging through the majority voting rule is capable of creating very accurate committees of decision trees [23]. Data sampling by half&half bagging focuses on most often miss-classified data points from training data set. We used this sampling technique when training members of the committee.

From the previous investigation in classifiers fusion [4, 8, 9] we found, that the Choquet integral based combination method with the λ-fuzzy measure is competitive with other schemes exploiting more sophisticated fuzzy measures.

The main issue investigated in this paper is the feature space partitioning scheme aiming to create neural network committees specific for each region of the partitioned space.

The paper is organised as follows. In the second section, the databases used are described. The idea of half&half sampling is briefly described in the third section. Section four presents background on data space partitioning and classifier selection approach. The λ-fuzzy measure and fuzzy integral are described in the fifth section. Section six presents the combination schemes involved. The experimental procedures and the results of the experiments are described in section seven. Finally, section eight presents conclusions of the work.

2. DATA

The *ESPRIT* Basic Research Project Number 6891 *(ELENA)* provides databases and technical reports designed for testing both conventional and neural classifiers. All the databases and technical reports are available via anonymous ftp: *ftp.dice.ucl.ac.be* in the directory *pub/neura-netl/ELENA/databases*. From the *ELENA* project we have chosen one data set representing artificial data (*Clouds*), and two sets representing real applications, *Phoneme* and *Satimage*.

The data sets used are summarised in tabl. 1. The errors presented in the tabl. 1 are taken from the *ELENA* project. The errors presented are the

average errors obtained in the *ELENA* project when using MLP with two hidden layers of 20 and 10 units, respectively.

Table 1. Summary of the data sets used.

Data Set	Clouds	Phoneme	Satimage
# classes	2	2	6
# features	2	5	5
# samples	5000	5404	6435
Error %	12.3	16.4	11.9
Bayes %	9.66	---	---

3. HALF&HALF SAMPLING

It has been demonstrated that the AdaBoost algorithm [24] generates committees of low generalisation error [25]. The AdaBoost is a complex algorithm. Breiman has recently proposed a very simple the so-called half&half bagging approach [23]. When tested on decision trees the approach was competitive with the AdaBoost algorithm.

The basic idea of the half&half sampling is very simple. It is assumed that the training set contains N data points. Suppose that k classifiers have been already constructed. To obtain the next training set, randomly select a data point x. Present x to that subset of k classifiers, which did not use x in their training sets. Use the majority vote to predict the classification result of x by the subset of classifiers. If x is misclassified, put it in set MC. If not, put x in set CC. Stop when the sizes of both MC and CC are equal to M, where $2M \le N$. In [23], $M=N/4$ has been used.

4. SPACE PARTITIONING AND CLASSIFIER SELECTION

Let $Z=\{z_1,z_2,...,z_L\}$ is a set of classifiers and $\Omega=\{q_1,q_2,...,q\}$ is a set of class labels. Each classifier assigns an input vector $x \in \Re^n$ to a class from Ω, i.e., Z_k: $\Re^n \rightarrow \Omega$.

4.1. The Most Often Used Space Partitioning

The basic idea of space partitioning (*SP*) is very simple, we just have to divide the entire feature space $\Re^n$ into $K>1$ regions with the reference point v_i representing the i th region. The reference points are found by performing *c-means* or *frequency-sensitive competitive learning* clustering technique. The regions obtained are then denoited: $\Re_1$, $\Re_2$,..., $\Re_K$ [4,20].

4.2. Proposed Space Partitioning

Usually, defining the regions the training data set is taken as is. These regions are not related to the classification regions, nor do they need to have a specific shape or size. We found that the classification performance is improved when only so called the *"hard boundary points"* for classification are used to find the reference points v_i.

Breiman defined the *"hard boundary points"* as data points laying on or near the boundary between classes and being consistently miss-classified by a constructed committee [23]. We define our *"hard boundary points"* as the data points miss-classified by any member in the committee. These points are filtered out from the training data set by members of the committee and are used to find the reference points v_i. The data point is labelled as *"miss-classified"* if at least one member miss-classifies the given training data point. The regions and the reference points are found by using a data clustering technique.

4.3. Classifiers Selection (Design Phase)

When the space is partitioned into the regions $\Re_i$, $(i=1,...,K)$, then we should select which classifier from the $Z=\{z_1,z_2,...,z_L\}$ will operate in $\Re_i$. Using the data points from the $\Re_i$, we estimate the classification accuracy of each member in the committee. The classifier $z_{j(k)}$ with the highest accuracy is nominated for data classification in the region $\Re_i$.

The presumption in classifier selection is that each classifier is "an expert" in some local area of data space. Space partitioning and classifier selection gives at least the same classification accuracy on the training data as the best member in the committee Z [20].

4.4. Classifier Selection for Data Classification

The classifier selection (*CS*) method for the data classification works in the following way. For any sample $x \in \Re^n$, find the nearest cluster centre from $v_1,...,v_K$. The index of the region is determined in the following way:

$$k = \arg \min_{i=1,...,K} d(\mathbf{x}, \mathbf{v}_i) \qquad (1)$$

with $d(x,v_i)$ being the Euclidean distance between the data point x and vector v_i representing the i th region. Use $z_{j(k)}$ to label x in region k.

The number of classifiers L is not necessarily equal to the number of regions K. Some classifiers might never be selected. Often, even the classifier

with the highest average accuracy over the whole data space might be never selected into the final set. On the other hand, one classifier might be nominated for more than one region.

5. Background on λ-Fuzzy Measure and Fuzzy Integral

Definition 1. A set function $g: 2^Z \to [0,1]$ is a *fuzzy measure* if

1. $g(0)=0$; $g(Z)=1$,
2. *if* $A, B \subset 2^Z$ *and* $A \subset B$ then $g(A) \leq g(B)$,
3. *if* $A_n \subset 2^Z$ for $1 \leq n \leq \infty$ and the sequence $\{A_n\}$ is monotone in the sense of inclusion, then

$$\lim_{n\to\infty} g(A_n) = g(\lim_{n\to\infty} A_n).$$

In general, the fuzzy measure of a union of two disjoint subsets cannot be directly computed from the fuzzy measures of the subsets. Sugeno [26] introduced the decomposable so called λ-fuzzy measure satisfying the following additional property

$$g(A \cup B) = g(A) + g(B) + \lambda g(A) g(B) \qquad (2)$$

for all $A, B \subset Z$ and $A \cap B = 0$, and for some $\lambda > -1$.

Let $Z=\{z_1, z_2, \ldots z_L\}$ be a finite set (a set of committee members in our case) and let $g^i = g(\{z_i\})$. The values g^i are called the densities of the measure. The value of λ is found from the equation $g(Z)$, which is equivalent to solving the following equation:

$$\lambda + 1 = \prod_{i=1}^{L} (1 + \lambda g^i) \qquad (3)$$

When g is the λ-fuzzy measure, the values of $g(A_i)$ can be computed recursively as follows

$$g(A_1) = g(\{z_1\}) = g^1 \qquad (4)$$

$$g(A_i) = g^i + g(A_{i-1}) + \lambda \; g^i g(A_{i-1}), \quad for \;\; 1 < i \leq L \qquad (5)$$

Fuzzy measures are significantly more expressive comparing to classical measures. Let us assume that we are given a fuzzy measure space Z, 2^Z, g) and sets $A, B \subset 2^Z$ such that $A \cap B = 0$ and $A \cup B \subset 2^Z$. A fuzzy measure g is then capable of capturing any of the following possible situations:

1. $g(A \cup B) > g(a) + g(B)$, which expresses some inherent *complementary* or *positive synergy* between A and B with respect to the property measured by g;
2. $g(A \cup B) < g(a) + g(B)$, which expresses some *redundancy* or *negative synergy* between A and B with respect to the property measured by g;
3. $g(A \cup B) = g(a) + g(B)$, which expresses the fact that there is *no interaction* between A and B in terms of the property measured by g.

The probability theory, which is based on the classical measure theory, is capable of capturing only the situation (*c*). This illustrates that the fuzzy measure theory provides a broader framework than the classical measure theory. However, to utilise this broad framework, we need to construct fuzzy measures that express the actual interaction among sets with respect to properties of interest.

Definition 2. Let g be a fuzzy measure on Z. The *discrete Choquet integral* of a function $h: Z \to \Re^+$ with respect to g is defined as

$$C_g\{h(z_1), \ldots, h(z_L)\} = \sum_{i=1}^{L} \{h(z_i) - h(z_{i-1})\} g(A_i) \qquad (6)$$

where indices i have been permuted so that

$$0 \leq h(z_1) \leq \ldots \leq h(z_L) \leq 1, \;\; A_i = \{z_i, \ldots, z_L\}$$

and $h(z_0)=0$.

There are a number of interpretations on the meaning of fuzzy integral. A fuzzy integral can be understood as a fuzzy expectation, the maximal grade of agreement between two opposite tendencies [27], or the maximal grade of agreement between the objective evidence and the expectation [16]. In this paper, a fuzzy integral is considered as a maximum degree of belief for an object to belong to a certain class.

6. Combination Schemes Used

In our investigations we used four combination schemes, namely the majority vote, averaging and, combination by the Choquet integral with the λ-fuzzy measure. Next, we briefly describe the combination schemes used.

6.1. Majority Vote

The correct class is the one most often chosen by different classifiers. If all the classifiers indicate different classes, then the one with the overall maximum output value is selected to indicate the correct class.

6.2. Averaging

This approach simply averages the individual classifier outputs. The output yielding the maximum of the averaged values is chosen as the correct class q:

$$q = \arg \max_{j=1,\ldots,Q} \left(\bar{y}_j = \frac{1}{L} \sum_{i=1}^{L} y_{ji} \right) \qquad (7)$$

where Q is the number of classes, L is the number of classifiers and y_{ji} represents the j-th output of the i-th classifier.

6.3. Combination by Choquet Integral

We assume that committee members have Q outputs representing Q classes, and data point x needs to be assigned into one of the classes. The class label c for the data point x is then determined as follows:

$$c = \arg \max_{q=1,\ldots,Q} C_g(q) \qquad (8)$$

where $C_g(q)$ is the Choquet integral for the class q with respect to the λ-fuzzy measure g. The densities of the measure and the values of λ are determined through minimisation of the classification error for the training data. For this purpose the random search procedure was invoked [28]. The values of the function $h(z)$ that appear in the Choquet integral are given by the output values of members of the committee (an evidence provided by the members).

6.4. Combination by Choquet Integral With Data Dependent Densities

This approach assumes that data space is partitioned into K regions with reference point v_i representing the i-th region. The densities of the measure and the values of λ are determined in each region.

Assume that the committee members have Q outputs representing Q classes and data point x needs to be assigned into one of the classes. Then the class label c for the data point x is determined as follows:

$$c = \arg \max_{q=1,\ldots,Q} C_{gk}(q) \qquad (9)$$

where $C_gk(q)$ is the Choquet integral for class q calculated in region k. The index of the region is determined in the following way:

$$k = \arg \min_{i=1,\ldots,K} d(\mathbf{x}, \mathbf{v}_i) \qquad (10)$$

with $d(x,v_i)$ being the Euclidean distance between the data point x and vector v_i representing the i th region.

7. Experimental Testing

All comparisons between different classifiers in the *ELENA* project have been done using the Holdout method with equal training and testing parts of the data. To make the comparisons feasible, we have also used equally sized training and test data sets.

In all the tests presented here, we train a set of one-hidden layer MLPs with 10 sigmoidal hidden units using the Bayesian inference technique [29, 30] – to obtain regularised networks. We run each experiment seven times, and the *mean* errors presented are calculated from these seven trials. In each trial, the data set used is randomly divided into training and testing parts of the same size. In the half&half sampling approach, the size of the data sets *MC* and *CC* was set to $M=N_{learn}/4$, where N_{learn} is the size of the learning set.

Tabl. 2-4 summarise the results of the experiments. In the first set of experiments we trained 20 neural networks with half&half sampling technique and combined them using the majority vote, the averaging, and the Choquet integral based technique with one common for the entire data space λ-fuzzy measure. For this round of tests, the data space was not partitioned (column named K=1). The following notations are used in the tables: *Mean* stands for the percentage of the average classification error, *Std* is the standard deviation of the error, *The Best* stands for the single neural network with the best average performance, *MV* means majority vote, *AV* stands for the averaging, and *CI* means the Choquet integral with the λ fuzzy measure.

The *MV* and *AV* treat classifiers equally without considering their differences in performance and do not require any special training. A half of data from both learning and test data sets was used to estimate the densities of λ-fuzzy measure in *CI*. The fuzzy measure has been learned through the minimisation of the classification error.

As can be seen from the tables, there is an obvious improvement in classification accuracy when combining networks. The achieved performances can be compared with the error rates obtained in the *ELENA* projects (tabl. 1). The AV is always slightly better than the MV combination method however the difference is not statistically significant. The *CI* provided the best overall performance.

For the *Clouds* data set the classification error is quite close to the theoretical one. Therefore, any improvements are hardly achieved.

Table 2. Performance of the neural network committees for the *Clouds* data.

	K=1		K=10		K=20		K=50		K=100	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
The best	10.77	0.18								
MV	10.42	0.20								
AV	10.38	0.26								
CI	9.85	0.13								
SP&CS			10.48	0.33	10.36	0.42	10.12	0.18	9.89	0.25
SPmc&CS			10.34	0.11	10.17	0.10	9.92	0.18	9.79	0.18
CI+SP			9.78	0.11	9.75	0.11	9.72	0.10	9.77	0.15
CI+SPmc			9.76	0.08	9.72	0.04	9.70	0.06	9.71	0.08

Table 3. Performance of the neural network committees for the *Phoneme* data.

	K=1		K=10		K=20		K=50		K=100	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
The best	15.07	0.86								
MV	11.20	0.20								
AV	11.10	0.11								
CI	10.83	0.31								
SP&CS			14.00	0.39	13.53	0.48	12.10	0.31	10.82	0.40
SPmc&CS			13.53	0.31	12.54	0.26	10.61	0.29	10.25	0.20
CI+SP			9.36	0.30	9.10	0.21	8.95	0.20	9.07	0.29
CI+SPmc			9.22	0.22	8.84	0.12	8.72	0.14	8.74	0.16

Table 4. Performance of the neural network committees for the *Satimage* data.

	K=1		K=10		K=20		K=50		K=100	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
The best	11.87	0.21								
MV	10.37	0.19								
AV	10.23	0.18								
CI	9.80	0.22								
SP&CS			11.60	0.24	11.13	0.21	10.70	0.22	10.35	0.25
SPmc&CS			11.12	0.19	10.44	0.19	10.18	0.19	9.09	0.13
CI+SP			9.09	0.12	8.79	0.17	8.48	0.09	8.37	0.19
CI+SPmc			8.92	0.21	8.53	0.15	8.45	0.10	8.31	0.17

With the second set of experiments we were aiming to investigate the possibility to increase the classification accuracy with already trained 20 neural networks using the space partitioning techniques. The whole data space was partitioned into K=10, 20, 50, and 100 regions. In the tests the *fuzzy c-means* clustering technique was used.

The following additional notations are used in the tables: *SP&CS* - usual space partitioning with combination by classifier selection approach; *SPmc&CS* – the proposed space partitioning obtained by using miss-classified data points and classifier selection approach; *CI+SP* – the combination by the Choquet integral with data depended densities estimated in the space regions partitioned by *SP*; *CI+SPmc* – the combination by the Choquet integral with data depended densities estimated in the space regions partitioned by proposed *SPmc*.

As can be seen from tabl. 2 and 3, the classification by classifier selection approach with new space partitioning algorithm (*SPmc*) outperformed the classification by classifier selection with standard space partitioning algorithm *SP*. The result is explained by examining the *Clouds* data set in fig. 1 and 2. The fig. 1 shows the 10 regions with reference points (pentagrams) distributed over the whole feature space by *SP* and fig. 2, shows the 10 regions and reference points obtained by using the miss-classified data points. It is obvious, if all classifiers agree in some regions then this region is out of interest. The SP distribute the reference points v_i over the entire data space without considering class labels and a part of v_i stack in the space parts where appearance of data points from more than one class is very rare.

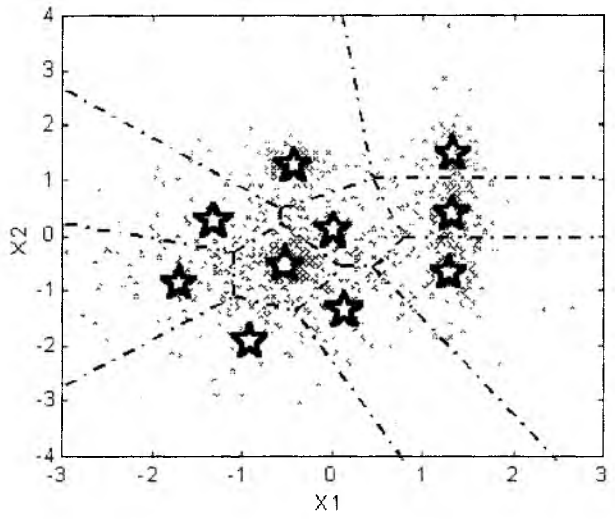

Figure 1. The *Clouds* data set with 10 reference points ("pentagrams") obtained by using the whole training data set (*SP*).

For the *CI* combination method with data dependent densities (*CI+SP* and *CI+SPmc*) when K is very high (K=100), the evaluation of densities becomes very problematic. For the 20 members in the committee the 20 values of λ-fuzzy measure need to be estimated in the regions. In some regions the training data set reduces to 15 data points or even less. The miss-estimated densities degrade the combination capabilities of CI with data dependent densities. This situation is observed in tabl. 2 and 3 comparing columns K=50 with K=100.

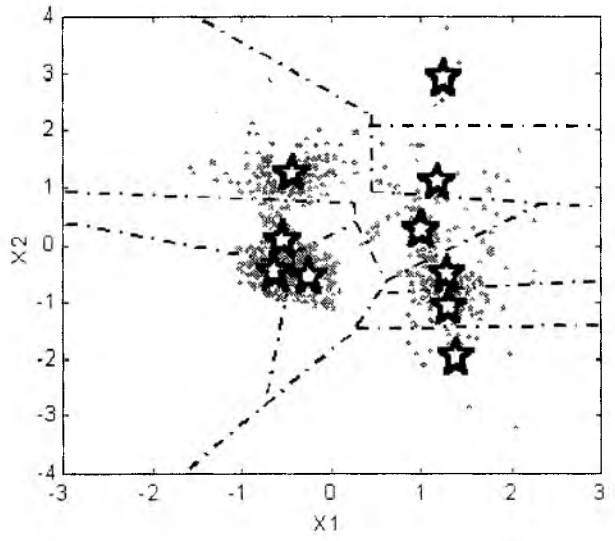

Figure 2. The *Clouds* data set with 10 reference points ("pentagrams") obtained by using only the miss-classified data points (*SPmc*)

Comparing the *CI+SP* with the *CI+SPmc*, the *CI+SPmc* is always slightly better than the *CI+SP* combination method however the difference is not statistically significant. The *CI+SPmc* provided the best overall performance.

8. Conclusion

In this paper we studied two different classifiers fusion algorithms exploiting the combination weights expressed over the entire data space and the combination with data dependent weights. The combination schemes with data dependent weights always outperformed classifiers fusion algorithms with the ordinary combination weights common in whole feature space. We have also shown that the space partitioning obtained by using the miss-classified data points can improve the data dependent combination schemes.

References

[1]. Nilsson, N.J. 1965, Learning Machines: Foundations of Trainable Pattern-Classifying Systems, *McGraw Hill*, NY.

[2]. Xu, L., Krzyzak, A., Suen C.Y., Methods for combining multiple classifiers and their applications to handwriting recognition. IEEE *Trans. Systems, Man, and Cybernetics* 22(3), 1992: pp. 418-435.

[3]. Lam, L., and Suen, C.Y., Optimal combination of pattern classifiers, *Pattern Recognition Letters* 16, 1995: pp. 945-954.

[4]. Verikas, A., Lipnickas, A., Malmqvist, K., Bacauskiene, M., Gelzinis, A., Soft combination of neural classifiers: A comparative study. *Pattern Recognition Letters* 20, 1999: pp. 429-444.

[5]. Kittler, J., Hatef, M., Duin, R.P.W., Matas, J., On combining classifiers, IEEE *Transaction on Pattern Analysis and Machine Intelligence*, vol. 20, **3**, 1998: pp. 226 –239.

[6]. Perrone, M.P., Cooper, L.N., When networks disagree: Ensemble method for neural networks. In: Mammone, R.J. (ed). *Neural Networks for Speech and Image* Processing, Chapman-Hall, 1993.

[7]. Verikas, A., Signahl, M., Malmqvist, K., and Bacauskiene, M., Fuzzy committee of experts for segmentation of colour images, In Proceedings of *5th European Congress on Intelligent Techniques and Soft Computing*, Aachen, Germany, 1997, vol. 3, pp. 1902-1906.

[8]. Verikas, A., Lipnickas, A., Malmqvist, K., Fuzzy measures in Choquet integral based neural networks fusion, In B.

Chandrasekaran, M.D. Levine, C.H. Chen (eds) Recent Research Developments in *Pattern Recognition Research*, Transworld Research Network, vol. I, Part I, 2000:pp.: 119-135.

[9]. A.Verikas, A. Lipnickas, K. Malmqvist, Fuzzy measures in neural networks fusion. Proceedings of the *7th International Conference on Neural Information Processing*, ICONIP-2000, Taejon, Korea, 2000, pp.: 1152 -1157.

[10]. Ho, T.K., Hull, J.J., and Srihari, S.N., Decision combination in multiple classifier systems, IEEE *Trans. on Pattern Analysis and Machine Intelligence* 16(1), 1994: pp. 66-75.

[11]. Kuncheva, L.I., An application of OWA operators to the aggregation of multiple classification decisions, In: Yager, R., Kacprzyk, J. (eds). *The Ordered Weighted Averaging Operators. Theory and Applications*, Kluwer Academic Publishers, 1997: pp. 330-343.

[12]. Tumer, K., and Ghosh, J., Linear and order statistics combiners for pattern classification, In: A.J.C. Sharkey (ed). *Combining Artificial Neural Nets: Ensemble and Modular Multi-Net Systems*, Springer-Verlag, 1999: pp. 127-162.

[13]. Ceccarelli, M., and Petrosino, A. 1997, Multi-feature adaptive classifiers for SAR image segmentation, *Neurocomputing* 14, 345-363.

[14]. Chen, W., Gader, P.D., Shi, H., Improved dynamic programming-based handwritten word recognition using optimal order statistics. In: Proceedings of the *International Conference Statistical and Stochastic Methods in Image Processing II*, San Diego, 1997: pp. 246-256.

[15]. Grabisch, M. 1995, Fuzzy integral in multicriteria decision making, *Fuzzy Sets and Systems* 69, 279-298.

[16]. Tahani, H., Keller, J.M., Information fusion in computer vision using the fuzzy integral. IEEE *Trans Systems, Man and Cybernetics* 20(3), 1990: pp.: 733-741.

[17]. Huang Y.S., Suen C.Y., A method of combining multiple experts for the recognition of unconstrained handwritten numerals. IEEE *Transaction on Pattern Analysis and Machine Inteligence*, vol. 17, **1**, 1995: pp. 90-94.

[18]. Woods K., Kegelmeyer W.P., Bowyer K. Combination of Multiple Classifiers Using Local Accuracy Estimates. IEEE *Trans. on Pattern Analysis and Machine Intelligence*, vol. 19, **4**. 1997: pp.405-410.

[19]. Giacinto G., Roli F., Adaptive selection of image classifiers. *ICIAP'97, 9th ICIAP*, Florence, Italy, Sept 17-19. Lecture Notes in

[20]. Computer Science 1310, Springer Verlog Ed., 1997: pp. 38-45.

[21]. Kuncheva L.I., Clustering and selection model for classifier combination. In Proc. *Knowledge-Based Intelligent Engineering System and Allied Technologies*, Brighton, UK, 2000. (submitted).

[22]. Optitz, D.W., and Shavlik, J.W., Generating accurate and diverse members of a neural-network ensemble, In: Touretzky, D.S., Mozer, M.C., Hasselmo, M.E. (eds). *Advances in Neural Information Processing Systems* 8, MIT Press, 1996: pp. 535-541.

[23]. Maclin, R., J.W. Shavlik,. Combining the predictions of multiple classifiers: Using competitive learning to initialize neural networks, Proceedings of the *14th International Conference on Artificial Intelligence*. 1995.

[24]. Breiman L., Half&Half bagging and hard boundary points. Technical report 534, Statistics Departament, University of California, Berkeley. 1998.

[25]. Freund Y., Schapire R. E., A decision-theoretic generalization of on-line learning and an application to boosting, *Journal of Computer and System Sciences* 55, pp. 119-139, 1997.

[26]. Freund Y., Schapire R. E.,, Experiments with a new boosting algorithm, In: Proceedings of the *Thirteenth International Conference "Machine Learning"*, pp. 148-156, 1996.

[27]. Sugeno, M., Fuzzy measures and fuzzy integrals: A survey. In: Automata and Decision making, Amsterdam, North Holland, 1977: pp. 89-102.

[28]. Pham, T.D., and Yan, H., Color image segmentation using fuzzy integral and mountain clustering, *Fuzzy Sets and Systems* 107, 1999: pp. 121-130.

[29]. Verikas, A., Gelzinis, A. Training neural networks by stochastic optimisation, *Neurocomputing* 30, 2000: pp.: 153-172.

[30]. Bishop, C.M. 1996, *Neural Networks for Pattern Recognition*, Clarendon Press, Oxford.

[31]. MacKay, D.J., Bayesian interpolation, *Neural Computation*, 4, 1992: pp. 415-447.