We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS BOOK CITATION INDEX INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

1

A New Divide and Conquer Based Classification for OCR

Hamid Parvin, Hosein Alizadeh and Behrouz Minaei-Bidgoli
Iran University of Science and Technology, Tehran, Iran

1. Introduction

Recognition systems have found many applications in almost all fields these years (Parvin et al., 2008a,c). Improving the recognition performance is one of the most challenging tasks in pattern classification (Parvin et al., 2009a). However, Most of classification algorithms have obtained good performance for specific problems; they have not enough robustness for other problems. Therefore, recent researches are directed to the combinational methods which have more power, robustness, resistance, accuracy and generality (Kuncheva, 2005). Ensemble learning algorithms train multiple base classifiers and then combine their predictions. Since the generalization ability of an ensemble could be significantly better than a single classifier, ensemble learning has been a hot topic during the past years (Dietterich, 2002). It has been established firmly as a practical and effective solution for difficult classification problems. It appeared under numerous names: hybrid methods, decision combination, multiple experts, mixture of experts, classifier ensembles, cooperative agents, opinion pool, decision forest, classifier fusion, combinational systems and so on. For a good coverage on ensemble learning the reader is referred to (Dietterich, 1997; Minaei et al., 2004; Jain et al., 2000).

A large number of researches for improving the performance of classification methods have been done. Most of the late researches are about ensemble methods of classification. Although they improve significantly the performance, they cannot reach an effective way in the case that the number of classes are fairly large. This paper suggests a new method which transforms a multiclass problem to pairwises classes ones. To do this operation it uses confusion matrix to determine which of pairwise classes can be better distinguished. The confusion matrix determines the error distribution on different classes (Parvin et al., 2008d). The entry a_{ij} from confusion matrix determines that how many samples of class c_j are classified as class c_i. In order to achieve this matrix, we have to examine the classifier on validation set.

To optimize the division of classes, the proposed method employs genetic algorithm. This optimization phase is applied each time for dividing a metaclass (a set of classes) into two smaller metaclasses optimally. This method is similar to creation of a binary tree. Each node is equal to one classifier that distinguishes the classes of the left and right nodes. This process continues until each group of classes contains just one class in leafs. This new method is applied on a very large OCR dataset (Khosravi et al., 2007) which is a challenging problem in Farsi languages.

Source: Convergence and Hybrid Information Technologies, Book edited by: Marius Crisan,
ISBN 978-953-307-068-1, pp. 426, March 2010, INTECH, Croatia, downloaded from SCIYO.COM

To better understand the proposed method one should take an overview on the following subsections 1.1 to 1.4. These subsections include a brief coverage of the related methods.

1.1 Decision tree

A common and obvious way for classifying an instance is from a sequence of questions, so that next question is asked with regard to this current question. Using trees are the most common representation way for theses question-answers. Decision tree is used to create a classifier ensemble, expansively. Also, they are used for the application of data mining and clustering. Their functionality is understandable for human. Besides, unlike other methods such as ANN, they are very quick. It means their learning phase is quicker than other methods (Duda, 2001). Different structures of decision trees are described in (Breiman, 1984; Duda, 2001). One of the most important specifics of them is that each node asks a question only on one feature. In this paper, a new classification method is proposed which operates like decision trees, however it makes decision over all features.

1.2 K-Nearest Neighbor

Nearest Neighbor techniques are simple but powerful non-parametric classification systems (Darasay, 1991). The simplest version of them is Single Nearest Neighbor. It bypasses the problem of probability densities completely and simply classifies an unknown sample as belonging to the same class as the most similar or nearest sample point in the training set of data. Nearest can be taken to mean of the smallest Euclidian distances in n-dimensional feature space. Although Euclidian distance is probably the most commonly used distance function or measure of dissimilarity between feature vectors, it can be used from other metrics like: Manhattan or Maximum distances.

A more general version of the Single Nearest Neighbor technique bases the classification of an unknown sample on the "votes" of K of its nearest neighbor rather than on only its Single Nearest Neighbor. The K-Nearest Neighbor classification procedure is denoted by KNN. If the costs of error are equal for each class, the estimated class of an unknown sample is chosen to be the class that is most commonly represented in the collection of K nearest neighbor (Gose et al., 1996). In this paper, the KNN with the Euclidian distance is used as a base classifier.

1.3 Neural Network

The Artificial Neural Network or ANN algorithms are the commonly used as base classifiers in classification problems (Roli et al., 2001). The first wave of interest in neural networks emerged after the introduction of simplified neurons by McCulloch and Pitts in 1943. These neurons were presented as models of biological neurons and as conceptual components for circuits that could perform computational tasks.

The elements of the ANNs are input vectors, output vectors, target vectors, weight, transfer function and bias. There are different forms to connect the neurons, and then the result is different network topologies (Sanchez et al., 2006).

Each unit of neural network performs a relatively simple job: receive input from neighbors or external sources and use this to compute an output signal which is propagated to other units. Apart from this processing, a second task is the adjustment of the weights. The system is inherently parallel in the sense that many units can carry out their computations at the same time. Within neural systems it is useful to distinguish three types of units: input units

which receive data from outside the neural network, output units which send data out of the neural network, and hidden units whose input and output signals remain within the neural network. During operation, units can be updated either synchronously or asynchronously. With synchronous updating, all units update their activation simultaneously; with asynchronous updating, each unit has a probability of updating its activation at a time t, and usually only one unit will be able to do this at a time. In some cases the latter model has some advantages.
In most cases we assume that each unit provides an additive contribution to the input of the unit with which it is connected. The total input to unit k is simply the weighted sum of the separate outputs from each of the connected units plus a bias or offset term θ_k as equation 1:

$$s_k(t) = \sum_j w_{jk}(t) y_j(t) + \theta_k(t) \tag{1}$$

The contribution for positive w_{jk} is considered as an excitation and for negative w_{jk} as inhibition. In some cases more complex rules for combining inputs are used, in which a distinction is made between excitatory and inhibitory inputs. We call units with a propagation rule sigma units. Generally, some sort of threshold function is used: a hard limiting threshold function, a linear or semi-linear function or a smoothly limiting threshold.
A neural network has to be configured such that the application of a set of inputs produces the desired set of outputs. Various methods to set the strengths of the connections exist. One way is to set the weights explicitly, using a priori knowledge. Another way is to train the neural network by feeding its teaching patterns and letting it to change its weights according to some learning rule like Gradient Descent (Haykin, 1999). In this paper, the Multi Layer Perceptron is used as a base classifier. As mentioned, the KNN and MLP are two base classifiers used in this paper and we compare our proposed method with them, too.

1.4 Genetic Algorithm

Genetic Algorithm is a random search technique based on natural genetic. It has been shown to be an effective tool to use in data mining and pattern recognition (De Jong et al., 1993). There are two different approaches to applying GA in pattern recognition: apply a GA directly as a classifier and use a GA as an optimization tool. In this paper we will focus on the second approach and use a GA to minimize the classification error.
A Genetic Algorithm can be considered as a composition of three essential elements: first, a set of potential solutions called individuals or chromosomes that will evolve during a number of Generations. A chromosome contains a sequence of genes. The number and type of genes of each chromosome depend on the problem. This set of solutions is also called population. Second, an evaluation mechanism that allows assessing the quality or fitness of each individual of the population. And third, an evolution procedure that is based on some genetic operators such as selection, crossover and mutation. The crossover takes two individuals to produce two new individuals. The mutation consists in modifying randomly a gene of an individual. The basic steps of GAs, which are also followed in the proposed classification method, are shown in Fig. 1.
In the rest of this paper, in section 2, the proposed approach will be described. Section 3 says that how Genetic algorithm optimizes the ensemble tree construction. Section 4 contains its

results on Farsi OCR dataset. It also compares them with previous methods. Section 5 concludes the study.

```
Begin
        1. t=0
        2. Initialize population P(t)
        3. Compute fitness of P(t)
        4. t=t+1
        5. If termination criterion achieved go to step 10
        6. Select P(t) from P(t-1)
        7. Crossover P(t)
        8. Mutate P(t)
        9. Go to step 3
        10. Output best and stop
End.
```

Fig. 1. The original GA which is used in the proposed method

2. Proposed method

The main idea of the presented method is dividing the classification problem into smaller problems. Suppose that a metaclass is a subset of classes (minus null). Also, suppose that all classes are in a one large metaclass. So, in each level, there is a classifier to divide a metaclass into two smaller metaclasses. Indeed, this method is like divide-and-conquer method.

For this method, the dataset is partitioned into three sets: training, evaluation and test sets. Our approach contains several steps. In the first step, by using a base classifier, we do a primary classification and extract the confusion matrix from the evaluation data set. Table 1 shows this confusion matrix on evaluation data. In this step a multiclass classifier is trained on training dataset. Then, the confusion matrix is made by using the results of this classifier on evaluation data.

	0	1	2	3	4	5	6	7	8	9
0	1943	0	0	6	1	30	2	0	0	2
1	6	1985	4	0	4	10	7	3	1	25
2	2	1	1957	38	17	2	8	7	0	2
3	0	0	18	1918	23	1	4	4	0	3
4	6	2	6	32	1945	5	6	2	0	4
5	29	1	0	0	2	1948	5	0	0	1
6	3	9	7	1	5	0	1942	9	2	9
7	8	0	3	2	2	1	1	1975	0	0
8	2	1	0	1	0	3	2	0	1991	1
9	1	1	5	2	1	0	23	0	6	1953

Table 1. The confusion matrix corresponding to Farsi OCR dataset

This matrix contained important information about functionality of classifiers. Also, close and error prone classes are recognized using this matrix. In fact, the confusion matrix determines error distribution on different classes. Item a_{ij} from confusion matrix determines how many instances from class c_j are recognized as class c_i. In the second step, we use a classifier ensemble more or less like decision tree. We train one classifier correspond to each node that divides the data into two metaclasses. Each of metaclasses can contain several classes. This categorization is done based on error rate of confusion matrix. For example, suppose that the first level classifier divides data in two metaclasses. One is contained classes 0-4 and the other classes are in metaclass two. We will have 214 error based on the confusion matrix. Also, if the metaclass 1 contains classes 0,1,2,4 and 5 and metaclass 2 contains the other classes, we will have 245 errors. As shown in the table 1, metaclasses 1 and 2 are highlighted by gray levels such that the metaclass 1 is lighter than 2. Thus, all white cells are counted as errors between metaclasses. Also, in this step, each of these metaclasses is divided in two new smaller metaclasses. This procedure continues, until there is one class in each node. The optimal selection of these metaclasses is executable by different methods, such as recursive methods and GAs.

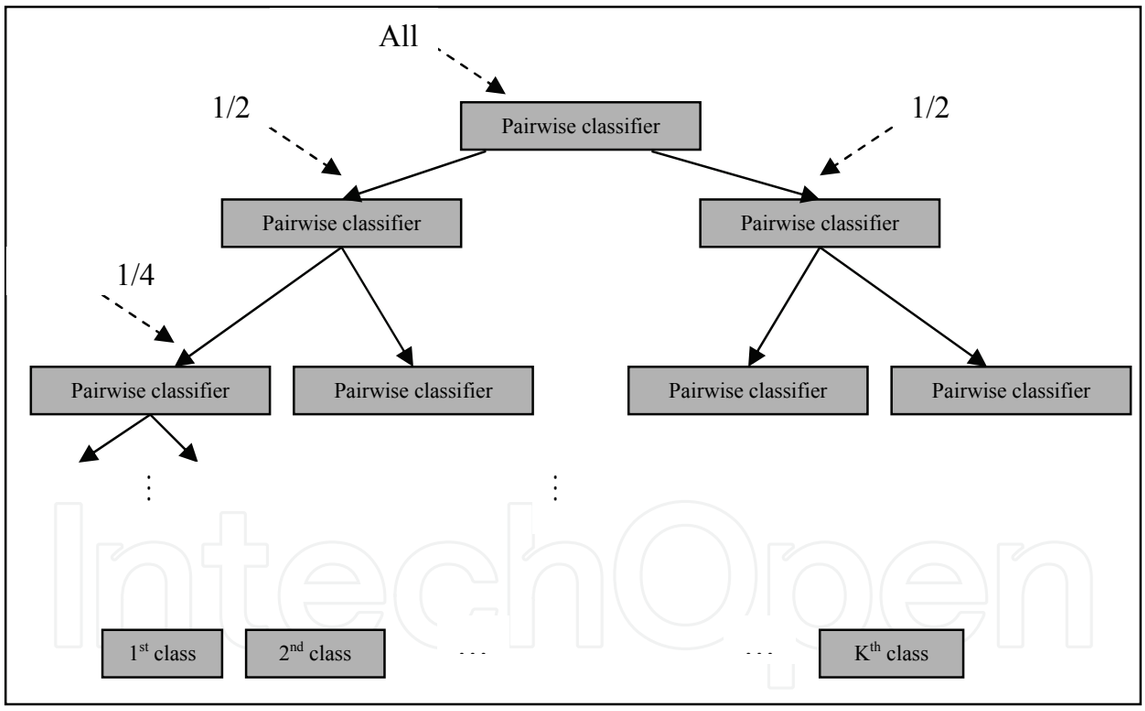

Fig. 2. The structure of proposed method for multiclass

There is a question that does this method like decision trees? Decision tree does comparison on only one feature in each node. This is implemented with only a simple thresholding or comparison, whereas, the classification in new method is done on the total feature space by MLPs or KNNs. As Deviparikh's definition of classifier fusion in (Parikh et al., 2007), the new proposed method cannot be a classifier fusion, too. In fact, it is a new kind of classifier ensemble.

3. Tree construction

Bandyopadhyay and Muthy in (Bandyopadhyay et al., 1995) have used GA as an effective tool in pattern recognition. Most applications of GAs in pattern recognition optimize some parameters in the classification process. The searching capability of genetic algorithms is used in this paper for the purpose of appropriately deriving the optimal tree. In fact, GA is used to optimize the selection of classifiers. The fitness function for GA is the total error of all nodes and the main problem of GA is the total error minimization. The total error is the sum of errors in all levels. So, it seems that when the constructed tree is balanced, the error is less. Fig. 2 depicts our proposed approach. As shown in this figure, for each node there is an MLP or a KNN as base classifier.

Each chromosome in GA contains 26 genes. The value 1 for each gene means that the corresponding class is on the right side of the tree; otherwise, it is on the left side of the tree. Genes 1-10 are used for first level. They determine the left and right sub trees. In the next level, genes 11-15 are used for left side. Similarly, genes 16-20 are used for right side and so on.

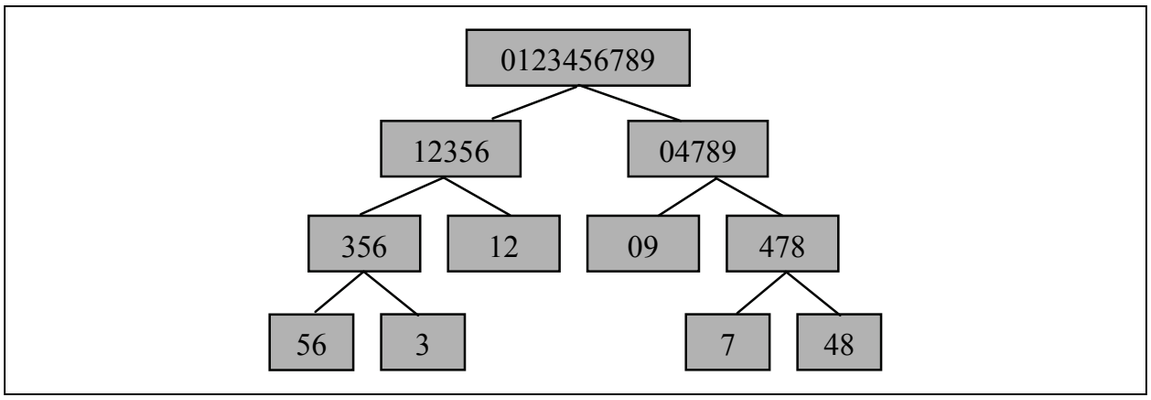

Fig. 3. The Corresponding tree of chromosome 1000100111,11000,01110,100,101.

For example, suppose that there is a chromosome like 1000100111,11000,01110,100,101. The first ten genes means that the metaclass in the first level is divided into classes 0,4,7,8 and 9 in the right node and other classes in the left node at next level. See the corresponding tree of this chromosome which is shown in Fig. 3.

4. Experimental results

In this section, experimental results of the study are shown. After introducing dataset, the used parameters and results are explained.

4.1 Dataset

There is a long time that offline handwritten OCR recognition system is known as an important topic. Recently, a large part of researches have been focused on improving accuracy of character recognition system. We evaluate our method on Farsi digits OCR. We use a large handwritten dataset of Farsi digits, named "Hoda" (Khosravi et al., 2007). We divide the data into 3 parts: training, evaluation and test sets which contain 40,000, 20,000 and 20,000 instances, respectively. The validation data set acts as pseudo-testing for

obtaining fitness of each chromosome as it was explained above. In this study, the 106 extracted features from this data are utilized which are described in (Khosravi et al., 2007). Some examples of instances of this dataset are depicted in Fig. 4.

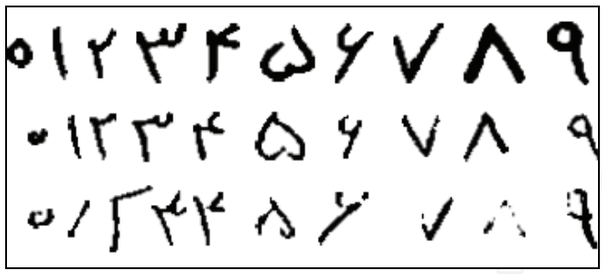

Fig. 4. Some instances of Farsi digits OCR dataset

4.2 Used parameters

In this empirical study, a two hidden layer Perceptron is used as a base classifier. Also, the KNN with K=3 is another base classifier. The confusion matrix is obtained from these classifiers. After that, GA is used to determine the optimal tree. We use Gaussian and Scattered operators respectively for mutation and crossover. The Scattered crossover function creates a random binary vector and selects the genes where the vector is a 1 from the first parent, and the genes where the vector is a 0 from the second parent, and combines the genes to form the child. The Gaussian mutation on each entry of the parent vector follows a Gaussian distribution. For GA optimization, we use 200 individuals in our population, running the GA over 500 generations. Fitness function for GA is the error ratio obtained from confusion matrix.

4.3 Experiments

Table 2 shows the results of classification performance by previous and proposed method. We ran the program ten times and got the averages. As shown in Table 2, the recognition

	Obtained Accuracy by Base Classifier	
	MLP	KNN
Simple Classifier	95.7	96.66
Full Ensemble	96.01	-
Unweighted Static Classifier Selection	97.11	-
Weighted Static Classifier Selection	97.15	-
Proposed method	97.20	96.86

Table 2. The results of proposed ensemble method

ratio has improved approximately 1.42%, by applying the proposed method by MLP. Also, we arrived to 0.2% improvement by KNN.

5. Conclusion and future works

We proposed a new method for improving the performance of multiclass recognition system. This method uses population based approaches, in special GA, to obtain best way to conquer classification the classes from each other. Also the method is based minimizing error on evaluation set without learning the evaluation. This is because of necessitated speed in evaluating the chromosome. Applying the proposed approach leads to more accurate logical results than the simple classification, on handwritten digits dataset. We used this method on two classifier ensemble systems. Also, we used one kind of base classifier per experiment. It yields to a better result in both cases. We divided the classes as balanced as possible, in each level. As future work, it can be worked on as unbalanced division. However, hereby, our optimization has been done globally; we can further, focus on the level based optimization as well on future.

6. References

Alizadeh H., Minaei-Bidgoli B., & Amirgholipour S.K. (2009). A New Method for Improving the Performance of K Nearest Neighbor using Clustering Technique, *International Journal of Convergence Information Technology, JCIT,* (*DBLP Indexed*), ISSN: 1975-9320.

Bandyopadhyay S. & Muthy C. A. (1995). "Pattern Classification Using Genetic Algorithms", Pattern Recognition Letters, (1995).Vol. 16, pp. 801-808.

Breiman L., Friedman J., Olshen R & C. Stone (1984) Classification and Regression Trees, *Wadsworth International*, Belmont, California.

Darasay B.V. (1991). *Nearest Neighbor pattern classification techniques*, Las Alamitos, LA: IEEE Computer Society Press.

De Jong K.A., Spears W.M. and Gordon D.F. (1993). Using genetic algorithms for concept learning. *Machine Learning* 13, pp. 161-188.

Dietterich T.G. (1997). "Machine-learning research: four current direction," *AI Magazine*, 18, 4, pp. 97-135.

Dietterich T.G. (2002). "*Ensemble learning," in The Handbook of Brain Theory and Neural Networks, 2nd edition*, M.A. Arbib, Ed. Cambridge, MA: MIT Press.

Duda R. O., Hart P. E., and Stork D. G.(2001). Pattern Classification, 2nd ed. John Wiley & Sons, NY.

Gose E., Johnsonbaugh R. & Jost S. (1996) *Pattern Recognition and Image Analysis*, Prentice Hall, Inc., Upper Saddle River, NJ 07458.

Haykin S. (1999). "*Neural Networks, a comprehensive foundation", second edition,* Prentice Hall International, Inc. ISBN: 0-13-908385-5.

Jain A.K., Duin R.P.W. & Mao J. (2000). "Satanical pattern recognition: a review," *IEEE Transaction on Pattern Analysis and Machine Intelligence, PAMI-22*, 1, pp. 4-37.

Khosravi H., Kabir E.(2007). "Introducing a very large dataset of handwritten Farsi digits and a study on the variety of handwriting styles", *Pattern Recognition Letters*, vol 28 issue 10 pp:1133-1141.

Kuncheva L. I. (2005) "*Combining Pattern Classifiers,*" *Methods and Algorithms*. New York: Wiley.

Minaei-Bidgoli B., Kortemeyer G. & Punch W.F. (2004) "Optimizing Classification Ensembles via a Genetic Algorithm for a Web-based Educational System", (*SSPR /SPR 2004), Lecture Notes in Computer Science (LNCS)*, Volume 3138, Springer-Verlag, ISBN: 3-540-22570-6, pp. 397-406.

Parikh D. and Polikar R. (2007). "An Ensemble-Based Incremental Learning Approach to Data Fusion," IEEE Trans. Syst., Man, Cybern. C, Appl. Rev., vol. 37, no. 2.

Parvin H., Alizadeh H., Minaei-Bidgoli B. & Analoui M. (2008a). "CCHR: Combination of Classifiers using Heuristic Retraining", 4th *Int. Conf. on Networked Computing and advanced Information Management, NCM08*, Korea, pp.302-305, by IEEE CS, ISBN: 978-0-7695-3322-3.

Parvin H., Alizadeh H. & Minaei-Bidgoli B. (2008b). "A New Approach to Improve the Vote-Based Classifier Selection", 4th *Int. Conf. on Networked Computing and advanced Information Management, NCM08*, Korea, pp.302-305, by IEEE CS, ISBN: 978-0-7695-3322-3.

Parvin H., Alizadeh H., Moshki M., Minaei-Bidgoli B. & Mozayani N. (2008c). "Divide & Conquer Classification and Optimization by Genetic Algorithm", *3rd Int. Conf. on Convergence and hybrid Information Technology, ICCIT08*, pp.858-863, ISBN: 978-0-7695-3407-7, by IEEE CS, Korea.

Parvin H., Alizadeh H. & Minaei-Bidgoli B. (2009a). Using Clustering for Generating Diversity in Classifier Ensemble, *International Journal of Digital Content: Technology and its Application, JDCTA,* (*DBLP Indexed*), ISSN: 1975-9339, Vol. 3, Num. 1, pp. 51-57.

Parvin H., Alizadeh H. & Minaei-Bidgoli B. (2009b). A New Method for Constructing Classifier Ensembles, *International Journal of Digital Content: Technology and its Application, JDCTA,* (*DBLP Indexed*), ISSN: 1975-9339.

Parvin H., Alizadeh H. & Minaei-Bidgoli B. (2009c). "Validation Based Modified K-Nearest Neighbor", *Book Chapter in IAENG Transactions on Engineering Technologies, Vol. II– Special Edition of the World Congress on Engineering and Computer Science, AIP Conference Proceedings, Volume 1127, pp. 153-161.*

Parvin H., Alizadeh H., Minaei-Bidgoli B. & Analoui M.(2008d) "A Scalable Method for Improving the Performance of Classifiers in Multiclass Applications by Pairwise Classifiers and GA", 4th Int. Conf. on Networked Computing and advanced Information Management *(NCM 2008)*, Korea, pp.137-142, by IEEE CS, ISBN: 978-0-7695-3322-3.

Roli F., Giacinto G. & Vernazza G. (2001) Methods for designing multiple classifier systems. In J. Kittler and F. Roli, editors, *Proc. 2nd International Workshop on Multiple Classifier Systems*, Vol. 2096 of Lecture Notes in Computer Science, Cambridge, UK, Springer-Verlag, pp. 78–87.

Sanchez A., Alvarez R., Moctezuma J.C. & Sanchez S. (2006) Clustering and Artificial Neural Networks as a Tool to Generate Membership Functions, *Proceedings of the 16th IEEE International Conference on Electronics, Communications and Computers.*

Convergence and Hybrid Information Technologies

Edited by Marius Crisan

ISBN 978-953-307-068-1

Hard cover, 426 pages

Publisher InTech

Published online 01, March, 2010

Published in print edition March, 2010

Starting a journey on the new path of converging information technologies is the aim of the present book. Extended on 27 chapters, the book provides the reader with some leading-edge research results regarding algorithms and information models, software frameworks, multimedia, information security, communication networks, and applications. Information technologies are only at the dawn of a massive transformation and adaptation to the complex demands of the new upcoming information society. It is not possible to achieve a thorough view of the field in one book. Nonetheless, the editor hopes that the book can at least offer the first step into the convergence domain of information technologies, and the reader will find it instructive and stimulating.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen