Proceedings of the 17th World Congress
The International Federation of Automatic Control
Seoul, Korea, July 6-11, 2008

A Novel Multi-Source Classification Approach Based on Evidence Combination

Deqiang Han* Chongzhao Han*
Yi Yang*

**Institute of Integrated Automation, Xi'an Jiaotong University, Xi'an, Shaanxi, 710049 China (Tel: 86-029-82668775; e-mail: digital_king@263.net)*

Abstract: In this paper, a novel classifiers combination approach based on evidence theory is proposed. In classifiers combination, the diversity among member classifiers is known to be a necessary condition for improving classification performance. In our implementation of classifiers combination, we generate the diversity by utilizing different feature spaces to implement member classifiers. And by using different types of classifiers selectively and dynamically according to their expert environments, the diversity can be further enlarged. Thus better combined classification performance can be achieved. In the experiments, the approach proposed shows its efficacy and rationality.

1. INTRODUCTION

Individual classification methods are recently challenged by the approaches of multiple classifiers combination. The combination of several complementary classifiers can improve the performance of individual classifiers (Kittler *et al*., 1998). The multiple classifiers combination has shown wonderful performance in handwriting recognition (Xu *et al*, 1992) and text categorization (Bi *et al*, 2004), etc.

Classifiers combination methods contain two major types: classifiers selection and classifiers fusion (Kuncheva, 2002). The selection algorithms are based on a proposition that each classifier is the expert or the dominator (Kuncheva, 2002) over various local area of the total dataset or various feature sub-spaces of total dataset. Selection algorithms mainly include static selection and dynamic selection. Classifiers fusion supposes that different classifiers are not only competitive but also complementary or cooperative. The fusion algorithms mainly include "Sum" rule, "Product" rule (Kittler *et al*, 1998), majority voting (Franke *et al*, 1992), behavior knowledge space (BKS) (Huang *et al*, 1993), fuzzy integral (Cho *et al*, 1995), decision template (DT) algorithm (Kuncheva, 2002), maximum entropy method based on Bayes frame (Saerens *et al*, 2004), Dempster-Shafer theory (Shafer, 1976) and the methods such as boosting (Schapire, 1999) and bagging (Breiman, 1996).

It is intuitively clear that the combination of identical classifiers will be no better than a single member classifier thereof. If there exists "the best" or "the perfect" member classifier, then no combination is needed (Kuncheva, 2005). Creating diversity among classifiers, i.e., generating complementary member classifiers is one of the keys to success in multiple classifiers combination. But it is hard to create diversity. And the understanding of diversity is still incomplete. How to create and enlarge the diversity relies on specific applications.

In this paper, we implement member classifiers based on different feature spaces of the same dataset. The information provided by different feature spaces may always be complementary. Then the diversity can be created. Different types of classifiers have their own expert application environment. For the classifiers such as neural network and Support Vector Machine (SVM), they are accomplished in constructing the complicated decision planes or boundaries in training set. And they focus on the global information but might always be weak near the decision plane or boundaries, i.e., in the ambiguous area. For the classifiers such as nearest neighbor and *k*-nearest neighbors (*k*-NN), they are adept in discovering the class distribution near the given test sample. And they focus on local area information while do not think much of the global information. Thus, the two different types of classifiers both have their advantages and drawbacks. If we utilize the two types of classifiers selectively and dynamically according to specific environments to implement classifiers combination, their drawbacks may be counteracted and their advantages may be reinforced. Then the diversity can be enlarged further and better classification accuracy can be expected. Based on such ideas and according to Dempster rule of combination, we implement the classifiers combination. Two types of individual classifiers are used including Generalized Regression Neural Network (GRNN) and *k*-NN. The two types of classifiers are chosen dynamically according to their expert environment. In the use of Dempster rule of combination, mass functions are needed, which represent the evidences provided by different information sources. But the generation of mass functions is always difficult. Given a test sample, for the GRNN classifier, we generate mass functions based on the measurement values of the GRNN classifier's output; for the *k*-NN classifier, the mass functions are generated based on class distribution on given test sample's *k* nearest neighbors. Experimental results provided show the novel multi-source classification approach proposed is reasonable and effective.

2. MULTIPLE CLASSIFIERS COMBINATION

Multiple classifiers combination has created a lot of excitement in the machine learning and pattern classification accuracy (Banfield *et al*, 2005). The model of multiple classifiers combination is shown in Fig. 1.

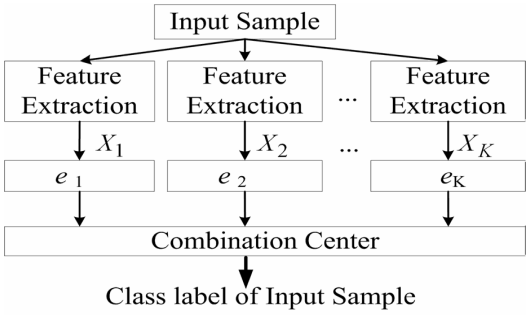

Fig. 1. Model of Multiple Classifiers Combination

In Fig. 1, each e_k is a member classifier, where $k = 1, 2, ..., K$, and K is the total number of member classifiers. $X_1, X_2, ..., X_K$ are the different feature vectors extracted from input sample for each member classifiers. Based on preliminary decision results and output information of member classifiers, the combined decision can be derived by using different fusion algorithms in the decision combination center.

For the member classifiers, given a test sample $\mathbf{x} \subseteq R^d$, there are totally M kinds of possible classes in the class space: $C_1 \cup \ldots \cup C_i \cup \ldots C_M$, where $C_i \ \forall i \in \Lambda$, $\Lambda = \{1, \ldots, M\}$. Based on the output types of member classifiers, the basic output types can be summarized into three categories (Xu *et al*, 1992):

1) Abstract Level: where the classifier only produces a unique class label. Given K classifiers e_k, $k = 1, 2, ..., K$, each classifier assigns a class label j_k to sample $\mathbf{x}$, i.e., $e_k(\mathbf{x}) = j_k$.

2) Ranked Level: where classifiers rank all possible labels in the mutually exclusive set in a sequence $R_k \subseteq \Lambda$ with the label at top being the first choice.

3) Measurement Level: where the classifier attributes to each label a measurement value (such as *a posterior* probability or membership function value) which represents the degree to which a sample $\mathbf{x}$ belongs to that class. For sample $\mathbf{x}$, each member classifier brings out an output vector $Me(k) = [\omega(C_1), \omega(C_2), ..., \omega(C_M)]$.

There are lots of combination algorithms, which can be utilized in combination center. Dempster rule of combination is one of the most effective approaches. In Dempster-Shafer theory (Shafer, 1976; Duan, 1991), elements in the frame of discernment Θ are mutually exclusive. Define $m: 2^{\Theta} \to [0,1]$ as the basic probability assignment (BPA, also called mass function), which satisfies:

$$\sum \{m(A) \mid A \subseteq \Theta\} = 1, m(\varnothing) = 0 \tag{1}$$

Define the belief function and plausibility function as follows:

$$Bel(A) = \sum\nolimits_{B \subseteq A} m(B) \tag{2}$$

$$pl(A) = \sum\nolimits_{A \cap B \neq \varnothing} m(B) \tag{3}$$

Let $m_1, m_2, ..., m_n$ be n mass functions, the new combined evidence can be derived based on Dempster rule of combination as follows ($\oplus$ denotes the orthogonal sum):

$$\begin{aligned} m(A) &= m_1 \oplus m_2 \oplus \cdots \oplus m_n \\ &= \begin{cases} 0, & A = \varnothing \\ \dfrac{\sum\limits_{\bigcap A_i = A} \prod\limits_{1 \le i \le n} m_i(A_i)}{\sum\limits_{\bigcap A_i \neq \varnothing} \prod\limits_{1 \le i \le n} m_i(A_i)}. & A \neq \varnothing \end{cases} \end{aligned} \tag{4}$$

3. MEMBER CLASSIFIERS

Besides the combination algorithms, the type of the member classifier is also important to the multiple classifiers combination. In our research, we adopt two types of classifiers: the GRNN classifier and the *k*-NN classifier.

The GRNN is able to approximate complex nonlinear mappings directly from the input-output data with a simple topological structure. Unlike the BP neural network, which requires a large number of iterations in training to converge to a desired solution, GRNN needs only a single pass of learning (Bhatti *et al.*, 2004). It has advantages such as simplicity, robustness, fast computing speed and optimal approximation. It can approximate any arbitrary function with linear or non-linear relationships between input and output variables. It is widely used in pattern classification, nonlinear curve fitting, etc. The output of the GRNN for an input vector $\mathbf{x} = [x_1, x_2, ..., x_d]^T \in X \subseteq \mathcal{R}^d$ is given by:

$$Y(\mathbf{x}) = \frac{\sum\limits_{k=1}^{n} Y^k \exp\left(-\left\|\mathbf{x} - \mathbf{x}_k\right\|^2 / \sigma^2\right)}{\sum\limits_{k=1}^{n} \exp\left(-\left\|\mathbf{x} - \mathbf{x}_k\right\|^2 / \sigma^2\right)} \tag{5}$$

where d is the dimension of input observation space $X \subseteq \mathcal{R}^d$; T represents the transpose operation of vectors; σ is the kernel width; $\mathbf{x}_k$ and Y^k are inputs and outputs of

training samples. The larger σ is, the smoother the function approximation will be. Too large σ means a lot of neurons will be required to fit a fast changing function. Too small σ means many neurons will be required to fit a smooth function and the network may not generalize well. To fit data very closely, use σ smaller than the typical distance between input vectors (Wasserman, 1993).

For the application of classification, $C = \{C_1, C_2, ..., C_M\}$ is a pattern space, which consists of M mutually exclusive and exhaustive sets of specified patterns (classes). We set M output nodes of GRNN. For training the GRNN classifier, we set the target output of input samples according to their class labels. If a training sample belongs to class C_j, where $j = 1, ..., M$, then its target output will be the vector of $Y_{\text{target}} = [Y(C_1), ..., Y(C_i), ..., Y(C_M)]$, where $Y(C_i) = 0, i \neq j$ and $Y(C_i) = 1, i = j$. For a given test sample, the output of GRNN classifier will be $Y_{\text{out}} = [\omega(C_1), ...\omega(C_i), ..., \omega(C_M)]$, where $\omega(C_i) \in [0,1]$, which can be considered to be the membership function for the given test sample belonging to class C_i. Then the classification decision can be made as follows (τ is the classification decision threshold.):

$$Class\ of\ \mathbf{x} \equiv \max_i(\omega(C_i)) > \tau \tag{6}$$

k-nearest neighbor (*k*-NN) is a nonparametric approach for classification which is simple but effective in many practical applications. It does not need the priori knowledge. It is an instance-based classification method. Various distance functions can be used e.g. the Euclidean and Mahalanobis distance and so on. The result of *k*-NN is a sub optimal, yet popular in practice, nonlinear classifier (Theodoridis *et al*, 2006). We consider the problem of classifying samples into M categories or classes. The set of classes is denoted by $C = \{C_1, C_2, ..., C_M\}$. The available information is assumed to consist of a training set $\tau = \{(\mathbf{x}^{(1)}, C^{(1)}), ..., (\mathbf{x}^{(N)}, C^{(N)})\}$ of N n-dimensional samples $\mathbf{x}^{(j)}, j = 1, ..., N$ and their corresponding class labels $C^{(j)}, j = 1, ..., N$, which takes values in C (We assume for simplicity the class of each training vector to be known definitely). A certain distance function $d(\cdot,\cdot)$ is defined to represent the similarity between samples. Given an unknown feature vector $\mathbf{x}$, then (Theodoridis *et al*, 2006):

1) Out of the N training vectors, identify the k nearest neighbors, irrespective of class label. k is chosen to be odd for a two class problem, and in general not to be a multiple of the number of classes M.

2) Out of these k samples, identify the number of vectors k_i, which belong to class $C_i, i = 1, ..., M$. Obviously: $\sum_i k_i = k$.

3) Assign $\mathbf{x}$ to class C_i with the maximum number k_i of samples. When tie happens, $\mathbf{x}$ can be assigned the class label randomly and it can also be rejected to be classified.

4. NOVEL MULTI-SOURCE CLASSIFICATION APPROACH BASED ON EVIDENCE COMBINATION

To obtain better classification performance, we propose a novel selective classifiers combination approach, which can enlarge the diversity among member classifiers. Based on different feature spaces of the same training dataset, we can design multiple classifiers by utilizing either GRNN or *k*-NN. We utilize Dempster rule of combination to implement classifiers combination, so the mass functions are required. The generation of mass functions is always the most difficult step in practical use of evidence combination. Some approaches (Xu *et al*, 1992; Zhang *et al*, 2002) generate mass functions based on the member classifiers' classification performances. In (Bi *et al*, 2004), the mass functions are constructed based on the output at measurement level. The approaches to generating mass functions are dependent on the output type of member classifiers and specific applications.

4.1 Generating mass functions for GRNN

The GRNN classifier should belong to the classifier at measurement level. Suppose that there are L different feature spaces. Given a test sample $\mathbf{x}$, the output at feature space l ($l = 1, ..., L$) of the GRNN classifier is $Y = [\omega(C_1), ...\omega(C_i), ..., \omega(C_M)]$, where M represents the number of all the possible classes. Given a test sample $\mathbf{x}$, we can generate mass functions as follows (Bi *et al*, 2004):

$$m_l(\{i\}) = \omega_l(C_i) / \sum_{j=1}^{M} \omega_l(C_j) \tag{7}$$

The classification decision of member classifier l based on:

$$class\ of\ x \equiv \max_i m_l(\{i\}) > \tau_l \tag{8}$$

4.2 Generating mass functions for k-NN

In our previous research (Han *et al*, 2007), we proposed a novel mass function generation approach for the *k*-NN classifier. *k*-NN classifier's direct output is just the class label. But when the class label is obtained, for a given sample based on a *k*-NN classifier, we can obtain the class distribution on given sample's *k*-nearest neighbours. Suppose that there are L different feature spaces. Given a sample $\mathbf{x}$, according to feature space l ($l = 1, ..., L$), its feature vector is $\mathbf{x}_l$. Based on

k-NN, in feature space l, we can identify k_l nearest neighbors of $\mathbf{x}_l$, irrespective of class label. Out of these k_l samples, we can identify k_{li} samples belonging to class $C_i, i = 1,...,M$. We can generate BPA as follows (Han *et al*, 2007):

$$m_l(\{C_i\}) = k_{li} / \sum_{j=1}^{M} k_{lj} \tag{9}$$

For example, given a test sample $\mathbf{x}$, on one feature space, the class distribution on its k nearest neighbors ($k = 13$) can be described in Fig. 2. In the k ($k = 13$) nearest neighbors, the class distribution is 7 samples of Class 1, 3 samples of Class 2 and 3 samples of Class 3. In this case, the mass function should be $m(\{1\}) = 7/13;\ m(\{2\}) = 3/13;\ m(\{3\}) = 3/13;$

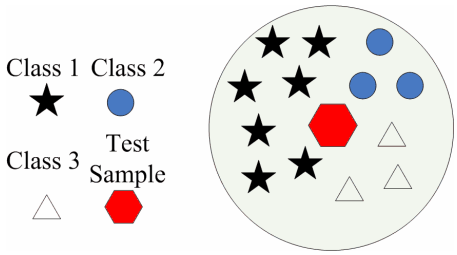

Fig. 2. Class distribution on a test sample's neighborhood

4.3 The member classifiers' selective combination approach

Different types of member classifiers may have their different expert environment. For GRNN classifiers, they are accomplished in constructing the complicated decision plane or boundaries in training samples. And they focus on the global information but they might always be weak near the decision plane, i.e., in the ambiguous area. For k-NN classifiers, they are adept in discovering the class distribution near the given test sample. And they focus on local area information while do not think much of the global information. We utilize the two types of classifiers selectively and dynamically according to specific environments to implement classifiers combination. Their drawbacks may be counteracted and their advantages may be reinforced. The member classifiers' selective combination is implemented as follows:

1) We train L GRNN classifiers according to L different feature spaces. We input all the training samples to the trained GRNN classifiers again to evaluate the classification performance on training samples and to derive all the misclassified training samples. The collection of the misclassified training samples in each feature space is denoted as E_l, where $l = 1,...,L$.

2) For a given test sample, calculate its k_l nearest neighbors in each of the L feature spaces, where $l = 1,...,L$. The collection of the k_l nearest neighbors is denoted as T_l.

3) If element number of $E_l \cap T_l$ is no less than a threshold w_l, we generate mass function for the given test sample based on the approach for k-NN which is described in 4.2, otherwise we generate corresponding mass function based on the approach for GRNN which is described in 4.1.

4) For all the test samples, by repeating the steps from 1 to 3, we generate their corresponding mass functions for each feature space. Then, based on Dempster rule of combination, the combined mass functions for each test sample can be derived. Final combined decision can be made as follows (τ_c is the combined decision threshold.):

$$class\ of\ x \equiv \max_i m(\{i\}) > \tau_c \tag{10}$$

5. EXPERIMENT

The experiment is based on artificial data (3 classes), which has 1200 samples totally (400 samples per class). The artificial dataset has two different feature spaces: feature space I (illustrated in Fig. 3) and feature space II (illustrated in Fig. 4). In feature space I, Class 2 is not linear separable with other two classes. Class 1 and Class 3 are linear separable. In feature space II, Class 3 is not linear separable with other two classes. Class 1 and Class 2 are linear separable. The information for classification provided by the two different feature spaces is complementary. Randomly select 200 samples from each class (totally 600 samples) for training. The remainders (600 samples) are reserved for testing. Two GRNN member classifiers are designed based on feature space I and feature space II respectively. The target outputs of the GRNN classifiers are listed in Table 1.

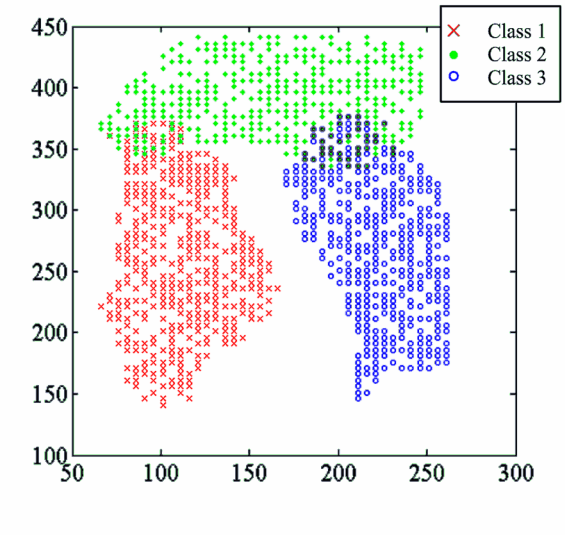

Fig. 3. Feature space I of the artificial dataset

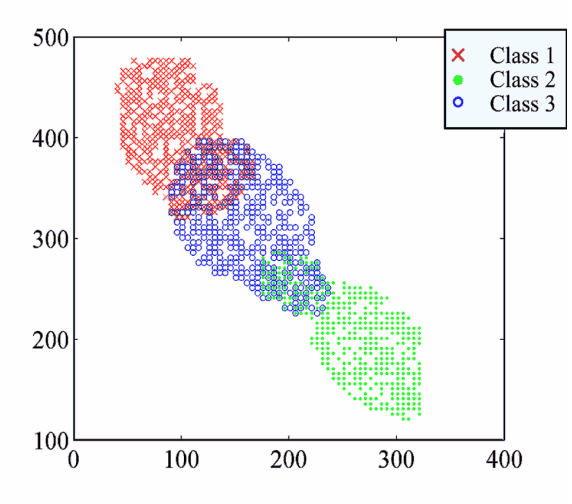

Fig. 4. Feature space II of the artificial dataset

Table 1. Target Outputs

Class	Target output
Red (Class1)	[0.95 0.05 0.05]
Green (Class2)	[0.05 0.95 0.05]
Blue (Class3)	[0.05 0.05 0.95]

The target output values, e.g. for class1, is not set as [1 0 0] but [0.95 0.05 0.05] for convenience of computational convergence. The classification performances on training samples and test samples for the two GRNN classifiers are listed in Table 2. Kernel width are: $\sigma_I = 5$, $\sigma_{II} = 6$. Decision threshold is 0.5. "Total" means "average".

Table 2. GRNN's Classification Accuracy

Training Samples				
Classifier	Class1	Class2	Class3	Total
Feature Space I	98.5%	96%	95.5%	96.7%
Feature Space II	93%	93%	79.5%	88.5%
Test Samples				
Classifier	Class1	Class2	Class3	Total
Feature Space I	92.5%	90%	90.5%	91%
Feature Space II	82%	82.5%	63.4%	76.5%

Two *k*-NN member classifiers are also designed based on feature space I and II respectively. The corresponding classification performances are listed in Table 3. Set $k = 23$. (For three-class classification, $k = 23$, which is not the multiple of 3, cannot bring ties for 3 classes)

Table 3. *k*-NN's Classification Accuracy

Training Samples				
Classifier	Class1	Class2	Class3	Total
Feature Space I	97.5%	96%	91.5%	95%
Feature Space II	92%	89%	69.5%	83.5%
Test Samples				
Classifier	Class1	Class2	Class3	Total
Feature Space I	93%	96%	89.5%	93%
Feature Space II	94%	89.5%	63%	82.2%

Based on mass function generation method introduced in 4.1, by using (4), we can derive the combined GRNN classification accuracy listed in Table 4 (Decision threshold: 0.8).

Table 4. Combined GRNN Classification Accuracy (Test Samples)

Classifier	Class1	Class2	Class3	Total
Combined	95%	97.5%	88.5%	93.7%

Based on mass function generation method introduced in 4.2 and by utilizing (4), we can derive combined *k*-NN classification accuracy listed in Table 5 (Decision threshold: 0.8).

Table 5. Combined *k*-NN Classification Accuracy (Test Samples)

Classifier	Class1	Class2	Class3	Total
Combined	100%	96%	91.5%	95.8%

Base on the novel selective combination approach proposed (set $w_I = 4$ and $w_{II} = 2$), we can derive combined classification accuracy listed in Table 6 (Decision threshold: 0.8). w_I is set greater than w_{II} , for GRNN, the accuracy on feature space I is better than that of feature space II.

Table 6. Selective Combined Classification Accuracy (Test Samples)

Classifier	Class1 wise	Class2 wise	Class3 wise	Total
Selective combined	100%	98.5%	92.5%	97%

From the experimental results, it can be concluded that based on classifiers combination (such as the combined GRNN, the combined *k*-NN, the selective combined method proposed in this paper), the classification accuracy is superior to the best member classifier's accuracy. When we use the selective combined classification method proposed in this paper, classification accuracy is superior to the combined GRNN and the combined *k*-NN classification accuracy. It is because that by using different types of classifiers selectively according to their expert environment, the diversity can be enlarged further. If there exist more diversity among dataset's different feature spaces or there exist effective ways to enlarge the diversity, better accuracy for combined classification can be derived. Assigning the member classifier to its expert environment is an effective idea to further improving the classification accuracy based on classifiers combination.

6. CONCLUSIONS

Diversity among member classifiers is crucial to improving the classification accuracy based on classifiers combination. To achieve better combined classification accuracy, we should enlarge the diversity. In this paper, we implement member classifier based on different feature spaces of the same dataset and propose a novel multiple classifiers combination method based on evidence combination by using different types of member classifiers (*k*-NN and GRNN) selectively and dynamically according to their expert environment. Experimental results show that the approach proposed in this paper is effective and reasonable. The approach proposed also has its drawbacks such as the problem of optimal selection of parameters used. It always depends on experiences and specific application. This work is just a preliminary attempt and lots of further works are still needed.

ACKNOWLEDGMENT

This work is supported in part by the National Natural Science Foundation of China, No.60574033 and the Grant for State Key Program for Basic Research of China (973) No.2007CB31006.

REFERENCES

Banfield, R.E., L.O. Hall, K.W. Bowyer and W.P. Kegelmeyer (2005). Ensemble diversity measures and their application to thinning. *Information fusion*, **Vol. 6**, **Issue 1**, pp. 49-62.

Bhatti, M.W., Y.J. Wang and L. Guan (2004). A neural network approach for human emotion recognition in speech. In: *Proc. Circuits and systems, ISCAS '04*, **Vol. 2**, pp.181-184, Vancouver, Canada.

Bi, Y.X., D. Bell, H. Wang, G.D. Guo, and G. Kieran (2004). Combining multiple classifiers using Dempster's rule of combination for text categorization. In: *Lecture notes in artificial intelligence*, **Vol. 3131**, pp. 127-138, Barcelona, Spain

Breiman, L. (1996). Bagging predictors. *Machine learning*, **No. 26**, pp. 123-140.

Cho, S.B. and J.H. Kim (1995). Combining multiple neural networks by fuzzy integral for robust classification. *IEEE Trans on systems, man and cybernetics*, **Vol. 25**, **Issue 2**, pp. 380-384.

Duan, X.S. (1991). *Evidence theory and decision, artificial intelligence (in Chinese)*, RenMin University of China Press, Beijing.

Franke, J. and E. Mandler (1992). A comparison of two approaches for combining the votes of cooperating classifiers. In: *Proc. 11th IAPR Int'l Conf. Pattern recognition, conf. B: Pattern recognition methodology and systems*, **Vol. 2**, pp.611-614, Hague.

Han, D.Q., C.Z. Han and Y. Yang (2007). Multiple k-NN classifier fusion based on evidence Theory. In: *Proc. IEEE International conference on automation and logistics*, *(ICAL) 2007*, Jinan, Shandong, China.

Huang, Y.S. and C.Y. Suen (1993). The behavior-knowledge space method for combination of multiple classifiers. *Proc. IEEE Conference on computer vision and pattern recognition*, pp. 347-352, New York.

Kittler, J., M. Hatef, R.P.W. Duin and J. Matas (1998). On combining classifiers. *IEEE Trans on pattern analysis and machine intelligence*, **Vol. 20**, **No. 3**, pp. 226-239

Kuncheva, L.I. (2005). Diversity in multiple classifier system. *Information Fusion* **Vol. 6**, **Issue 1,** pp. 3- 4.

Kuncheva, L.I. (2002). Switching between selection and fusion in combining classifiers: An experiment. *IEEE Trans on systems, man and cybernetics*, **Vol. 32**, **No. 2**, pp.146-156.

Saerens, M. and F. Fouss (2004). Yet Another Method for Combining Classifiers Outputs: A Maximum Entropy Approach. In: *Proc. The 5th international workshop on multiple classifier systems*, pp. 82-91, Cagliari, Italy.

Schapire, E.R. (1999). A brief introduction to boosting. *Proc. 16th Int. Joint Conf. AI*, pp.1401-1406, Stockholm.

Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton University Press, Princeton.

Theodoridis, S. and K. Koutroumbas (2006). *Pattern recognition (3rd Edition)*, pp. 48-50, Elsevier, USA.

Wasserman, P.D. (1993). *Advanced methods in neural computing*, pp. 155-161 and pp. 35-55, Van Nostrand Reinhold, New York.

Xu, L., and A. Krzyzak (1992). Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE Trans on systems, man and cybernetics*, **Vol. 22**, **No.3**, pp. 418- 435.

Zhang, B. and S.N. Srihari (2002). Class-wise multi-classifier combination based on Dempster-Shafer theory. In: *Proc. The seventh international conference on control, automation, robotics and vision*, Singapore.