Analysis of Multiple Classifier System Using Product and Modified Product Rules

Mohammed Falih Hassan
Electrical and Computer Engineering Department
Western Michigan University
Kalamazoo, Michigan 49008-5329
Email: Mohammedfalih.hassan@wmich.edu,
mufalh@yahoo.com

Ikhlas Abdel-Qader, Ph.D., P.E. Professor
Electrical and Computer Engineering Department
Western Michigan University
Kalamazoo, Michigan 49008-5329
Email: ikhlas.abdelqader@wmich.edu

Abstract— **One of the key factors in designing a successful multiple classifier system (MCS) is choosing an appropriate combining rule. Many theoretical and experimental efforts have been focused on estimating the probability of classification error for different combining rules. In this work, assuming N classifiers and two independent and identically distributed classes, we investigate using product and modified product rules and derive formulas to estimate their classification error probability under two class distributions, Gaussian and uniform. We also validate our derivations with computer simulations. The performance results of product, modified product, average, and majority vote rules are compared. The comparisons are done in term of probability of classification error as a function of class variance and number of classifiers. The results show that the modified product rule outperforms others while the product rule ranks last.**

Keywords—Multiple classifier systems; Combining rules; probability of classification error.

I. INTRODUCTION

In some applications, single “monolithic” classification algorithm suffers from two issues. The first is when we have small training data size i.e., when the available training data describes a small part of its distribution [1, 2], such a training produces a weak classifier with a large bias error. The second issue is when training a single classifier with large data size results in the opposite type of error called variance error. In some literature the bias and variance errors are called under training and over training errors. Combining Multiple Pattern Classifiers (MPCs) (sometimes called Ensemble Classifiers) are proposed to solve these problems and others as described in [3].

Two of the major research areas in Combining pattern classifiers are the ensemble diversity and methods used to combine the classifiers output. Combining a number of base classifiers that have same knowledge about feature space will not be beneficial. Therefore a level of diversity is needed to improve the ensemble performance [4, 5].

On other hand, combining methods are used to fuse or merge classifiers output for the purpose of improving the system recognition rate. There are two methods to combine classifiers output, one is use class label and another use continuous classifiers output. Many different combining rules are suggested for purpose of improving the ensemble performance [1]. In spite of successful usage of classifier combining in many applications, their improvements in classification lack foundation theory [3]. However many theoretical works are suggested in order to evaluate the interrelated effect of using combining rules on ensemble classification rate [6, 7, 8].

In [6] a closed form of classification error probability is derived for six fusion strategies (minimum, maximum, average, median, majority vote and oracle). Results show that the performance of different combining rule varies significantly and depends on output class posterior probabilities distributions (Gaussian or uniform). While in [7], a performance comparison is made between two combining rules, the sum and majority vote, in two classes problems. Their results show that for independent and identically distributed Gaussian posterior probabilities, sum always outperforms vote. The work presented in [8], gave a theoretical and experimental assessment for weighted average linear combiner. Their results show that the overall performance of classifier ensemble depends on the performance of individual classifiers and the correlation between their outputs.

In this work, we propose a mathematical model that estimates the performance of multiple classifier system using product and modified product rules as a combining methods. Our derivations are based on two class posterior probability distributions, Gaussian and uniform. Then we validate our derivations with computer simulations that show agreement with our theoretical results. We also evaluate the ensemble performance based on product and modified product rules and compare them with other combining rules.

978-1-4799-8802-0/15/$31.00 ©2015 IEEE

II. MATHMATICAL DEFINITIONS

The typical model of MCS is shown in Fig. 1. In this figure, x_k represents the k^{th} feature in space R^n, $x_k \in R^n$. The feature x_k feeds into N classifiers which already trained to recognize new data. Each classifier provides an estimation of the posterior probabilities of M classes $(p_i(\omega_j/x))$, where $i = 1,2,\ldots.,N$ and $j = 1,2,\ldots.,M$, and $p_i(\omega_j/x)$ represents a random variable takes continuous values between [0,1].

L. I. Kuncheva [9] defines a compact form in term of a matrix that contains all posterior probabilities for all classifiers and classes, this matrix called Decision Profile Matrix (DPM).

$$DPM = \begin{bmatrix} p_1(\omega_1/x) & p_1(\omega_2/x) & \ldots & p_1(\omega_M/x) \\ p_2(\omega_1/x) & p_2(\omega_2/x) & \ldots & p_2(\omega_M/x) \\ \vdots & \vdots & \vdots & \vdots \\ p_N(\omega_1/x) & \ldots & \ldots & p_N(\omega_M/x) \end{bmatrix} \qquad (1)$$

Each column in the decision profile matrix represents a support from each classifier to the ω_j class while rows represent the support from a single classifier to M classes.

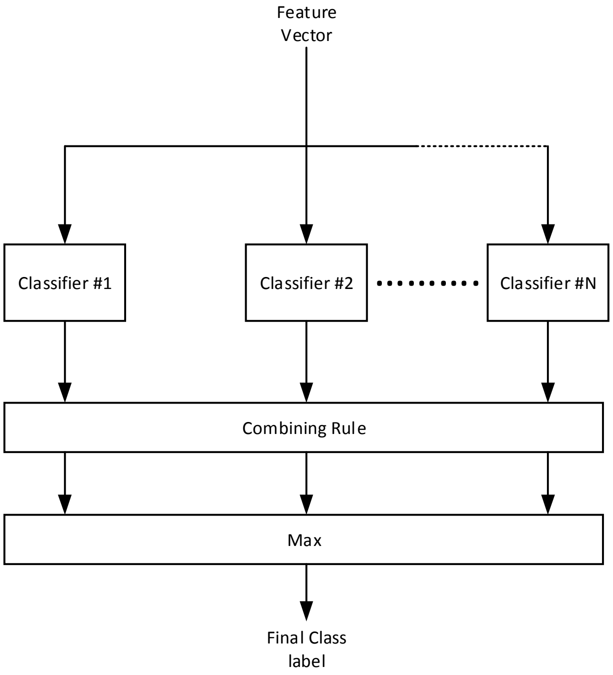

Fig. 1. Typical model for MCS.

Columns of DPM matrix of equation (1) are fed into the combiner or fusion rule which gives an estimate of the overall posterior probability $\hat{p}(\omega_j/x)$ for the class ω_j as:

$$\hat{p}(\omega_j/x) = F(\,p_1(\omega_j/x)\,, p_2(\omega_j/x), \ldots, p_N(\omega_j/x)), \qquad (2)$$

$$where\; j = 1,2,\ldots,M$$

where F is the fusion rule that used to combine base classifiers outputs. The maximum value among $\hat{p}(\omega_j/x)$ is selected to decide to what class a feature vector belongs. In the rest of paper, we assume that each classifier output $p_i(\omega_j/x)$ is a random variable with two classes that have two distributions. One is Gaussian distribution with mean m_g and variance σ_g^2 , and another is uniform distribution with mean m_u and variance σ_u^2 . The subscripts under each symbol are referring to the distribution type.

III. PROBABILITY OF CLASSSIFICATION ERROR

In this section, a closed form expression is derived for classification error probability resulting from using product and modified product rules. Assuming two class distributions (Gaussian and uniform), N base classifiers $(i = 1,2,\ldots,N)$, and two classes $j = 1,2$, we will have $p_i(\omega_1/x) = 1 - p_i(\omega_2/x)$. For the purpose of simplifying expressions, we set;

$$p_i = p_i(\omega_1/x)\; and\; \bar{p}_i = p_i(\omega_2/x) \rightarrow p_i = 1 - \bar{p}_i$$

A. Product Rule

The typical formula of product rule is defined as:

$$\hat{p} = \prod_{i=1}^{N} p_i \qquad (3)$$

where $\hat{p}$ is the overall posterior class probability.

The aim here is to estimate the probability density function for $\hat{p}$ and its moments. By taking the natural logarithm of both sides of (3), the multiplication process between random variables is converted into addition, that means:

$$\log(\hat{p}) = \sum_{i=1}^{N} \log(p_i) \qquad (4)$$

We are not interested in estimating the probability density function of $\log(p_i)$ but rather in its first and second moments. From the statistic theory [10], if there is a function $f(x)$ of random variable X provided that $f(x)$ is differentiable and the moments of X are finite, then the moments of $f(x)$ is approximated as:

$$E[f(X)] \approx f(m_x) + \frac{\ddot{f}(m_x)}{2}\sigma_x^2 \qquad (5)$$

$$VAR[f(X)] \approx (\dot{f}(m_x))^2\sigma_x^2 \qquad (6)$$

where m_x and σ_x^2 are the mean and variance of the random variable X. Definitions $\dot{f}(x)$ and $\ddot{f}(x)$ are the first and second derivatives of $f(x)$ respectively. In the following, we derive two formulas for classification error, one for Gaussian distribution and another for uniform distribution.

All random variables (p_i) are Gaussian, independent and identically distributed, then from (5) and (6), the moments of each one is define as;

$$E[log(p_i)] = log(m_g) - \frac{\sigma_g^2}{2m_g^2} \quad (7)$$

$$VAR[log(p_i)] = \frac{\sigma_g^2}{m_g^2} \quad (8)$$

According to the central limit theorem, the probability density function of the sum of k random variables approaches Gaussian distribution as k become large. Then the resulting mean and variance of the random variable ($\log(p_i)$) are;

$$E[log(\hat{p})] = N[log(m_g) - \frac{\sigma_g^2}{2m_g^2}] \quad (9)$$

$$VAR[log(\hat{p})] = E[(log(\hat{p}) - E[log(\hat{p})])^2]$$

$$= \frac{N\sigma_g^2}{m_g^2} \quad (10)$$

To find the distribution of $\hat{p}$, we take the exponent of both sides of (4). From the probability theory, if X and Y are random variables where Y $= log$ (X) and Y has a Gaussian distribution with mean m and variance σ^2 , then the random variable X has a lognormal distribution with probability density function (pdf) defined as:

$$f_{\hat{p}}(x) = \frac{1}{x\,\sigma\sqrt{2\pi}}\,exp\left[-\frac{1}{2}\left(\frac{log(x)-m}{\sigma}\right)^2\right] \quad (11)$$

$where \;\; 0 < x < \infty ,$ With

$$E[X] = e^{m+\sigma^2/2} \quad (12)$$

$$VAR[X] = e^{2m+\sigma^2}\,(e^{\sigma^2} - 1) \quad (13)$$

The cumulative distribution is given by;

$$F_{\hat{p}}(x) = \Phi\left(\frac{log(x)-m}{\sigma}\right) \quad (14)$$

Then the probability of classification error is calculated when we are uncertain about $\hat{p}$ using:

$$p_e = F(\hat{p} < 0.5)$$

$$= \Phi\left(\frac{log(0.5)-N[log(m_g)-\frac{\sigma_g^2}{2m_g^2}]}{\frac{\sqrt{N}\,\sigma_g}{m_g}}\right) \quad (15)$$

We also assumed that the posterior classifier probabilities have uniform distribution with mean (m_u) and variance $\sigma_u^2 = w^2/3$, where w is the width of the uniform probability density function. The rest of the previous assumptions and derivations also hold for uniform distribution with minor changes. Then the probability of classification error for uniform distribution is can be written as:

$$p_e = F(\hat{p} < 0.5)$$

$$= \Phi\left(\frac{log(0.5)-N\,[log(m_u)-\frac{W^2}{6m_u^2}]}{(\sqrt{N}W)/(\sqrt{3}m_u)}\right) \quad (16)$$

B. Modified Product Rule

A closer look at (9) and (10) reveals that the mean and variance of $\log(p_i)$ grows linearly with N. That means the performance of product rule degrades rapidly with increasing N. If we divide the right side of (4) by N, which makes (9) independent on N as well as reduces the variance as defined in (10) by a factor of $1/N$. Therefore the modified version of product rule becomes:

$$\hat{p} = \left(\prod_{i=1}^{N} p_i\right)^{1/N} \quad (17)$$

Equation (17) is usually referenced as the geometric mean. In parallel steps of the derivations from (4) to (10), we get the following:

$$E[log(\hat{p})] = [log(m_g) - \frac{\sigma_g^2}{2m_g^2}] \quad (18)$$

$$VAR[log(\hat{p})] = \frac{\sigma_g^2}{Nm_g^2} \quad (19)$$

The results of derivations defined in (18) and (19) confirmed our conclusions in the previous section. The probability of classification error for Gaussian distribution is;

$$p_e = \Phi\left(\frac{log(0.5)-[log(m_g)-\frac{\sigma_g^2}{2m_g^2}]}{\frac{\sigma_g}{\sqrt{N}\,m_g}}\right) \quad (20)$$

and for the uniform distribution is:

$$p_e = \Phi\left(\frac{log(0.5)-[log(m_u)-\frac{W^2}{6m_u^2}]}{W/(\sqrt{3N}m_u)}\right) \quad (21)$$

The formulas defined in (15), (16), (20) and (21) are very valuable since it help us in predicting the performance of product and modified product rules versus class mean and class variance as well as understanding the impact of variation the number of base classifiers.

IV. RESULTS AND DISUSSION

In order to validate our derivation for estimating the probability density function (see (11)), we generated a computer program that uses 10 classifiers, each classifier gives an estimate of the posterior probability density (p_i). Each estimated p_i is considered as a random variable with Gaussian distribution that has $m_g = 10$ and σ_g^2=*2*. We implemented the product rule and estimated the overall posterior probability ($\hat{p}$) by multiplying the individual random variable probabilities p_i for each classifier and computed the pdf of the result. Fig. 2 shows the two density functions from the results of the simulation program and from the mathematical derivations (11), the x-axis is normalized for the purpose of clarity. The similarity between the empirical and theoretical results is clearly evident. There are noticeable small difference between the two graphs are expected since equations (5) and (6) that were used in the derivation are just an approximation to exact value.

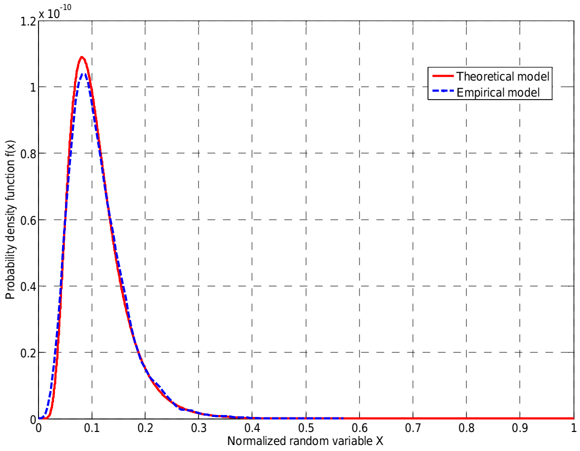

Fig. 2. A comparison of probability density functions of ($\hat{p}$) between theoretical model and computer simulations $m_g = 10, \sigma_g = \sqrt{2}$ and *N=10*.

Another computer experiment is designed to verify the derivation of the probability of classification error as defined in (15). The set up in this experiment is similar to the previous one, except that there are 20 classifiers and each has a Gaussian distribution with $m_g = 1$ and σ_g as a variable. Fig. 3 shows the probability of classification error as a function of σ_g. The figure again shows that the computer simulation and theoretical model are in agreement. The small difference between theoretical and practical results again is due to the approximations made in (5) and (6) as well as the limited data distribution generated by a simulation program. We also attempted to evaluate the performance of the product rule. According to our models, the most influential factors are mean and variance of posterior class probabilities. Fig. 4 displays a two dimensional plot between m_g and σ_g as a function of classification error for 9 classifiers.

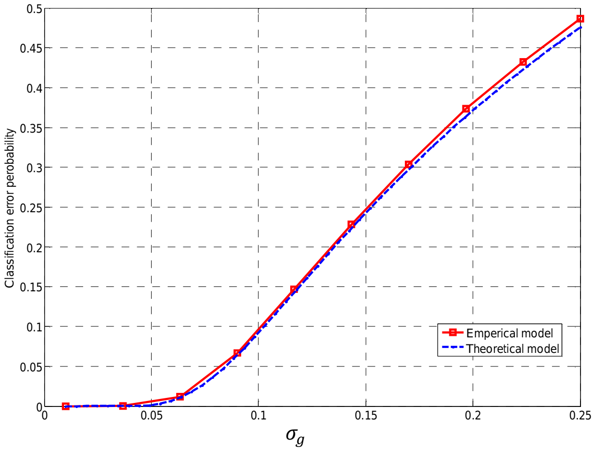

Fig. 3. A comparison in term of the probability of classification error against σ_g between theoretical model and computer simulation, $m_g = 1$ and *N=20*.

As shown the operating characteristic with low classification error probability is limited to a small region { $m_g > 0.93$ & $\sigma_g < 0.1$}. Also a carful investigation on the m_g axis, shows that the abruptly changed at $m_g \approx 0.93$ while exhibit a smooth change on σ_g axis. It is now clear that the behavior of the product rule is very sensitive to changes in m_g and less to σ_g variations. This is due to the fact that the probability density function of $\hat{p}$ (as shown in Fig. 2) is concentrated into small region, therefore a small change in m_g results in a large change in the mean as well as in variance of the random variable $\hat{p}$ (see (12) and (13)).

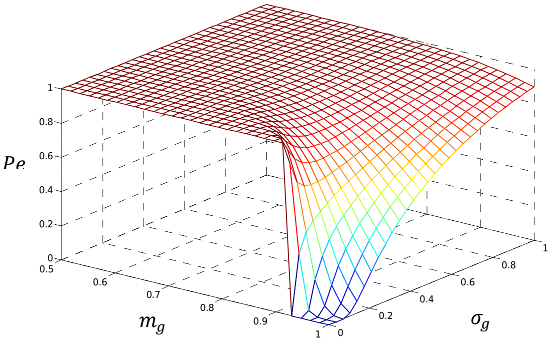

Fig. 4. Probability of classification error of product rule as a function of σ_g , m_g and *N=9*.

Such a sensitive relation can cause a significant degradation or improvement in the system performance abruptly. If we remove the condition that all random variables have same mean and variance, and assign different values to each one, then we should expect a more robust performance. Fig. 5 shows a two dimensional plot for probability of classification error using modified product rule as a function of m_g and σ_g for 9 classifiers. It is clear that modified product rule exhibits better performance than

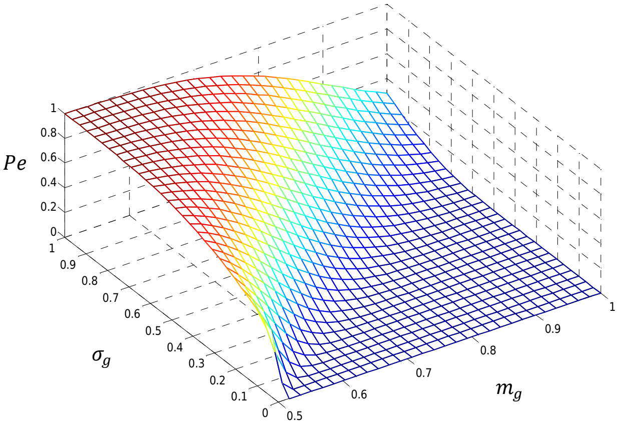

Fig. 5. Probability of classification error for modified product rule as a function of σ_g , m_g and *N=9*.

product rule. Since, it displays a smoother behavior against changes in m_g and σ_g as well as it has a larger region with low classification error compared to the product rule performance. Fig. 6 and Fig. 7 show the performance of the product, modified product, average and majority vote rules in term of classification error as a function of σ_g and w respectively. (Formulas for average and majority vote rules are taken from [6]), for $m_g = m_u = 1\ and\ N = 7$. The comparison included two distributions, Gaussian and uniform. As shown in Fig. 6 and Fig. 7, the product rule exhibits poor performance for σ values of 0.1 and higher, and its overall performance ranked last among others combing rules. While modified product rule outperforms others, notably in uniform distributions. These results were expected, since Fig. 4 suggested that the low classification error region of the product rule is limited to $m > 0.93\ and\ \sigma < 0.1$.

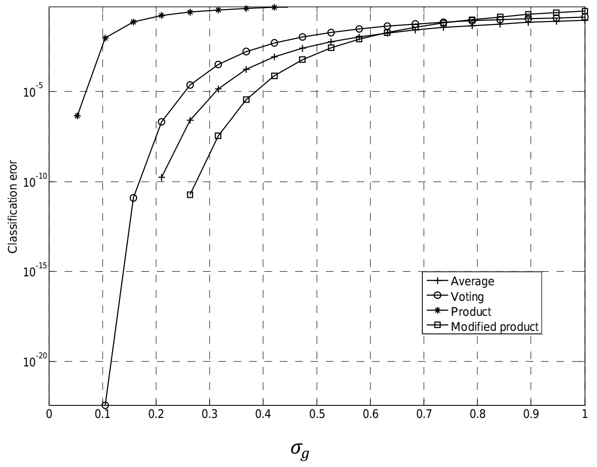

Fig. 6. Classification error for different combining rules as a function of σ_g for Gaussian distribution, $m_g = 1$ and *N=7*.

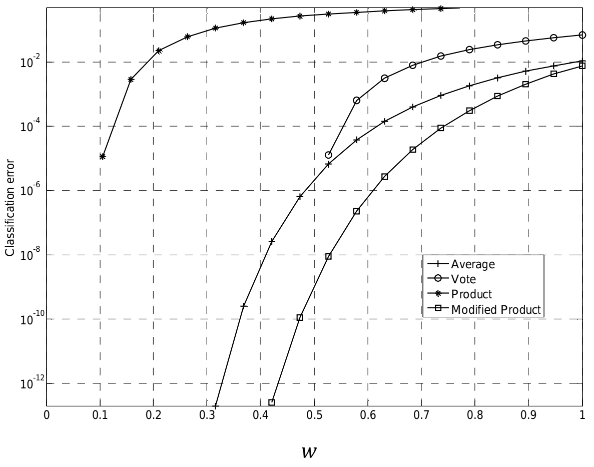

Fig. 7. Classification error for different combining rules as a function of w for uniform distribution, $m_u = 1$ and *N=7*.

Finally, we show in Fig. 8 a comparison of the performance of modified product, average and vote rules as a function of classifier numbers for $m_g = 0.8$ and $\sigma_g = 0.3$. It is clearly evident that the modified product rule gives superior performance compared to others. Product rule is not considered in the comparison because its performance degrades exponentially with the increase in classifiers number. This behavior results from the fact that the total class variance of product rule increases linearly with the increase of the number of classifiers causing exponential performance degradation.

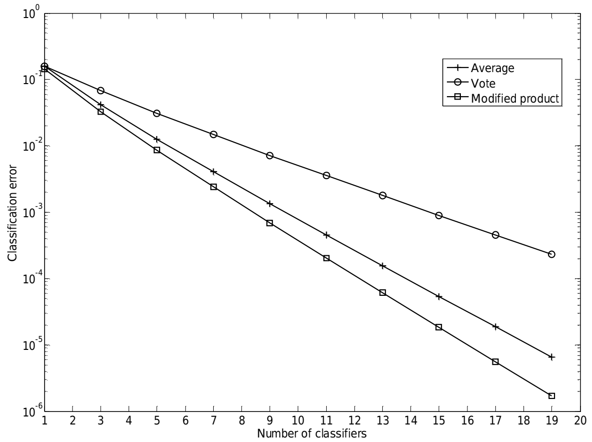

Fig. 8. Classification error for different combining rules as a function of number of classifiers for $m_g = 0.8$ and $\sigma_g = 0.3$.

V. CONCLUSIONS

In this work, assuming N classifiers and two classes, we derived a mathematical model for estimating the classification error probability of ensemble of classifiers operating under product and modified product rules and under the assumption that the posterior class probability is independent and identically distributed. We also considered two class distributions, Gaussian and uniform. Our derivations were verified using computer simulations. The system performance in terms of classification error as a function of mean and variance of posterior class probabilities was investigated. We also addressed the impact of posterior class variance and number of classifiers on the probability of classification error. Results show that product rule ranks last among other combining rules, while the modified product rule outperforms them. Our future work is focused on investigating the performance of MCS using real data to validate the predictions of our theoretical model.

REFERENCES

[1] Polikar, R., "Ensemble based systems in decision making," IEEE Circuits and Systems Magazine, vol. 8, pp. 21-45, 2006.

[2] Wozniak, M., "Combining pattern recognition algorithms chances and limits," Pro. 2nd International Conference on Computer Engineering and Technology (ICCET), pp. 111-115, 2010.

[3] L. I. Kuncheva, Combining Pattern Classifiers: Methods and Algorithms. Second edition, New Jersey, NJ: Wiley, 2014.

[4] Bagheri, M.A. ; Qigang Gao ; Escalera, S., "A Framework towards the Unification of Ensemble Classification Methods, " Proc. 12th International Conference on Machine Learning and Applications (ICMLA), pp. 351 – 355, 2013.

[5] Dara, R.A, Makrehchi, M, Kamel, M.S, "Filter-Based Data Partitioning for Training Multiple Classifier Systems," IEEE Transactions on Knowledge and Data Engineering, Vol. 22, No. 4, 2010.

[6] Kuncheva, L.I, "A theoretical study on six classifier fusion strategies," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, pp.281-286, 2002.

[7] Kittler, J and Alkoot, F.M, "Sum versus vote fusion in multiple classifier systems," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 25, pp.110-115, 2003.

[8] Fumera, G and Roli, F, "A theoretical and experimental analysis of linear combiners for multiple classifier systems," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, pp.924-956, 2005.

[9] L.I. Kuncheva, J.C. Bezdek, and R. Duin, "Decision templates for multiple classifier fusion: An experimental comparison," Pattern Recognition, vol. 34, no. 2, pp. 299–314, 2001.

[10] Haym Benaroya, Seon Mi Han, and Mark Nagurka, Probability Models in Engineering and Science. CRC Press, 2005.