A theoretical analysis of the limits of majority voting errors for multiple classifier systems

Dymitr Ruta and Bogdan Gabrys

Applied Computational Intelligence Research Unit
Division of Computing and Information Systems, University of Paisley,
High Street, Paisley PA1 2BE, United Kingdom
Tel: +44 (0) 141 848 {3284, 3752}; Fax: +44 (0) 141 848 3542
E-mail: {ruta-ci0, gabr-ci0}@paisley.ac.uk

Abstract

A robust character of combining diverse classifiers using a majority voting has recently been illustrated in the pattern recognition literature. Furthermore, negatively correlated classifiers turned out to offer further improvement of the majority voting performance even comparing to the idealised model with independent classifiers. However, negatively correlated classifiers represent a very unlikely situation in the real-world classification problems and their benefits usually remain out of reach. Nevertheless, it is theoretically possible to obtain 0% majority voting error using a finite number of classifiers at the error level lower than 50%. We attempt to show that structuring classifiers into relevant multistage organisations can widen this boundary as well as the limits of majority voting error even more. Introducing discrete error distributions for analysis, we show how majority voting errors and their limits depend on parameters of a multiple classifier system with hardened binary outputs (correct/incorrect). Moreover, we investigate sensitivity of boundary distributions of classifier outputs to small discrepancies modelled by the random changes of votes and propose new more stable patterns of boundary distributions. Finally, we show how organising classifiers into different structures can be used to widen the limits of majority voting errors and how this phenomenon can be effectively exploited.

Running head: Combining classifiers using majority voting

Keywords: Combining classifiers, majority voting, multistage organisations, generalisation, error distribution, margin.

1. Introduction

An increasing scientific effort dedicated to pattern recognition problems is currently directed towards the design and analysis of multiple classifier systems (MCS). For a number of applications combining classifiers has been shown to outperform the traditional, single-best classifier approach [1,2,3,4,5]. Furthermore, there are already strong theoretical foundations explaining potential benefits of combining classifiers [6,7]. The notion of combining evidences as means of improving performance and reliability of a system is not new and has been exploited in a variety of fields. Combining economic forecasts [8] or multi-version software [9] are just two of a large number of examples of combining applications. Especially wide applicability combining has found in the domain of artificial neural networks [1] and classification, where through its direct applications for pattern recognition, the problem is usually referred to as a pattern recognition using multiple classifier system [2].

The main motivation for combining classifiers is based on the assumption that different classifiers using different data representations, different concepts and modelling techniques are likely to arrive at classification results with different patterns of generalisation [10,11,12]. A visible evidence of such multi-dimensional diversity among classifiers takes the form of different occurrences of classification errors for different classifiers reported over a set of input instances. As most combination functions benefit from disagreement to errors of individual classifiers, the greater this disagreement, the lower the impact of individual errors on the final decision and effectively the lower combined classification error.

The level of exploitation of the potential benefits of combining classifiers depends on the combination function applied. One of the simplest combiners operating on binary classification outputs (correct/incorrect) is the majority voting (MV). Due to its simplicity, MV can be applied to classifiers producing different types of outputs as they all can be converted to the uniform binary outputs: 1/0 (correct/incorrect). Applications of MV for pattern recognition have already been studied in detail in [13,14,15,16]. Lam and Suen[14] studied MV performance for both odd and even number of independent classifiers supported by conditions of beneficial addition of one and two classifiers to the MCS. However, all considerations shown in [14] were constrained by a naive assumption of independent errors and complex calculations if the assumption of equal performances is relaxed. Kuncheva et al.[16] uncovered the limits of achievable MV performance. Although shown for a small number of equally performing classifiers the results shown in [16] confirmed an intuitive fact of a large potential improvement of MV performance for the case of negatively correlated classifiers.

This work is concerned with a deeper investigation of the behaviour of majority voting errors and their limits for both independent and dependent classifiers relaxing the assumption of equal performances. Assuming in general no prior knowledge about the data, we base our analysis fully on the binary classifier outputs (correct/incorrect) and a freedom in distributing the outputs for fixed classifier performances. Given the above assumptions, the limits of majority voting errors are investigated in absolute terms ignoring all potential limitations originating from the data or classifiers.

Exploiting the ideas presented in [16,17] we propose to apply discrete and approximated continuous error distributions for an effective description of errors of MV combiner and their limits for large number of voters. For both independent and dependent classifiers we consider a beneficial extension of MCS and for the independent classifiers we derive a simple extendibility condition based on the introduced error distribution. For the dependent system of classifiers we reformulate the limits of majority voting error defined in [16] in terms of the introduced discrete error distributions. Furthermore, it turns out that patterns of boundary error distribution represent very unstable and unreliable solutions, severely degrading a generalisation ability of a system. This instability can be explained in terms of margins of a combined classifier [18], quantitatively expressing the minimal level of changes in the class support outputs needed to misclassify a sample. The system with binary outputs distributed according to the boundary patterns shown in [16] results in the smallest margins, corresponding to the least confidence possible. We propose new more stable patterns of boundary error distribution, which result in the largest margins possible, so that MCS representing such a distribution is likely to have a better generalisation ability.

Another purpose of this paper is to show that the theoretical limits of majority voting errors can be substantially widened or even virtually removed for a large number of classifiers. Namely, we show theoretically that by structuring classifier outputs into relevant multistage organisations, a 100% combined performance can be obtained from a large number of weak classifiers with individual performances close to 0%. In other words, we attempt to show that by structuring classifiers one can reduce the influence of individual errors on the final error of the combined system providing that specific requirements on distribution of classifiers' outputs are fulfilled. Consequently, in such systems, the role of diversity among classifiers increases.

The remaining of this paper is organised as follows. Section 2 provides a general probabilistic analysis of classifier errors and their relation to errors of majority voting under the assumption of independence. Starting from the simplified, strongly constrained Bernoulli model we gradually relax limiting assumptions and show the tools appropriate for dealing with these cases. Finally, we show a simple form of beneficial extendibility condition for MCS, making use of the introduced error distributions. Next section illustrates the nature of majority voting errors and their limits in a realistic dependent MCS. In section 4, we present the idea of structuring classifiers into multistage organisations as a way of widening the limits of MV errors. Section 5 provides the results from enumerative end empirical experiments. Finally, summary and conclusions are given in section 6.

2. Independent multiple classifier system

Independence is a notion correlated to a certain degree with diversity and usually represents a desirable state of MCS [14]. Generally, classifiers are considered independent if they produce independent errors. Under this assumption, classifiers' errors can be modelled by random variables and their distributions. Assuming initially equal probabilities of errors for all classifiers, we start our analysis from the simplest and the most appropriate for this case Bernoulli model.

2.1. Bernoulli model

We consider a system of M independent classifiers: $D=\{D_1,...,D_M\}$, each of which arrives at the correct classification with the same likelihood $1-e$, where e is a probability of error. Under such conditions classifier outputs can be considered as a Bernoulli variable: $X_i=\{[0,1-e_i],[1,e_i]\}$ realising the Bernoulli distribution with parameters: $\mu_i=e_i$ and

$v_i = e_i(1-e_i)$, where μ_i and v_i stand for the mean and variance of the ith variable. Majority voting error for such case can be easily calculated from the Bernoulli's formula:

$$E_{MV}(M) = \sum_{i=k}^{M} \binom{M}{i} e^i (1-e)^{M-i} \qquad k = \lceil M/2 \rceil \tag{1}$$

where $\lceil X \rceil$ is a ceil operator of rounding to the nearest higher integer. It can be proven by induction that provided $e < 0.5$, the function $E_{MV}(M)$ is monotonically decreasing in M for separately odd and even values of M. It means that if classifiers make independent errors of the same level and they more often classify correctly than incorrectly, then the more classifiers used, the lower MV error, with the limit going to zero for infinite M:

$$\lim_{M \to \infty} E_{MV}(M) = 0 \tag{2}$$

2.2. Relaxation of the equal performances assumption

More realistically, different classifiers in general make errors with different probabilities and calculations of the MV error for such a system is no longer as easily tractable. Namely, it requires probability calculations for all combinations of classifiers being in error in MV sense. This process is computationally very complex (complexity of the order of 2^M) and calculations of majority voting errors very quickly tend to be intractable even for small values of M.

Let now assume that each of M independent classifiers arrives at the correct classification with the likelihood $1-e_i$, where e_i is the probability of error produced by the i^{th} $(i = 1,...,M)$ classifier. Looking for possibilities of resolving the problem of calculation intractability, one can notice that each classifier realise the binomial distribution of a Bernoulli variable:

$$X_i = \{[0,1-e_i],[1,e_i]\} \qquad \mu_i = e_i \qquad v_i = e_i(1-e_i) \tag{3}$$

The parameters of this distribution: mean μ_i and variance v_i are additive with respect to the distribution summing. Therefore, the sum of Bernoulli variables of all classifiers forms a new random variable X with binomial distribution, which is no longer Bernoulli distribution:

$$X = \{[0,p_x(0)],[1,p_x(1)],...,[M,p_x(M)]\} \qquad \mu = \sum_{i=1}^{M} e_i \qquad v = \sum_{i=1}^{M} e_i(1-e_i) \tag{4}$$

where $p_X(i)$ refers to the probability of observing exactly i errors at the output of MCS. The mean μ and the variance v become important parameters of MCS. For simplicity, we refer to such a discrete distribution of a random variable X as a discrete error distribution (DED). The majority voting error, given the DED, can be shortly represented as:

$$E_{MV} = \sum_{i=k}^{M} p_X(i) \qquad k = \lceil M/2 \rceil \tag{5}$$

According to the central limit theorem [22,23], a sum of a large number of binomial distributions converges to the normal distribution with the same parameters μ and v shown in (4). Thus, for a large number of classifiers, the DED can be approximated by the normal distribution defined algebraically as:

$$X_A(x) = \frac{1}{\sqrt{2\pi v}} exp\left(\frac{-(x-\mu)^2}{2v} \right) \quad x = \{0,1,...,M\} \tag{6}$$

A calculation of such a distribution is incomparably faster than the respective DED. Not being constrained by computational complexity any longer, one can analyse different phenomena concerning the combination of many independent classifiers. An example of the approximated DED is shown in Figure 1. To gain independence of the distribution shape with respect to different discrete numbers of classifiers, one can apply a normalised continuous error distribution (CED) holding the same integration properties as the original DED with the assumption of applicability of the normal distribution approximation. The continuous error distribution can be obtained by scaling the parameters of approximated DED given by (4) according to the following formulas:

$$\mu_N = \frac{1}{M}\sum_{i=1}^{M} e_i \qquad v_N = \frac{1}{M^2}\sum_{i=1}^{M} e_i(1-e_i) \tag{7}$$

Normalised continuous error distribution of the form:

$$X_{NA}(x) = \frac{1}{\sqrt{2\pi v_N}} exp\left(\frac{-(x-\mu_N)^2}{2v_N} \right) \tag{8}$$

can be used as an elegant tool for visualisation and interpretation of majority voting errors behaviour for a large number of classifiers. An example of the CED is shown in Figure 2. For a system with a large number of classifiers for which the CED can be approximated by (8), the error of majority voting can be calculated by simple integration:

$$E_{MV} = \int_{k/M}^{1} X_{NA}(x)dx \tag{9}$$

which pertains to the MV error (4) for the original DED. The precision of calculating a majority voting error using (9) increases with an increasing number of classifiers, and its value depends only on the parameters of distribution X_{NA}. The normalised mean μ_N, denoting simply the mean classifier error plays an important role, as it defines the position of the normal distribution peak. Adding worse performing classifiers causes an unfavourable shift of the CED towards the threshold of MV error. However, assuming similar performances of combined classifiers, μ_N does not change substantially. In such a case, the MV error can be reduced only due to decreasing normalised variance v_N with an increasing number of classifiers as shown in Figure 3. It conforms to the Bernoulli's law of large numbers [22,23]. However, as a consequence of the normalisation (7), the decreasing variance is only beneficial for $\mu_N < 0.5$, which is again the same condition of decreasing majority voting error as for the Bernoulli's model discussed in Section 2.1. An interesting question remaining is to which extent worse performances of classifiers are acceptable to keep reducing errors of MV by adding new classifiers to the MCS. Essentially, it is a question of beneficial extendibility of multiple classifier system.

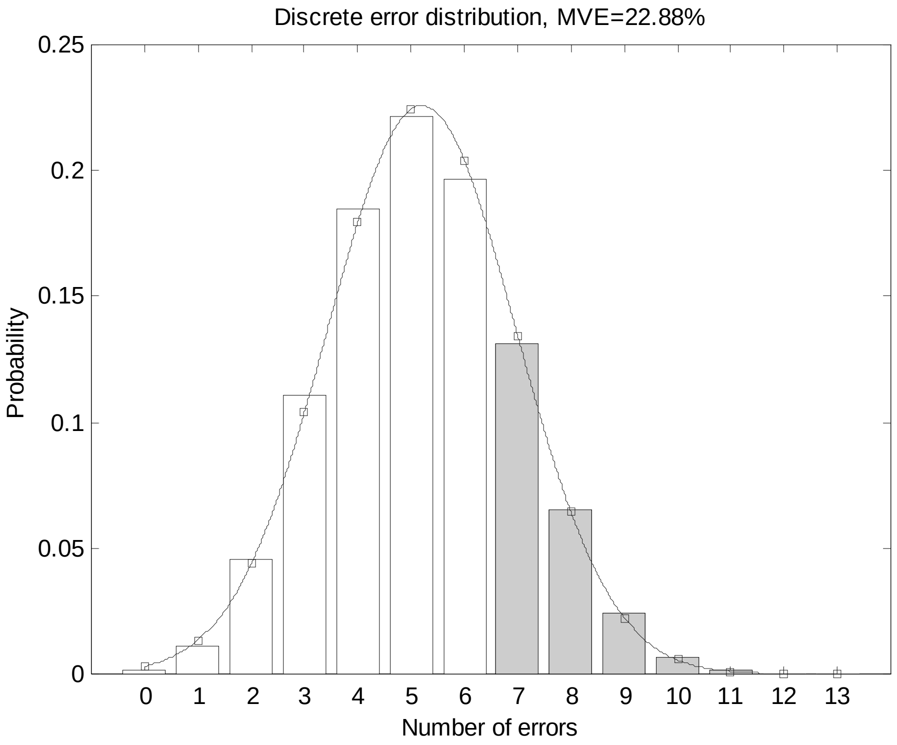


FIGURE 1. Discrete error distribution with normal distribution approximation. Parameters: 13 independent classifiers with 40% error each. Shaded bars refer to errors in majority voting sense. The MV error rate corresponds to the sum of all shaded bars.

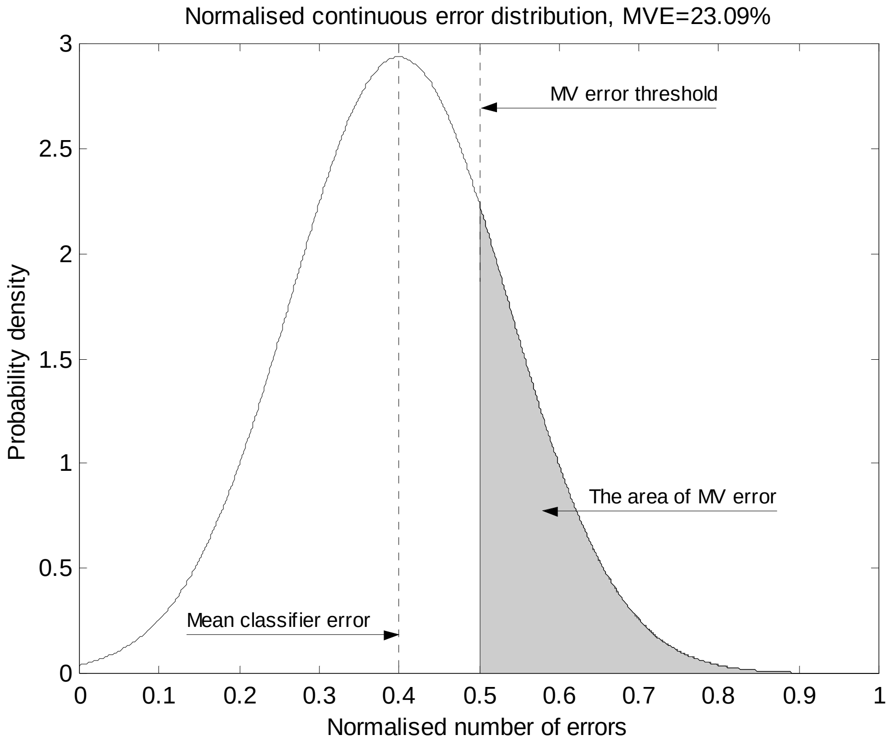


FIGURE 2. Normalised continuous error distribution. Parameters: 13 independent classifiers with 40% error each. Shaded area refers to majority voting error rate.

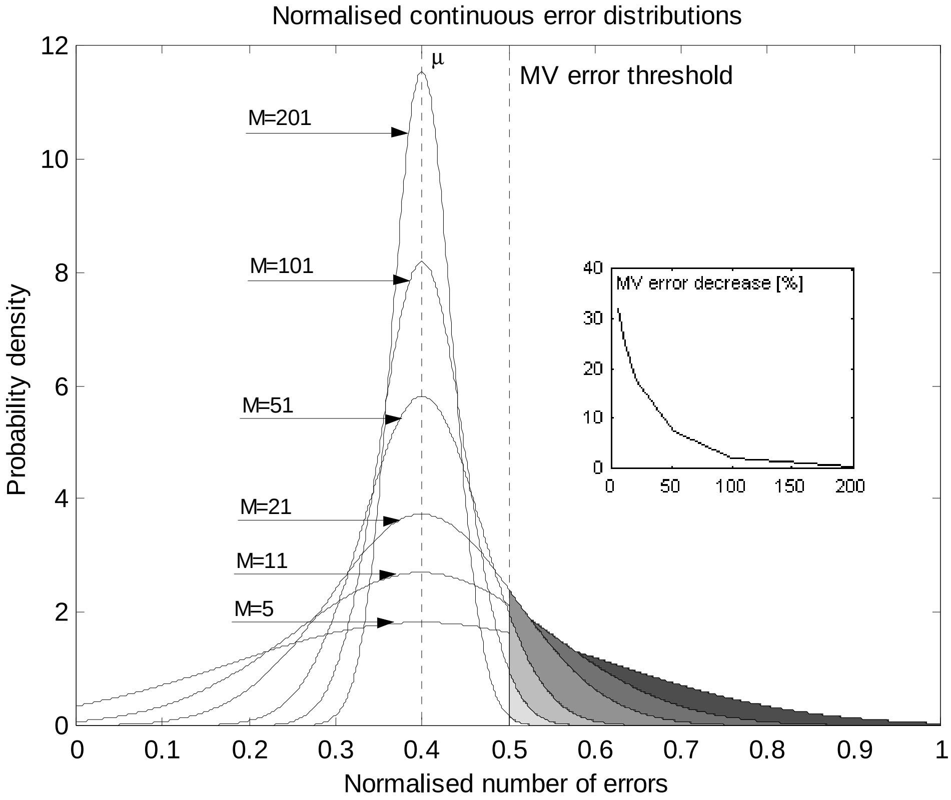


FIGURE 3. A family of normalised continuous error distributions for increasing number of classifiers with the same individual errors of 40%. Decreasing shaded area corresponds to reducing majority vote error. The relation between this error reduction and the number of independent classifiers is shown in the subfigure.

2.3. Extendibility condition

Extendibility of an independent MCS is an important issue from the perspective of classifier selection, as it potentially allows applying an effective selection algorithm. The problem can be generally formulated as: what properties should a classifier or a group of classifiers have, which if added to the MCS, would cause a reduction of MV error. The problem due to its discrete nature needs to be considered within a discrete platform of analysis. With MV combiner applied, it makes sense to keep only odd numbers of classifiers. Thus extending MCS at a single step should concern exactly two classifiers. Given MCS with M classifiers such a system can be described by the DED X defined by (4) as:

$$X = \{[0, p_X(0)], [1, p_X(1)], ..., [M, p_X(M)]\} \tag{10}$$

The objective is to reduce the error of majority voting given by (5) by adding a pair of classifiers described by the DED Y:

$$Y = \{[0, p_Y(0)], [1, p_Y(1)], [2, p_Y(2)]\} \tag{11}$$

Let Z denote the error distribution of the extended MCS:

$$Z = \{[0, p_Z(0)], [1, p_Z(1)], ..., [M+2, p_Z(M+2)]\} \tag{12}$$

The probabilities $p_Z(i)$, $i = 1,...,M+2$ can be derived from known values of $p_X(j)$, $j = 0,1,...,M$ and $p_Y(q)$, $q = 0,1,2$ as shown in the following relation:

$$p_Z(i) = \sum_{\substack{j=0,\dots,M \\ q=0,1,2 \\ j+q=i}} p_X(j) p_Y(q) \tag{13}$$

Given $k = \lceil M/2 \rceil$ and (5), the likelihood of the extended MCS being in MV error is then:

$$E_{MV}(M+2) = \sum_{i=k+1}^{M+2} p_Z(i) = \sum_{i=k+1}^{M+2} \sum_{\substack{j=0,\dots,M \\ q=0,1,2 \\ j+q=i}} p_X(j) p_Y(q) \tag{14}$$

Making use of the normalisation: $\sum_{i=0}^{2} p_Y(i) = 1$ equation (14) can be rewritten as:

$$E_{MV}(M+2) = p_X(k-1) p_Y(2) + p_X(k)[p_Y(1) + p_Y(2)] + \sum_{i=k+2}^{M} p_X(i) \tag{15}$$

Placing (10) into (15) leads to the following:

$$E_{MV}(M+2) = E_{MV}(M) + p_X(k-1) p_Y(2) - p_X(k) p_Y(0) \tag{16}$$

The extendibility condition can be expressed as a positive reduction of MV error, thus:

$$E_{MV}(M) - E_{MV}(M+2) = p_X(k-1) p_Y(2) - p_X(k) p_Y(0) > 0 \tag{17}$$

which after reordering takes the following concise form of the extendibility condition:

$$\frac{p_Y(2)}{p_Y(0)} < \frac{p_X(k)}{p_X(k-1)} \tag{18}$$

It is interesting that we did not make any assumptions concerning the level of error made by any classifier. This proves the general character of the above extendibility condition. More specifically, denoting errors of a pair of classifiers to be added by: e_{M+1}, e_{M+2} we can express (18) directly by these errors as in the following:

$$e_{M+2} < \frac{1 - e_{M+1}}{1 + e_{M+1}[p_X(k-1) - p_X(k)]/p_X(k)} \tag{19}$$

Example: Extending MCS with $M = 1$ classifier

Let the error of a classifier in the MCS be e_1. The extendibility condition (17) can be rewritten for this case as follows:

$$e_3 < \frac{1 - e_2}{1 + e_2(1 - 2e_1)/e_1} \tag{20}$$

Figure 4 shows a contour plot of MV error as a function of both errors e_2 and e_3, with e_1 fixed at the level of 0.1. The vectors illustrate a negative gradient of the majority voting errors pointing in the directions of its maximum decrease. The curled lines corresponding to the extendibility conditions, depict another interesting phenomenon shown in Figure 4. Namely, provided the extendibility condition is met and assuming a fixed mean error rate of a pair of classifiers, the reduction of the majority voting error will be the highest for maximally different performances i.e. it is better to add a pair of classifiers with error rates

$e_2 = 0.3, e_3 = 0.1$ than $e_2 = 0.2, e_3 = 0.2$. Moreover, the greater the difference in the classifier performances, the more the change of the majority voting error is influenced only by the better performing one. This is especially evident for low error rates e_1. Figure 5 shows how the boundary of beneficial extension of MCS changes for different values of e_1. Triangular areas limited by dashed lines and corresponding extension curves represent the areas of beneficial extension for the realistic case when added classifiers both perform worse than the actual classifier to be extended. For increasing error e_1, the area of beneficial extension considerably decreases. It means that it is increasingly difficult to beneficially extend MCS by two classifiers with errors $e_2, e_3 \geq e_1$. For $e_1 > 0.5$ it is not possible for any e_2, e_3.

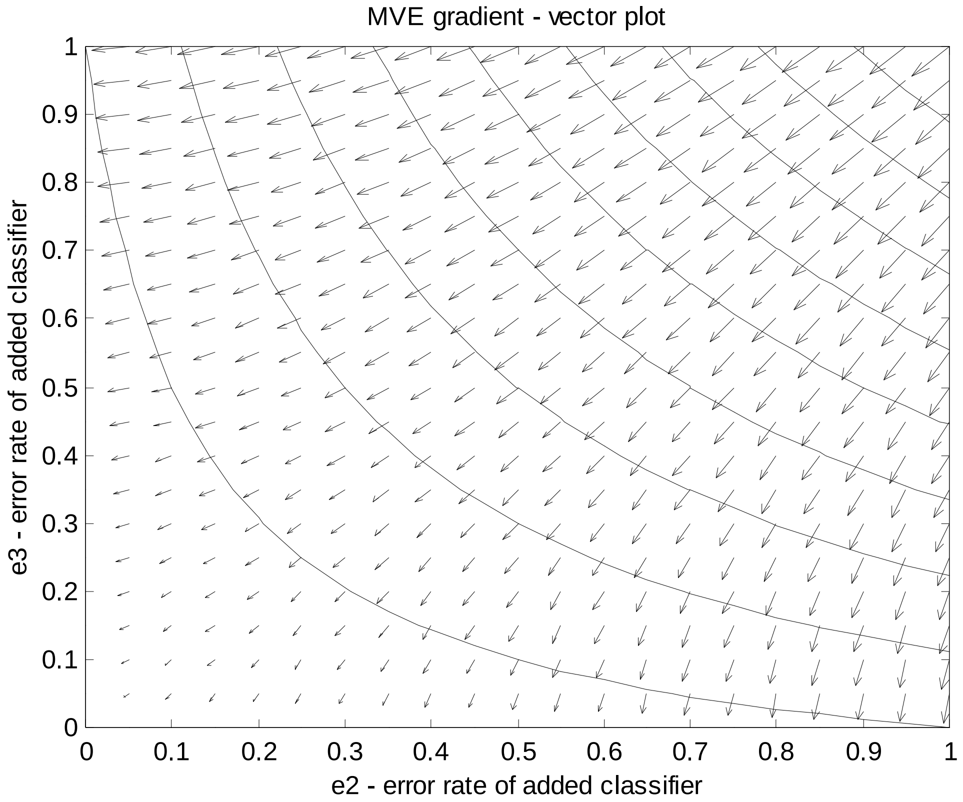


FIGURE 4. Majority voting error as a function of individual errors of a joining pair of independent classifiers e_2, e_3. Parameters: MCS with 1 classifier at error level $e_1 = 10\%$. Solid lines correspond to the same combined MV error out of all 3 classifiers. The vectors visualising gradient of the combined error, point in the directions of the maximal reduction of MV error.

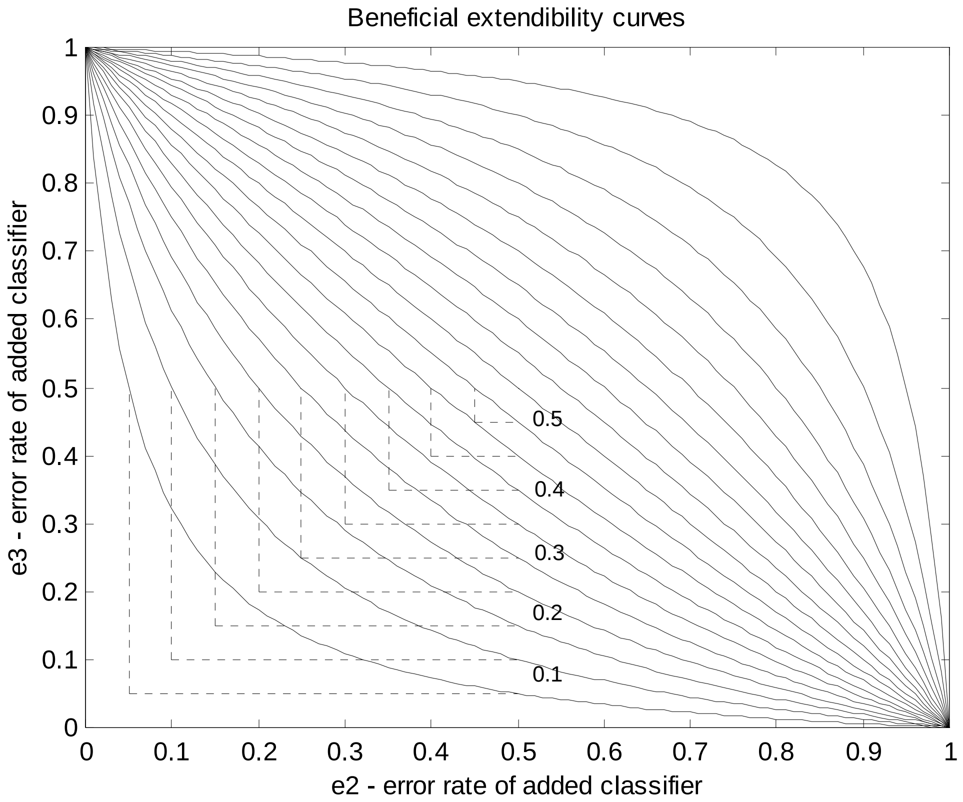


FIGURE 5. Extendibility curves for different errors of a single classifier. Dashed lines limit the area corresponding to individual errors of joining classifiers e_2, e_3 greater than error e_1 but producing MV error lower than e_1.

3. Dependent Multiple Classifier System

Error independence in MCS is a far too deep simplification of the real-world classification problems. It is true that classifiers act independently but their errors are usually strongly dependent through the common character of the data they are trained on. Unavoidable error correlations will be observed, which reflects the fact that for hard cases or outliers, classifiers tend to agree on errors. The realistic dependence among classifiers can be captured by the discrete error distribution, shape of which is diametrically different than for corresponding independent classifiers. As shown in Figure 6-A,B,C for 3 different datasets, the DED for realistic dependent classifiers tends to decay sharply (Figure 6-A,B) but not completely with increasing error level, or in even worse scenario, shown in Figure 6-C, the DED is raising slowly for higher error levels. Strong dependency is evident in the form of increasing number of cases for which majority or all classifiers produce errors. Figures 6-D,E,F show corresponding discrete error distributions under the assumption of error independence. The visible difference is that for independent classifiers the probability mass of DED is more concentrated around the mean classifier forming the Gaussian-like shape, whereas the dependence of classifiers is reflected by a polarisation of the probability mass present mostly at both ends of the DED. In terms of majority voting error the independence offers huge improvement comparing to realistic dependent classifiers. The question arises: does dependence among classifiers always cause harmful effects in terms of combining classifiers? As shown in [16,17], there are two forms of the dependency: negative and positive. Only the latter causes dramatic loss of the performance comparing to independent system. The negative correlation, if achieved, leads to further reduction of MV error even comparing to the independent MCS.

An important assumption made at this point and the consecutive analysis is that the only information available is given in the form of binary classifier outputs reported for the validation dataset. Any other prior information related to the data or classifiers is not available. Such assumptions allow considering the limits of majority voting error in absolute terms subject only to the distribution of the outputs among the samples. It means for instance that even though the data hold intrinsic limitations like noise or Bayes error, this information is not available at the abstract level of our considerations and we assume the absolute limits of majority voting errors as potentially achievable.

To have a clear picture of what can be gained and lost by introducing dependence into MCS, the boundary distributions of classifier outputs have to be investigated.

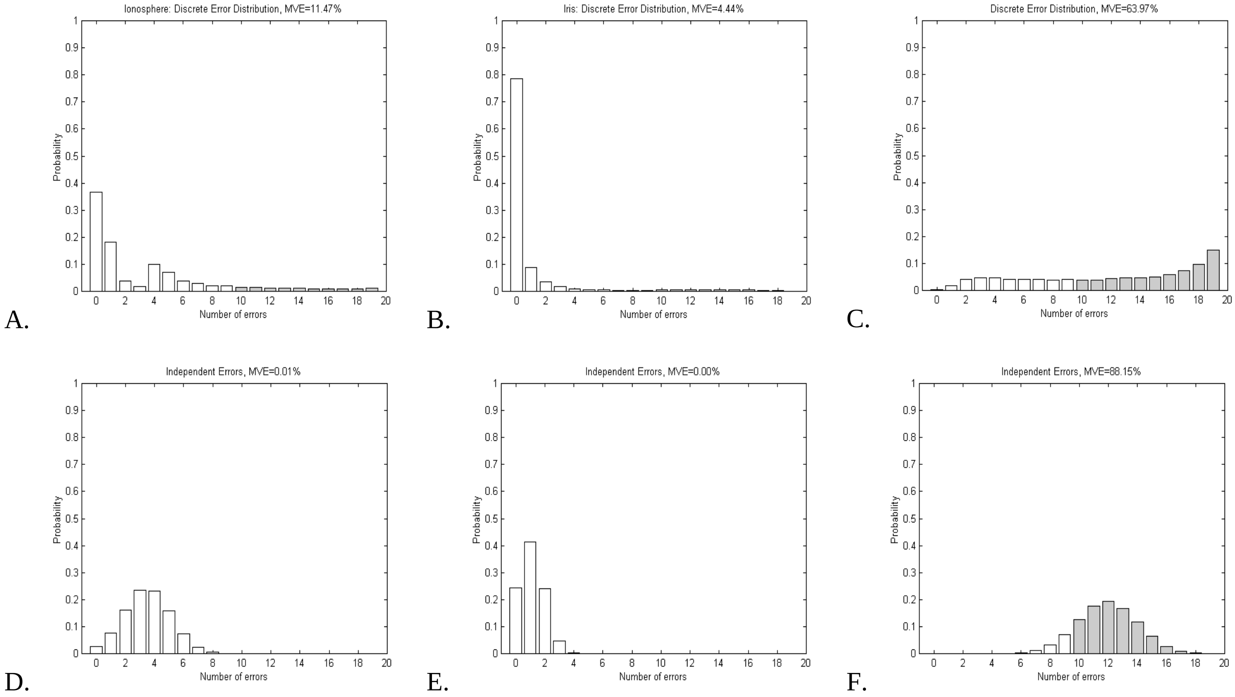


FIGURE 6. Discrete error distributions for 3 realistic datasets: A: iris (ELENA Database), B: *biomed* (Tumour diagnosis) and C.: *chromo* (UCI Repository) together with the equivalent error distributions D,E,F respectively under the assumptions of independence. 19 different classifiers available in PRTOOLS 3.0 (http://www.ph.tn.tudelft.nl/~bob/PRTOOLS.html) have been used for classification. Shaded bars correspond to errors in majority voting sense.

3.1. The patterns of boundary distributions of errors

Having a pool of M (odd number) trained classifiers: $D = D_1,...,D_M$, described by a discrete error distribution: $X = \{[0, p_X(0)],[1, p_X(1)],...,[M, p_X(M)]\}$. Kuncheva et al.[16] defined specific boundary distributions of classifier outputs corresponding to the pattern of success and pattern of failure. These patterns represent cases, for which correct votes are maximally (minimally) exploited to produce the lowest (highest) MV errors. MV is a specific fusion method, for which the most successful combination of classifier outputs is not the one with all correct outputs but the combination containing minimal number of correct outputs sufficient for correct MV. The best distribution of classifier outputs for a single input sample in MV sense is the one with the presence of exactly $k = \lceil M/2 \rceil$ correct outputs and $k-1$ errors, which is the maximum number of individual misclassifications that can be reduced for a single data entry. Similarly, the worst case is not the one with all errors but the combination of k errors and $k-1$ correct votes, which are completely wasted. Applying these extreme partitions of votes for many input samples leads to the formulation of pattern of success and failure.

Pattern of success (PS)

As mentioned above the most efficient partition of votes assumes the presence of exactly k correct votes and $k-1$ errors. Correct outputs are for such case maximally exploited while errors maximally reduced. Denoting sum of all mean classifier errors by $s_e = \sum_{i=1}^{M} e_i$, where e_i is a mean error rate of the i^{th} classifier, one can consider two cases of optimal distribution strategy.

For $s_e <= k-1$, (Figure 7-A,B) all individual classifier errors have to be distributed at the most efficient DED level of $k-1$ errors. All remaining input samples should indicate 0 errors. To calculate partition between the two levels l_1 and l_2 of DED the following set of two equations with two unknowns has to be solved:

$$\begin{cases} p_X(l_1) + p_X(l_2) = 1 \\ p_X(l_1)l_1 + p_X(l_2)l_2 = s_e \end{cases} \tag{21}$$

Writing (21) for levels 0 and $k-1$ gives the following solution:

$$\begin{cases} p_X(0) = (k-1-s_e)/(k-1) \\ p_X(k-1) = s_e/(k-1) \end{cases} \tag{22}$$

For $s_e > k-1$, (Figure 7-C) not all individual classifier errors can be covered by $k-1$ level of DED and for some entries the number of errors produced by MCS has to be greater than $k-1$. To sacrifice as few data entries as possible, it is optimal to keep the excess of errors at the highest DED level of M errors. Solving again (19) for DED levels of: $k-1$ and M errors gives the following partitions between these two levels:

$$\begin{cases} p_X(k-1) = (M-s_e)/(M-k+1) \\ p_X(M) = (s_e-k+1)(M-k+1) \end{cases} \tag{23}$$

A complete definition of PS can be formulated as follows:

Definition 1. Let $X = \{[0, p_X(0)],[1, p_X(1)],\ldots,[M, p_X(M)]\}$ be a DED describing MCS with M classifiers. Let s_e denote the sum of all mean classifier errors: $s_e = \sum_{i=1}^{M} e_i$ and $k = \lceil M/2 \rceil$. The distribution X is called a pattern of success producing the smallest possible MV error if the following holds:

$$\begin{cases} s_e <= k-1 \\ p_X(0) = (k-1-s_e)/(k-1) \\ p_X(k-1) = s_e/(k-1) \\ p_X(i) = 0 \quad i \neq \{0, k-1\} \end{cases} \vee \begin{cases} s_e > k-1 \\ p_X(k-1) = (M-s_e)/(M-k+1) \\ p_X(M) = (s_e-k+1)/(M-k+1) \\ p_X(i) = 0 \quad i \neq \{k-1, M\} \end{cases} \tag{24}$$

Pattern of failure (PF)

As mentioned above the worst partition of votes for one data entry assumes the presence of exactly $k-1$ correct votes and k errors. The correct outputs are for such a case maximally wasted whilst errors maximally exploited. Keeping the same denotations as for pattern of success one can similarly consider two cases of the worst distribution strategy.

For $s_e <= k$, (Figure 7-D,E) all individual classifier errors have to be distributed at the most inefficient DED level of k errors. All remaining input samples contribute to the level of 0 errors. Solution of (19) for the levels k and 0 gives the following partition:

$$\begin{cases} p_X(0) = (k-s_e)/k \\ p_X(k) = s_e/k \end{cases} \tag{25}$$

For $s_e > k$, (Figure 7-F) not all individual classifier errors can be covered by the k^{th} error level of DED and for some entries, the number of errors produced by MCS has to be greater than k. To waste as few entries as possible, it is optimal to keep the excess of errors at the highest DED level of M errors. Distribution of votes between these two levels obtained from the solution of (21) gives the following result:

$$\begin{cases} p_X(\mathrm{k}) = p_X(\mathrm{k}) = (M - s_e)/(\mathrm{M}-\mathrm{k}) \\ p_X(\mathrm{M}) = p_X(\mathrm{M}) = (s_e - \mathrm{k})/(\mathrm{M}-\mathrm{k}) \end{cases} \tag{26}$$

A complete definition of PF can be formulated as follows:

Definition 2. Let $X = \{[0, p_X(0)], [1, p_X(1)], ..., [M, p_X(M)]\}$ be a DED describing MCS with M classifiers. Let s_e denote the sum of all mean classifier errors: $s_e = \sum_{i=1}^{M} e_i$ and $k = \lceil M/2 \rceil$. Distribution X is called a pattern of failure producing the largest possible MV error if the following holds:

$$\begin{cases} s_e <= k \\ p_X(0) = (k-s_e)/k \\ p_X(k) = s_e/k \\ p_X(i) = 0 \quad i \neq \{0, k\} \end{cases} \vee \begin{cases} s_e > k \\ p_X(k) = (M - s_e)/(M-k) \\ p_X(M) = (s_e-k)/(M-k) \\ p_X(i) = 0 \quad i \neq \{k, M\} \end{cases} \tag{27}$$

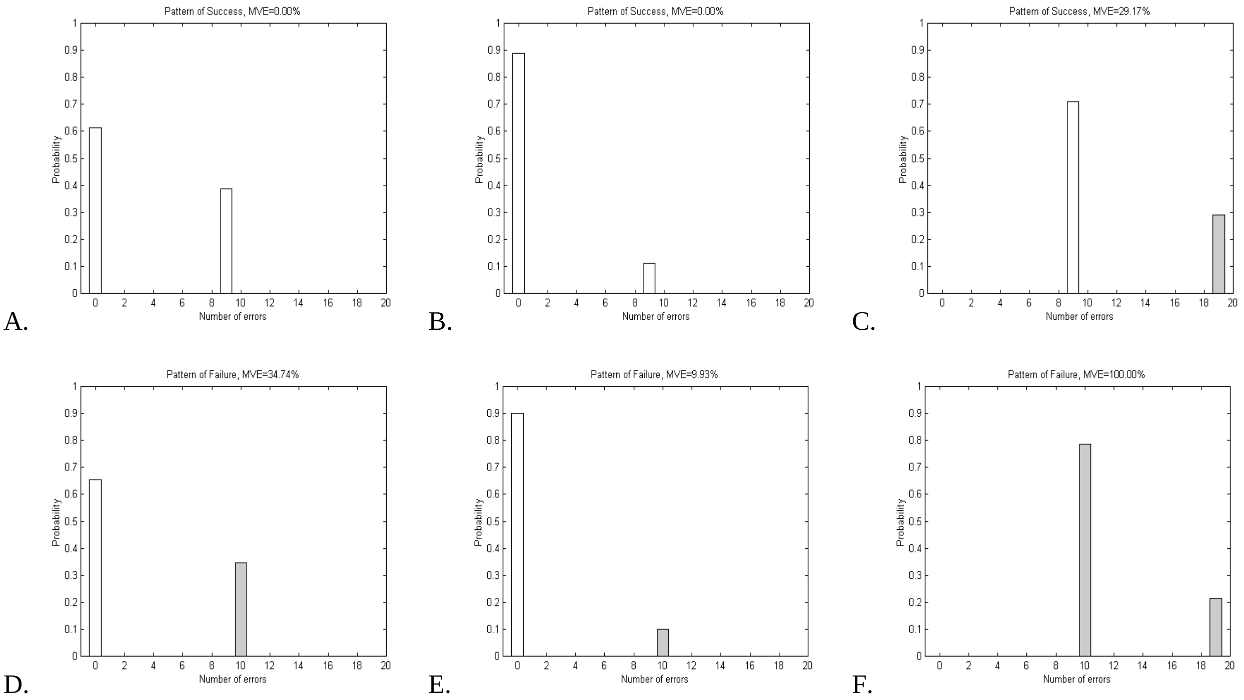


FIGURE 7. Patterns of Success: A,B,C and Failure: D,E,F for the corresponding datasets presented in Figure 6.

3.2. The patterns of stable boundary distribution of errors

Any potential practical usability of the boundary distributions of outputs would have to face small differences in individual errors and the distributions of outputs expected to occur between validation and testing set. This could lead to the situation where limits of majority voting errors corresponding to these boundary distributions of outputs measured for the validation set are no longer the limits for the testing set. This stability of majority voting performance is usually associated with the concept of margin [18] representing the difference between the weight assigned to the correct label and the maximal weight assigned to the single incorrect label. Taking the values within the range (-1,1) the margin represents directly a measure of stability or confidence of the classification output. Values of the margin close to the unit in both positive and negative direction correspond to greater confidence and stability of the output being either correct or incorrect. The margins close to 0 reflect the highest confusion about the output observed when two or more highest ranked classes show similar scores.

In our case as only binary information (correct/incorrect label) is available, the margin for a single example can be associated with a difference between correct and incorrect votes. The margin d defined that way would take the integral values between $-M$ and M with a step 2, or after normalisation through division by M :

$$d = \left\{ -1, \frac{2-M}{M}, \frac{4-M}{M}, \ldots, -\frac{1}{M}, \frac{1}{M}, \ldots, \frac{M-2}{M}, 1 \right\} \tag{28}$$

As it appears from (28) for considered in this work odd number of classifiers, the minimal margins are $\pm 1/M$ and they correspond to the most unstable state where just a single flip of an output may change majority voting output to the opposite. This is exactly the case of patterns of success and failure discussed in the previous section.

As a response we define new patterns of distribution holding the same values of error limits of majority voting shown for pattern of success and failure in previous sections, but keeping the margins for each sample as large as possible for the fixed individual classifier error rates. In this strategy, the priority is to distribute errors as far from MV decision boundary as possible, but still under constrain of preserving the limits of majority voting errors derived from PS and PF.

Stable pattern of success (SPS)

Let us consider the binary outputs given by M classifiers. Keeping the same denotations as in previous sections, for $s_e = \sum_{i=1}^{M} e_i$ and $k = \lceil M/2 \rceil$, the following two cases have to be considered:

For $s_e <= k-1$, (Figure 8-A,B) the stable pattern of success can be achieved by distributing all individual classifier errors at the lowest levels of DED, for which all samples are misclassified by up to $k-1$ classifiers. There are in general two levels of DED, at which these requirements can be fulfilled: $\lceil s_e \rceil - 1$ and $\lceil s_e \rceil$. Distribution of errors between these two levels of DED obtained from the solution of (19) takes the following form:

$$\begin{cases} p_X(\lceil s_e \rceil - 1) = \lceil s_e \rceil - s_e \\ p_X(\lceil s_e \rceil) = s_e - \lceil s_e \rceil + 1 \end{cases} \tag{29}$$

In this strategy, the highest priority is to keep all samples at the level of DED as far to the left from MV decision boundary of $k-1$ errors as possible. Comparing to the original pattern of success, for which a part of samples has been misclassified by exactly $k-1$ classifiers, and the rest of data was classified correctly by all classifiers, here more samples (or provided $s_e >= 1$ all samples) are misclassified by some classifiers but also fewer samples are misclassified at the boundary level of $k-1$ errors. Effectively in such a distribution, the lowest margins of some samples are increased for the price of reduction of large positive margins of another samples.

For $s_e > k-1$ (Figure 8-C), SPS remains consistent with PS and is defined by (23).

Full definition of SPS can be thus formulated as follows:

Definition 3. Let $X = \{[0, p_X(0)], \ldots, [M, p_X(M)]\}$ be a DED describing MCS with M classifiers. Let $s_e = \sum_{i=1}^{M} e_i$ denote the sum of all mean classifier error rates and $k = \lceil M/2 \rceil$. Distribution X is called the stable pattern of success if the following is true:

$$\begin{cases} s_e <= k\text{ - }1 \\ p_X(\lceil s_e \rceil - 1) = \lceil s_e \rceil - s_e \\ p_X(\lceil s_e \rceil) = s_e - \lceil s_e \rceil + 1 \\ p_X(i) = 0 \quad i \neq \{\lceil s_e \rceil - 1, \lceil s_e \rceil\} \end{cases} \vee \begin{cases} s_e > k\text{ - }1 \\ p_X(k-1) = (M - s_e)/(M - k + 1) \\ p_X(M) = (s_e - k + 1)/(M - k + 1) \\ p_X(i) = 0 \quad i \neq \{k-1, M\} \end{cases} \tag{30}$$

Stable pattern of failure (SPF)

Similarly to the stable pattern of success, two cases depending on the value of s_e are considered.

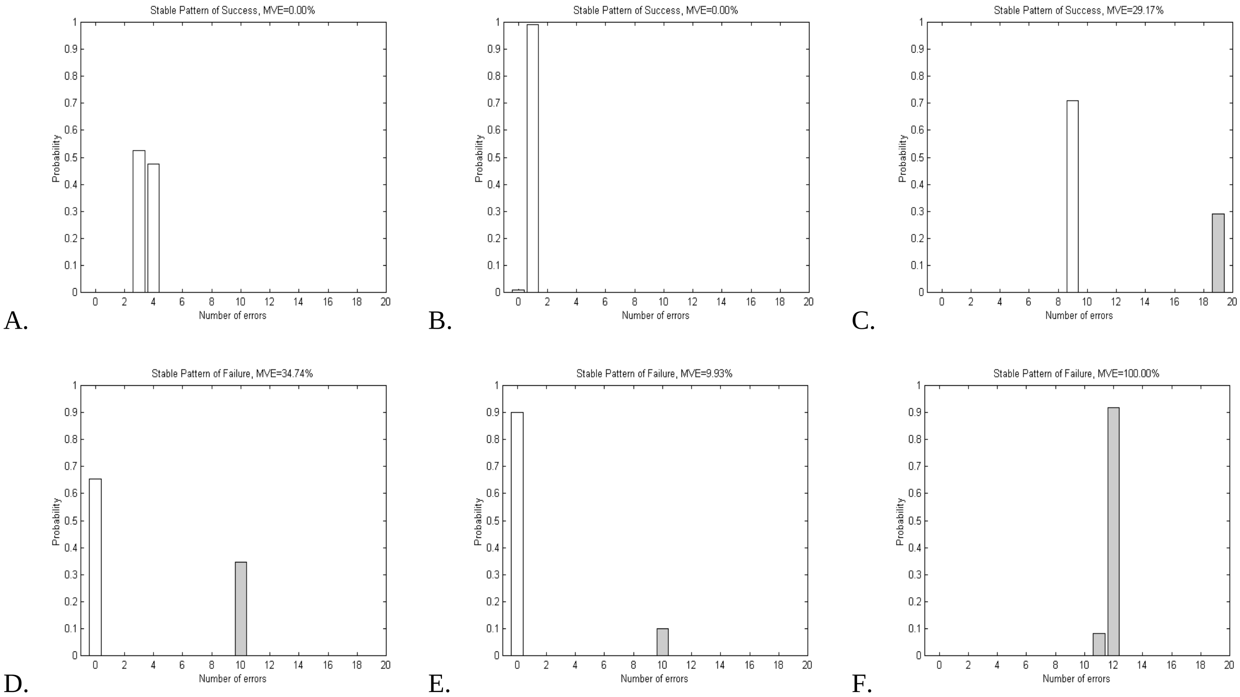
For $s_e <= k$, (Figure 8-D,E) SPF remains consistent with PF and is defined by (25).

For $s_e > k$, (Figure 8-F) the priority of SPF is to keep all the samples at the level of DED as far to the right from MV decision boundary of k errors as possible. This can be achieved by distributing all individual classifier errors at the highest levels of DED, for which all samples are misclassified by at least $\lceil M/2 \rceil$ classifiers. There are two levels of DED, at which these requirements can be fulfilled: $\lceil s_e \rceil - 1$ and $\lceil s_e \rceil$, and this remains consistent with the case of SPS for $s_e <= k$. Partition between these two levels of DED is therefore defined by (26).

SPF can be fully defined as follows:

Definition 4. Let $X = \{[0, p_X(0)], \ldots, [M, p_X(M)]\}$ be a DED describing MCS with M classifiers. Let $s_e = \sum_{i=1}^{M} e_i$ denote the sum of all mean classifier error rates and $k = \lceil M/2 \rceil$. Distribution X is called the stable pattern of failure if the following is true:

$$\begin{cases} s_e <= k \\ p_X(0) = (k - s_e)/k \\ p_X(k) = s_e/k \\ p_X(i) = 0 \quad i \neq \{0, k\} \end{cases} \vee \begin{cases} s_e > k \\ p_X(\lceil s_e \rceil - 1) = \lceil s_e \rceil - s_e \\ p_X(\lceil s_e \rceil) = s_e - \lceil s_e \rceil + 1 \\ p_X(i) = 0 \quad i \neq \{\lceil s_e \rceil - 1, \lceil s_e \rceil\} \end{cases} \tag{31}$$



A. B. C.

D. E. F.

FIGURE 8. Stable Patterns of Success: A,B,C and Failure: D,E,F for the corresponding datasets presented in Figure 6.

3.3. The limits of majority voting error

Both patterns of boundary error distributions PS (PF) and SPS (SPF) provide the same limits of majority voting error. Let us emphasise once more, that these error distributions are very distant from realistic distributions. This is reflected in consecutive Figures 6,7,8. In Figure 6 the realistic distributions of errors are shown for 3 real-world datasets using 15 different classifiers and confronted with their corresponding independent equivalents. The following two figures present visualisation of error distributions corresponding to patterns of success and failure, and their stable alternatives. Using (5) and definitions introduced in section 3.1 and 3.2, the limits of majority voting error can be expressed directly as a function of a sum of individual errors s_e and summarised in the following form:

$$E_{MV}^{min} = max\left\{0, \frac{s_e - k + 1}{M - k + 1}\right\} \qquad E_{MV}^{max} = min\left\{\frac{s_e}{k}, 1\right\} \tag{32}$$

where as before: $s_e = \sum_{i=1}^{M} e_i$ and $k = \lceil M/2 \rceil$. Figure 9-A presents a 3-D plot of the limits of majority voting errors together with independent MV, shown as a function of both: mean classifier error and the number of classifiers. The 2-D projections for the selected numbers of classifiers are shown in Figure 9-B. First to be noticed is the fact, that the problem of MV errors is symmetrical with respect to the middle point $(0.5, 0.5)$, which reflects equivalence between error distributions for small mean error e to the distribution of correct outputs under large e. Another conclusion from this symmetry is that the pattern of failure can be extracted from the pattern of success applied for opposite classier outputs and vice versa. Curves for independent classifiers can be interpreted as boundaries between negative and positive dependence among classifier outputs. Using this terminology, as it appears from Figure 9, even for positive dependence majority voting offers substantial improvement of performance provided the mean error rate is lower than half. However, the challenge is to target the area under the curve of independence corresponding to negative dependence for which as the limits show the improvement may be dramatic.

A decrease of majority voting errors for the negatively dependent classifiers is large even comparing to the independent classifiers. Quite interesting is the fact that having only 3 classifiers with 33% error each, applying majority voting may lead to the complete reduction of error. For a larger number of classifiers one can still obtain a total reduction of combined errors using classifiers with individual errors up to 50% for the number of classifiers M going to infinity. Moreover, even having many extremely poor voters at the error rate of 90%, if they are negatively dependent, majority voting can still decrease the error by up to 10%, while independence implies combined error close to 100%. Another conclusion coming from Figure 9 is that the more classifiers in MCS the harder it is to achieve negative dependence among the classifiers as it represents highly unlikely statistical state.

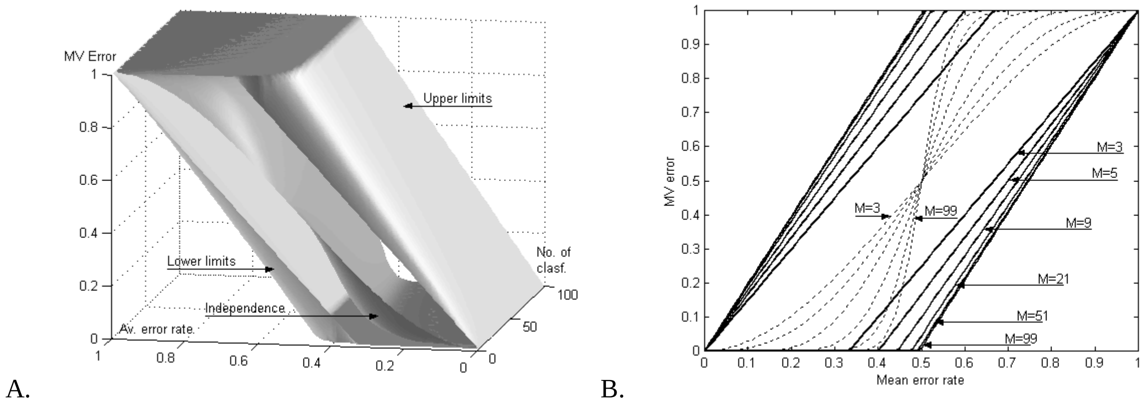


FIGURE 9. Majority voting error limits shown as a function of the number of classifiers and the mean classifier error. A. 3D plot including surface corresponding to MV errors for independent classifiers. B. 2D projections for the selected numbers of classifiers.

4. Multistage organizations

The analysis of the majority voting limits presented so far has been based on the assumption that classifier outputs are combined in parallel at single step. However, this is in fact the simplest structure in a general class of multistage organizations where majority voting can be applied separately for groups of outputs on different levels. There is wide evidence of multistage classification in the literature [19,20]. Good example is a system proposed by Wolpert[19] where the classifiers at consecutive stage were trained on the outputs from the classifiers at previous stage. In our case, as assumed in previous sections the binary classifier outputs represent the only information available. Majority voting applied for the binary outputs (correct/incorrect) produces single output exactly of the same type and meaning. That way perceived majority voting acts as an aggregation operator which can be in general applied to groups of outputs separately on each layer and produce outputs that can be further grouped and combined in the same way in the next layer. This type of systems was only briefly mentioned by Ho and Hull[21] but to the best knowledge of the authors multistage organisations have not been treated in any other publication. Such system is in fact concerned with two processes of selection and combining following each other. In the first stage, the outputs have to be organised in groups and majority voting is applied for each group separately producing single binary outputs, forming the next layer. In the next and each consecutive layer exactly the same way of grouping and combining are applied with the only difference that the number of outputs in the next layer is reduced to the number of groups formed in the current layer. This repetitive process is continued until the final single output of the system is reached. Here, we refer to such a system as a multistage organisation with majority voting (MOMV) since the decision at each consecutive node is given by majority voting. Figure 10 shows the example of MOMV for 15 classifiers.

Assuming odd number of votes as a requirement for majority voting such a system becomes highly constrained but this does not affect the generality of analysis related to the error limits of MOMV considered in this section.

For further simplicity, we constrain the system by the following assumptions:

1. At all layers only groups with odd number of members are valid
2. At each layer all groups contain the same number of members.

3. Individual error rates of classifiers are the same and equal to e

Using these assumptions the structure of MOMV with M classifiers, denoted by S_M can be defined as:

$$S_M = (g_1, g_2, ..., g_L) \quad M = \prod_{i=1}^{L} g_i \tag{33}$$

where g_i, $i = \{1,...,L\}$ denotes the size of groups at i^{th} layer. In addition to the structure, a unique MOMV should also contain information about initial allocation of outputs to different groups. Assuming ordered grouping of outputs according to their places at each layer, this information is fully contained by the permutation of outputs P_M at the first layer. The unique MOMV with M classifiers can be thus formally defined by a pair of the structure S_M and the permutation P_M as follows:

$$MOMV = (S_M, P_M) \tag{34}$$

Using definitions presented above, an example of MOMV shown in Figure 10 is thus described by the structure: $S_{15} = (5,3)$ and $P_{15} = (3,9,5,7,1,4,15,6,2,11,14,8,10,13,12)$.

Let us now consider the error limits of such a system. In this case there are three factors conditioning the final performance: the distribution of outputs, the structure, and the permutation. All can be considered individually to pursue their boundary properties in terms of the error limits of MOMV.

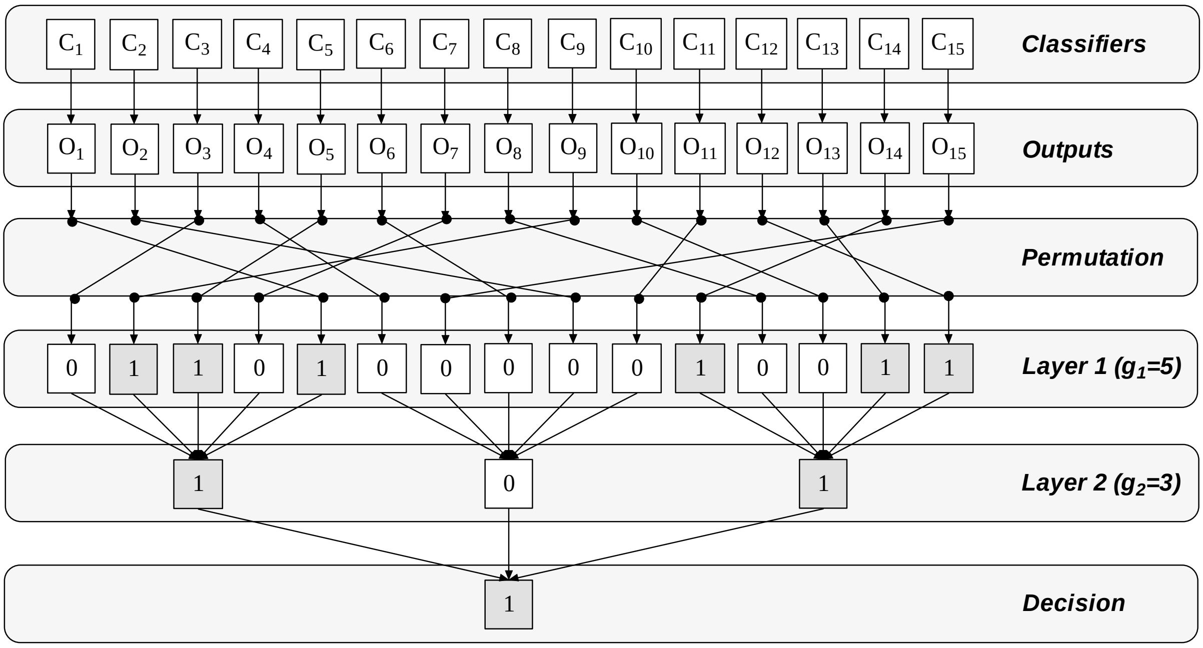


FIGURE 10. The scheme of the multistage organisation system with 15 classifiers, structure $S_{15} = (5,3)$ and permutation $P_{15} = (3,9,5,7,1,4,15,6,2,11,14,8,10,13,12)$. Trained classifiers $C_1,...,C_{15}$ produce binary outputs $O_1,...,O_{15}$ (1-correct, 0-incorrect), which after permutation P_{15} are subsequently grouped and aggregated by majority voting at each layer according to the structure S_{15}. The final output is reached at the decision layer.

4.1. Optimal distribution of outputs for MOMV

Given a system (S_M, P_M) with a fixed structure $S_M = (g_1, ..., g_L)$ and permutation P_M the objective is to find such distribution of errors so that the final error of the whole system is minimal. Notably, within each group regardless of the layer, the standard majority voting is applied in parallel. Thus, locally within each group, the boundary distributions of outputs discussed in section 3.1 and 3.2 are fully applicable. Specifically, they apply to the group of g_L outputs located at the top layer L, at which the final output of MOMV is produced. From the PS definition (24) applied for this layer, exactly $\lceil g_L/2 \rceil$ correct outputs is a minimum sufficient for producing correct output of a system. Following the bottom-up propagation of the pattern of success applied for each group leads to the optimal distribution of outputs within the structure. Examples of such distribution are shown in both Figure 10 and Figure 11 for the structures $S_{15} = (5,3)$ and $S_{27} = (3,3,3)$ respectively. Note that for instance each correct vote at the second layer in Figure 10 derives from a group of $M' = 5$ outputs at the first layer, exactly $\lceil M'/2 \rceil = 3$ of which are correct. The same applies to all layers grater than one in both figures. Generally one can say that each correct vote at layer $L-k$ should be the majority voting output from a group of outputs at layer $L-k-1$, for which locally we observe the outputs distributed according to PS. Simultaneously, no correct output can be wasted, resulting in the incorrect group voting output, so ideally each incorrect output at layer $L-k$ should be the majority voting output from a group of all incorrect outputs at layer $L-k-1$. This second condition is also held throughout the structures in Figure 10 and Figure 11. The optimal distribution of outputs in Figure 11 proves that only 8 correct outputs out of 27 are sufficient to produce correct system output for the structure $S_{27} = (3,3,3)$. This is large difference comparing to the equivalent value of 14 correct outputs needed to produce correct vote in the standard parallel structure $S_{27} = (27)$. By symmetry, it can be also shown that for the same system, only 8 errors may result in the error of the whole MOMV, which would produce correct vote for structure $S_{27} = (27)$. Essentially, by applying the structure $S_{27} = (3,3,3)$ instead of $S_{27} = (27)$ for 27 classifiers, the most efficient correct (incorrect) MOMV is observed for a greater (smaller) number of incorrect votes. Effectively, the limits of errors produced by structured system are expected to be stretched apart comparing to the equivalent parallel system. Extending considerations to a general case of S_M is a more complex problem. However, it could be noticed that the top-down decomposition of MV outputs gives the minimal number of correct votes k_S, for which the whole system is still correct. This can be given by the following formula:

$$k_S = \prod_{i=1}^{L} \lceil g_i/2 \rceil \tag{35}$$

This agrees with the example of a system shown in Figures 10 and 11, where for the given structures: $S_{15} = (5,3)$ and $S_{27} = (3,3,3)$, exactly $k_S = 6 = 3 \cdot 2$ and $k_S = 8 = 2 \cdot 2 \cdot 2$ correct votes respectively are the minimum sufficient to produce the correct output from the system.

Generally, it means that for a given structure, up to $M - k_S$ individual classifier errors can be reduced by applying MOMV. Similarly it could be shown that only k_S individual errors is sufficient to produce incorrect output of MOMV, which implies that up to $M - k_S$ correct individual outputs can still result in error of MOMV. This phenomenon becomes quite striking for a large number of classifiers, e.g. for $M = 243$, where introducing the structure $\mathrm{S}_{243} = (3,3,3,3,3)$ may potentially lead to the reduction of 211 errors, so that only 32 correct outputs (13.2%) may ensure correct system output.

Pattern of success for MOMV (PS$_{MOMV}$)

Let us consider $MOMV = (S_M, P_M)$ with a fixed structure $S_M = (g_1,...,g_L)$ and permutation P_M. Let s_e denote the sum of all individual classifier errors: $s_e = \sum_{i=1}^{M} e_i$ and the boundary number of correct individual votes is given by (35), two cases of optimal distribution strategy can be considered.

For $s_e <= M - k_S$, PS$_{\mathrm{MOMV}}$ can be achieved by distributing all individual classifier errors at the level of DED of exactly $M - k_S$ and keeping all remaining samples with all correct classifier outputs, i.e. at the level 0 of DED. Partition between these two levels of DED solved using (21) gives the following:

$$\begin{cases} p_X(0) = (M - k_S - s_e)/(M - k_S) \\ p_X(M - k_S) = s_e/(M - k) \end{cases} \tag{36}$$

For $\mathrm{s}_\mathrm{e} > M - k_S$, not all individual classifier errors can be covered by $M - k_S$ error level of DED and for some entries the number of errors produced by MCS has to be greater than $M - k_S$. To waste as few entries as possible, the excess of errors is kept at the highest DED level of M errors. Partition between these two levels of DED obtained from (21) is defined as follows:

$$\begin{cases} p_X(M - k_S) = (M - s_e)/k_S \\ p_X(M - k_S) = (s_e - M + k_S)/k_S \end{cases} \tag{37}$$

The definition of PS$_{\mathrm{MOMV}}$ can be thus interpreted as:

$$\begin{cases} s_e <= M - k_S \\ p_X(0) = (M - k_S - s_e)/(M - k_S) \\ p_X(M - k_S) = s_e/(M - k) \\ p_X(i) = 0 \quad i \neq \{0, M - k_S\} \end{cases} \vee \begin{cases} s_e > M - k_S \\ p_X(M - k_S) = (M - s_e)/k_S \\ p_X(M) = (s_e - M + k_S)/k_S \\ p_X(i) = 0 \quad i \neq \{M - k_S, M\} \end{cases} \tag{38}$$

Pattern of failure for MOMV (PF$_{MOMV}$)

Assuming the same conditions as in PS$_{\mathrm{MOMV}}$ again, two cases of optimal distribution strategy can be considered:

For $s_e <= k_S$, PF$_{\mathrm{MOMV}}$ can be achieved by distributing all individual classifier errors at the level of DED of exactly k_S errors and keeping all remaining samples with all correct classifier outputs, i.e. at the level 0 of DED. Partition between these two levels is:

$$\begin{cases} p_X(0) = (k_S - s_e)/(k_S) \\ p_X(k_S) = s_e/k_S \end{cases} \tag{39}$$

For $s_e > k_S$, not all individual classifier errors can be covered by k_S error level of DED and similarly as before the excess of errors has to be kept at the highest DED level of M errors. Partition between these two levels of DED results in:

$$\begin{cases} p_X(k_S) = (M - s_e)/(M - k_S) \\ p_X(M) = (s_e - k_S)/(M - k_S) \end{cases} \tag{40}$$

Thus the final definition of PF$_{\text{MOMV}}$ takes the form:

$$\begin{cases} s_e <= k_S \\ p_X(0) = (k_S - s_e)/k_S \\ p_X(k_S) = s_e/k_S \\ p_X(i) = 0 \quad i \neq \{0, k_S\} \end{cases} \vee \begin{cases} s_e > k_S \\ p_X(k_S) = (M - s_e)/(M - k_S) \\ p_X(M) = (s_e - k_S)/(M - k_S) \\ p_X(i) = 0 \quad i \neq \{k_S, M\} \end{cases} \tag{41}$$

4.2. Optimal permutation

Definitions (38) and (41) define only the optimal number of errors produced by MCS for a number of samples. They represent the necessary condition of optimality but are still not sufficient to reach the limits of MOMV for a given structure. The optimality of the distribution can be obtained provided the specific properties of the outputs propagation are observed in the structure S_M:

p1) Each correct vote at layer $L - i$ $(i = 0, ..., L - 2)$ should be the majority voting output from a group at layer $L - i - 1$, within which the outputs are distributed locally according to PS.

p2) Each incorrect output at layer $L - i$ $(i = 0, ..., L - 2)$ should be the majority voting output from a group of all errors at layer $L - i - 1$.

Assuming correct outputs of MOMV, backward propagation of these rules enforces specific permutations of outputs at the first layer, required to take advantage of optimal patterns discussed in previous section. There are however many of such permutations, which reflects the fact that there are many ways of distributing correct votes within each group. Figure 11 shows the examples of the permutation of 8 correct votes and 19 errors to reach optimality of the structure $S_{27} = (3,3,3)$.

Rather than formally define any specific optimal permutation of outputs, we associate the optimal permutation with the properties p1 and p2. In other words given the structure $S_M = (g_1, ..., g_L)$, and outputs distributed according to PS$_{\text{MOMV}}$ the permutation of the classifier outputs at the first layer is considered optimal if at each node of the structure one of the properties p1 or p2 is always satisfied.

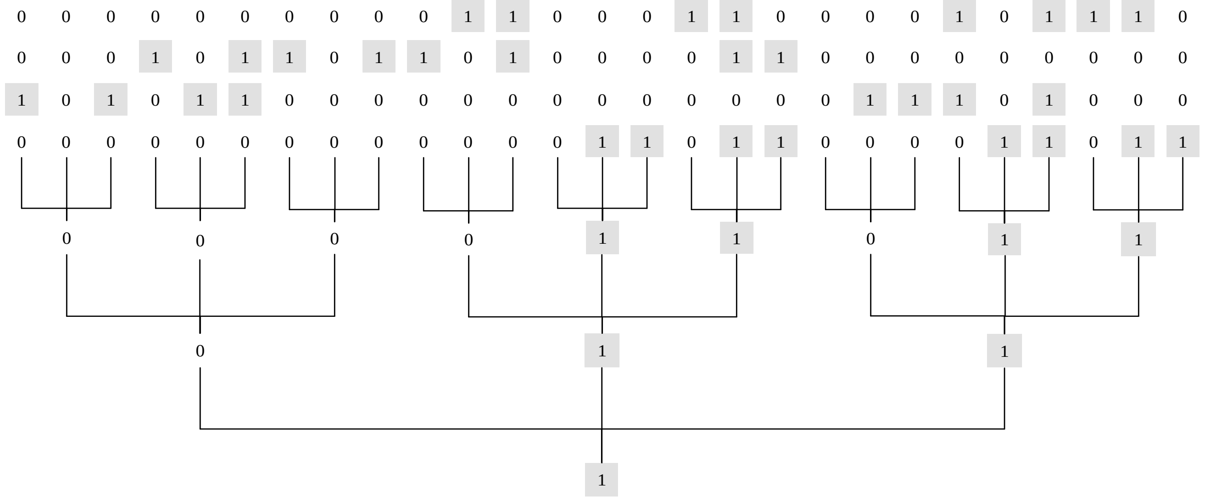
4.3. Optimal structure

Optimality of the structure of MOMV is an ambiguous term, which requires further definition. By optimal structure, we mean the structure, which for a given number of classifiers M gives the possibility of the largest stretching of the limits of errors produced by MOMV in comparison to standard MV error limits. Given M classifiers let associate optimality of a structure S_M with the minimisation of the number of correct votes k_S, still sufficient for correct output of MOMV. Exploiting the assumption of odd values of g_i at each stage, equation (35) can be rewritten as:

$$k_S = \prod_{i=1}^{L} \left(\frac{g_i + 1}{2} \right) \tag{42}$$

It can be shown that given the constraint: $M = \prod_{i=1}^{L} g_i = const$, k_S is monotonically decreasing function of L, which corresponds to the smallest possible sizes of groups g_i. Thus k_S becomes minimal for the largest possible number of layers L, which provided fixed M corresponds to the smallest groups of classifiers at each stage. As the minimal odd number of classifiers for which majority voting makes sense is 3, the optimal structure should contain groups of 3 outputs wherever it is possible. The structure S_M is thus optimal for $M = 3^L$ as for such expression: $g_i = 3$ can be true for all L layers. The structure $S_{27} = (3,3,3)$ analysed in the previous section and shown in Figure 11 is an example of such optimal structure.

Being constrained by many assumptions, optimality of solution $M = 3^L$ can be practically observed for very few cases of a reasonable number of classifiers: $M_{opt} = \{9,27,81,...,3^L\}$. More realistically, $M \neq 3^L$ enforces some differences in the size of groups at different layers. Nevertheless, from the above analysis it is suggested that all groups in the structure should be constructed using as small (odd) number of classifiers as possible.



0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 1 1 0 0 0 0 1 0 1 1 1 0

0 0 0 1 0 1 1 0 1 1 0 1 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0

1 0 1 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 0 1 0 0 0

0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 1 1 0 0 0 0 1 1 0 1 1

0 0 0 0 1 1 0 1 1

0 1 1

1

FIGURE 11. Optimal multistage organisation with 27 classifiers structured by optimal structure: $S_{27} = (3,3,3)$. Optimality of error distribution enforces the presence of exactly 8 correct votes for every data sample. Examples of optimal permutations are shown above the structure.

4.4. The limits of MOMV errors

As mentioned above the limits of MOMV error depend on three key properties of such a system of classifiers: the distribution of errors, permutation, and the structure. Boundary distributions of errors presented as patterns of success and failure solve the problem only locally – within each group, whereas permutation only validates this process. Moreover, the error limits of MOMV with the simplest structure: $S_M = (M)$ are equal to the limits for the non-structured system discussed in the section 3.3 and given by (32). The structure of MOMV appears to be the main factor driving the expansion of the error limits of the system. Using this criterion of optimality the structure $S_M = (M)$ is the worst possible as no shift is observed. The MOMV with the optimal structure defined in Section 4.3 offers the voting system with maximally widened limits of error, provided the individual classifiers errors are distributed according to PS_{MOMV} or PF_{MOMV} and holding additional conditions of the optimal permutation discussed in the section 4.2. By analogy to (32), based on (38) and (41) the absolute limits of MOMV errors can be defined by:

$$E_{MOMV}^{min} = max\left\{0, \frac{s_e - M + k_S}{k_S}\right\} \qquad E_{MOMV}^{max} = max\left\{\frac{s_e}{k_S}, 1\right\} \tag{43}$$

These limits have been shown graphically for different mean classifier errors and number of classifiers in Figure 12. Figure 12-A represents a 3-dimensional visualisation of the error limits of MOMV presented as a function of both, mean classifier error and the number of classifiers. Figure 12-B shows the 2-D projections for selected numbers of classifiers. Comparing to the error limits shown for the parallel system in Figure 9, one difference is very clear. Namely, unlike in simple majority voting where total reduction of error was possible up to the mean classifier error at the level of 0.5, for MOMV this boundary can be easily overcome or even virtually removed assuming large number of classifiers. For 27 poor experts, each giving only 30% of correct output, it is still possible to attain 100% performance of MOMV structured by $S_{27}(3,3,3)$. However, similarly to the plot in Figure 9, the symmetry with respect to the point $(0.5,0.5)$ is preserved, which means that by analogy one can construct an extremely bad performing MOMV made of relatively well performing classifiers.

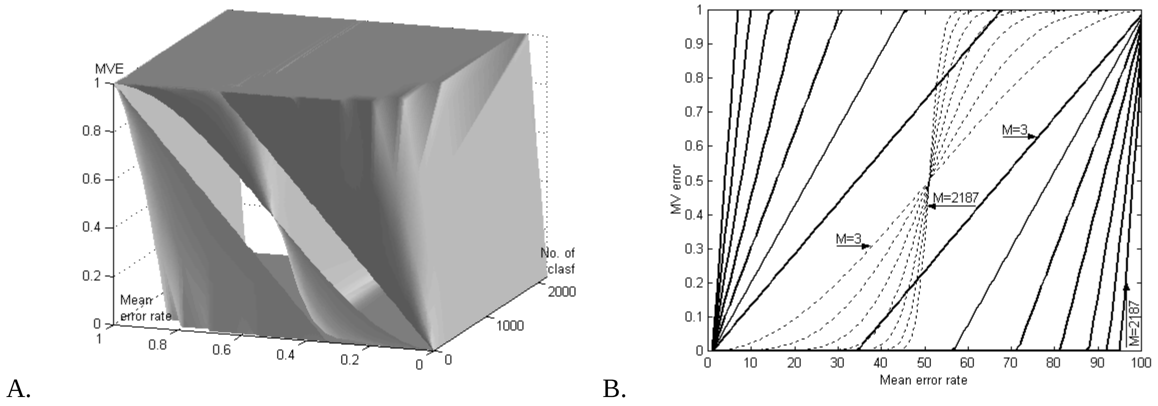


FIGURE 12. The error limits of MOMV shown as a function of the number of classifiers and the mean classifier error. A. 3D plot including surface corresponding to MOMV errors for independent classifiers. B. 2D projections for the selected numbers of classifiers.

5. Stability of boundary distribution of errors

The patterns of success and failure introduced by Kuncheva et al.[16] have been primarily used as an algorithmic description of how to achieve the majority voting limits given a mean classifier error. We have argued in the previous sections that if a realistic multiple classifier system was to be described by a PS it would represent a potentially unstable solution with a bad generalisation performance. This is due to very small classification margins represented by the original PS. This led to the introduction of SPS in section 3.2, which are meant to represent a more stable solution. In order to verify out claims, we have run extensive experiments examining the sensitivity of all presented patterns of error distributions to random flips of outputs, which could be thought of as simulating differences between validation and testing sets commonly present in the real-world classification problems.

For this purpose special functions have been implemented in Matlab, which given mean classifier error generate matrices of classifiers outputs distributed according to the analysed patterns: PS, PF, SPS, SPF, and PS_{MOMV}, PF_{MOMV} providing all other conditions of optimality discussed in the section 4. For all boundary patterns, we used 1000×27 classification matrices corresponding to 27 classifiers applied for 1000 data samples. The sensitivity of different patterns of error distribution and respective change in the majority voting error have been examined by generating new different classification matrices from the original matrices by the individual classifier output mutation with increasing mutation rates. The mutation rate is defined as a probability of a change of each bit in the classification matrices from 0 to 1 and vice versa. All the experiments were carried out for 50 values of mean classifier error and 50 different mutation rates spread equally within the range (0,1). To avoid random irregularities each experiment was repeated 10 times and the results averaged.

The results are presented in a form of a set of 3-D plots in Figure 13 and Figure 14. The impact of both the mutation rate and the mean classifier error on the majority voting error for different patterns is shown in Figure 13. The following Figure 14 depicts the differences in the previous plots and provides clear interpretations.

The most visible is much higher stability of SPS and SPF (Figure 13-C,D) in comparison to PS and PF (Figure 13-A,B) respectively. This is reflected by quite substantial region of flat area at 0% combined error. The larger margins evident by errors distributed far from the boundary of majority voting change impose greater stability of the overall system error effectively remaining at the lower level. The magnitude of improvement achieved by SPS and SPF depends on the combination of mean classifier error and the mutation rate as shown in Figure 14-A,B. For 27 classifiers, using SPS, SPF can prevent up to 25% changes of the majority vote in comparison to PS and PF under the same conditions. The plots corresponding to PS_{MOMV} and PF_{MOMV} (Figure 13-E,F) fully reflect the fact of expanded error limits in comparison to PS and PF. The lines corresponding to the same height (MOMV error) of the plot become extremely dense for the whole range of mean classifier error, as the mutation rate is close to 0. This explains two phenomena. Firstly, the limits of MOMV errors have really been substantially extended. Secondly, one can notice that a high steepness of the plot for the mutation rate close to 0 means that PS_{MOMV} and PF_{MOMV} are very unstable and sensitive to just single flips of outputs. Effectively due to a poor generalisation, ability the performance is dropping sharply with increasing mutation rate. Denoting mean classifier error by $\bar{e}$, this can be taken to extreme for $\bar{e} = k_S / M$ and $\bar{e} = (M - k_S)/M$ for PS_{MOMV} and PF_{MOMV} respectively. For these boundary levels of mean classifier error, the steepness is the largest, and these boundary solutions represent very unstable states. Reflection of this fact can be also found in Figure 14-C,D,E,F showing the difference of errors between PS_{MOMV}, PF_{MOMV} and PS, PF, and SPS, SPF. The only area where this improvement is considerable is for small mutation rates and mean classifier error close to the boundary of $\bar{e} = k_S / M$ and $\bar{e} = (M - k_S)/M$ for PS_{MOMV} and PF_{MOMV} respectively. These boundary values of $\bar{e}$, for which the absolute limits of MOMV error of 0% and 100% for PS_{MOMV} and PF_{MOMV} respectively are still preserved, are substantially extended in comparison to equivalent boundary values of $\bar{e}$ for PS and PF. For small mutation rate these shifts are the sources of considerable gain of PS_{MOMV} and PF_{MOMV} observed around these boundary values of $\bar{e}$ in Figure 14-C,D. However, for larger mutation rates the performance gain quickly reduces to 0. Comparison with the stable patterns depicted in Figure 14-E,F shows even the areas where MOMV becomes noticeably worse strategy. However, it has to be remembered that MOMV is presented in the form, which is not optimised in terms of stability. This problem of the stability of MOMV remains a complex and open issue beyond the scope of this paper.

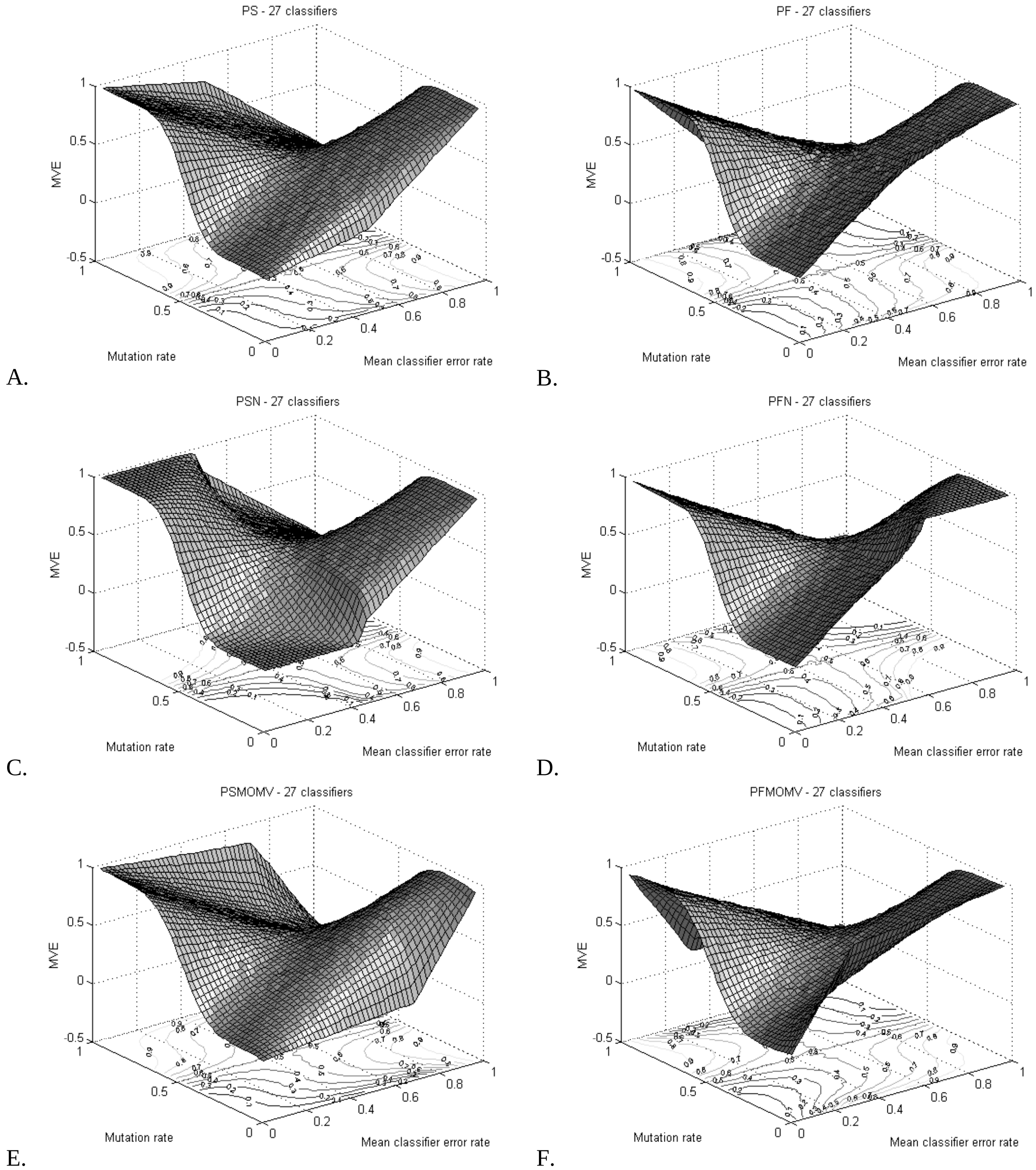


FIGURE 13. 3D plot of MV error as a function of mean classifier error and mutation rate for 27 artificial classifiers generating 1000 outputs each according to analysed patterns. Subfigures A,B,C,D,E,F correspond to the PS, PF, SPS, SPF, PS_{MOMV}, PF_{MOMV} respectively.

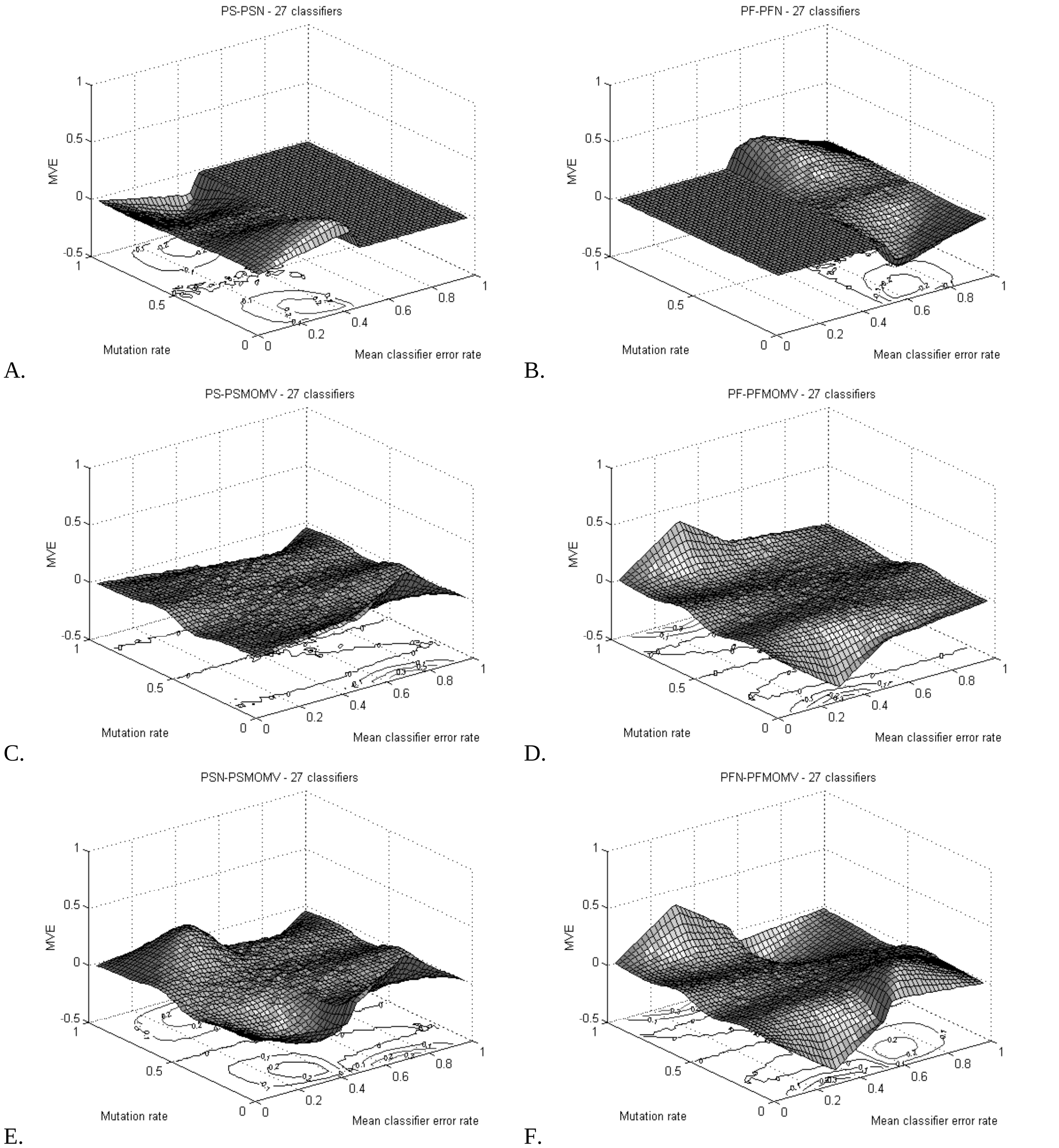


FIGURE 14. Comparison of different patterns from Figure 13 showing 3D plots of the error differences as a function of mean classifier error and mutation rate. Subfigures A,B,C,D,E,F show the differences: PS-PSN, PF-PFN, PS-PS$_{\mathrm{MOMV}}$, PF-PF$_{\mathrm{MOMV}}$, PSN-PS$_{\mathrm{MOMV}}$, PFN-PF$_{\mathrm{MOMV}}$.

6. Conclusions

In this paper we thoroughly investigated the problem of the limits of majority voting errors. Starting the analysis from an idealised multiple classifier system with independence assumption, we showed how the majority voting error of such a system depends on the mean classifier error and the number of classifiers. Introducing discrete error distribution and its normalised-continuous extension for the analysis, the behaviour of the combined errors was shown for a less constrained system consisting of a large number of classifiers with different performances, which normally becomes intractable due to exponential complexity of probability calculations. Exploiting the simplicity of the concept of error distribution, the condition of beneficial extendibility of a system by a pair of classifiers was derived in a new concise form. The improvement of the overall system performance with independent errors was reported to have its limits for mean classifiers error of 50%.

Relaxing the assumption of independent classifier outputs, we showed that in addition to the mean classifier error and the number of classifiers, the distribution of classifier outputs establishes new complex parameter strongly affecting the combined error. The limits of the majority voting corresponding to the pattern of success or failure, have been shown as a function of other parameters of the system. We noticed that these boundary distributions of classifier outputs, though most efficiently exploiting the correct/incorrect classifier outputs, represent extreme and potentially very sensitive solutions. When analysed in a context of the classification margins, the PS and PF represent cases with the smallest possible margins (i.e. 1 vote). Since for a real-world classification problems large positive margins are associated with a better generalisation performance of a classification system, an MCS representing the PS (PF) could therefore have very weak generalisation ability.

As a response we proposed new boundary distributions of classifier outputs called stable patterns of success and failure: SPS, SPF designed to preserve the absolute theoretical limits of majority voting errors corresponding to PS and PF, but at the same time substantially improving the classification margins for as many data samples as possible. The improvement of the stability of patterns has been illustrated both numerically and visually. The expected improvement of the generalisation ability of a system characterised by SPS (SPF) has been experimentally confirmed showing substantial gain of the performance in comparison to the system with PS (PF). The limits of majority voting errors showed the possibility of achieving 100% classification performance for the mean classifier errors up to 50% if the number of classifiers is large. One of the major contributions of this paper was showing that this boundary could be extended even further. Namely, we noticed that a standard system of classifiers voting in parallel represents only the simplest case in a general class of the multistage organisations with majority voting. Optimising all the parameters of such a system results in a substantially expanded error limits. The shift of the error limits was shown in both theoretical derivations and experimental observations. Increasing the number of classifiers leads to further expansion of MOMV errors up to virtually removal of any limits for large number of classifiers. Individual performances of classifiers in such a system are no longer of crucial importance. Increasing role is assigned to the relevant distribution of errors and appropriate allocation of errors within the structure. The generalisation analysis of the presented PS_{MOMV}, PF_{MOMV} revealed however quite high instability of the limits of MOMV errors. This can be explained by small classification margins of different classifier groups, which is due to small number of classifiers in these groups as prescribed by the optimality of the system in terms of the majority voting limits. It appears that in multistage systems, to improve the confidence of generated outputs leads to unavoidable partial loss of the optimality of error limits as considered in this paper. The stability of such systems remains a complex and open problem.

Nevertheless, for small mutation rate, the boundary distributions of classifier outputs: PS_{MOMV}, PF_{MOMV} considerably outperformed the patterns corresponding to simple majority voting systems in particular for mean individual errors greater than 50%.

REFERENCES

[1] Sharkey AJC. Combining artificial neural nets: ensemble and modular multi-net systems. Springer-Vorlag, Berlin, 1999.

[2] Bezdek JC. Fuzzy models and algorithms for pattern recognition and image processing. Kluwer Academic, Boston, 1999.

[3] Zhilkin PA, Somorjai RL. Application of several methods of classification fusion to magnetic resonance spectra. Connection Science 1996; 8(3,4): 427-442.

[4] Rogova G. Combining the results of several neural network classifiers. Neural Networks 1994; 7(5): 777-781.

[5] Sharkey AJC, Sharkey NE. Combining diverse neural nets. The Knowledge Engineering Review 1997; 12(3): 231-247.

[6] Kittler J. Combining classifiers: a theoretical framework. Pattern Analysis & Applications 1998; 1: 18-27.

[7] Hashem S. Optimal linear combinations of neural networks. Neural Networks 1997; 10(4): 599-614.

[8] Granger CWJ. Combining forecasts – twenty years later. Journal of Forecasting 1989; 8(3): 167-173.

[9] Littlewood B, Miller DR. Conceptual modeling of coincident failures in multiversion software. IEEE Transactions on Software Engineering 1989; 15(12): 1596-1614.

[10] Geman S, Bienstock E, Doursat R. Neural networks and bias/variance dilemma. Neural Computation 1992, 4: 1-58.

[11] Partridge D, Griffith N. Strategies for improving neural net generalisation. Neural Computing & Applications 1995; 3: 27-37.

[12] Partridge D. Network generalization differences quantified. Technical Report 291, Department of Computer Science, University of Exeter, 1994.

[13] Lam L, Suen CY. A theoretical analysis of the application of majority voting to pattern recognition. In Proceedings of the International Conference on Pattern Recognition, Jerusalem 1994, pp 418-420.

[14] Lam L, Suen CY. Application of majority voting to pattern recognition: an analysis of its behaviour and performance. IEEE Transactions on Systems, Man, and Cybernetics 1997; 27(5): 553-568.

[15] Battiti R, Colla AM. Democracy in neural nets: voting schemes for classification. Neural Networks 1994; 7(4): 691-707.

[16] Kuncheva LI, Whitaker CJ, Shipp CA, Duin RPW. Limits on the majority vote accuracy in classifier fusion. Accepted for the Pattern Analysis and Applications, available at: *http://www.bangor.ac.uk/~mas00a/papers/lkpami.ps.gz.*

[17] Kuncheva LI, Whitaker CJ. Measures of diversity in classifier ensembles (submitted), available at: *http://www.bangor.ac.uk/~mas00a/papers/lkml.ps.gz.*

[18] Schapire RE, Freud Y, Bartlett P, Lee WS. Boosting the Margin: A New Explanation for the Effectiveness of Voting Methods. The Annals of Statistics 1998; 26(5): 1651-1686.

[19] Wolpert DH, Stacked Generalization. Neural Networks 1992; 5: 241-259.

[20] Pudil P, Novovicova J, Blaha S, Kittler J. Multistage Pattern Recognition with Rejection Option. In Proceedings of the 11th IAPR International Conference on Pattern Recognition B, II, 1992, pp. 92-95.

[21] Ho TK, Hull JJ, Srihari SN. Decision combination in multiple classifier systems. IEEE Transactions on Pattern Analysis and Machine Intelligence 1994; 16(1): 66-75.

[22] Maisel L. Probability, statistics and random processes. Simon and Schuster, New York, 1971.

[23] Hamburg M. Statistical analysis for decision making. Harcourt Bruce & World, New York, 1970.