Analysis of the correlation between majority voting error and the diversity measures in multiple classifier systems

Dymitr Ruta, Bogdan Gabrys

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

Abstract

Combining classifiers by majority voting (MV) has recently emerged as an effective way of improving performance of individual classifiers. However, the usefulness of applying MV is not always observed and is subject to distribution of classification outputs in a multiple classifier system (MCS). Evaluation of MV errors (MVE) for all combinations of classifiers in MCS is a complex process of exponential complexity. Reduction of this complexity can be achieved provided the explicit relationship between MVE and any other less complex function operating on classifier outputs is found. Diversity measures operating on binary classification outputs (correct/incorrect) are studied in this paper as potential candidates for such functions. Their correlation with MVE, interpreted as the quality of a measure, is thoroughly investigated using artificial and real-world datasets. Moreover, we propose new diversity measure efficiently exploiting information coming from the whole MCS, rather than its part, for which it is applied.

1. Introduction

An increasing scientific effort dedicated to pattern recognition problems is currently directed to combining classifiers. For a number of applications, combining classifiers has been shown to outperform the traditional single-best classifier approach [1-7]. According to the current state of knowledge there are two main factors deciding about the performance of MCS. Individual performances of classifiers are one of these factors but they are not sufficient for evaluation of MCS performance. Inadequate character of taking only individual performances as an indicator of the effective combination has been already confirmed in [6], which agrees with the results of [7] and many more related publications reviewed in [1] and [2]. It turned out that specific dependencies among classifier outputs strongly influence MCS performance and thus have to be taken into account in addition to individual performances of classifiers [6]. Capturing the precise formula describing the impact of these two factors on MCS performance is usually very difficult and is subject to combination methods used.

Majority voting is one of the simplest combining methods operating on binary classification outputs (correct/incorrect) [8-12]. However, quite often its performance turned out to be comparable with more advanced combiners [5,13,15]. Moreover, the simplicity of MV can be efficiently exploited in many ways. Specifically, it can be applied for a large number of classifiers, which using more advanced combination method would inevitably lead to intractability. MV can also be easily used at different levels of hierarchical structure, which as it was illustrated in [12], offers further potential reduction of MVE. These features make MV quite an attractive combination method. Despite the simplicity of MV itself, at certain stage, the problem of complexity may arise anyway. This is the case when a subset of classifiers to be combined is to be selected in some optimal way from a larger pool of classifiers. Such selection involves a searching procedure, which in the simplest form is of the exponential complexity. For such cases, the possible solution is to model the MV behaviour by means of less complex function operating on classifier outputs. Essentially, the aim is to find a function simpler than MV but still well correlated with MV performance. The diversity measures are primary candidates for this purpose.

Diversity among classifiers is the notion describing the level to which classifiers vary in data representation, concepts, strategy etc. Consequently, this should be reflected in different classifiers making errors for different data samples. As shown in many papers, such phenomenon of disagreement to errors is highly beneficial for MV [16,17] and combining purposes in general [18-20]. The measures of the diversity cover a large variety of statistical measures including

correlation, similarity, disagreement, and many others [17,21]. For MV purposes, only diversity measures operating on binary classification outputs remain under the scope of considerations in this paper. They may potentially provide information whether classifiers are worth combining or not, obtained at lower computational cost. The relevance of such decision is subject to the relationship between MVE and the diversity measure. The quality of such relationship can be evaluated by means of the correlation coefficient calculated between the two functions.

The analysis of the correlation between MVE and various diversity measures reported in the literature is the main issue studied in this work. For this purpose, we design a new diversity measure, optimised for the use with MV combiner, which exploits the whole information space provided by the classifier outputs.

The remaining of this paper is organised as follows. In section 2, the theoretical foundations of MV and its errors are briefly presented. Next section provides the description of pairwise and non-pairwise diversity measures operating on binary classifier outputs. Section 4 provides experimental results highlighting the degrees of correlation between MVE and presented diversity measures obtained for several standard datasets and classifiers, followed by thorough interpretation of the results. Finally, conclusions including discussion and suggestions for applications are given in section 5.

2. Majority voting errors

Majority voting is a simple combination method operating on binary inputs (1-correct/0-incorrect). It can be applied practically for any classifiers as their outputs can always be mapped, if necessary, to the binary representation. Given a system of M classifiers: $D = \{D_1,...,D_M\}$, let $y_j(\mathbf{x}_i)$ $i = 1,...,N$ $j = 1,...,M$ denote the output of the j^{th} classifier for the i^{th} multidimensional input sample $\mathbf{x}_i$. Given the binary outputs from M classifiers for a single input sample, the decision of MV y_i^{MV} can be obtained according to the following formula:

$$y_i^{MV} = \begin{cases} 0 \text{ if } \sum_{j=1}^{M} y_j(\mathbf{x}_i) \le \lfloor M/2 \rfloor \\ 1 \text{ if } \sum_{j=1}^{M} y_j(\mathbf{x}_i) > \lfloor M/2 \rfloor \end{cases} \quad (1)$$

A more detailed definition of MV including the rejection rule observed for $\sum_{j=1}^{M} y_j(\mathbf{x}_i) = M/2$ when M is even can be found in [9]. However, this work is not concerned with a detailed study of MV itself and in further analysis, without any loss of generality, we assume odd M. In [12] we carried out an extensive analysis of MV errors and its limits. We summarise briefly its relevant findings. Let Z refer to the discrete error distribution (DED) built on the outputs from MCS and defined as:

$$\mathbf{Z} = [z_0, z_1,..., z_M] \quad (2)$$

where z_j $(j = 0,..., M)$ denotes the ratio of the number of samples misclassified by exactly j classifiers to the total number of samples. Given (2) the error of majority voting (MVE) can be defined as:

$$E_{MV}(M) = \sum_{i=\lceil M/2 \rceil}^{M} z_i \quad (3)$$

For the fixed mean classifier error $\bar{e}$, MVE can be decreased by conditional extension of MCS by a pair of classifiers. For a given number of classifiers the MVE can vary substantially, which is subject to a distribution of individual errors over the input dataset. The limits of achievable MVE correspond to the specific patterns of boundary error distribution [12] and for given $\bar{e}$ can be expressed by:

$$E_{MV}^{min} = \max\left\{0, \frac{M\bar{e} - \lceil M/2 \rceil + 1}{M - \lceil M/2 \rceil + 1}\right\} \qquad E_{MV}^{max} = \min\left\{\frac{M\bar{e}}{\lceil M/2 \rceil}, 1\right\} \quad (4)$$

This work is concerned with the possibility of modelling the specific dependencies among the outputs of classifiers in order to establish a reliable relationship with MVE while having significantly reduced complexity associated with evaluation of all combinations of classifiers in MCS.

3. Diversity measures

Diversity among classifiers is the notion describing the level to which classifiers vary in data representation, concepts, strategy etc. That way perceived multi-dimensional diversity has many faces but its effects observed at the outputs of classifiers are the same: occurrences of errors for different input data. Reflection of this fact can be found in a number of definitions of the diversity measures, the core of which is usually a simple disagreement rule [17]. The phenomenon of disagreement to errors was shown to improve combining performance [16-20]. Majority voting is not the exception to this rule, conversely, as it was shown in [16,17] it benefits explicitly from classifiers disagreement to errors. Most of the diversity measures presented in this paper has already been studied for artificial data by Kuncheva in [17]. In this work, we investigate an extended number of diversity measures applied to a number of real-world datasets and using realistic classifiers.

3.1. Pairwise diversity measures

Diversity measure starts to be meaningful if it is applied for a group of the minimum of two classifiers. In the

simplest case, a measure can be applied for examining diversity between exactly two classifiers. Such measures are usually referred to as pairwise diversity measures (PDM). For more than two classifiers, PDM is typically obtained by averaging the PDM's calculated for all pairs of classifiers from the considered pool of classifiers. A strong point of such measures is substantially reduced system complexity (quadratic order). However, simplicity of PDM's is achieved for the price of vaguer relationship with MVE if applied for a group of more than two classifiers. This is due to applying averaging by which the information about error coincidences for more than two classifiers is lost.

For simplicity, we introduce the following notation for PDM. We consider a system of M parallely voting classifiers: $D = \{D_1,...,D_M\}$ producing binary outputs (1-correct, 0-incorrect) for each of N input samples $\mathbf{x}_i\,(i = 1,...,N)$. Let $N^{ab}\ a,b = \{0,1,*\}$ denote a number of input samples, for which the considered pair of classifiers produce sequence of outputs: $\{a,b\}$. The star denotes any of the outputs: $* = 0 \text{ or } 1$. Note that $N = N^{**}$.

The correlation measure ρ

The correlation measure between two classifiers $\{c_i, c_j\}$ is probably the most intuitive and can by calculated by:

$$\rho_{i,j} = \frac{N^{11}N^{00} - N^{01}N^{10}}{\sqrt{N^{1*}N^{0*}N^{*1}N^{*0}}} \tag{5}$$

Please notice that the measure is symmetrical with respect to the change of outputs: $0 \leftrightarrow 1$.

Product-moment correlation measure r

This measure was used by Sharkey and Sharkey[16] as a guidance for selection of the most diverse neural network classifiers. Adopting this measure to the binary representation of errors it can be defined by:

$$r_{i,j} = \frac{N^{00}}{\sqrt{N^{*0}N^{0*}}} \tag{6}$$

The Q statistics

Q statistics was used by Kuncheva[17] for assessing the level and sign of dependency between a pair of classifiers with binary outputs, where –1 means full negative dependence, +1 full positive dependence. The measure is defined by:

$$Q_{i,j} = \frac{N^{11}N^{00} - N^{01}N^{10}}{N^{11}N^{00} + N^{01}N^{10}} \tag{7}$$

This measure is also symmetrical with respect to the change of outputs: $0 \leftrightarrow 1$.

The disagreement measure S

The disagreement measure was used by Skalak[22] to determine the diversity between two classifiers in a form of a ratio between the number of samples for which classifiers disagreed, to the total number of observations. This can be written as:

$$S_{i,j} = \frac{N^{01} + N^{10}}{N} \tag{8}$$

This measure is symmetrical with respect to the change of outputs: $0 \leftrightarrow 1$.

The double-fault measure F

The measure was used by Giacinto and Roli[23] to create a matrix of pairwise dependencies used for selecting the least related classifiers. The measure estimates the probability of coincident errors for a pair of classifiers, which is:

$$F_{i,j} = \frac{N^{00}}{N} \tag{9}$$

3.2. Non-pairwise diversity measures

Pairwise diversity measures if applied for more than two classifiers, suffer from loosing some information about error relations. Non-pairwise diversity measures (NDM) try to avoid this disadvantage by using different representations based on the whole group of classifiers for which it is applied. From informational point of view, if NDM is applied for the whole MCS it uses the complete available information coming from the outputs of individual classifiers. However, if not, there is still unexploited information given in a form of outputs from unexamined classifiers but still describing the same dataset and therefore indirectly related to the examined classifiers. To the best knowledge of the authors, there is no evidence in the literature of including this additional information into the definition of any diversity measure.

For simplicity, we introduce the following notation for NDM. We consider a system of M parallely voting classifiers: $D = \{D_1,...,D_M\}$ producing binary outputs (1-correct, 0-incorrect) for each of N input samples $\mathbf{x}_i\,(i = 1,...,N)$. Let $m(\mathbf{x}_i)$ denote the number of classifiers producing error for the input sample $\mathbf{x}_i$. It can be expressed by:

$$m(\mathbf{x}_i) = M - \sum_{j=1}^{M} y_{i,j} \tag{10}$$

where $y_{i,j}$ is the binary output (1-correct, 0-incorrect) from the j^{th} classifier for the i^{th} input sample. Let $\bar{e}$ denote the mean classifier error defined as:

$$\bar{e} = \frac{1}{NM}\sum_{i=1}^{N} m(\mathbf{x}_i) \tag{11}$$

The entropy measure H

This measure was used by Kuncheva[17] and shows the level of disagreement of outputs from a set of classifiers:

$$H = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{M - \lceil M/2 \rceil} \min\{m(\mathbf{x}_i), M - m(\mathbf{x}_i)\} \tag{12}$$

The entropy measure becomes maximal: $H = 1$ for the highest disagreement, which is the case of observing $\lfloor M/2 \rfloor$ votes with identical value (0 or 1) and $M - \lfloor M/2 \rfloor$ with the alternative value. The lowest entropy $H = 0$ is observed if all classifier outputs are identical. This measure is symmetrical with respect to change of outputs: $0 \leftrightarrow 1$.

The measure of "difficulty" θ

This measure originates from a study of Hansen and Salomon[24] and was developed by Kuncheva and Whitaker[17] for the case of binary classification outputs. Given the set of M classifiers, the measure is built on the basis of discrete error distribution $\mathbf{Z}$ defined in (2). Namely, it measures the variance of $\mathbf{Z}$, which can be defined as:

$$\theta = \frac{1}{N}\sum_{i=0}^{M} (z_i - \bar{z})^2 \tag{13}$$

The measure of difficulty is symmetrical with respect to the change of outputs: $0 \leftrightarrow 1$.

Kohavi-Wolpert variance KW

This measure follows the similar strategy as the measure of difficulty. Namely, it measures the average variance from binomial distributions of outputs for each classifier [25]. The measure can be simply calculated by:

$$KW = \frac{1}{NM^2}\sum_{i=1}^{N} [m(\mathbf{x}_i)(M - m(\mathbf{x}_i))] \tag{14}$$

It can be shown that for independent classifiers: $KW = M\theta$, which derives from the definition of variance for joint binomial distribution. This measure is symmetrical with respect to the change of outputs: $0 \leftrightarrow 1$.

Interrated agreement measure κ

This measure was developed in [26] to measure the level of agreement while correcting the chance (see [26] for details). Using the notation presented above it can be expressed by:

$$\kappa = 1 - \frac{\sum_{i=1}^{N} m(x_i)(M - m(\mathbf{x}_i))}{NM(M-1)\bar{e}(1-\bar{e})} \tag{15}$$

This measure is symmetrical with respect to the change of outputs: $0 \leftrightarrow 1$.

The measure of "fault majority" ω

This idea of this diversity measure was inspired by the work of Kuncheva[17], and it extends the notion of discrete error distribution introduced by Ruta and Gabrys[12]. Let a pool of M classifiers be described by a discrete error distribution $\mathbf{Z}$ defined in (2). Distribution $\mathbf{Z}$ describes the system of classifiers as a whole and can provide diversity measure itself (like difficulty measure). However, in this form $\mathbf{Z}$ calculated for the whole pool of classifiers can not be exploited for description of subsets of classifiers forming MCS, and needs to be recalculated which implies the computational complexity of the same order as examining exhaustively MVE.

This disadvantage can be eliminated by introducing partial discrete error distributions (PDED). Such distributions could be assigned individually to each classifier referring to the degree to which different classifiers participate separately in different levels of $\mathbf{Z}$. Description of the MCS takes then the form of a set of M PDED's $\mathbf{Z}_i$, which can be defined by:

$$\mathbf{Z}_i = [z_{i,0}, ..., z_{i,M}] \tag{16}$$

The components $z_{i,j}$ of PDED Z_i denote the normalised number of cases, in which the i^{th} classifier was among j classifiers being in error over validation set. This can be expressed formally as:

$$z_{i,j} = \begin{cases} \sum_{k=1}^{N}[1 - y_{i,k} \mid m(\mathbf{x}_k) = 0]/N & \text{for } j = 0 \\ M\sum_{k=1}^{N}[1 - y_{i,k} \mid m(\mathbf{x}_k) = j]/Nj & \text{for } j \neq 0 \end{cases} \tag{17}$$

Such representation preserves specificities of individual classifiers, but still allows to reconstruct DED $\mathbf{Z}$ by simple aggregation: $\mathbf{Z} = \sum_{i=1}^{M}\mathbf{Z}_i/M$. The complexity of such a system is kept at the low quadratic level. Similarly to pairwise diversity measures, PDED's once calculated can be used for the description of any subset of classifiers from MCS. However, unlike for other NDM, described in (12)-(15) the components $z_{i,j}$ contain information about interactions of the i^{th}

classifier with all remaining classifiers from the whole pool of classifiers. An example of DED decomposed into 18 PDED's is shown in Figure 1.

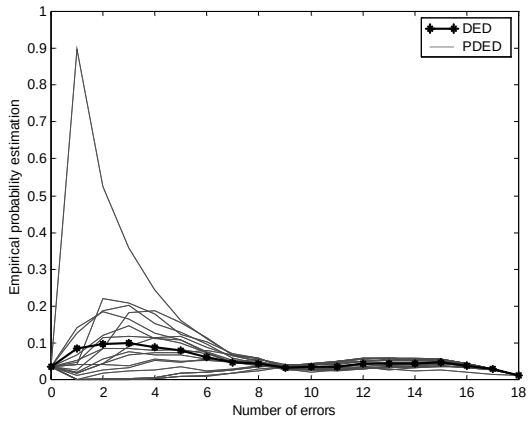

Figure 1. DED decomposed into individual PDED's corresponding to 18 different linear and nonlinear classifiers. The results have been obtained for the *Liver* dataset from UCI repository.

Representation of error relations within the system by means of PDED's offers a variety of potential diversity measures, which can be based on shape analysis or local minimisation of different PDED's. A strong point of this approach is that the design of the diversity measure can be flexibly adjusted to take into account the combination method used. For MV combiner, the crucial fact that needs to be taken into account is the decision boundary at $\lceil L/2 \rceil - 1$ where $L \le M$ denotes the number of classifiers in the examined subset from a pool of M classifiers. One of the simplest measures exploiting this fact can be associated with a sum of only those PDED components that could contribute to MVE for a considered subset of classifiers. Moreover, the sum should be calculated for the components coming from $\lceil L/2 \rceil$ locally best classifiers. By the locally best we mean the classifiers for which PDED components $z_{i,j}$ for a given j are the smallest. Formally, the measure can be defined as follows:

$$\omega = \sum_{j=\lceil L/2 \rceil}^{L} \sum_{i^*=1}^{\lceil L/2 \rceil} z_{i^*,j} \qquad (18)$$

where index i^* refers to the classifiers sorted according to their values of $z_{i,j}$ for the fixed j. Note that errors contributing to DED levels below the MV threshold of $\lceil L/2 \rceil$ errors for the whole MCS will not influence the measure as they surely will not produce MVE for any subset of classifiers. Moreover, the measure examines $\lceil L/2 \rceil$ locally best rather than all classifiers in the subset although at different levels different classifiers from the subset may contribute to the measure.

4. Experiments

The experiments have been performed with several standard datasets from UCI repository. Each dataset was randomly split into 3 equal subsets and used as training, validation and testing set respectively. Short description of considered datasets is shown in Table 1:

Datasets	#cases	#feat	#class
Iris	150	4	3
Biomed	194	5	2
Diabetes	768	8	2
Wine	178	13	3
Liver	345	6	2
Cancer	569	30	2
Vehicle	846	18	4
Ionosphere	351	34	2
Phoneme	5404	5	2
Satimage	6435	36	6

Table 1. The parameters of considered datasets.

Using a mixture of 15 different linear and non-linear classifiers trained on the training sets, binary matrixes of outputs (BMO) have been generated from hardened classification outputs obtained over validation set. This procedure was repeated for each dataset until BMO reached the size of 5000 instances. Given classification outputs in the form of BMO all analysed diversity measures have been examined for all k-element k={3,5,…,13} combinations from a pool of 15 classifiers. The relationship between DM and MVE has been evaluated by means of correlation coefficients calculated for series of k-elements combinations for each dataset. Complete results for the groups of three classifiers are shown in Table 2. This is accompanied by the evolution of correlation coefficients for increasing number of classifiers in the examined groups, shown for the *Cancer* dataset in Figure 2.

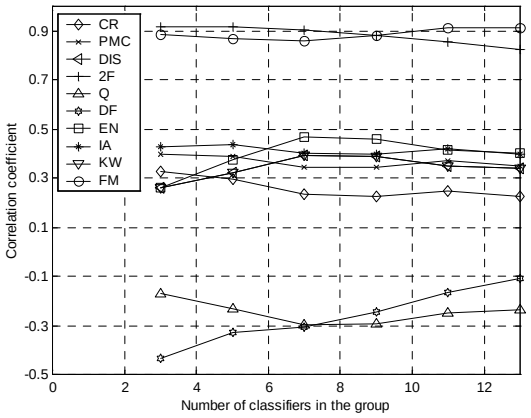

Figure 2. *Cancer* dataset. Evolution of correlation coefficients between analysed DM's and MVE for the groups of 3,5,…,13 classifiers.

The results very clearly show the outstanding performance of the non-pairwise measure of "fault majority". The only measure that holds comparable performance is the pairwise double-fault measure. The common feature that differs F_{AV} and ω from the rest of other measures (excluding r_{AV}) is the fact that they are non-symmetrical with respect to the change of outputs $0 \leftrightarrow 1$. Moreover, their definitions (9), (17) reflect an increased emphasis put on measuring error coincidences instead of a disagreement influenced equally by errors and correct outputs. To support our findings, we compare graphically relationships of F_{AV}, ω and a typical measure of disagreement: Q statistics, with MVE. This is shown in Figure 3. In the view of the above, coincidences of errors tend to model MVE better than disagreement of classifier outputs, which stays with the agreement to definition of MVE and its behaviour presented in [12]. What is more, Figure 3 also illustrates that "fault majority" measure tends to cope better with increasing number of classifiers than other diversity measures. This phenomenon can be explained by both, the strength of a fuller representation of the properly designed non-pairwise measure and an improved exploitation of the total information given by the outputs of all individual classifiers.

	Average Pairwise DM					Non-pairwise DM				
	ρ_{AV}	r_{AV}	S_{AV}	F_{AV}	Q_{AV}	θ	H	κ	KW	ω
Iris	-0.12	-0.08	0.44	0.80	-0.28	-0.56	0.44	0.01	0.44	0.93
Biomed	-0.32	-0.28	0.47	0.70	-0.46	-0.58	0.47	-0.29	0.47	0.92
Diabetes	-0.61	-0.59	0.62	-0.39	-0.59	-0.64	0.62	-0.59	0.62	0.66
Wine	-0.52	-0.39	0.68	0.94	-0.56	-0.67	0.68	-0.39	0.68	0.96
Liver	-0.04	0.04	0.08	0.30	-0.03	-0.19	0.08	-0.04	0.08	0.72
Cancer	0.34	0.41	0.27	0.93	-0.18	-0.44	0.27	0.44	0.27	0.88
Vehicle	-0.61	0.06	0.71	0.94	-0.69	-0.79	0.71	-0.50	0.71	0.98
Ionosphere	0.34	0.50	0.18	0.95	0.09	-0.48	0.18	0.30	0.18	0.96
Phoneme	0.50	0.51	-0.47	0.59	0.35	0.35	-0.51	0.51	-0.47	0.52
Satimage	0.00	0.08	0.16	0.68	-0.20	-0.35	0.16	0.04	0.16	0.91
MEAN	-0.10	0.03	0.31	0.64	-0.26	-0.44	0.31	-0.05	0.31	0.85

Table 2. Correlation coefficients between analysed diversity measures and majority voting error obtained for 10 standard datasets using groups of 3 out of 15 linear and non-linear classifiers.

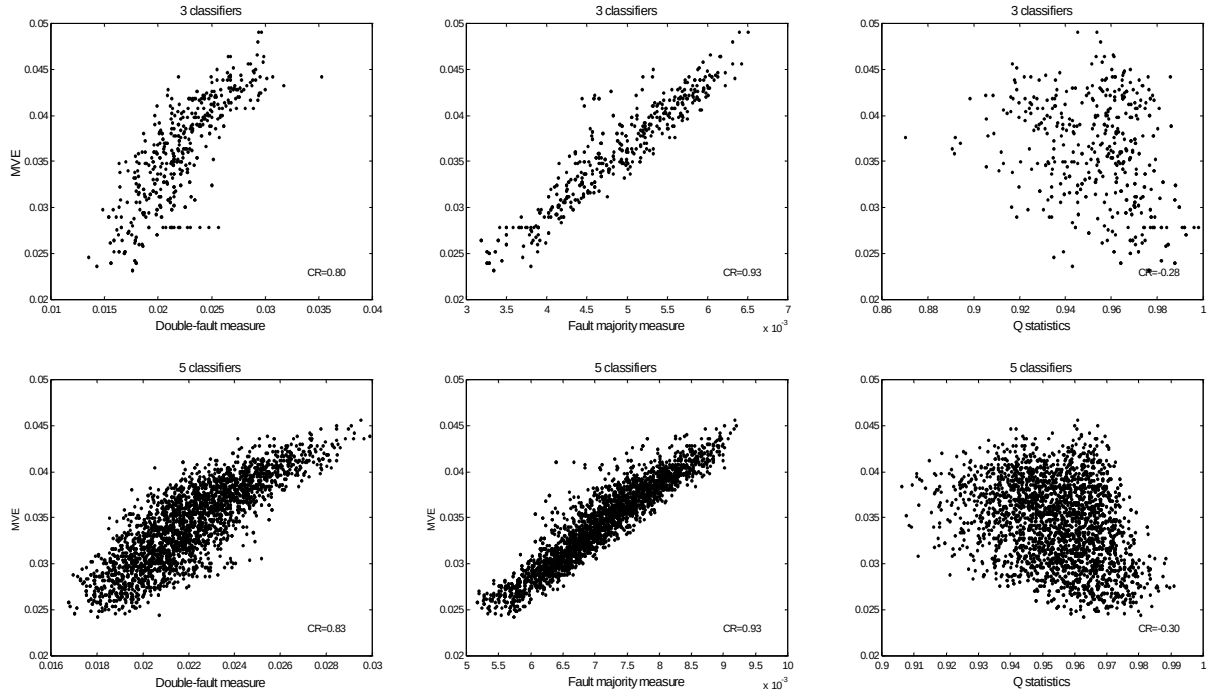

Figure 3. *Iris* dataset. Relationship between majority voting error and 3 diversity measures: Double-fault; Fault majority; Q statistics, for all combinations of 3 and 5 classifiers from 15 classifiers available in the complete pool of classifiers.

5. Conclusions

In this paper, we studied relationship between majority voting errors and diversity measures operating on binary classifier outputs. A number of pairwise and non-pairwise measures have been presented and examined for 10 standard datasets using 15 common linear and non-linear classifiers. Furthermore, we have developed an original diversity measure based on the discrete error distributions. The "fault majority" measure (ω) was designed with an explicit relation to the majority voting included in its definition and incorporating information coming from all the classifiers in a pool, even when only a subset was examined. Extensive experimental results showed the newly proposed non-pairwise diversity measure ω and the pairwise diversity measure F_{AV} to be the most consistently and highly correlated with majority vote error MVE, with ω outperforming the other measures for 8 out of 10 datasets and coming close second behind F_{AV} for the remaining 2 datasets. The strength of both measures has been identified as coming from asymmetry of their definitions with respect to the classifier outputs and greater emphasis put on measuring coincidences of errors. The disagreement evident in the definitions of the majority of the other measures and reflected by the symmetry of these measures with respect to the change of classifier outputs turned out to be inadequate for majority voting applied for realistic datasets. This was also the case for Q statistics, which as shown in [17] performed well for artificial data generated by similarly performing classifiers covering wide range of diversity. Another interesting phenomenon observed was an increased resistance of ω measure from deteriorating with increasing number of classifiers examined. This fact was also explained as the result of improved information exploitation, which comes in addition to a more appropriate representation of the properly designed non-pairwise diversity measure. The presented results allow considering ω and F_{AV} as recommended diversity measures when MV is used for combining classifiers. For other fusion methods the performance of FM measure is questionable, as it has been designed to work particularly well with the majority vote combiner. Potential good correlation between FM in its form given in this paper and other combiners would mean that MVE itself could be used as a good non-pairwise diversity measure for other combining methods.

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] Xu L, Krzyzak A. Methods of combining multiple classifiers and their applications to handwriting recognition. IEEE Transactions on Systems, Man, and Cybernetics 1992; 23(8): 418-434.

[5] Optimal combinations of pattern classifiers. Pattern Recognition Letters 1995; 16: 945-954.

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

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

[8] 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.

[9] 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.

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

[11] Kuncheva LI, Whitaker CJ, Shipp CA, Duin RPW. Limits on the majority vote accuracy in classifier fusion. IEEE Transactions on Pattern Analysis and Machine Intelligence. (submitted)

[12] Ruta D, Gabrys B. A theoretical analysis of the limits of majority voting errors for multiple classifier systems. Pattern Analysis and Applications (submitted).

[13] Cho SB, Kim JH. Combining multiple neural networks by fuzzy integral for robust classification. IEEE Transactions on Systems, Man, and Cybernetics 1995; 25(2): 380-384.

[14] Kang HJ, Kim K, Kim JH. A framework for combining of multiple classifiers at an abstract level. Engineering Applications of Artificial Intelligence 1997; 10(4): 379-385.

[15] Kuncheva LI, Bezdek JC. On combining classifiers by fuzzy templates. Proc. NAFIPS'98, Pensacola, FL, 1998, pp 193-197

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

[17] Kuncheva LI, Whitaker CJ. Measures of diversity in classifier ensembles. Machine Learning. (submited)

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

[19] Tumer K, Ghosh J. Error correlation and error reduction in ensemble classifiers. Connection Science 1996; 8(3,4): 385-404

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

[21] Afifi AA, Azen SP. Statistical analysis. A computer oriented approach. Academic Press, New York, 1979.

[22] Skalak DB. The sources of increased accuracy for two proposed boosting algorithms. Proc. AAAI'96, Integrating Multiple Learned Models Workshop, Portland, OR, 1996

[23] Giacinto G, Roli F. Design of effective neural network ensembles for image classification processes. Image Vision and Computing Journal 2000. (to appear)

[24] Hansen S, Salamon P. Neural network ensembles. IEEE Transactions on Pattern Analysis and Machine Intelligence 1990; 12(10): 993-1001.

[25] Kohavi R, Wolpert DH. Bias plus variance decomposition for zero-one loss function. Machine Learning: Proc. 13th Interntaional Conf., Morgan Kaufmann, 1996, pp 275-283.

[26] Fleiss JL. Statistical methods for rates and proportions. John Wiley & Sons, 1981.