

New Measure of Classifier Dependency in 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
{ruta-ci0, gabr-ci0}@paisley.ac.uk

Abstract. Recent findings in the domain of combining classifiers provide a surprising revision of the usefulness of diversity for modelling combined performance. Although there is a common agreement that a successful fusion system should be composed of accurate and diverse classifiers, experimental results show very weak correlations between various diversity measures and combining methods. Effectively neither the combined performance nor its improvement against mean classifier performance seem to be measurable in a consistent and well defined manner. At the same time the most successful diversity measures, barely regarded as measuring diversity, are based on measuring error coincidences and by doing so they move closer to the definitions of combined errors themselves. Following this trend we decided to use directly the combining error normalized within the derivable error limits as a measure of classifiers dependency. Taking into account its simplicity and representativeness we chose majority voting error for the construction of the measure. We examine this novel dependency measure for a number of real datasets and classifiers showing its ability to model combining improvements over an individual mean.

1. Introduction

In many recent works dedicated to pattern recognition the efforts are being shifted towards classifier fusion as a way of further improvement of the recognition rate [1]-[3]. Combining classifiers is now perceived as a universal and obvious advancement of the single-best strategy, often even paraphrased as “gather all and combine” [1]. Very quickly it turned out that “combining all” is expensive and very rarely optimal, which initiated the studies concentrating on the question of what makes some combinations of classifiers work better than others. Individual mean performance was the first indicator to start with. However it has been shown that well performing recognition systems can also be constructed out of weak individually, but strong as a team, classifiers [4]. The system property responsible for the team strength is known under many names in the literature including: disagreement, diversity, independence etc, all denoting certain characteristics among classifier outputs [1]-[9]. More detailed analysis revealed that specific distributions of outputs corresponding to maximum error dispersion account for the optimal performance of the voting systems [10], [11]. In

general, however, each combiner has its own characteristic features and phenomena explaining successful combination for one combiner may no longer be applicable for others. Nevertheless the classifier diversity, in addition to the individual performances, is believed to explain all the amazing successes of classifier fusion.

The problems with the diversity started to emerge with the attempts of measuring it. Focused mainly on the outputs disagreement, a majority of diversity measures investigated in [12]-[15] showed very weak correlations with combined performances as well as their improvements. In [14] Shipp and Kuncheva illustrated even an apparent conflict between diversity tendencies shown by some measures as falling and by others as rising among AdaBoost generated classifiers as training progressed. An interesting finding emerged from our recent diversity investigations [16]. A substantial gain in the correlation with majority voting error has been observed for measures operating on error coincidences with an asymmetry to change of the individual classifier outputs in their definitions. Our measure ('Fault Majority') built using this principle and additionally exploiting some characteristics of the combiner it was designed for, showed the best average correlation with the majority voting error. In our further investigations into the concept of diversity, the analysis of different error coincidence levels within a pool of classifiers to be combined showed further improvement in the correlation to majority voting error [17]. Consequently the closer the measure to the definition of the combined error, the better its correlation with the combined errors and the better use of the measure.

In this paper we attempt to take a full advantage of these findings and propose to use the normalized combiner error directly as a measure of dependency. This proposition may seem naive but if we confront the properties of such a measure with aims that diversity measures are trying to achieve, this "naive measure" shows superior quality in all aspects. Majority voting combiner is ideal for that purpose as measuring its error is often computationally cheaper than applying complex diversity measures [10]-[17]. We try to explain and illustrate that there is no point of using diversity measures to model majority voting error rather than just using this error itself. Moreover, we show that unlike diversity measures, a measure of combiner error normalized within its limits can be used for modelling the performance improvement of the system. We justify our claims by a number of experiments with majority voting and indicate potential applicability for other combiners.

The remainder of the paper is organized as follows. In Section 2 we raise some doubts about the usefulness of diversity in the context of combining classifiers. Section 3 provides the definition of the novel measure based on the majority voting error. In Section 4 we consider possibility of reducing complexity of the measure. The following section shows the experiments conducted with real datasets and classifiers evaluating the new measure. Finally summary and conclusion are given in Section 6.

2. Problems with Useful Diversity

Diversity perceived as a certain dependency among variables is a well-established statistical concept. More recently, rapid technology development led to rediscovery of

the diversity in its entirely new meaning and importance in the context of combining evidences [5]-[8]. In the software development domain, the successful strategy turned out to be designing systems composed of diverse multi-version software implementations preventing from coincident failures [5], [6]. In the pattern recognition domain, diversity is claimed to be the property of the multiple classifier systems deciding about their performance [7], [8]. In general terms diversity is a clear concept of variety, multiplicity and the hope was that diverse (here different) classifiers should produce more reliable and improved classification results. It turned out however that different classifiers are not necessarily diverse and to evaluate the diversity one needs an appropriate measure. We believe that to be called useful with relation to combining classifiers, diversity measure should be:

1. well correlated with combiner performance and/or its improvement;
2. appropriately normalized between extreme values of combined performance;
3. simple or at least less complex than calculating the combined performance.

With respect to these requirements a vast majority of diversity measures completely fail on the first point, some of them fulfil the second and third condition, which additionally depends on a combiner used. The most naive is probably the illusion that a simple measure based on outputs disagreement can be correlated well with a number of different sometimes quite complex combining methods. It is also reflected in many unsuccessful attempts to define a universally useful diversity measure indicating potential benefits of combining [12]-[15]. It is therefore our belief that any measure of diversity, which could be used as a certain criterion for selecting classifiers to be combined, or deciding whether one should use a combination of classifiers in the first instance, should be designed in close connection with the combination method (i.e. majority voting, fuzzy templates, mixture of experts etc.) as indicated in [16] and [17]. In an extreme case, the combination performance, which we advocate in this paper, could be used instead of some kind of “universally useful diversity measure”. Trivially, as we intend to show, the measure of combined performance does perfectly everything what diversity measures unsuccessfully try to do and often does it even cheaper, using the same evidence.

3. Majority Voting as Dependency Measure

Majority voting (MV) is an example of a simple fusion operator that can be applied to combine any classifiers as their outputs can always be mapped, if necessary, to the binary representation. Given a system of M trained classifiers: $D = \{D_1, \dots, D_M\}$ applied for N input data x_i , $i = 1, \dots, N$ we can represent the system outputs as a binary matrix of outputs $Y^{N \times M}$ (0-correct, 1-error). The decision of majority voting combiner y_i^{MV} for a single i^{th} data sample can be obtained by the following formula:

$$y_i^{MV} = \begin{cases} 0 & \text{if } \sum_{j=1}^M y_{ij} < \lceil M/2 \rceil \\ 1 & \text{if } \sum_{j=1}^M y_{ij} \geq \lceil M/2 \rceil \end{cases} \quad (1)$$

A more detailed definition of MV including the rejection rule, observed for equal number of opposite votes 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 [11] we presented the majority voting error limits assuming that all classifiers perform at the same mean level. Recalling these limits for a specific mean classifier error e we have:

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

where E_{MV}^{\min} and E_{MV}^{\max} stand for the lower and upper limit of MV error respectively.

Now what we propose is that measuring majority voting error (1) and normalizing it within the limits (2) we obtain a very informative measure expressing the relative position of the majority voting error between its limits and call it Relative Error (RE) measure. To get an intuitive norm and relevant interpretation of the specific points of the RE measure we propose the following definition:

$$RE = \begin{cases} (E_{MV} - E_0) / (E_0 - E_{MV}^{\min}) & \text{if } E_{MV} \leq E_0 \\ (E_{MV} - E_0) / (E_{MV}^{\max} - E_0) & \text{if } E_{MV} > E_0 \end{cases} \quad (3)$$

where E_0 is a specific value of the system error, for which the RE measure is equal to 0. The graphical interpretation of the measure with marked components is shown in Figure 1. We consider two possibilities for the value E_0 , which is the majority voting error assuming classifier independence or in the second version just mean classifiers error. Note that for both versions of the RE measure, its values range within the limits $\{-1,1\}$ where the worst case: $RE=1$ corresponds to the maximum majority voting error possible and in the best case: $RE=-1$ the combined error reaches its theoretical minimum. For the case of E_0 being independent error, the measure resembles in its values the concept of negative and positive correlations with the same meaning of the values as discussed in [10]. In the experimental section we evaluate its performance as a dependency measure using real datasets and classifiers.

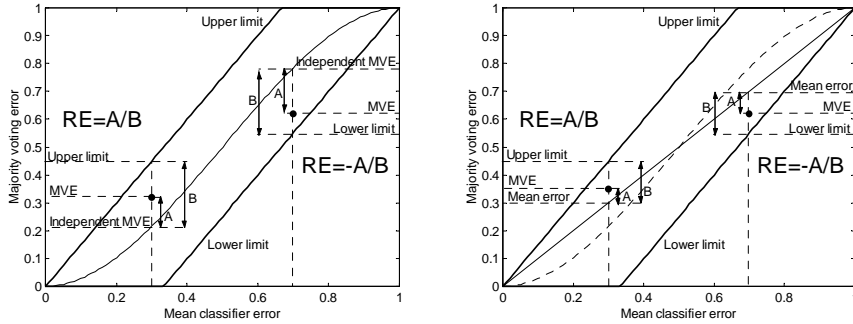


Figure 1. Graphical interpretation of the RE in two versions: with E_0 as independent majority voting error (left), E_0 denoting mean classifier error (right).

4. Complexity Reduction

The classifiers dependency measured by the performance of the combiner may be accused of being too complex and indeed it requires recalculating the majority voting error for different subsets of classifiers taken to the fusion. Applying the RE measure for all classifier combinations from a given pool imposes exponential complexity of the process equal to the exhaustive evaluation of the combiner performance. However the evaluative complexity of non-pairwise measures discussed in [12] is exactly of the same order and the cost of the individual measure is commonly higher. Moreover, seemingly simple pairwise measures in addition to their quadratic evaluative complexity have to add the complexity of measuring the diversity for all pairs within the combination to get the average value, which make them complex even using precalculated matrix of pairwise diversity values. Binary matrix representation of outputs from multiple classifiers is particularly useful for fast calculation of the majority voting. What is required to get the majority vote output for a particular sample is to sum all values within corresponding row of the matrix $Y^{N \times M}$ and check if this sum is greater or equal $\lceil M/2 \rceil$. Given matrix Y the error calculation is equally simple as it is just mean out of all combiner outputs, obtainable by just one line of code in Matlab notation:

$$E_{MV} = \text{mean}(\text{sum}(Y') > M/2);$$

Further potential reduction of the individual costs of the measure is available through the approximation of the majority voting error. In [17] we showed a decomposition of the majority voting error into levels of general or exclusive coincidences. The k^{th} level of general coincidence L_k^G is just the summing, for all the k -combinations of classifiers, of the sums of cases where all k classifiers fail. In the exclusive version L_k^E additional condition is that all the remaining classifiers should produce correct outputs. Using these definitions majority voting error can be decomposed as:

$$E_{MV} = \frac{1}{N} \sum_{i=\lceil M/2 \rceil}^M L_i^E = \frac{1}{N} \sum_{i=\lceil M/2 \rceil}^M (-1)^{i-k} \binom{i-1}{k-1} L_i^G \quad (4)$$

We also showed efficient ways of calculation of these coincidences within the framework of set analysis. Additional savings are potentially available provided it is possible to estimate any of the two types of coincidence levels (see [17] for details). Recalling the experimental results from [17], we observed a substantial loss in the correlation with majority voting error if just one of the required levels is missing. In other words approximation of a number of coincidence levels has to be extremely accurate to produce precise estimations of the overall MV error. A promising phenomenon was observed for the higher levels of general coincidences. Plotted in the logarithmic scale they showed a linear tendency as if the classifiers were independent starting from a certain order of dependency, but with substantially higher mean classifier error. Measuring just 2 points of this line would give the values of all other required levels and potentially the value of majority voting error (4). In the experimental section we examine potential applicability of this option and compare the approximated values with combiner errors.

5. Experiments

The experiments have been carried out in two groups. The first part provides comparative evaluation of the proposed RE measure in two versions discussed in Section 3. Applying a number of classifiers for real datasets, we examine the correlation between the RE and the improvement over individual mean error and compare it against representative diversity measures used so far. In the second experiment we examine the relevance of the components of RE measure. The reliability of the majority voting error limits is examined by comparing the true limits with the ones based just on the information of mean classifier errors. Relating to the majority voting error component of RE measure we show some examples of its modelling implementation discussed in Section 4. To maintain the generality of our findings we chose 11 commonly used classifiers and applied them for a classification of 4 real datasets taken mostly from the UCI Repository¹. The experiments have been carried out by applying 100 times random splitting into equally populated training and testing sets, the outputs obtained for classification of the testing sets have been hardened and stored in binary matrices.

5.1 Experiment 1

In this experiment both versions of the RE measure have been applied for all the combinations of 3, 5, 7, and 9 out of 11 classifiers. The same has been done with Q statistics measure and Double Fault measure discussed in [12]-[15] for comparison purposes. All these measures have been compared against the difference between the majority voting error and mean classifier error and the quality of a measure was evaluated by calculating correlation coefficients. Table 1 shows the comparative results for all examined datasets and Figure 2 depicts the meaning of the correlation coefficients for the combinations of 3 classifiers. The results clearly show the superior quality of the presented measures in comparison to Q statistics, which only for one dataset showed correlation coefficients reaching 0.9, and even worse Double Fault measure. The first version of the RE_{IMV} measure (with independent MV error as E_0) failed to capture the considered dependency for *Chromo* dataset, whereas in the second version, RE_{ME} (E_0 as a mean classifier error) showed an outstanding correlation with the difference between majority voting error and mean error for all considered datasets. Very high correlations commonly exceeding 0.98 suggest that the presented measures appropriately model performance improvement that can be obtained by majority voting comparing to mean classifier error. Moreover the quality of the measures does not fall for higher number of classifiers in a team. Nevertheless occasional failures (like for *Chromo* dataset with RE_{IMV}) may suggest that the measures do not deal well with large differences in classifier performances, which for *chromo* dataset exceeds 20% but further experiments are currently conducted to investigate this issue.

¹ University of California Repository of Machine Learning Databases and Domain Theories, available free at: <ftp.ics.uci.edu/pub/machine-learning-databases>.

Table 1. Correlations between the improvement of the majority voting error over the mean classifier error ($E_{MV}-E_{ME}$) and both versions of the RE measure compared against Q Statistics and Double Fault measures. The correlation coefficients were measured separately for the combinations of 3, 5, 7, and 9 out of 11 classifiers within each dataset.

#Clasf	Iris				Biomed				Satimage				Chromo			
	3	5	7	9	3	5	7	9	3	5	7	9	3	5	7	9
Q	32.8	32.8	30.7	47.4	90.1	95.8	95.4	93.9	79.9	80.7	78.7	76.5	-7.4	13.3	21.2	25.8
DF	27.3	34.7	40.8	38.8	28.8	26.5	23.5	21.0	43.1	48.2	51.9	56.1	32.4	63.4	76.6	78.5
RE _{IMV}	99.0	98.6	98.7	98.7	99.9	99.8	99.8	99.6	98.5	99.3	99.9	99.7	35.8	2.5	-5.9	-6.0
RE _{ME}	96.8	98.2	98.6	98.7	98.8	99.4	99.5	99.6	98.0	97.9	98.0	98.3	99.4	99.8	99.8	99.8

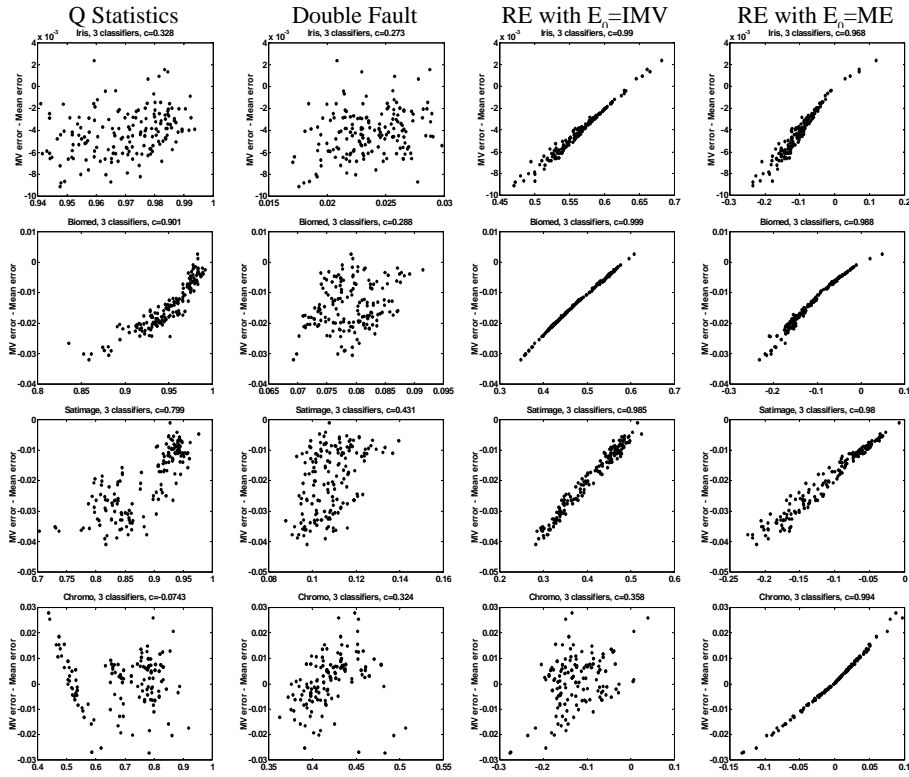


Figure 2. Visualization of correlations between the improvement of the majority voting error and the measures from Table 1. Coordinates of all points represent the measures examined for all 3-element combinations out of 11 classifiers for which the measures were applied.

5.2 Experiment 2

The definition of the RE measure consists of two major components: majority voting error and the estimates of its limits for given mean classifier error. The derivation of the limits is performed with an assumption that all classifiers to be combined have the

same classification performance. Realistically, classifiers perform differently, and there is a limited number of samples which means that realistic boundary distributions of errors in the binary matrix may result in slightly different or substantially different limits if the differences in performances are high. In the example of an extreme case of having 3 classifiers with individual error rates: {0.01, 0.5, 0.99}, as the mean is 0.5, definitions (2) result in the limits {0.25, 0.75} whereas in reality the limits are {0.49, 0.51}. However in realistic situations if the performance differences are high, the worst classifiers are rejected and consequently the final differences are never that extreme. The results in Table 2 showing extremely small absolute discrepancies from the real limits confirm our expectations. The limits component of the RE measure can be therefore considered as a fast and accurate estimation of the true MV error limits.

As discussed in Section 4, the majority voting error component can be calculated in a relatively fast way. Nevertheless definition of majority voting error (4) decomposed into sums of higher coincidence levels makes it tempting to model the error by approximation of the required coincidence levels. Following the observations of linear progression (in a logarithmic scale) of the general coincidence levels, we examine now the accuracy of the approximation based on information of just two boundary coincidence levels. Figure 3-A shows the evolution of the general coincidence levels in the logarithmic scale for four considered datasets. The crucial part of the higher levels is visibly quite close to the linear trend. For all datasets we applied linear regressions for the levels influencing majority voting error defined by (4), resulting in the lines appearing in Figure 3-B. Based on the approximated values of all general coincidence levels we then calculated the majority voting errors according to (4) and compared them with the real error values as shown in Table 3. Although the plots in Figure 3-B show quite satisfactory match for each coincidence level individually, small deviations from the true values result in big differences between the real and approximated majority voting error. It seems that approximation of the MV error by means of error coincidence levels introduced in [17] may not be suitable due to high sensitivity of the combined error to small deviations from the real values of coincidence levels. These results agree in a sense with our findings from error coincidence analysis performed in [17]. Nevertheless the results provide a new interesting characteristic of the multiple classifier system, which is the slope of the falling general coincidence levels. The slope coefficient m , where $f(x)=mx+b$, has a direct equivalence with the error rate of equally performing team of independent classifiers: $p=\exp(m)$. For both cases each consecutive coincidence level falls by multiplying the previous level by the equivalent independent error rate p .

Table 2. Average absolute relative abbreviations (in %) from the real upper limits of majority voting error imposed by applying definitions (2) based only on mean classifier error. Average abbreviations are calculated separately for each dataset out of all combinations of 3 classifiers.

Dataset	Iris	Biomed	Satimage	Chromo
Av. Abbrev. [%]	0.44	0.19	0.14	0.05

Table 3. Comparison between the real and approximated values of the majority voting error for all datasets and applying all 11 classifiers. The error rates are shown in percentages.

Dataset	Iris	Biomed	Satimage	Chromo
Real MVE [%]	3.64	9.97	14.76	56.57
Approx. MVE [%]	1.97	6.40	8.01	36.02

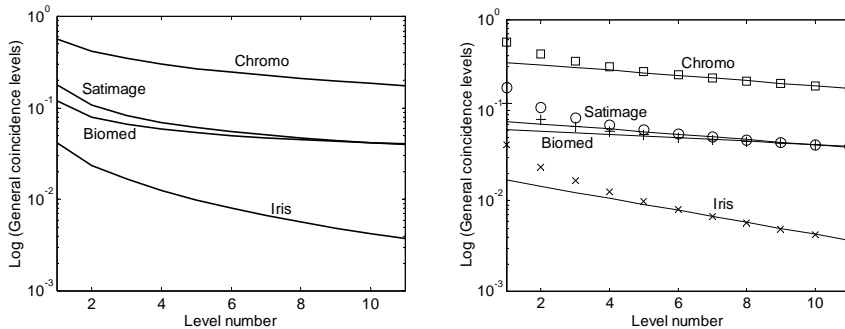


Figure 3. Linear regression of the normalized higher levels of general coincidence. A (left): Evolution of the values for increasing levels in logarithmic scale. B (right): Lines matched in the logarithmic scale to the higher levels (6:11) of general coincidence.

6. Summary and Conclusions

In this paper we attempted to show a new strategy for constructing the classifier dependency measure directly exploiting characteristics of the combining method applied to the classifier team. Inspired by the unsuccessful attempts of finding the relationship between the diversity measures and combining performance, we claim that at the moment it is cheaper and more accurate to measure the combining performance directly rather than using any of a number of existing diversity measures.

As a result we proposed a Relative Error measure which uses majority voting error scaled within accurately approximated error limits. We showed that all components of the RE measure are easy to calculate and its complexity remains of the same order as established diversity measures. Showing two versions of the RE we proved experimentally their very high correlations with the improvement of the majority voting performance over mean classifier error, the indicator commonly used to evaluate the quality of the diversity measures. Both versions have been scaled between -1 and 1 where: -1 indicates a maximum possible improvement while 1 indicates a maximum possible degradation of the performance. The difference between two versions is the meaning of 0 value, which in the first version corresponds to the majority voting error with the assumption of classifier independence, while for the second version means equal values of the majority voting error and average individual classifier error. While the performance of the second version of the measure was extremely good for all four data sets, the first version has failed quite spectacularly for the *Chromo* dataset. The reasons for this failure are not quite clear and require further investigations.

We have also attempted to model the majority voting error using approximated error coincidences which turned out to be inaccurate in predicting the error value though the error coincidences individually are approximated relatively well. However, what is very interesting is the fact that the majority voting error was consistently underestimated for all datasets which might indicate a possibility of using these estimates as lower limits for a given set of classifiers and data sets.

References

1. Sharkey A.J.C.: Combining Artificial Neural Nets: Ensemble and Modular Multi-net Systems. Springer-Verlag, Berlin Heidelberg New York (1999).
2. Gabrys B.: Combining Neuro-Fuzzy Classifiers for Improved Generalisation and Reliability. To be presented at the WCCI'2002 Congress, Honolulu, USA, (2002).
3. Gabrys B. Learning Hybrid Neuro-Fuzzy Classifier Models From Data: To Combine or not to Combine?. In Proceedings of the EUNITE'2001 Conference, Tenerife, Spain (2001).
4. Rogova G.: Combining the results of several neural network classifiers. *Neural Networks* 7(5) (1994) 777-781.
5. Partridge D., Krzanowski W.J.: Software diversity: practical statistics for its measurements and exploitation. *Information & Software Technology* 39 (1997) 707-717.
6. Littlewood B., Miller D.R.: Conceptual modeling of coincident failures in multiversion software. *IEEE Transactions on Software Engineering* 15(12) (1989) 1596-1614.
7. Sharkey A.J.C., Sharkey N.E.: Combining Diverse Neural Nets. *The Knowledge Engineering Review* 12(3) (1997) 231-247.
8. Partridge D., Griffith N.: Strategies for improving neural net generalisation. *Neural Computing & Applications* 3 (1995) 27-37.
9. Lam L., Suen C.Y.: Application of majority voting to pattern recognition: an analysis of its behaviour and performance. *IEEE Trans. on Sys., Man, and Cyber.* 27(5) (1997) 553-568.
10. Kuncheva L.I., Whitaker C.J., Shipp C.A., Duin R.P.W.: Limits on the majority vote accuracy in classifier fusion. *Pattern Analysis and Applications*, accepted, available at: <http://www.bangor.ac.uk/~mas00a/papers/lkpaa.ps.gz>.
11. Ruta D., Gabrys B.: A theoretical analysis of the limits of majority voting errors for multiple classifier systems. To appear in the journal of *Pattern Analysis and Applications*.
12. Kuncheva L.I., C.J. Whitaker.: Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy, submitted, available at: <http://www.bangor.ac.uk/~mas00a/papers/lkml.ps.gz>.
13. Shipp C.A., Kuncheva L.I.: Relationships between combination methods and measures of diversity in combining classifiers. *Information Fusion*, accepted, available at: <http://www.bangor.ac.uk/~mas00a/papers/csif.ps.gz>.
14. Shipp C.A, Kuncheva LI.: An investigation into how ADABOOST affects classifier diversity, submitted, available at: <http://www.bangor.ac.uk/~mas00a/papers/csIPMU02.ps.gz>
15. Whitaker C.J., Kuncheva L.I.: Examining the Relationship Between Majority Vote Accuracy and Diversity in Bagging and Boosting, submitted, available at: <http://www.bangor.ac.uk/~mas00a/papers/cjwst.ps.gz>.
16. Ruta D., Gabrys B.: Analysis of the Correlation Between Majority Voting Errors and the Diversity Measures in Multiple Classifier Systems. In Proceedings of the International Symposium on Soft Computing, Paisley, Scotland (2001).
17. Ruta D., Gabrys B.: Set analysis of coincident errors and its applications for combining classifiers, to appear in the upcoming book: *Pattern Recognition and String Matching*.