Set Analysis of Coincident Errors and Its Applications for Combining Classifiers

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. Although addressed in many papers, classifier dependency is still not well defined. Continuously being described by variety of statistical models from conditional probability to diversity measures, dependency among classifier outputs was recently shown to have a crucial impact on the performance of multiple classifier system. However, individual classifier performances still represent competitive and simple information clearly related to the performance of the combined system. In this work we show that all the measures related to classifier outputs can be reformulated to represent just different forms of the same information of error coincidences. Applying set analysis for the representation and description of error coincidences we define collection of classifier sets decomposed into two complementary types of coincidence levels. Furthermore we illustrate a high flexibility of using the coincidence levels, which supported be a simple algebra cover many established dependency measures including combining error in case of majority voting. Moreover we show that in the sets-collection representation of error coincidences a specific inclusion relation results in a quicker and more effective handling of dependency information under different complexity conditions. In the experimental section we examine relations of the introduced error coincidence levels with majority voting combiner using real datasets and classifiers and indicate further potential applications of the presented concepts.

1. Introduction

There is a common agreement in many recent publications related to pattern recognition that dependency among classifier outputs plays a key role in combining classifiers [1-8]. Diversity, independence, disagreement and most recently negative dependency are the terms often used to express a desirable relation among classifiers to ensure the maximum improvement of the fusion system [4-8]. In this variety of concepts the idea is the same: how to measure relationship among classifiers from their outputs so that it is possible to say something about the combined classifier performance?

Recent investigations indicate that error coincidences seem to be the most valuable information in this pursuit [7-10].

Coincident events are traditionally described by probabilistic models. For the case of independent events, their coincidence can be easily calculated from the product of the individual probabilities of events. However in the realistic pattern recognition situations it is very naive to assume independence among classifier outputs or even more negative diversity [9,10]. Moreover it is not uncommon that when different classifiers are applied for the same pattern recognition task they turn up to be strongly dependent and so are their errors showing similar patterns of misclassification [9]. To be complete, the probabilistic analysis of the dependent events of error occurrences would require exponentially complex calculations of conditional probabilities of all combinations of events and becomes unmanageable even for small number of classifiers [11]. What is more, probability estimations of higher order dependencies can be unreliable due to sparsity of examples unless vast amount of data is provided. To avoid these problems some studies consider simplifications in a form of using only lower order dependencies, possibly applying them to approximate higher order dependencies and completely ignoring or assuming independence for the highest order dependencies [11].

Alternatively, coincident errors can be represented by means of sets [12]. In this approach the errors from a single classifier are mapped into corresponding set of indices of misclassified samples. If more than one classifier misclassifies a particular sample then the index of this sample becomes an element in the intersection of sets corresponding to misclassifying classifiers. Using such representation, all available information related to error coincidences can be visualized as a complex architecture of overlapping sets resembling Venn Diagrams comprehensively discussed in [13]. Such collection of classifier sets represents a simple and coherent source of complete information about error coincidences. It encapsulates all conditional probability measures but at the same time preserves information about individual indices of misclassified samples, which would have been lost in the probabilistic representation of error coincidences.

Moreover, as we show in this work, applying a simple algebra on the cardinalities of different subsets of the collection of classifier sets leads to the derivation of various measures of dependency derived independently in the past. This relates for instance to the long-lasting debate of which information to choose: mean classifier error or pairwise diversity measure (i.e. 'Double Fault' measure discussed in [7,8]) to optimally select classifiers for a combining method [14]. We show that both types of information stem from common relation apparent in the set representation of coincident errors and for that reason instead of being competitive, they can be supportive as they both bring some new descriptive information.

We further consider potential applications of the set representation of error coincidences for majority voting operating on binary outputs (correct/incorrect), proved as simple yet quite powerful combining method [15-18]. We show a novel definition of majority voting error expressed by the predefined coincidence levels and investigate

its relation with cardinalities of lower and higher order error coincidences and their combinations.

The remainder of the paper is organized as follows. Section 2 gives a theoretical basis of a set-collection representation of error coincidences including definitions of two types of coincidence levels presented in the subsection 2.1, and an optimized method of collection generation and operations on error coincidences briefly described in section 2.2. In section 3 we show how the set analysis of error coincidences relates to the majority voting combiner and its error. Section 4 provides the results from the experiments with a number of real datasets and classifiers showing some numerical properties of the collection representation and correlations between majority voting error and different combinations of coincidence levels. Finally, summary, conclusions and further potential applications of the presented ideas are given in section 5.

2. Set Representation of Coincident Errors

Given a system of M trained classifiers $D = \{D_1,..., D_M\}$ applied for the classification of N data samples let $y_i = (y_{i1}, y_{i2},...y_{iM})$ denote a joint output of a system for the i^{th} input sample $\mathbf{x}_i$, where y_{ik} refers to individual output of the k^{th} classifier for that sample and $i = 1,..., N$, $k = 1,..., M$. In this work we consider only binary outputs assuming $y_{ik} = 1$ as an error and $y_{ik} = 0$ as a correct classification. The outputs form a binary matrix of outputs $Y^{N\times M}$, which is a starting point of our considerations.

For the purpose of combining, there is no need to keep the information about all the outputs but only about errors and their distribution among all the classifiers [8]. This fact makes sets analysis particularly attractive for the description of errors and the way the errors are shared by different classifiers. From the set analysis standpoint each classifier D_k can be associated with a set S_k, containing the indices of misclassified data samples. We call such sets the *classifier sets*. In a very common situation of more than one error for the same sample, its index has to be shared by corresponding classifier sets. We denote a complex system of overlapping classifier sets obtained in this way by $S = \{S_1,..., S_M\}$ and call it shortly a *collection*. The set representation of coincident errors can be now formally expressed as a mapping from a binary matrix of outputs into a collection:

$$Y \to S \quad \Leftrightarrow \quad Y^{N\times M} \to \{S_1,..., S_M\} \tag{1}$$

The collection can be visualized by means of Venn Diagrams [13]. Figure 1 shows an example of such diagram for given outputs from 3 classifiers. For a larger number of classifier sets, visualization of coincident errors is more complicated as it is difficult to visually represent all combinations of classifiers given in the form of exclusive subsets of classifier sets. Figure 2 shows some examples of Venn Diagrams for more than 3 classifier sets. Venn Diagrams and their construction represent a complex mathematical problem on its own and some further related details can be found in [13]. For the pur-

pose of this work only some properties of the Venn Diagrams and graphs associated with them will be used.

#	D_1	D_2	D_3
1	1	0	0
2	0	1	1
3	1	1	1
4	0	0	1
5	1	1	1
6	1	1	0
7	0	1	0
8	0	1	1
9	1	0	1

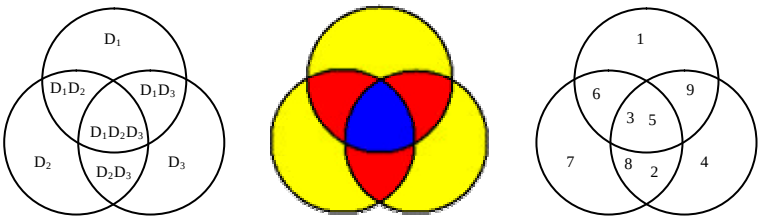


Figure. 1 Visualization of a set representation of coincident errors. A – binary outputs from 3 classifiers (0-correct, 1-error). B, C – Venn Diagrams showing all mutually exclusive subsets. D – Venn Diagram with the indices of samples put in the appropriate subsets positions.

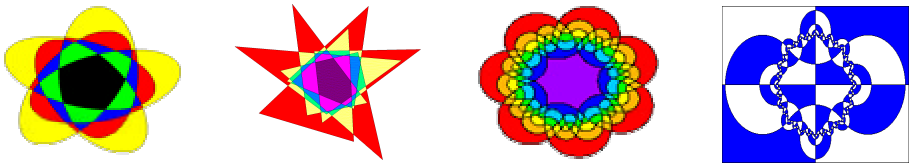


Figure. 2 Venn Diagrams for more than 3 classifiers. A: 5 congruent ellipses. B: 6 triangles. C: 7 symmetrical sets - Grunbaum construction. D: bipartite plot of 8 sets – Edward's construction. See [13] for further detail.

2.1 General and Exclusive Coincidences

Given a collection of classifier sets, the coincidences of errors from a specific combination of classifiers $\{D_{i_1}, D_{i_2}, ..., D_{i_k}\}$ can be viewed as subsets representing intersections of corresponding classifier sets.

$$C^G\left(\{D_{i_1}, D_{i_2}, ..., D_{i_k}\}\right) = C^G_{i_1 i_2 ... i_k} = \bigcap_{j=1}^{k} S_{i_j} \tag{2}$$

where $i_1 ... i_k$ represent indices of classifiers and:

$$\left|\{D_{i_1}, D_{i_2}, ..., D_{i_k}\}\right| = k \text{ [1]} \tag{3}$$

We call such type of coincidence the *k^{th}-order general coincidence* of the errors from classifier subset $\{D_{i_1}, D_{i_2}, ..., D_{i_k}\}$. We call it general as it does not depend on the other classifier sets, which may also overlap the considered intersection subset. On the other hand much different but not less important information related to the same subset of classifiers is given if considered intersection subset is not overlapped by any

[1] The symbol $|S|$ denotes the cardinality of a set S and means the number of elements in this set

other classifier set from the collection S. We refer to such intersection as the *k^{th}-order exclusive coincidence* and define it formally by:

$$C^{E}_{i_1 i_2 \ldots i_k} = \bigcap_{j=1}^{k} S_{i_j} - \bigcup_{j=k+1}^{M} S_{i_j} \tag{4}$$

Both coincidences are graphically shown in Figure 3-a for a simple example of 3 classifier sets.

To be able to consider relationships between these two types of coincidences effectively, we define a *k^{th}-level of coincidence* as a sum of all different coincidences of the same order and type as shown in the following formula:

$$L_k = \left|C_{i_1 i_2 \ldots i_k}\right| + \left|C_{i_1 i_2 \ldots i_{k-1} i_{k+1}}\right| + \ldots + \left|C_{i_{M-k+1} i_{M-k+2} \ldots i_M}\right| \tag{5}$$

Note that the number of coincidences to be summed is:

$$\left|\left\{C_{i_1 i_2 \ldots i_k}, \ldots, C_{i_{M-k+1} i_{M-k+2} \ldots i_M}\right\}\right| = \binom{M}{k} \tag{6}$$

An important fact about the introduced coincidences is that they both represent complete information about error relations among classifiers and in a sense complement each other. In different applications however one type of coincidence may be more useful than another for different reasons.

We consider now the relations between general and exclusive coincidence levels. Each k^{th}-level of one type of coincidence can be decomposed into a sum of equal and higher levels of the other type of coincidences as defined by a pair of formulas:

$$L^{G}_k = \sum_{i=k}^{M} \binom{i}{k} L^{E}_i \qquad\qquad L^{E}_k = \sum_{i=k}^{M} (-1)^{i-k} \binom{i}{k} L^{G}_i \tag{7}$$

An obvious consequence of the definition of exclusive coincidence, is that all their levels sum up to the cardinality of the union of classifier sets:

$$\sum_{i=k}^{M} L^{E}_i = \left|\bigcup_{i=k}^{M} S_i\right| \tag{8}$$

From (2) and (4) it is also clear that any exclusive coincidence is always smaller or equal to the corresponding general coincidence of the same order. Consequently the same relation applies to coincidence levels:

$$\left|C^{E}_{i_1 \ldots i_k}\right| \le \left|C^{G}_{i_1 \ldots i_k}\right| \quad \overset{(5)}{\Rightarrow} \quad L^{E}_k \le L^{G}_k \tag{9}$$

Note that for the highest level where $k = M$, both levels are equal: $L^{E}_M = L^{G}_M$. For the application purposes we consider a sum of higher levels of exclusive coincidences. It can be shown that adding L^{E}_i from certain $i = k$ to the highest level ($i = M$) gives the following simple expression:

$$\sum_{i=k}^{M} L_i^E = \sum_{i=k}^{M} \sum_{j=i}^{M} (-1)^{j-i} \binom{i}{j} L_j^G = \sum_{i=k}^{M} (-1)^{i-k} \binom{i-1}{k-1} L_i^G \quad (10)$$

Effectively, sequential summing and subtraction of consecutive general coincidence levels can easily represent the sum of exclusive coincidence levels.

Example: 3 classifier sets

Given 3 classifier sets $\{S_1, S_2, S_3\}$ we consider two scenarios, when different types of coincidence information is given and the other is to be retrieved.

1. *General coincidences are given.* In this case only cardinalities of different combinations of set intersections are necessary. This allows to calculate general coincidence levels according to (4) as shown below:

$$\begin{aligned} L_1^G &= |C_1^G| + |C_2^G| + |C_3^G| = |S_1| + |S_2| + |S_3| \\ L_2^G &= |C_{12}^G| + |C_{13}^G| + |C_{23}^G| = |S_1 \cap S_2| + |S_1 \cap S_3| + |S_2 \cap S_3| \\ L_3^G &= |C_{123}^G| = |S_1 \cap S_2 \cap S_3| \end{aligned} \quad (11)$$

   For the example shown in Figure 1, general coincidence levels would give the following values: $L_1^G = 5+6+6 = 17$, $L_2^G = 3+3+4 = 10$, $L_3^G = 2$.

   From (7) and given general coincidence levels (11), the sums of exclusive coincidence levels can be retrieved as follows:

$$\begin{aligned} L_1^E &= L_1^G - 2L_2^G + 3L_3^G \\ L_2^E &= L_2^G - 3L_3^G \\ L_3^E &= L_3^G \end{aligned} \quad (12)$$

   Again for the example in Figure 1, the values for exclusive coincidence levels would be: $L_1^E = 17 - 2 \cdot 10 + 3 \cdot 2 = 3$, $L_2^E = 10 - 3 \cdot 2 = 4$, $L_3^G = 2$

2. *Exclusive coincidences are given.* Optimally they can be extracted directly from the binary matrix of outputs Y, introduced in section 2. As in previous case the respective coincidence levels can be calculated according to (5):

$$\begin{aligned} L_1^E &= |C_1^E| + |C_2^E| + |C_3^E| \\ L_2^E &= |C_{12}^E| + |C_{13}^E| + |C_{23}^E| \\ L_3^E &= |C_{123}^E| \end{aligned} \quad (13)$$

   For the example shown in Figure 1, exclusive coincidence levels would give the following values: $L_1^E = 1+1+1 = 3$, $L_2^E = 1+1+2 = 4$, $L_3^E = 2$.

   And further transformed into general coincidence levels according to (7):

$$\begin{aligned} L_1^G &= L_1^E + 2L_2^E + 3L_3^E \\ L_2^G &= L_2^E + 3L_3^E \\ L_3^G &= L_3^E \end{aligned} \qquad (14)$$

with values: $L_1^G = 3+2\cdot 4+3\cdot 2=17$, $L_2^E = 4+3\cdot 2=10$, $L_3^G = 2$ for the example in Figure 1.

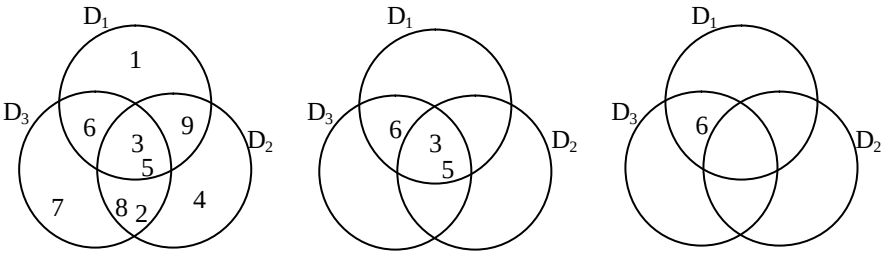


Figure. 3 Two types of error coincidences for the classifier D_1 and D_3 of the ensemble $\{D_1, D_2, D_3\}$. A: Example of error indices distribution. B: General coincidences $C_G(\{D_1, D_3\}) = \{3,5,6\}$. C: Exclusive coincidences $C_E(\{D_1, D_3\}) = 6$.

2.2 Collection generation

The question might be asked: is the collection representation of error coincidences any better than binary matrices? We show that the answer to this question a is very decisive *yes*. First immediate advantage of the collection is that it stores only valuable information related to errors. For the purpose of combined error modeling, the superiority of information related to errors rather than correct outputs was recently confirmed in [8] and formulated as asymmetry of the diversity measure with respect to the output change. Furthermore, whatever the purpose of using error coincidences, the information regarding any specific coincidence has to be accessible as fast as possible and the collection of classifier sets offers further advantages in comparison to the binary matrices. Rather than extracting all different combinations of classifier set intersections as it happens for the binary matrix of outputs, collection offers much simpler access to any coincidence information, exploiting inclusion relation among different coincidences. Inclusion properties can be very effectively implemented for both quick generation of collection data structure and extraction of any information in a graph propagation manner.

Generation of the collection of classifier sets is not a straightforward process and it requires a specific data structure. As shown in previous section general coincidences of higher orders are non-increasing in value. In practice they are decreasing very quickly as long as the classifier errors are not strongly correlated. Exploiting this fact we show a simple mechanism for rapid generation of all general coincidences. One

observation is here important. Namely, general coincidence of the k^{th} order can be most effectively obtained from the intersection of two general coincidences of the $k-1$ order as they have the most narrowed set of elements that we have to deal with. Starting from the lowest coincidence levels it is possible to devise a recurrent algorithm rapidly generating all general coincidence levels. Both the algorithm and an associated graph are shown in Figure 4.

Assuming that we have easier access to exclusive coincidences, we can immediately retrieve the general coincidences according to (7). Graphs associated with Venn Diagrams can be effectively used for an illustration of this procedure as shown in Figure 5. General coincidence of the k^{th} order is found as a sum of all exclusive coincidences (nodes) found on all the paths of ordered graph propagation starting from the highest order exclusive coincidence down towards the respective exclusive coincidence of the k^{th} order included in the considered general coincidence.

```
function gc=GenSetCon(c)
if size(coll)==2 then    %if 2 sets
   ints = c(1)∩c(2)  %intersection
   gc= lintsl          %cardinality
else
   M=size(coll);      %number of
sets
   gc=[];
   for i=1:M-1
      ints=[];
      for j=i+1:M
         ints=[ints,
c(i)∩c(j)];
      end:
```

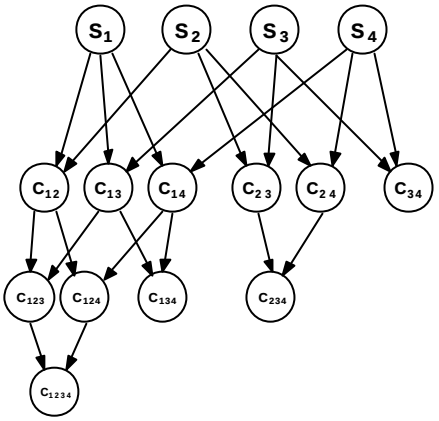
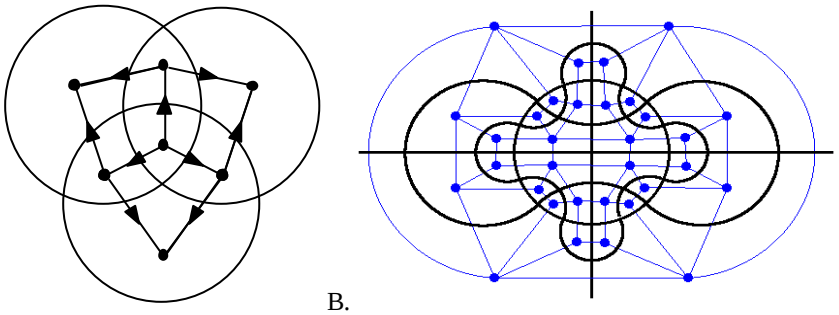


Figure. 4 Collection generation. A. Algorithm. B. Visualization of the collection generation process.



A. B.

Figure. 5 Graphs associated with Venn Diagrams. A. An ordered graph of exclusive coincidences for 3 classifier sets. B. Unordered graph for Edward's construction of 5 sets, to order the graph, all vertices have to be directed towards lower order coincidence.

3. Majority Voting Errors

Majority voting is an example of a simple fusion operator that can be applied for combining multiple classifiers with both soft and binary outputs. Moreover, 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: $D = \{D_1,..., D_M\}$ applied for N input data x_i, $i = 1,..., N$ we can obtain the binary matrix of outputs $Y^{N\times M}$ (0-correct, 1-error) introduced in Section 2. The decision of majority voting combiner y_i^{MV} for a single i^{th} data sample can be obtained according to the following formula:

$$y_i^{MV} = \begin{cases} 0 \text{ if } \sum_{j=1}^{M} y_{ij} \geq \lceil M/2 \rceil \\ 1 \text{ if } \sum_{j=1}^{M} y_{ij} < \lceil M/2 \rceil \end{cases} \quad (15)$$

For the whole binary matrix the estimate of the combined error would be then:

$$e_{MV} = \sum_{i=1}^{N} y_i^{MV} / N \quad (16)$$

A more detailed definition of MV including the rejection rule observed for $\sum_{j=1}^{M} y_{ij} = M/2$ when M is even can be found in [15]. 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. What is interesting about majority voting with respect to the presented set representation of error coincidences is that the definition of the MV error can be explicitly expressed by means of both types of coincidence levels as shown in the following relations:

$$e_{MV} = \frac{1}{N}\sum_{i=\lceil M/2 \rceil}^{M} L_i^E = \frac{1}{N}\sum_{i=\lceil M/2 \rceil}^{M} (-1)^{i-k}\binom{i-1}{k-1} L_i^G \quad (17)$$

Not surprisingly, only higher levels of coincidences decide about the level of majority voting. The question is however, whether a single coincidence level may convey a meaningful information about the performance of majority voting. In many recent publications the relation between various diversity measures and the performance of majority voting combiner has been studied [6-8]. In our work [8] we showed that two diversity measures: 'double fault' and 'fault majority' had especially high correlation to the majority voting error. The correlations were also better than for a measure based on a mean classifier error. The strength of both measures has been identified as resulting from the asymmetry of their definitions with respect to the classifier outputs and greater emphasis put on measuring coincidences of errors. The set analysis of error coincidences presented in this paper can be used to generalize the problem of error focused diversity measures and provide further tools to construct other effective diversity measures. For instance it can be noticed that the mean classifier error is an equivalent to the 1st order general coincidence level and the 'double fault' measure is

nothing else but a normalized 2nd order general coincidence level in our set representation presented above. Moreover, the most successful 'fault majority' measure has an explicit relation with exclusive coincidence levels and the whole concept of discrete error distributions shown in [8,16] on which it is based, represent normalized exclusive coincidence levels introduced in section 2.1. The usefulness of higher order coincidences for modeling of combiner performance is rather unclear and rarely discussed in the past mainly due to high complexity of such analysis. Nevertheless in the experimental section we examine in detail the relationship between different variations of coincidence levels and majority voting performance. The results of these experiments indicate significant potential applications of the set-based analysis of coincident errors for variety of purposes related to majority voting combiner.

4. Experiments

The experiments have been carried out in two groups. In the first simple experiment we intend to evaluate the performance of the collection data structure in terms of the time of accessing information about error coincidences and compare it against traditional but optimized method of coincidences retrieval from the binary matrix of outputs. The purpose of a second experiment is to comprehensively examine the relationship between majority voting error and different levels of error coincidences. In particular we want to reevaluate the significance of mean classifier error and 'double fault' diversity measure (here 1st and 2nd order coincidence levels) in relation to the majority voting performance and learn if the higher order coincidence levels may bring some new valuable information improving this relation. To maintain the generality of our findings we chose 11 out of 19 commonly used classifiers and applied them for a classification of 8 real datasets taken mostly from the UCI Repository of Machine Leaming Databases[2]. Short description of classifiers and datasets is provided in Table 1 and Table 2 respectively.

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

Table 1. Description of classifiers used in the experiments.

Classifier name	Description
Klclc	Linear classifier by KL expansion of common cov matrix
Kljlc	Linear classifier by KL expansion on the joint data
Loglc	Logistic linear classifier
fisherc	Minimum least square linear classifier
Ldc	Normal densities based linear classifier
Nmc	Nearest mean linear classifier
Nmsc	Scaled nearest mean linear classifier
Perlc	Linear classifier by linear perceptron
Persc	Linear classifier by non-linear perceptron
Pfsvc	Pseudo-Fisher support vector classifier
Qdc	Normal densities based quadratic classifier
Udc	Uncorrelated normal densities based quadratic classifier
Knnc	k-nearest neighbour classifier
Parzenz	Parzen density based classifier
Treec	Binary decision tree classifier
Lmnc	Feed forward neural network by Levenberg-Marquardt rule
Rbnc	Radial basis neural network classifier
Rnnc	Random neural network classifier
Bpxnc	Feed forward neural network classifier by backpropagation

Table 2. Description of datasets used in the experiments.

Dataset	#cases	#features	#classes
Wine	178	13	3
Iris	150	4	3
Thyroid	215	5	3
Texture	5500	40	11
Biomed	194	5	2
Liver	345	6	2
Satimage	6435	36	6
Chromo	1143	8	24

4.1 Experiment 1

To prove the efficiency of set representation of error coincidences we compared the time of extracting the cardinalities of all possible general coincidences from the collection data structure and from the binary matrix of outputs. For that purpose we generated a series of artificial binary matrices of outputs Y^{10000k} where $k = \{3,5,7,9,11\}$ equivalent to the outputs from $k = 3,...11$ classifiers with mean error of 40%. For each of these matrices we generated the respective collection data structure and for both matrix and collection we measured the time needed to calculate cardinalities of all possible combinations of general error coincidences extracted from

the binary matrix of outputs and from collection according to the algorithm shown in Figure 4. The results are shown in the Table 3.

Table 3. Comparison of the time needed to extract cardinalities of all general coincidences from a binary matrix of outputs and a collection for different number of classifiers.

# Classifiers	Binary Matrix – Time [s]	Collection – Time [s]
3	0.02	0.01
5	0.16	0.12
7	0.84	0.58
9	4.49	2.96
11	22.77	14.58

Although complexity of the process remains the same, the time of extracting information about all general coincidences is shorter for the collection representation of error coincidences. The time savings become more significant for the larger number of classifiers. The results confirm an advantage of the set representation of error coincidences in terms of accessibility and managebility of vital information about error coincidences.

4.2 Experiment 2

In this experiment we examined relation between majority voting error and individual and combined coincidence levels of both types. First step towards this goal was to prepare binary matrices of classification outputs. Initially we applied 19 different classifiers for 8 real datasets split 100 times randomly into equally populated training and testing set. Descriptions of both classifiers and datasets used are provided in Table 1 and Table 2 respectively. The classifier outputs obtained for only testing set were hardened and stored in large binary matrices of outputs. It turned very quickly that error coincidence analysis of 19 classifiers is very expensive computationally and virtually impossible to perform using available resources. For this reason only 11 classifiers have been chosen guided by the minimum average 'double fault' measure and the binary matrices reduced accordingly.

In the next step, out of these 11 classifiers we considered all combinations of 3, 5, 7 and 9 classifiers for which all exclusive and general coincidence levels have been calculated and stored together with the majority voting error associated with each combination. We measured dependency between majority voting error and different coincidence levels separately for different sizes of the classifier team. Effectively for each series of all k-element combinations of classifiers we obtained k correlation coefficients between majority vote errors and k series of 1^{st} to k^{th} order coincidence levels measured separately. Figure 6 shows the evolution of the correlation coefficients along increasing levels of general error coincidences for all examined datasets. Clearly there is greater relation of the lower levels of general coincidences. There is however no consistent rule on which of the lower levels is the most informative in terms of correlation

to the majority voting error. For the extreme example of *Wine* dataset 1st level of general coincidence is surprisingly completely uncorrelated with majority voting but already the 2nd level reaches very high value of the correlation coefficient above 0.9. On the other hand there is the example of *Liver* dataset for which the 1st level of general coincidence is the most correlated with combined error and the correlations for the following levels are falling dramatically and by the 3rd level remain in the completely uncorrelated state with correlation coefficients close to zero. Taking into account all the datasets from Figure 6, 1st and 2nd general coincidence levels share the position of most correlated information to the majority voting error. However, it does not have to be general rule as we can see for the *Texture* dataset where for the combinations of 9 classifiers the maximum correlation seems to be at the 3rd level and the shapes of the correlation curves suggest that for a larger number of classifiers this maximum could possibly be shifted towards higher coincidence levels. The tendency of increased correlation coefficient for the fixed level of general coincidence but rising number of classifiers is also evident for all datasets. The exception to this rule is observed only for the 1st level where this tendency is opposite but gets reversed between 1st and 2nd coincidence level. Effectively the emerging conclusion could be that for small number of classifiers the 1st coincidence level (equivalent to the mean classifier error) has a greater chance to be better correlated with majority voting than for large number of classifiers where the 2nd and possibly higher levels are more likely to reach the maximum correlation.

The equivalent relations between majority voting error and exclusive coincidence levels, are shown for just 2 datasets in Figure 7. The plots represent two patterns of correlation curves that we persistently observed for all datasets. For some datasets the correlation curves tend to peak around the middle level and fall on both sides as shown for the *Texture* dataset, whereas for other datasets the correlation curves were completely shapeless oscillating chaotically around zero or slightly above zero correlation. The first pattern is easy to explain, as the exclusive coincidence levels are the ones taken directly to the sum forming majority voting error definition. The middle levels are positioned at the decision boundary of majority voting and are the first taken to the error definition. The second pattern, very commonly observed among the datasets, proves however that exclusive coincidence levels on their own represent rather insufficient information for effective modeling of majority voting error.

Furthermore, apart from the relation to individual coincidence levels in the same manner we also examined the MV error relation to the three types of sums of coincidence levels:

1. k^{th} sum of type 1 of coincidence levels from 1 to k out of M classifiers.
2. k^{th} sum of type 2 of coincidence levels from k to M out of M classifiers.
3. k^{th} sum of type 3 of coincidence levels from $\lceil M / 2 \rceil$ to M out of M classifiers according to the formula (10).

These sums have been designed as simple examples of the combinations of coincidence levels and are intended to indicate whether the joint coincidence information could improve the correlation with majority voting error. Figure 8 shows the results for type 1 sums for general coincidence levels illustrated for 4 representative datasets. For comparison the correlation curves for individual general coincidence levels (shown in

Figure 6) are also shown here in thin lines. What is striking for the *Liver* dataset in the fact that the sum of the first two general coincidence levels is better than the 1st level individually despite drastic fall of correlation for the 2nd level individually. In other words the pairwise coincidence information brings such vital additional information to just mean classifier errors that jointly this information is the most meaningful in terms of majority voting error. If correlation curves are not sloping down that fast the sum of more than two first levels may be optimal. For *Satimage* dataset the sum of the first 3 and for the *Thyroid* the sum of the first 5 general coincidence levels gave the maximum correlation with the majority voting error. In the extreme cases like for the *Wine* dataset the sum of coincidence levels is not optimal due to disastrous contribution of a completely uncorrelated 1st level. Nevertheless addition of further general coincidence levels substantially raises correlation coefficients. Summarizing, the sum of type 1 of the general coincidence levels may substantially improve correlation with majority voting error comparing to the correlation with individual coincidence levels. Furthermore it prevents to a certain degree, situations of completely uncorrelated measure that can happen for individual general coincidences as shown for *Liver* and *Wine* datasets. The gain in the correlation to the majority voting error is likely to continue by adding further general coincidence levels when correlation curves for individual coincidence levels are falling slowly.

Individual examples of the sums of type 2 and type 3 are shown in Figure 9 but they did not bring any significant results. This is somewhat surprising as in the special cases they become exactly the definition of majority voting error represented by exclusive and general coincidence levels respectively. The unit correlations between type 2 sums of exclusive coincidence levels are observed for summing the higher levels starting from the middle ($\lceil M / 2 \rceil$ for M classifiers) as shown for *Satimage* dataset in Figure 9; the unit correlations between type 3 sums of general coincidence levels are observed for summing the higher levels starting from the middle according to (10) as shown for *Biomed* dataset in the same figure. However only small variations from these precisely defined sums result in a total decorrelation of their values with the majority voting error. In that sense, additionally due to costly information of highest levels of coincidences they require, their significance is rather weak.

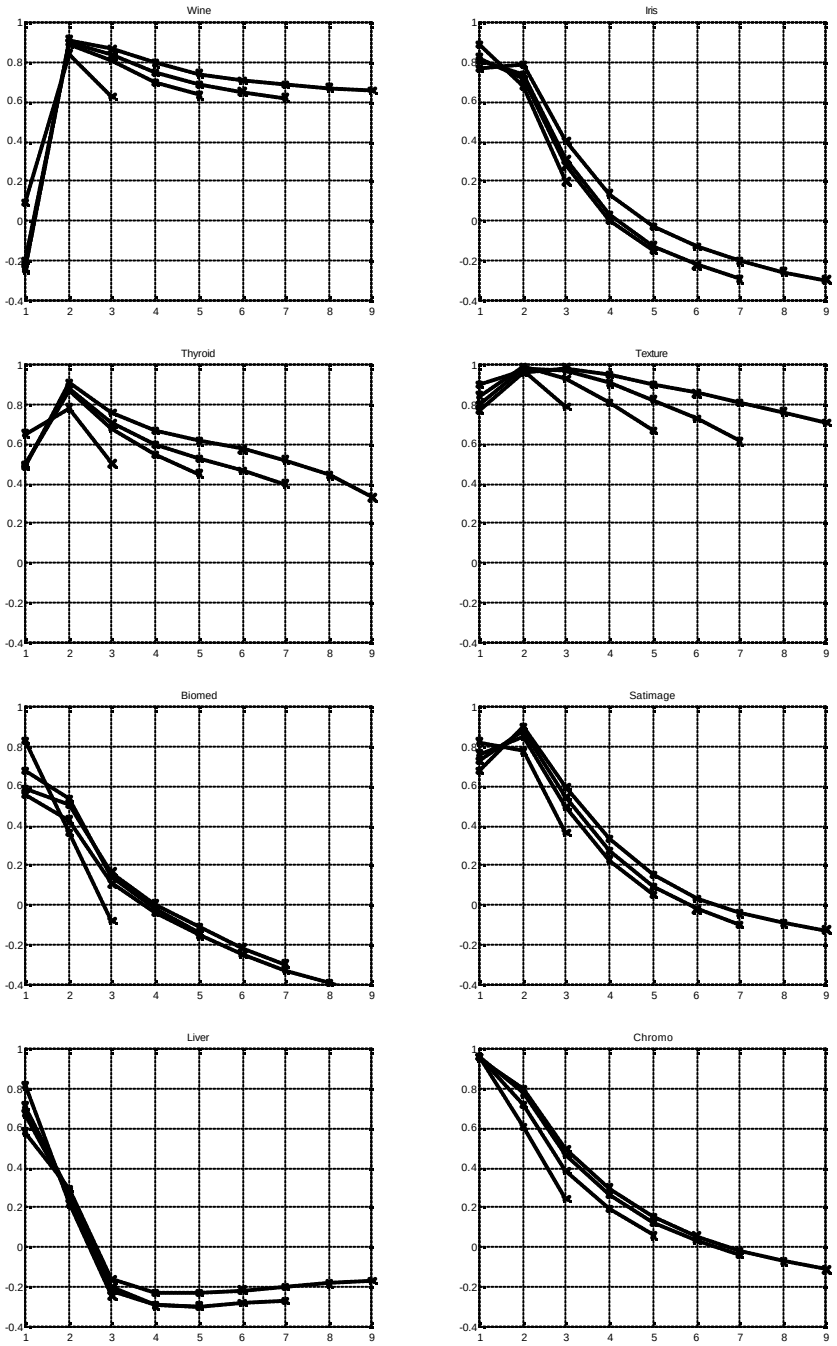


Figure. 6 Correlation coefficients between majority voting error and individual general coincidences grouped in series of 3, 5 ,7 ,9 out of 11 classifiers for 8 considered datasets as marked.

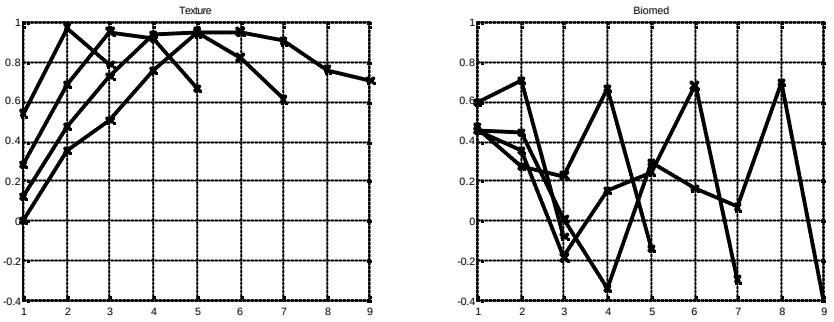


Figure. 7 Correlation coefficients between majority voting error and individual exclusive coincidences grouped in series of 3, 5, 7, 9 out of 11 classifiers for 2 representative datasets showing two patterns of the relationship observed for all 8 examined datasets.

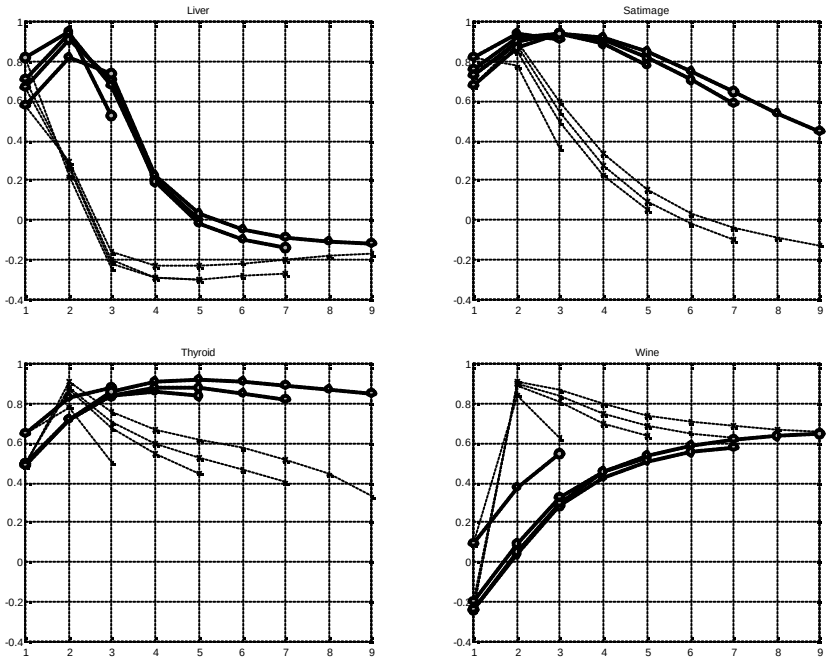


Figure. 8 Correlation coefficients between majority voting error and type 1 sum (from 1st to kth level) of general coincidence levels here showed in bold lines. For comparison, correlation curves of the individual general coincidences are also shown in thin lines. 4 out of 8 examined datasets with most representative patterns of the relationship are presented are presented.

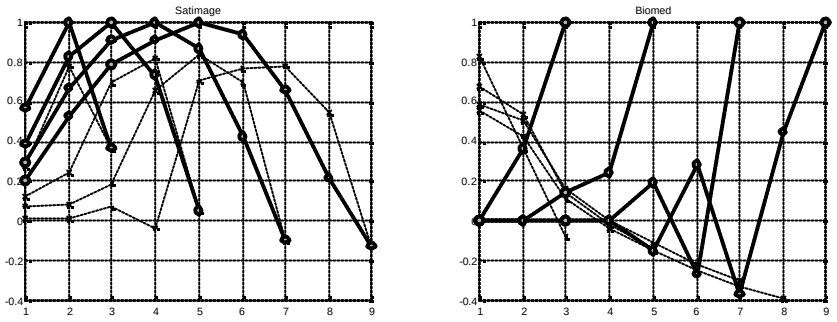


Figure. 9 Correlation coefficients between majority voting error and: LEFT: type 2 sum (from k^{st} to M^{th} level) of exclusive coincidence levels here showed in bold lines. For comparison, correlation curves of the individual exclusive coincidences are also shown in thin lines, RIGHT: type 3 sum of general coincidence levels with correlation curves of the individual general coincidences shown for comparison in thin lines.

5. Summary, Conclusions and Future Work

In this paper we attempted to show that relations among errors of multiple classifiers could be very effectively modeled by a collection of overlapping sets holding indices of misclassified samples for each classifier. Comparing this set representation of error coincidences against complete binary matrix of classifier outputs we proved experimentally that the collection of classifier sets offers faster access to the information about error coincidences. This important advantage has been achieved by excluding redundant information related to the correct classification outputs and efficiently exploiting inclusion relation among coincident errors for the algorithm of fast coincidence retrieval. Comparing against probabilistic models of error coincidences our set representation covers all possible probabilistic measures with the additional advantage of keeping indices of individual samples misclassified by particular classifiers. On top of that the error coincidences represented by a collection of sets can be to a certain degree visualized by Venn Diagrams, examples of which have been shown in this work, contributing to a greater understanding of the complexity of the error relations.

Enormous number of subsets arising from the collection of overlapping sets inspired us with the definition of the two complementary types of coincidences: exclusive and general. Exclusive coincidences expressing joint error incidence of selected classifiers with the presence of all correct outputs from the other members of the team represents more constrained information than general coincidence completely ignoring the background information. To handle such distinct information arising from a vast number of subsets and also for the application purposes we defined respective coincidence levels numerically describing the sum of cardinalities of all coincidence subsets of the same number of classifier sets. By performing a simple algebra on the coinci-

dence levels we derived definitions of a number of diversity measures operating on error coincidences. Not only mean classifier error (1st general coincidence) and 'double fault' measure (2nd general coincidence level) but also complex 'fault majority' measure and even majority voting error itself is shown as a quantity derivable from simple operations on coincidence levels.

Developing the findings from our previous work [8] related to the successful asymmetric diversity measures, we thoroughly investigated the relations between majority voting errors and individual and combined error coincidence levels examined by the correlation coefficients. For a number of real datasets we showed experimentally that general coincidences represent the information much better related to the majority voting errors than exclusive coincidences. What is even more precious the lowest levels of general coincidences, which represent the least complex measures, appear to be the most informative in terms of the relation to the majority voting error. Unfortunately, the linearly complex 1st level of general coincidence representing simply mean classifier error is not always the best choice. It appears that the 2nd level of general coincidence ('double fault') would show the best average correlation with the majority voting error, which confirms the findings from [8]. Further analysis revealed that for some datasets the 3rd general coincidence level exhibited the higher correlation with the majority voting error and for a larger number of classifiers the optimal level is quite likely to be shifted towards higher coincidence levels. Another novel finding is that combining general coincidence levels in a form of simple sums results very often in further improvement of the correlation with majority voting error observed as rising even if added level is on the falling individual tendency. This proves that different coincidence levels bring some portions of a unique valuable information related to the majority voting, which jointly can be used to achieve better correlated measures. Only for the cases when the first level is much less correlated than the second level, the sum of the first 2 general coincidence levels is not the optimal choice. For more steady individual correlation curve, the sum of even more first levels shows optimal. The a priori prediction of the evolution of the correlation curves is however difficult and possibly needs additional information related to the dataset and classifiers.

Summarizing, lower levels of general coincidences both individually and in simple combinations show high correlations commonly exceeding 0.9 with the majority voting errors. This information is very vital for a variety of applications related to the majority voting and possibly other combining methods. The above set analysis of error coincidence provides a good starting point for construction of simple yet well correlated with majority voting error diversity measure. The high correlation with combiner error means that it is possible to use the measure for selection of the best combination of classifiers or more securely a number of best combinations as shown in Figure 10.

Furthermore we intend to apply the above set analysis of error coincidences for the low cost approximation of the value of the majority voting error according to the novel definition (17), which just requires estimation of any type of coincidence levels. The information about error coincidences can be also applied for the analysis of the limits

of majority voting errors. Specifically, interesting is the level to which the combined error limits shrink with incoming information of error coincidences.

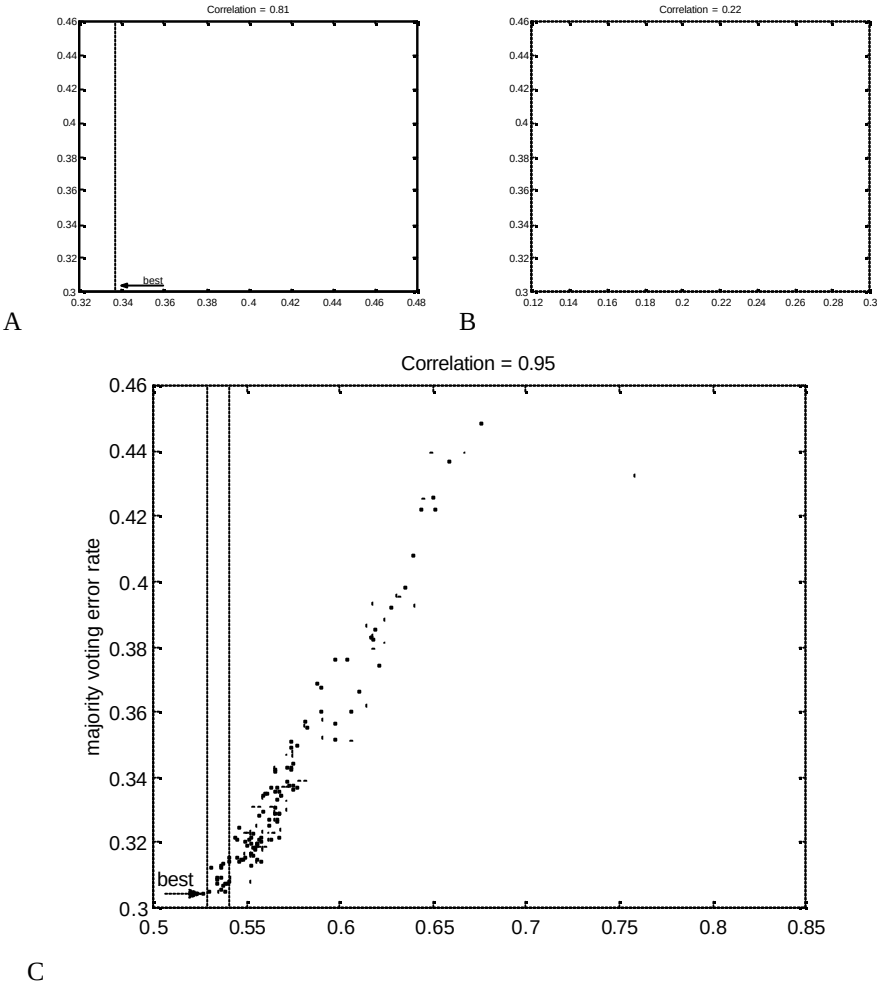


Figure. 10 Illustration of importance of correlation coefficient for classifier selection in the example of the dependence between general coincidence levels of 3 out of 11 classifiers and the majority voting error for the *Liver* dataset. A. Relation of the first general coincidence levels. B. Relation of the second general coincidence levels. C. Relation of the sum of the first and second general coincidence level with majority voting error.

References

1. Sharkey A.J.C. Combining Artificial Neural Nets: Ensemble and Modular Multi-net Systems. Springer-Verlag, Berlin Heidelberg New York (1999).
2. Sharkey A.J.C., Sharkey N.E.: Combining Diverse Neural Nets. The Knowledge Engineering Review 12(3) (1997) 231-247.
3. Rogova G. Combining the results of several neural network classifiers. Neural Networks 1994; 7(5): 777-781.
4. Partridge D, Griffith N. Strategies for improving neural net generalisation. Neural Computing & Applications 1995; 3: 27-37.
5. Kuncheva L.I., Whitaker C.J.: Measures of Diversity in Classifier Ensembles. Submitted to Machine Learning.
6. Shipp C.A. and L.I. Kuncheva, Relationships between combination methods and measures of diversity in combining classifiers, *Information Fusion*, accepted.
7. Kuncheva L.I., C.J. Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy, submitted.
8. Ruta D., Gabrys B.: Analysis of the Correlation Between Majority Voting Errors and the Diversity Measures in Multiple Classifier Systems. International Symposium on Soft Computing, Paisley (2001).
9. Petridge D., Krzanowski W.J. Software diversity: practical statistics for its measurements and exploitation. Information & Software Technology 39 (1997) 707-717.
10. Littlewood B, Miller DR. Conceptual modeling of coincident failures in multiversion software. IEEE Transactions on Software Engineering 1989; 15(12): 1596-1614.
11. Kang H.-J., Kim K., Kim J.H. Optimal approximation of discrete probability distribution with kth-order dependency and its application to combining multiple classifiers. Pattern Recognition Letters 18 (1997) 515-523.
12. Devlin K. Joy of sets: fundamentals of contemporary set theory. 2nd ed. Springer-Verlag New York (1993).
13. Ruskey F. A Survey of Venn Diagrams. The Electronic Journal of Combinatorics. Ed March (2001), available at http://www.combinatorics.org/Surveys/ds5/VennEJC.html
14. Giacinto G., Roli F. An approach to the automatic design of multiple classifier systems. Pattern Recognition Letters 22 (2001) 25-33.
15. 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.
16. Ruta D, Gabrys B. A theoretical analysis of the limits of majority voting errors for multiple classifier systems. Pattern Analysis and Applications, accepted.
17. Battiti R, Colla AM. Democracy in neural nets: voting schemes for classification. Neural Networks 1994; 7(4): 691-707.
18. 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, accepted.