Diversity measures for majority voting in the spatial domain

Andras Hajdu, Lajos Hajdu,
Laszlo Kovacs, and Henrietta Toman

Faculty of Informatics, University of Debrecen
Egyetem ter 1, 4010 Debrecen POB 12, Hungary
hajdu.andras@inf.unideb.hu,hajdul@math.klte.hu
{kovacs.laszlo.ipgd,toman.henrietta}@inf.unideb.hu

Abstract. The classic majority voting model can be extended to the spatial domain e.g. to solve object detection problems. However, the detector algorithms cannot be considered as independent classifiers, so a good ensemble cannot be composed by simply selecting the individually most accurate members. In classic theory, diversity measures are recommended that may help to explore the dependencies among the classifiers. In this paper, we generalize the classic diversity measures for the spatial domain within a majority voting framework. We show that these measures fit better to spatial applications with a specific example on object detection on retinal images. Moreover, we show how a more efficient descriptor can be found in terms of a weighted combination of diversity measures which correlates better with the accuracy of the ensemble.

Keywords: classifier combination, majority voting, spatial domain, diversity measures, biomedical imaging

1 Introduction

In decision making, the accuracy of the decision can be increased by composing ensembles from individual classifiers. In our previous work [1], we generalized the classical majority voting model to be applicable in the spatial domain. Namely, we introduced the terms $0 \leq p_{n,k} \leq 1$ describing the probability of a correct decision if k correct votes are present among the total number of n. This generalization was motivated by object detection problems in digital images, where image processing algorithms (detectors) are the members of the ensemble. Each individual algorithm votes in terms of a single pixel/region as its candidate for the center/object in the image domain. The region matching the geometry of the object with maximal number of candidates included is considered as the decision. Only the votes falling inside a proper region can vote together for the object. A good decision can be made even if the false candidates have majority, while bad decision is made only when a subset of false candidates with larger cardinality than the number of correct ones can be covered by a region matching the geometry of the object.

In classic majority voting, only the correctness of the votes influences the decision. However, in the object detection scenario, the spatial behavior of the votes are also important. Majority voting can be applied in the generalized model with further geometrical constraints (e.g. the spatial closeness of the candidates) that can be described by the terms $p_{n,k}$.

We applied the generalized model for the detection of the optic disc (OD) in retinal images. The OD is a bright region with circular shape having diameter d_{OD} (clinically predetermined constant). For the output of each detector for the OD center we consider the minimal bounding circles for all subgroups of the candidates. The circle with maximal number of candidates, having diameter less than or equal to d_{OD} is chosen for the OD as it is illustrated in Fig. 1.

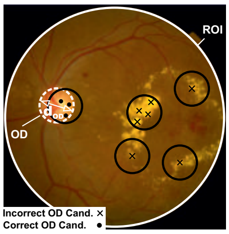


Fig. 1. Optic disc detection using spatial majority voting, where the black circles show the possible hotspots containing different number of OD candidates. The black dots and the black crosses represent the true and false OD candidates respectively.

In our application, the participating OD detector algorithms have individual respective accuracies p_1 = 0,6472; p_2 = 0,9765; p_3 = 0,3205; p_4 = 0,7593; p_5 = 0,3153; p_6 =0,2276; p_7 = 0,9582; p_8 = 0,7671, [2] on the Messidor dataset [3] containing 1200 retinal images with resolution 2240*1488 pixels. All the quantitative results presented later in the paper correspond to these.

In the literature of classic majority voting, several results are achieved for independent voters, but in object detection problems, the individual algorithms can hardly be expected to be independent. Besides the individual accuracies of the detector algorithms, the dependencies among them should also be taken into consideration, when an ensemble is composed from them.

In decision making theory, a possible simple approach to estimate dependency of the members is to consider diversity measures. These measures are defined between classifiers in [4]. In [5], it is proposed that we can reach optimally performing classifier combination by making up classifiers with high individual accuracies and sufficient level of diversity at the same time. Several earlier works (e.g. [6, 8]) confirmed that neither individual performances nor diversity alone can guarantee that the ensemble outperforms all the individual classifiers. Recent works (e.g.[4]) have been focused on finding suitable diversity measures, when majority voting is considered as a decision rule. Our motivation is to check the

reliability of these diversity measures in the spatial domain and to generalize them for better performance. The generalization is done in a natural way: we follow similar principles here that are considered also in the generalization of majority voting to the spatial domain.
The rest of the paper is organized as follows. In section 2 we list the diversity measures recommended in classic theory. Section 3 is dedicated to the generalization of the classic diversity measures. In section 4, we compare the performance of the classic and generalized measures in our spatial application. Section 5 introduces a novel methodology to derive a combined diversity measures from the individual ones. Finally, in section 6, we draw some conclusions.

2 Diversity measures in classic voting theory

Depending on whether it assesses the pairwise or groupwise dissimilarity, two types of diversity measures are considered. If a system of M classifiers $D = \{D_1, \dots, D_M\}$ is given, let y_{ij} denote the classifier output of the j-th classifier for the i-th input sample. Let $\mathbf{y}_i = [y_{i1}, \dots, y_{iM}]^T$ denote the joint output of a system for the i-th input sample x_i. Assuming that the output has binary form, $y_{ij} = 1$ means correct, while $y_{ij} = 0$ means incorrect classification. As the measures are mainly based on simple binary algebra, the following simplifications can be introduced, if we compare two classifiers with a diversity measure. Let N^{ab}, $a, b \in \{0, 1, *\}$ denote the number of input samples, where $*$ stands for any of the output 0 or 1. Here a belongs to the first classifier and b to the second one; i.e. N^{ab} and N^{ba} are different. The number of classifiers producing error on the input sample x_i $(i = 1, \dots, n)$ is denoted by $m(x_i)$ which can be expressed as $m(x_i) = \sum_{i=1}^{N} (1 - y_{ij})$. Finally the error rate of the j-th classifier can be calculated as $e_j = \frac{1}{N} \sum_{i=1}^{N} (1 - y_{ij})$.

In the literature (see [4, 8]) the following diversity measures are defined: minimum individual error, mean error, majority voting error, majority voting improvement, correlation coefficient, product-moment correlation measure, Q-statistics, disagreement measure, double-fault measure, entropy measure, measure of difficulty, Kohavi-Wolpert variance, interrater agreement measure, fault majority measure. Now we give a brief overview of some diversity measures from those can be considered for generalization to our spatial model.

- *The correlation coefficient $C2$:* it is a well known and frequently used statistical measure. For binary classifier output its definition takes the form:

$$C2_{ij} = \frac{N^{11}N^{00} - N^{01}N^{10}}{\sqrt{N^{1*}N^{0*}N^{*1}N^{*0}}}.$$

- *The disagreement measure $D2$:* it depends on the number of samples for which the classifiers disagreed and the total number of observations. It is calculated as:

$$D2_{ij} = \frac{N^{01} + N^{10}}{N}.$$

- *Mean error ME:* this measure takes the average of individual classifier error rates within the ensemble and can be defined by the following formula:

$$\bar{e} = \frac{1}{M} \sum_{i=1}^{M} e_j.$$

- *Interrater agreement measure IA:* this measure characterizes the level of agreement. With the notation presented above it can be expressed as:

$$IA = 1 - \frac{\sum_{i=1}^{N} m(x_i)(M - m(x_i))}{NM(M-1)\bar{e}(1-\bar{e})}.$$

3 Generalized diversity measures for the spatial domain

The diversity measures in section 2 give useful information on how to select the members to achieve the highest ensemble performance. More specifically, we can consider the correlation between the diversity measures and the system accuracy. In the literature, the case when the classifier decision making method is not the majority rule is rarely examined. These measures can be modified to the non-majority voting case in the spatial domain. In many image processing applications, more algorithms are used to detect the same object in the image. These algorithms can be considered as classifiers in a fusion method and the output of the algorithms (pixels/regions) as votes. In this voting method the good votes have to fulfill some geometrical constrains. Good decision can be made even in that case when the number of good votes is less than the half of the total votes. To achieve the best performance, the algorithms not making coincident error have to be combined. It can be proved that the modified diversity measures for the non-majority case can provide the possibility for better classifier selection. The aim of modifying the diversity measures is to reach higher correlation between them and the system accuracy. We could modify the calculation of the classic measures so that the original coherence in our specific environment is described. In this way, the generalized diversity measures consider the geometrical constraints, adopting them with getting close to each other.

This modification is logical, since close votes outside the good area cause the main problem. Some other interpretations of the pairwise diversity measures were investigated, as well. In some cases all, the variants correlated more with the system accuracy than the original diversity measure. The following formulas correlated most with the system accuracy are introduced here as generalized diversity measures for the spatial domain:

- *The generalized correlation coefficient $C2'$:*

$$C2'_{ij} = \frac{N^{11}N^{0'0'} - N^{01}N^{10}}{\sqrt{N^{1*}N^{0*}}N^{*0}}.$$

To handle spatial behavior of votes, now we consider also the notation $N^{0'0'}$. This figure stands for the number of cases, where for a pair of classifiers both of them made bad decision and these votes also fulfill the geometric constraint (that is, close to each other). Similarly, N^{00} means that though both algorithms give bad candidates, it does not mean a problem, because the geometrical constraints are not satisfied, so the chance for a final bad decision is not increased.

The modification of the other diversity measures, defined between two classifiers, can be interpreted in the same way. For the disagreement measure and that numbers describing the whole system of classifiers, (e.g. the interrater agreement measure), the generalization for our model needs some further modifications.

- *The generalized disagreement measure* $D2^{'}$*:* it depends on the number of samples for which the classifiers disagreed and the total number of observations. In this case all possible disagreement situations have to be described in the modified formula. It can be written as:

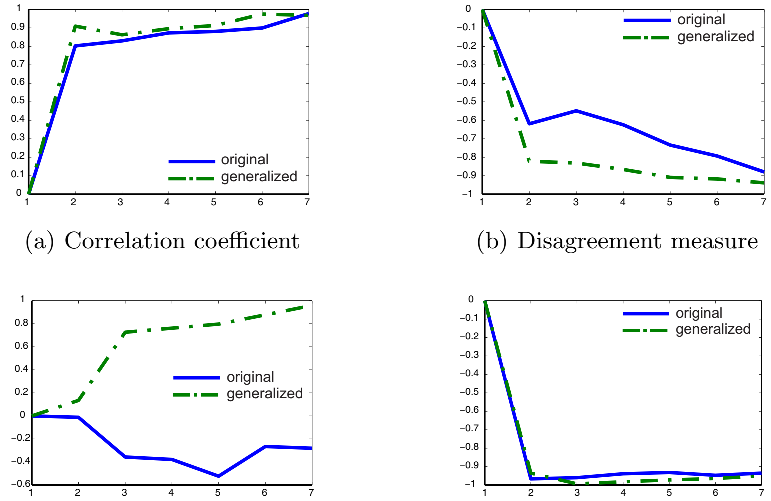
$$D2^{'}_{ij} = \frac{N^{01} + N^{10} + N^{0'1} + N^{10'} + N^{0'0} + N^{00'}}{N},$$

  where for example $N^{0'1}$ describes the number of the situations where one of the classifiers give bad vote fulfilling the geometrical constraints and the other give good vote.
- *The interrater agreement measure* $IA^{'}$*:* this measure characterizes the level of agreement. With the notation presented above it can be expressed as:

$$IA^{'} = 1 - \frac{\sum_{i=1}^{N} m'(x_i)(M - m'(x_i))}{NM(M-1)\bar{e}(1-\bar{e})}.$$

  In the classic formula $m(x_i)$ is the number of classifiers producing error for the input sample, and $m'(x_i)$ expresses the number of bad votes which are relevant in making the final decision, so the bad candidates fulfill the geometrical constraints, as well.

The plots in Fig. 2 (a), (b), (c) and (d) show examples about the effectiveness of the generalized diversity measures. Each dashed line shows the correlation between the system accuracy and the modified diversity measures. It can be observed that after modification this correlation is increased for each diversity measure.
Another interesting fact is that in the spatial domain we can handle ensembles consisting of an even number of voters, as well. Namely, in most of the classic studies the results are presented only for odd number of classifiers. The reason is that in classical majority voting, adding a new classifier can drop the system accuracy, so we cannot guarantee to achieve better performance because the parity of the number of the classifiers changes. This phenomena can be observed by the correlation curve of the diversity measures, as well.



(a) Correlation coefficient

(b) Disagreement measure

(c) Interrater Agreement measure

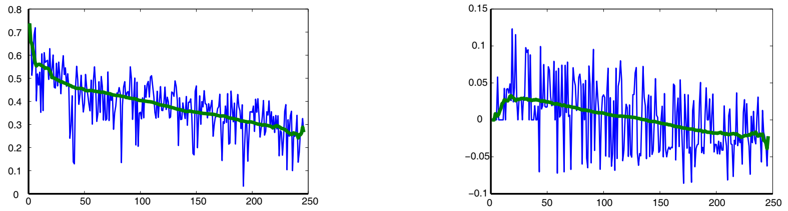
(d) Ensemble mean error

Fig. 2. Comparison of the generalized (dashed line) and the original (solid line) pairwise (a),(b) and non-pairwise (c),(d) diversity measures. The x-axis represents the number of the classifiers in the ensemble, while the y-axis the correlation value. The higher the absolute value of the correlation is, the better the effectiveness of the diversity measures is.

4 Distortion of the ensemble members

For generating the most accurate ensemble, the distortion of the algorithms is a relevant issue for applications. The distortion can be described as the difference between the optimal (real), and the actual output of the algorithms. If in such cases, the reason or the magnitude of the distortion is known, the inverse distortion vector can be calculated. By the help of this vector, the deviation can be reduced and the actual output can get closer to the optimal (real) value. The diversity measures can be built in our generalized model which is used for optic disc detection as an application. Using the inverse distortion, the achieved performance of the ensemble system is relevantly higher than the original (distorted) one. In this section we show, that for the diversity measures the inverse distortion step cannot be ignored.

The main problem with the diversity measures for a majority voting system is the amount of available training data. The high performance of the ensemble-based system generates few amount of data regarding bad votes, but the most of the diversity measures are built upon this information. For instance, the most important situation for our application is when the bad votes fulfilling the geometrical constraints may cause wrong final decision. Without appropriate number of such situations, the diversity measures generate incorrect values, which results in high distortion and low correlation with the system accuracy. By low correlation, the recommendation for the ensemble system is not satisfied. While the main motivation of the usage of diversity measures is to find the system with the best performance, sufficient number of special situations is not available, but can be interpolated. In our proposed model and in the application, all the diversity measures are smoothed to suppress the lack of data. Fig. 3 (a), (b) show the result of the smoothing step. This step is required not just for our modified diversity measure, but for the original ones, as well.



(a) Ensemble mean error interpolation for the original diversity measure.

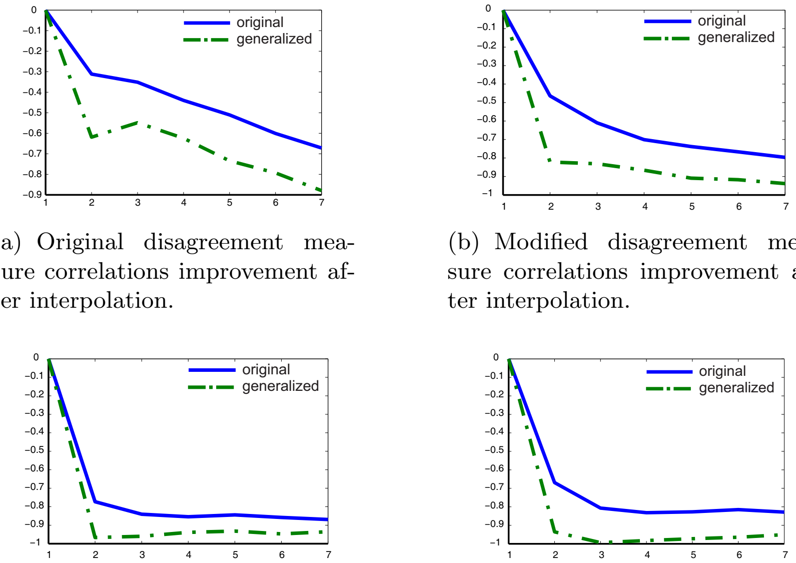
(b) Disagreement measure interpolation for the modified diversity measure.

Fig. 3. Comparison of the diversity measures before and after interpolating the missing cases. After interpolation, the curves are strongly smoothed i.e. abnormal values are removed. The x-axis represents the number of combinations of classifiers (247 different situations exist regarding 8 classifiers, where the diversity can be measured), while the y-axis the value of the diversity measure.

Fig. 4 (a) and (b) show, that after the interpolation step, the correlation between the system accuracy and the diversity measures is increased dramatically in both cases. The dashed lines show that after applying the interpolation, the correlation values are considerably increased independently whether a modification was applied or not. In case of non-pairwise diversity measure, Fig. 4 (c), and (d) show the similar results as Fig. 4 (a), and (b).

5 A weighted combination of diversity measures

While in most cases the dependencies between the assembled classifiers are unknown (e.g. between the algorithms in our OD application), by generating an ensemble from the classifiers having the highest accuracies the optimal performance may not be achieved. Although the diversity measures suggested by the literature are extended successfully in section 3, and their performance is improved by applying the interpolation, it cannot be guaranteed to choose the ensemble having the best accuracy regarding diversity measures. For solving



(a) Original disagreement measure correlations improvement after interpolation.

(b) Modified disagreement measure correlations improvement after interpolation.

(c) Ensemble mean error value of interpolation for the original diversity measure.

(d) Ensemble mean error value of interpolation for the modified diversity measure.

Fig. 4. Comparison of the diversity measures before and after interpolating the missing cases. After interpolation, the absolute value of the correlations are strongly higher, which is better for effectiveness of the diversity measures, with the system accuracy. The x-axis represents the number of the classifiers in the ensemble, while the y-axis the correlation value.

this problem the diversity measures can be considered as feature selectors and a weighted linear combination scheme can be applied for them. That is suppose that M classifiers and I diversity measures are given and the aim is to compose a system from the classifiers with the best performance regarding the diversity measures. This problem can be formulated as:

$$GD_j = \sum_{i=1}^{I} \alpha_{ij} d_{ij},\ j = 1, \dots, \binom{M}{k},\ k = (1, \dots, M),$$

where $\alpha_{ij} \in \mathbb{R}_{\geqslant 0}$ are the weights, d_{ij} are the values of the diversity measures, and GD_j is the value describing how good the specified system is considered as the diversity measures. In this case, the system with the maximal GD_j value will be choose:

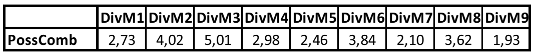
$$GD = \max_j (GD_j) = \sum_{i=1}^{I} \alpha_i d_i.$$

The appropriate selection of the weights α_i are well-known from the literature for independent feature selectors. Namely, the optimal weights can be determined from the individual accuracies of the feature selectors [8]. In this special case, the correlation values show the performance of the diversity measures as feature selectors. If we consider independent feature selectors $(D_1, D_2, \ldots, D_I)$ with accuracies $(p_1, p_2, \ldots, p_I)$, then GD can be maximized by assigning the following weights

$$\alpha_i = \ln \frac{p_i}{1 - p_i}, (i = 1, \ldots, I).$$

In our application, the accuracy p_i is the average correlation of the i-th diversity measure with the system accuracy regarding all possible assembled systems having the same number of members.

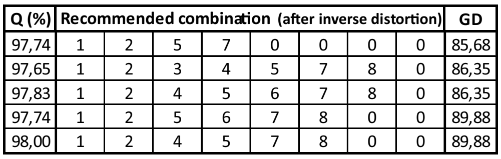
As an example for a special case because of the size of the full table, the optimal weights for the first nine diversity measures are shown in Fig. 5.



	DivM1	DivM2	DivM3	DivM4	DivM5	DivM6	DivM7	DivM8	DivM9
PossComb	2,73	4,02	5,01	2,98	2,46	3,84	2,10	3,62	1,93

Fig. 5. The applied weights in optimal weighted linear combination for the OD detection problem where the weights α_i were calculated as mentioned above. Every column contains a weight for a diversity measure (DivM) regarding a special case (PossComb).

In Fig. 6. the recommended combinations of algorithms for the weighted linear combination of the diversity measures are shown. It can be observed that the combined diversity measure(GD) well correlates with the system accuracy(Q).



Q (%)	Recommended combination (after inverse distortion)								GD
97,74	1	2	5	7	0	0	0	0	85,68
97,65	1	2	3	4	5	7	8	0	86,35
97,83	1	2	4	5	6	7	8	0	86,35
97,74	1	2	5	6	7	8	0	0	89,88
98,00	1	2	4	5	7	8	0	0	89,88

Fig. 6. The recommended combinations of the algorithms (expressed by sequential numbers of the algorithms in the middle of the table) using weighted linear combination of inverse distorted diversity measures. The first column (Q) shows the system accuracy, while the last column (GD) is the weighted combination of the diversity measures.

The ensemble system of the OD detector algorithms having the best accuracies can be found by applying our proposed method, and the selection can be made by GD value. The proposed weighted linear combination of diversity measures is novel for our extended model in the spatial domain.

6 Conclusion

In this paper the diversity measures introduced in classical majority voting are generalized for our voting model in spatial domain. We tested the generalized diversity measures for OD detection on the Messidor database of retinal images. Without having any information about the dependencies among the applied algorithms, the aim is to choose the best ensemble system having the highest accuracy. In case of missing training data, interpolation should be applied. Moreover, the generalized diversity measures outperform the classic ones, and the most accurate ensemble system can be found by an optimally weighted combination of diversity measures. We tested our proposed method on the Messidor database [3] because it is the largest public dataset, the others contain not enough images to evaluate these measures properly.

Acknowledgments. This research was realized in the frames of TÁMOP 4.2.4. A/2-11-1-2012-0001 ”National Excellence Program Elaborating and operating an inland student and researcher personal support system”. The project was subsidized by the European Union and co-financed by the European Social Fund.This work was supported in part by the János Bolyai grant of the Hungarian Academy of Sciences; the project TÁMOP-4.2.2.C-11/1/KONV-2012-0001 supported by the European Union, co-financed by the European Social Fund; the OTKA grants K100339 and NK101680; the project TÁMOP 4.2.1./B-09/1/KONV-2010-0007 implemented through the New Hungary Development Plan, co-financed by the European Social Fund and the European Regional Development Fund.

References

1. Toman, H., Kovacs, L., Jonas, A., Hajdu, L., Hajdu, A.: A Generalization of Majority Voting Scheme for Medical Image Detectors. Lecture Notes in Artificial Intelligence 6679, Vol. 2, 189–196 (2011)
2. Qureshi, R.J., Kovacs, L., Harangi, B., Nagy, B., Peto, T., Hajdu, A.: Combining Algorithms for Automatic Detection of Optic Disc and Macula in Fundus Images. Computer Vision and Image Understanding 116, 138–145 (2012).
3. Dataset MESSIDOR [Online]. Available: http://messidor.crihan.fr.
4. Ruta, D., Gabrys, B.: Classifier Selection for Majority Voting. Information Fusion 6, 63–81 (2005)
5. Sharkey, A.J.C., Sharkey, N.E.: Combining Diverse Neural Nets. The Knowledge Engineering Review 12, 231–247 (1997)
6. Ruta, D., Gabrys, B.: Analysis of the Correlation Between Majority Voting Error and the Diversity Measures in Multiple Classifier Systems. Proceedings of the 4th International Symposium on Soft Computing, 50–56 (2001)
7. Toman, H., Kovacs, L., Jonas, A., Hajdu, L., Hajdu, A.: Generalized Weighted Majority Voting with an Application to Algorithms Having Spatial Output. Lecture Notes in Artificial Intelligence 7209, Vol. 2, 56–67 (2012)
8. Kuncheva, L.I.: Combining Pattern Classifiers, Methods and Algorithms. John Wiley & Sons, Inc., New Jersey (2004)