Unsupervised Ensembles Techniques for Visualization

Bruno Baruque[1], Emilio Corchado[1], Bogdan Gabrys[2], Álvaro Herrero[1], Jordi Rovira[3], Javier Gonzalez[3]
1. Department of Civil Engineering, University of Burgos, Spain.
{bbaruque, escorchado}@ubu.es
2. Computational Intelligence Research Group, Bournemouth University, United Kingdom.
bgabrys@bournemouth.ac.uk
3. Department of Biotechonolgy and Food Science
jrovira@ubu.es

ABSTRACT: In this paper we introduce two unsupervised techniques for visualization purposes based on the use of ensemble methods. The unsupervised techniques which are often quite sensitive to the presence of outliers are combined with the ensemble approaches in order to overcome the influence of outliers. The first technique is based on the use of Principal Component Analysis and the second one is known for its topology preserving characteristics and is based on the combination of the Scale Invariant Map and Maximum Likelihood Hebbian learning. In order to show the advantage of these novel ensemble-based techniques the results of some experiments carried out on artificial and real data sets are included.

KEYWORDS: Exploratory Projections Model, Artificial Neural Networks, Ensemble Techniques, Outliers Identification.

INTRODUCTION

The main goal of all the visualization techniques is to bring the user a deeper understanding of a dataset by means of presenting some kind of graphical representation that can be easily inspected by humans, enabling to rapidly focus on the most interesting groups or clusters of data. In our case, we intend to extract useful information from high dimensional data sets by using visualization techniques that try to recognize the existence of outliers or cluster and identify interesting features, factors or directions in such data. One interesting way to accomplish those tasks is by performing dimensionality reduction to present some internal information which can be revealed by means of naked eye or based on different distance measures of the data space.

In this work we present two different approaches to carry out our goal of an easy interpretation of high-dimensional data: one is based on simple linear methods to obtain projections where as little information as possible is lost; while the other is based on the use of unsupervised Neural Topology Preserving models.

Exploratory Projections Pursuit (EPP) methods are those based on the identification of "interesting" directions in terms of any one specific index or projection. Such indexes or projections are, for example, based on the identification of directions that account for the largest variance of a data set (as Principal Component Analysis (PCA) [1], [2]). Having identified the interesting projections, the data is then projected onto a lower dimensional subspace in which it is possible to examine its structure visually, which normally involves plotting the projection in two or three dimensions. The remaining dimensions are discarded as they are mainly related to a very small percentage of the information or the data set structure. In that way, the structure identified within a multivariable data set may be easily analysed with the naked eye.

Another type of dimensionality reduction method, which can be implemented as an unsupervised neural network, is called a Maximum Likelihood Scale Invariant Map (MLSIM) [12]. It is similar to the Self-Organizing Map (SOM) [8] as it is a topology preserving technique but different in a sense that it is more appropriate to radial datasets [11], [12].

In this paper, an attempt is made to show the importance of using ensemble techniques combined with unsupervised models, in order to identify and remove outliers [3], [4] and to improve classification performance.

ENSEMBLE TECHNIQUE

The ensemble technique is utilized in this work to avoid the influence of outliers. As it frequently happens in nature, the aggregation of several individual models can outperform a single one. The "bootstrap aggregation" also known as "bagging" technique [13], is one of the methods for an ensemble construction that will be used in this paper.
This technique was initially used to improve the performance of classification trees. In the present study the idea is to employ the bagging technique [15] in combination with both the PCA analysis and the MLSIM network in order to have more than one independent analysis done over the same dataset, being in fact one analysis over each of several subsets of it. It is expected that, if any perturbation of the statistical characteristics of the dataset is produced only by a few of its components; this will be more evident in some runs than in others.

ROBUST DATA VISUALIZATION USING PCA

Principal Component Analysis (PCA) [1], [2] describes the variation in a set of multivariate data in terms of a set of uncorrelated variables each of which is a linear combination of the original variables. Its aim is to derive new variables, in decreasing order of importance, that are linear combinations of the original variables and are uncorrelated with each other. PCA can be implemented by means of some connectionist models [5], [6].
The disadvantage of this technique, both employing statistical or connectionist models is that this process is accomplished in a global way. In such a case every data sample situated far from the majority of the other samples of the dataset can influence the final result, as it introduces a high variance compared with the rest, even though it could be very small in number and could be considered as anecdotic or dispensable case. Almost in every mid-size non-artificial dataset a number of these outlier cases appear, distorting its variance and hence hindering its analysis [4].
The ensemble technique is extremely useful in dealing with this particular problem [17]. Firstly, it is necessary to obtain different subsets of the dataset. That is done by randomly selecting several cases from the dataset, and considering them as if they were a complete dataset. By doing this operation *n* times; *n* different datasets will be available, although they are really "subsets" of the main dataset. The next step consists of performing an individual PCA analysis on each one of the *n* subsets obtained by re-sampling the original one. If the whole dataset does not include elements that alter drastically its statistical properties (i.e. in this case the variance), the set of results obtained on the analysis of the different "subsets" should be similar by a small margin. On the contrary, if few cases that alter these statistical properties are included in the main dataset, it is expected to generate different results in terms of directions of the principal components obtained.

Combination of PCA Analyses

1. **Re-sampling and PCA Calculation.** Several entries from the analyzed dataset are randomly selected to form different subsets. The number of the entries selected depends on the experiment. Then a PCA analysis is performed over the selection from the dataset. These two steps are repeated several times in order to have several results to compare. In the case of this work we have always carried out 10 runs.
   The information obtained for each one of the tests is the following: a set of eigenvectors that represent the direction of the three principal components found and a set of measures that indicate "how much information" from the original dataset is represented by each one of the principal components found.
2. **Voting and Averaging.** To perform voting and averaging of directions to get the final principal components we have to identify if any direction is too deviated (due to the inclusion of outliers in that particular subset of data) to take part into the resolution of the final direction of each component. The criterion for considering similar directions is based on the following. We calculate the sum of all the vectors. Then we calculate the scalar product (the projection) of each one of the analyzed directions (vectors) in reference to the previous calculated sum. The vectors which are further away (their scalar product is under a certain threshold) are not taken into account in the following step. We calculate again the sum of the vectors, which corresponds now to an average for the directions that we estimated were “similar”, as they are no longer influenced by the more deviated ones.

ROBUST DATA VISUALIZATION USING MLSIM

In this part of our study, a different approach to data visualization is used. In this case, the visualization is accomplished by means of artificial neural network the main feature of which is its topological organization ability combined with the ensemble technique, as in the previous section.
Maximum Likelihood Scale Invariant Map (MLSIM) [11], [12], is an extension of the Scale Invariant Map (SIM) [10] based on the application of the Maximum Likelihood Hebbian Learning (MLHL) [9].
As done before with the Principal Component Analysis we intend to apply bagging in combination with MLSIM with the main objective of improving the classification performance of the ensemble of MLSIMs in comparison to individual MLSIM [16] and some other topology preserving maps (i.e. SOM), very commonly used for data visualization.

Combination of MLSIM

1. **Training the Ensemble.** When constructing MLSIM ensemble, first a subset of data is randomly drawn from the training dataset and used to train only one of the networks. For the next trained network the process is repeated. Thus, the networks of the ensemble are trained using slightly different datasets, giving as a result the desired diversity.
2. **Testing - the Classification.** We randomly divide the input dataset into ten subsets and perform training and testing processes ten times. Each time we iterate over this main algorithm we select a different part of the input dataset as a testing dataset, while the other nine parts are used for training. In this way the whole dataset is used to train and test the ensemble. At the end, we average the testing results obtained in the ten tests to achieve the final classification rate.
   Each of the ten times we perform the previously explained "outer loop", we take three steps:

a) Training. In this case the ensemble of MLSIMs is trained as it is explained in the previous section. We use the 9/10 of data considered as training data in the current iteration.

b) Labelling of output neurons. As MLSIMs use an unsupervised learning technique this step consists of presenting again the training dataset to the recently trained ensemble in order to label the output neurons with the most consistently recognised class label. As we know which class each of the training data belongs to through the given class label, by presenting the training data to the ensemble, we will consider that the output neuron is specialized in recognizing data from that class if it responded to training samples from that class as a winner in a majority of the cases.

c) Testing. In this step we present to the ensemble of MLSIMs the other 1/10 of data that was left out of the training process. The testing dataset is also labelled with its corresponding class labels, so we can compare the class the ensemble classifies a testing sample as with its real/given class label. In this way a measure of the accuracy of the ensemble can be obtained. To decide to which class an input belongs to, the MLSIM ensemble performs a majority voting among its composit MLSIMs [16]. The input is presented to each network individually, each one finds the winner neuron for that input and gives as an answer the class that this winner neuron is supposed to recognize better (as determined in step 2). The ensemble collects the answers from all its composit networks and returns the answer that was repeated the largest number of times (i.e. in the majority of the cases).

EXPERIMENTS AND RESULTS

PCA COMBINATION

Artificial Dataset

To test the capacity of the PCA ensemble technique, firstly a quite simple artificial data set is used in these series of experiments. The dataset is made of one cloud of points and several points spread far above the main cloud of points which are considered as outliers. The main cloud is an elongated cluster which moves within the axis delimited by the line defined by points in the following range: (1,1,0), (2,1.6,0), (3,2.2,0) and (4,2.8,0). The outlier points are spread over the same axis but displaced 5 units above in the vertical axis. There are 118 points in the main cluster and 8 outliers.

Dataset 1. In this set of experiments 10 subsets of 30 points randomly selected from the entire dataset (without replacement) are generated and PCA performed on each of them. Firstly, the method described above is applied to the dataset formed only by the main elongated data cluster (i.e. without the outliers). Studying the different vector directions the each one of the PCA analyses yields, it is easy to observe that the Re-PCA method has found almost the same direction for the first principal component, as it was expected. For the direction of the second and third principal components, they are clearly more dissimilar in the different tests, but still they all follow a consistent direction, except in one case.
The percentage of information (in form of the explained variance) that is represented by each one of the principal components is detailed in Table I.

Principal component	Percentage of information captured	
	max	*min*
First	72 %	68 %
Second	18 %	14 %
Third	14 %	11 %

Table I. Percentage of information captured by each of the principal components in the first part of the experiment (without outliers) including the maximum and minimum percentage of information (variance) from the analysed 10 subsets.

The results obtained performing exactly the same experiment but including now the 8 outliers in the sampled dataset are quite different. The distribution of the directions corresponding to the principal components, produced when outliers are taken into account, is much more "spread". This means that the direction found in each case is rather dissimilar to the other corresponding ones. We can even consider that in 3 cases out of 10 (30 % of the cases), the method has found opposed directions for the first and second principal components, as they are almost perpendicular to the directions obtained by the other 7 analyses. This three deviated directions will not be taken into account in the average calculation stage as the majority cluster consists of the 7 cases where the first principal component appears in the horizontal direction. The "percentage of information" that is represented by each of the principal components in this case is detailed in Table II.

Principal component	Percentage of information captured	
	max	*min*
First	69 %	49 %
Second	41 %	17 %
Third	13 %	8 %

Table II. Percentage of information captured by each one of the principal components in the second part of the experiment (including outliers) including the maximum and minimum percentage of information (variance) from the analysed 10 subsets.

The results presented in Table II are quite different from the results obtained without including the outliers. Comparing both tables (Table I and Table II), the influence of the presence or absence of outliers in a dataset in terms of the direction of the largest variance and relative difference between the *max* and *min* values for the principal components can be clearly seen.
The presence of these outliers makes the amount of information detected by the first component (Table II) to be inferior to the situation without outliers (Table I). In this case, due to the shape of the artificial dataset used, the amount of information represented by the second component (Table II) is a lot higher than in the case that does not include outliers (Table I). As it was expected, the inclusion of the outliers brings a great instability to the dataset, making different individual PCAs behave in an inconsistent way and resulting in very different results where really the analysis is made over subsets of the same dataset. The use of PCA ensemble in cases like this is of particular use as the 70% of the cases where "the true" principal component is found represents the majority which is selected and then further stability is added through averaging of the eigenvectors from these 70% of majority similar principal directions. The final averaged directions are shown in Fig 3.
Datasets 2 and 3. To generate datasets 2 and 3, again 10 subsets of 70 and 100 points respectively have been used to test the stability of the PCA analysis performed over them. The results obtained for data set 3 are shown in Fig. 1 and Fig. 2.

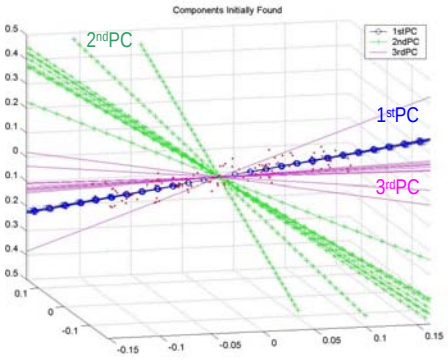

Figure 1: Projections of Re-PCA using 100 points (excluding outliers)

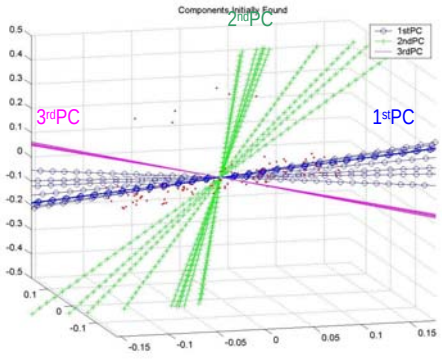

Figure 2: Projections of Re-PCA using 100 points (including outliers)

Principal component	Percentage of information captured	
	max	*min*
First	70 %	44 %
Second	44 %	15 %
Third	13 %	9 %

Table III. Percentage of information captured by each of the principal components (selecting 50 points and including outliers) including the maximum and minimum percentage of information from the analysed 10 subsets.

As it can be seen in all the above experiments, the more samples are included into the analysis, the more stable behaviour of the individual PCA. Comparing the results obtained for data sets 2 and 3 gives a proof of that, as the directions found using 100 points are slightly more consistent than when using only 30, 50 or 70 points. It can also be seen (Fig. 1 and Fig. 2) that including outliers in the analysed dataset brings a substantial degree of instability, giving as a result more spread "fans" (less consistent results) or even completely different directions for its principal components.

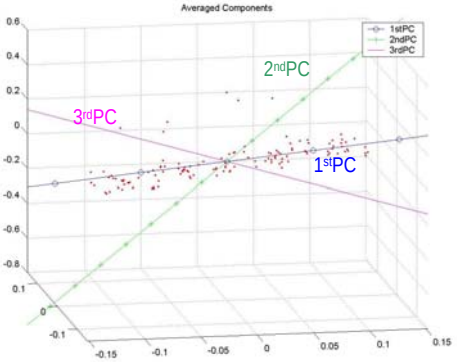

Figure 3: Resulting average for each of the principal components by voting between the directions obtained with 30 points for each analysis (excluding the 30% of the directions strongly influenced by the outliers).

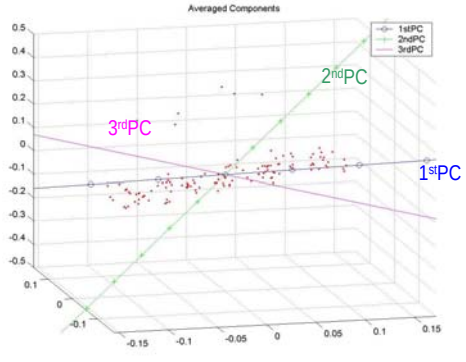

Figure 4: Resulting average for each of the principal components by voting and averaging of the directions shown in Fig 2 (using 100 points)

Calculating the average directions (Fig. 4) as explained above, we can obtain approximately the same main directions for the three principal components, as when we have only used 30 points for calculations (Fig. 3). This can be considered as an empirical proof of the robustness of the proposed Re-PCA method.

Real Dataset

A real life dataset has been used to test the performance of the PCA Ensemble technique. The dataset used comes from the food industry sector and consists of measurements obtained from several brands of seven types of Spanish cured ham.

Several samples of different parts of each type of ham were cut and measurements were taken over each of these samples, by an electronic nose α FOX 4000 (Alfa MOS, Toulouse, France) with a sensor array of 18 metal oxide sensors. Our final dataset consists of a total of 176 samples of ham, each of them composed of 18 different variables measured over it. 4 outlier samples have been spotted in the experiments detailed below.

Experiment: Performing a simple PCA analysis and projecting the data in the two axes determined by the principal components, gives the results showed in Fig. 5. The result presented in Fig. 6 is the one produced by the PCA ensemble method described above.

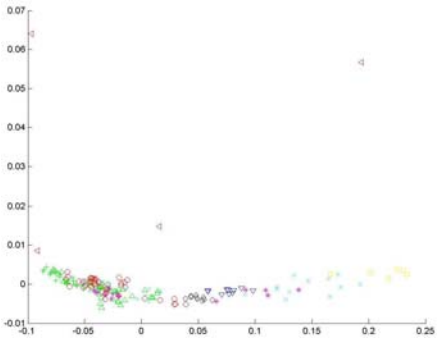

Figure 5: Projection of dataset (including outliers) over the 1st and 3rd Principal Components obtained from a Simple PCA analysis.

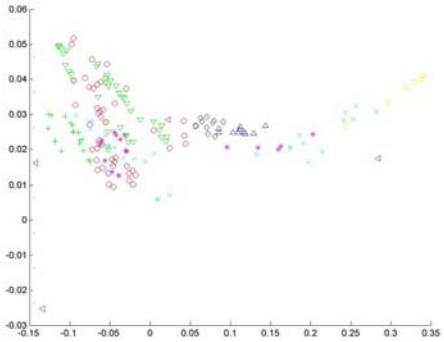

Figure 6: Projection of dataset (including outliers) over the 1st and 3rd Principal Components obtained from a Re-PCA analysis using 120 samples.

Certain interesting structure appears in the projection over the first and second Principal Components in both types of analysis. On the contrary, the image obtained by projecting data over the first and third principal components obtained by the classical PCA is completely distorted by the presence of the outliers, as it can be seen in Fig. 5. The result obtained by using the PCA Ensemble method is displayed in Fig. 6. It can easily be seen, that in this case interesting structure appears even to the naked eye, overcoming the influence of the outliers over the dataset. Table IV shows the percentage of information captured by the single PCA and the maximum and minimum percentage obtained in the ten composit PCA analysis forming the PCA ensemble.

Principal component	Percentage of information captured	
	Single PCA	*PCA Ensemble*
First	86.58 %	86.52 - 82.5 %
Second	8.74 %	9.473 - 8.06 %
Third	4.66 %	7.9 - 4.60 %

Table IV. Percentage of information captured by each of the principal components in the experiment (with outliers) including the results of a Single PCA and the maximum and minimum percentage of information (variance) from the analysed 10 subsets in the PCA Ensemble.

MLSIM COMBINATION

As stated in [12], the MLSIM is a more suitable technique for radial datasets. To test this feature a radial dataset has been generated by disposing six normal distributions of 2 dimensions in a radial way. Their centres are situated in points (3,2), (1,4), (-2,4), (-3,1), (-2,-4) and (1,-2) respectively. The number of samples corresponding to each distribution is as follows: 50, 100, 70, 50, 20 and 100. We have also included several outlier points to compare the results of different classification models when they are or are not present.

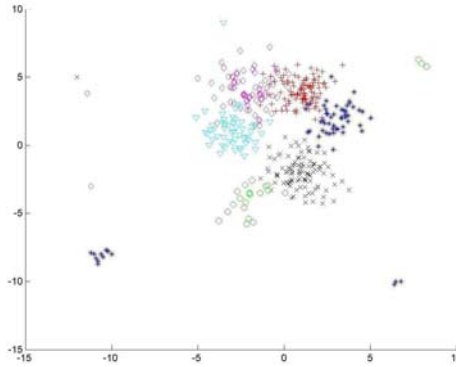

Figure 7: Samples of the radial artificial data used in this study.

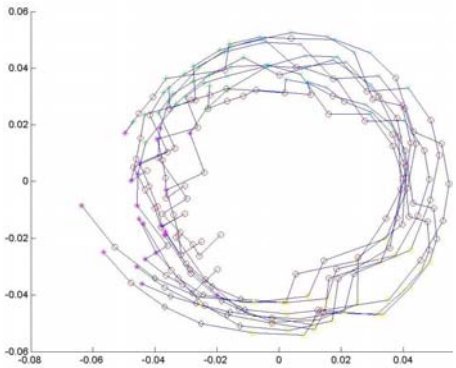

Figure 8: Ten MLSIMs trained on the bagged data. The weights of the first one were initialized to the Principal Components of the dataset. The following ones were initialized to the final weights of its predecessor.

The testing dataset is showed in Fig. 7, while Fig. 8 shows the results of training the ensembles of ten MLSIMs. As it can be seen, the MLSIM ensemble tries to expand and cover the whole dataset range by using a circular form. In this particular case of a radial form dataset, this approach should give better results, as that form fits better the form of the considered dataset.

	Accuracy of the Model (including outliers)		
	min	*max*	*average*
Single MLSIM	75.36%	83.17 %	79.6 %
Ensemble (10 MLSIMs)	82.19%	86.34%	84.15%

Table V. Classification accuracy of the two models applied to the data set from Fig. 7. The minimum, maximum and average accuracy from 10-fold cross validation testing runs are shown in the table. Both experiments were performed including 20 outlier points.

We have applied two classification models to the above described data set (including and without outliers). As expected the MLSIM ensemble model obtains better results than the single MLSIM, without and with outliers in the data (see Table V). In the case of a single MLSIM, when outliers (i.e. mislabelled data) are present in the studied data set; the variation between the best and worse accuracy results is close to 8%, meaning that the model exhibits certain instability. In the case of the model presented in this study, the MLSIM ensemble, we can see that the difference in this case is smaller: around 4% when outliers are present.
All these experiments have demonstrated how the MLSIM ensemble performs better than a single MLSIM model in the case of radial data sets with outliers.

CONCLUSIONS

In this work several experiments regarding the data visualization have been shown. As it is explained in [3] and [4], and experimentally verified in different examples of this study; the inclusion of outliers can hinder the analysis and visualization of a dataset in a significant way. The application of the ensemble based technique to the solution of this problem has proved to be valuable. As it can compensate the problems caused by the outliers by using ensembles can effectively stabilise non-stable and outlier sensitive approaches.

Acknowledgments
This research has been supported by the MCyT project TIN2004-07033 and the project BU008B05 of the JCyL.

REFERENCES

[1] Pearson, K. On lines and planes of closest fit to systems of points in space. Philosophical Magazine 2:559-572. (1901).

[2] Hotelling, H. Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology, 24:417-441,498-520. (1933).

[3] Cook, R. D. Detection of influential observations in linear regression. Technometrics 19, 15-18. (1977).

[4] Dixon, W. J. Analysis of extreme values, Ann. Math. Stat., 21, 488-506. (1950)

[5] Oja, E. Neural networks, principal components and subspaces. International Journal of Neural Systems 1(1):61-68. (1989).

[6] Sanger, D. Contribution analysis: A technique for assigning responsibilities to hidden units in connectionist networks. Connection Science, 1:115--138. (1989).

[7] Kohonen, T. Barna, G and Chrisley R. Statistical Pattern Recognition with Neural Networks. In Proceeding of International Joint Conference of Neural Networks (pp. 61-88), IEEE Press, 1988.

[8] Kohonen, T. The Self-Organizing Map. In Proceedings of the IEEE 78 (pp. 1464-1480), 1990.

[9] Fyfe, C. and Corchado, E. Maximum Likelihood Hebbian Rules. ESANN (European Symposium on Artificial Neural Networks), ISBN 2-930307-02-1, 2002.

[10] Fyfe, C. A Scale Invariant Map. Network: Computation in Neural Systems, 7 (pp 269-275), 1996.

[11] Corchado, E. and Fyfe, C. Maximum Likelihood Topology Preserving Algorithms. In Proceedings of the U.K. Workshop on Computational Intelligence, Birmingham, UK, 2002.

[12] Corchado, E. and Fyfe C. The Scale Invariant Map and Maximum Likelihood Hebbian Learning. International Conference on Knowledge-Based & Intelligent Information & Engineering System, IOS Press, 2002.

[13] Breiman, L. Bagging predictors. Machine Learning, 24:123–140. (1996).

[14] Schapire, R.E; Freud, Y; Bartlett, P. and Lee, W.S. Boosting the margin: a new explanation for the effectiveness of voting methods. The Annals of Statistics, 26(5):1651–1686, 1998.

[15] Kuncheva, L, Combining Pattern Classifiers: Methods and Algorithms. Wiley-Interscience, 2004.

[16] Ruta, D. and B.Gabrys, Classifier Selection for Majority Voting, Special issue of the journal of information fusion on Diversity in Multiple Classifier Systems, vol. 6, issue 1, pp. 63-81, 1 March 2005.

[17] Ruta, D. and B. Gabrys, A Theoretical Analysis of the Limits of Majority Voting Errors for Multiple Classifier Systems, Pattern Analysis and Applications, vol. 5, pp. 333-350, 2002.