A Rule for Combination of Spatial Clustering Methods

Danielly C. S. C. Holmes, Ronei M. Moraes, Rodrigo P. T. Vianna

Graduate Program in Decision Models and Health
Federal University of Paraiba
João Pessoa, Brazil
Email: daniellycristina9@gmail.com, ronei@de.ufpb.br, rodrigopissoa@gmail.com

***Abstract*—In the area of spatial analysis, spatial clustering methods use georeferencing information in order to identify significant and non-significant spatial clusters of the phenomenon in study in a specific geographical region. Several methods are available in the literature, such as scan statistic, Getis-Ord statistics, and the Besag and Newell method. In practical applications, all those methods are not able to produce results which can capture the real event with good accuracy. In this paper, we propose using the a combining classifier technique in order to provide better results for spatial clustering methods, using the majority voting rule for that combination. A study case was presented using epidemiological data of dengue fever from state of Paraba, Brazil, in the year of 2011 and the final results allowed to identify the priority and non-priority areas in the region of interest.**

***Keywords*–Majority vote rule; Spatial clustering methods; Statistical significance.**

I. Introduction

A classifier is defined as a function, whose domain is an attribute space in R^n and its co-domain is a set of class labels Ω with K elements, where $\Omega = \{w_1, ..., w_K\}$ [16][18]. Classification has been an area for research in the pattern recognition and machine learning communities [7]. The classification process can be performed using supervised classifiers and unsupervised classifiers. The supervised classifiers require a previous knowledge of the problem, which is translated by a training database that contains labeled samples. The unsupervised classifiers are performed using a database of unlabeled samples, i.e., samples for which their class are unknown. So, there is no previous knowledge about the real class labels [13].

In the scientific literature, several cases can be found, in which combining multiple classifiers provided an improvement of the results with respect to each individual classifier performance. So, that combination makes them more efficient [5][9][10][19][21][22].

Combining classifiers can be done using three architectures: in sequential (or linear) way, in parallel or hierarchically [24]. In order to provide the final decision, an architecture should be chosen, as well as a combination scheme of classifiers, which is called combiner [2]. One of these schemes is the static combiner, which performs combination using a predefined rule and no training is required over that architecture [24]. The architecture chosen and the combination scheme (including the combination rule) allow to create a new classifier [2].

In the past years, the number successfully applications combining classifiers is increasing in many areas, as, for instance, image classification, writing and character recognition; among others [5][17]. Several schemes can be found in the literature to combine classifiers, as voting (and its variations), sum, mean, median, product, minimum, maximum, adaptive weighting [2], logistic regression [5], Dempster-Shapher theory and mixture of experts [22]; among others [24].

In the area of spatial analysis, spatial clustering methods use georeferencing information in order to identify significant and non-significant spatial clusters of the phenomenon in study in a specific geographical region. Several methods are available in the literature, such as scan statistic [1][15], Getis-Ord statistics [3] and the Besag and Newell method [4][6]. In practical applications, all those methods are not able to produce results which can capture the real event with good accuracy. Each method works with different methodologies and provides different results with respect to the others. In addition to these issues, as there is no reference information about the real clusters, it is not possible to check the similarity between the results produced by one method and the true result. Thus, it is possible to use just indirect forms of evaluation, as for instance, maps of relative risk, in studies in public health [25]. These problems have do not have a good solution yet.

The first problem is quite similar to the classification problem which is solved by using combining classifiers. So, in this paper, we propose using the combining classifier technique in order to provide better results for spatial clustering methods.

This paper is organized as follows: the next section presents some theoretical aspects of spatial clustering methods and combining rules. Section 3 brings the new methodology proposed. On Section 4, the results obtained, as well as their analysis, are presented, followed by some considerations in Section 5.

II. Theoretical Aspects

In this section, some theoretical aspects are presented. Three methodologies of the spatial clustering are shown: Scan statistic, Getis-Ord statistic, and the Besag and Newell method. All these methods are used in order to identify significant and non-significant regions (binary information) in a geographic area of interest. It means, high values or small ones, which are statistically different to the others in a sub-region will be assigned on map with different signal of the others, which are not significant. The significant regions on the map are named spatial clusters. The main goal is to identify significant areas, visualize and describe the spatial patterns [12].

Statistical functions provide measure for the spatial associations and evaluate the statistical significance for it. They are divided into: global, local, and focused statistics. Global statistics identify the spatial structure which can be cluster autocorrelation, but not identifying the location of cluster or quantify the spatial dependency. Local statistics quantify spatial autocorrelation and clustering within the small areas in the geographical region of interest, i. e. they search for regions which are significantly different from the area where they are inserted. The focused statistics quantify clustering on certain specific location, which is named *focus*. In Spatial Epidemiology, these kind of tests can relate information about incidence of a disease and possible sources of contamination in the same region [12]. According to Knox [23], local statistics are generic tests and focused statistics are focused tests. In this paper, we use only generic tests.

A. Scan Statistic

The scan statistic does a survey of the study region, looking for the most likely significant events. The survey occurs in the following way: for each sub-region, a centroid ξ_i is associated, it contains a random variable X, which denotes the numbers of individuals that have the disease along with the population size on that sub-region. This method is based on circles positioned over each centroid, in which the radius r can be the greatest measure that involves a new neighboring centroid and within it a percentage of the population [15]; in other words, multiple circles are generated with different radius and different geographical localizations [11]. This process is finalized when all the centroids have been tested. The hypotheses are:

H_0: There is no spatial cluster in the geographical region.

H_1: There is at least one spatial cluster in the geographical region.

The hypotheses are tested by the Monte Carlo simulation [1]. As the circles are increaseding, a likelihood test is performed, in which we verify if the study region is a conglomerate. The test is based on the maximum likelihood method [18], assuming some probabilistic distributions, and the evaluation is done using the Monte Carlo simulation [1]. The Monte Carlo simulation is used to test if the clusters are statistically significant. The hypotheses test via Monte Carlo are generated simultaneously from simulated data multiple times under the null hypothesis and the p-value is $\frac{r}{(R+1)}$, in which the R is the number of occasional data repetition of the simulated data and r is the classification of the statistical test [1].

B. Getis-Ord Statistic

The Getis-Ord statistic measures the spatial association between the spatial dependencies functions. It performs the distance measurements only with the positive observations and with data that have a non-normal distribution [3].

The Getis-Ord statistics is estimated by groups of neighbors of the critical distance d of each area i. The critical distance is formed by a proximity matrix W, in which the elements are formed in function of the critical distance $w_{ij}(d)$. With that, two statistical functions were proposed: the Getis-Ord statistic evaluates the significance of the statistic generated. It is said to be significant if the p-value is lower than the adopted significance [3]. The global statistic $G(d)$ is equal to the traditional measures of spatial agglomeration with just one value $G(d)$. The global statistic is given by:

$$G(d) = \frac{\sum_i \sum_j w_{ij} x_i x_j}{\sum_i \sum_j x_i x_j} \tag{1}$$

in which $x_i \in X$ is a value observed in the position i and $w_{ij}(d)$ an element of the proximity matrix. The level of significance is defined as the probability of rejecting the null hypothesis (existence of spatial autocorrelation), if it is true. The p-value confronted with the adopted significance defining the significance of the Getis-Ord index generated. The analysis is based on the value of the index and its significance: the positive and significant value of $G(d)$ indicates spatial agglomeration of high values, the negative and significant values of $G(d)$ indicate spatial agglomeration of small values [3].

The local statistic G_i and G_i^* are measures of the spatial association for each area and they measure the association in each spatial unit for each observation i, in which G_i and G_i^*, shows the position which is surrounded by high or low values for the variable. The $G_i(d)$ equation for each observation i and distance d is shown in the following way [3]:

$$G_i(d) = \frac{\sum_{j, j \neq i} w_{ij} x_j}{\sum_j x_j}, \tag{2}$$

in which all positions j, except those ones where $j = i$, can be in the sum. This index is equal to the ratio of the sum of the values in the neighbouring positions by the sum of the values in the whole data series. However, in the statistic of G_i^*, all values of j, including those ones where $j = i$ are included in the sum [3].

$$G_i(d) = \frac{\sum_j w_{ij} x_j}{\sum_j x_j} \tag{3}$$

TABLE I. INTERPRETATION OF THE LOCAL INDEX SIGNIFICANCE.

Significance	*Statistic*	*p-value*
Negative***	Negative	p<0.005
Negative**	Negative	0,005<p<0,025
Negative*	Negative	0,025<p<0,05
Negative	Negative	p>0,05
Positive	Positive	p>0,05
Positive*	Positive	0,025<p<0,05
Positive**	Positive	0,005<p<0,025
Positive***	Positive	p<0,005

The Getis-Ord local index is interpreted as follows: the positive and significant standardized values (p-value less than 5%) means a spatial agglomeration with high values. The significant negative standardized statistical values (p-value less than 5%) indicates a spatial agglomeration with low values. According to Table 1, the interpretation is given in the following way: The smaller p-value implies the higher agglomeration and it does not matter whether is a positive or negative spatial agglomeration [3].

C. Besag and Newell Method

The Besag and Newell method [4] produces circular spatial clusters. The process is: a radius is determined in such a way that it contains a circle with at least p cases in its interior. The method starts with the circle radius equal zero. When the circle achieves p cases, the process stops; if that does not occur, the radius is increased, including a new centroid. The procedure is executed until at least p cases are found or when the number of centroids is finished.

Let C be the total number of cases in the study region and Y the total population exposed to the risk in the region. Let $C_j(i)$ and $M_j(i)$ be the number of cases and the population of the j areas closer to the centroid. The statistic of the test is based on the random variable A, defined as the minimum of areas next to the centroid [6]. So, we have:

$$A = \min_j \{C_j(i)\} \geq p \tag{4}$$

in each centroid is verified the existence of a spatial cluster. The cluster is said significant if the p-value is less the adopted significance. From the value a observed for A, the level of significance of the test is defined by $P(A \leq a)$, which tests the null hypothesis (absence of spatial clusters). The significance, denoted by $p_k(i)$ is calculated by the following equation [6]:

$$\begin{aligned} p_k(i) = P(A_i \leq a_i) = 1 - \sum_{j=1}^{k-1} \frac{(M_j(i)C/M)^j}{j!} \times \\ \times\, exp(M_j(i)C/M) \end{aligned} \tag{5}$$

in which M_j is the population observed in the area j.

D. Classifiers Combining Rule

As previously mentioned, there are several rules for combining multiple classifiers. In this section, two methods are presented and it is shown that they are equal when applied in the binary case [26].

The *majority voting* is the most popular rule for combining classifiers [8]. The majority voting rule defines the winner class as that one which obtained more than half of the total number of votes. If there is no class in this condition, then $x \in X$ did not receive a label (it works as a rejection option). Let $\Delta_{ji} \in 0, 1$ be the vote for the class j which was signed by the classifier i. Let H be a decision function which sign the final class for x, then:

$$H(x) = \begin{cases} j, & \text{i}f \sum_{i=1}^{D} \Delta_{ji}(x) = \\ & = \sum_{j=1}^{K} \sum_{i=1}^{D} \Delta_{ki}(x) \\ \text{r}ejection, & \text{o}therwhise. \end{cases} \tag{6}$$

where K is the number of classes in Ω and D is the number of classifiers [8][14].

Another kind of voting rule for combining classifiers is the *plurality voting*. In this case, the winner class is that one which receives the largest number of votes, i. e., it is not necessary achieve to get more than 50% of classifiers votes. Its equation is given by:

$$H(x) = j, \qquad if \sum_{i=1}^{D} \Delta_{ji} = \max_k \sum_{i=1}^{D} \Delta_{ki} \tag{7}$$

According to Zhou [26], in the case of binary classification, the *majority voting* and the *plurality voting* produce the same results.

III. Methodology

As mentioned before, spatial clustering methods use georeferencing information in order to identify significant and non-significant spatial clusters of a phenomenon in study in a geographical region of interest. All methods available in the literature are not able to produce results which can capture the real event with good accuracy and it is possible to use just indirect forms of evaluation of their results. In the applications using public health data, maps of relative risk can be used for this purpose. The measure of relative risk is defined as the probability of an individual to have a disease in a determined time divided by the cumulative incidence of the area of the interest [20].

The final result provided by a spatial clustering method is a georeferencing list of significant centroids, i.e., a database which contains the pair $(centroid, label)$. The methodology consists of applying an impair number of spatial clustering methods on the same area and data. From those applications, we obtained a number of georeferencing lists. Finally, over them is applied a voting rule in order to obtain the final class for each centroid in the region of interest.

In this paper, we applied that methodology using the three spatial clustering methods presented above, on the same area and data. From those applications, we obtained three georeferencing lists and we applied the majority voting on them. It is worth noting that the problem is a binary classification, then majority voting and plurality voting produce the same results.

IV. Results

The methodology designed and presented in the previous section was applied on epidemiological data of dengue fever from state of Paraba, Brazil, in the year of 2011. As dengue fever is a tropical disease, it is recurrent health problem in all country. Due to financial restrictions, it is important for the health authorities to know the areas in which the relative number of cases is significant larger than others, as well as areas where the relative number of cases is significantly lower than others. The first areas can be called priority areas and the second areas can be named protection areas.

Applications results for the three spatial clustering methods (Scan statistic, Getis-Ord statistic, and the Besag and Newell method) are presented in Figures 1-3. The relative risk map is presented in Figure 4 and the final decision map, obtained by the combination of those three spatial clustering methods using majority voting is presented in Figure 5. In the results below, we show the comparison of the decision map (result of the combination of spatial clustering methods from the majority vote rule) with all spatial clustering methods.

In the comparison of the final decision map (Figure 5) with the Scan statistic map (Figure 1), all the 16 cities on the final decision map (Figure 5) are in the scan statistic map. The Scan statistic map identified 53 cities with significant values and from those, 16 are present in the final decision map.

In the comparison of the final decision map (Figure 5) with the Getis and Ord map (Figure 2), of the 16 cities on the decision map, 10 are on the Getis-Ord map. According to the Getis-Ord map, 10 cities are in spatial clusterings of negative values and the rest of them are not significant.

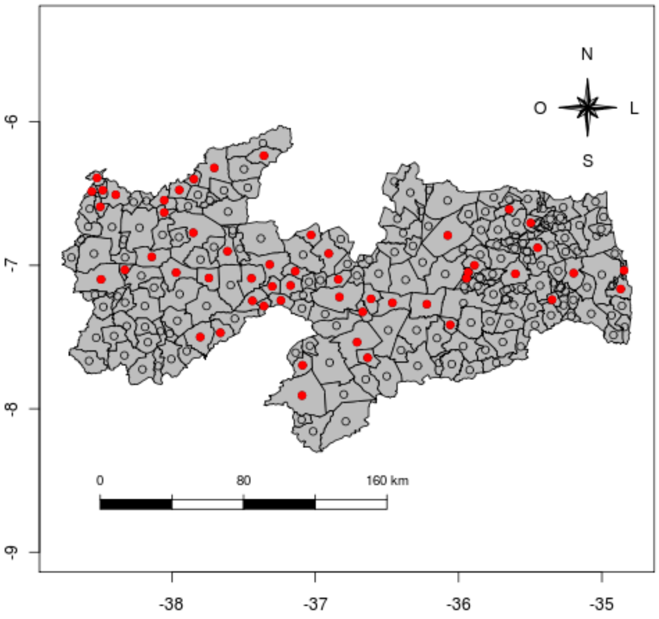

Figure 1. Scan satistic map of dengue fever for the state of Paraba in 2011.

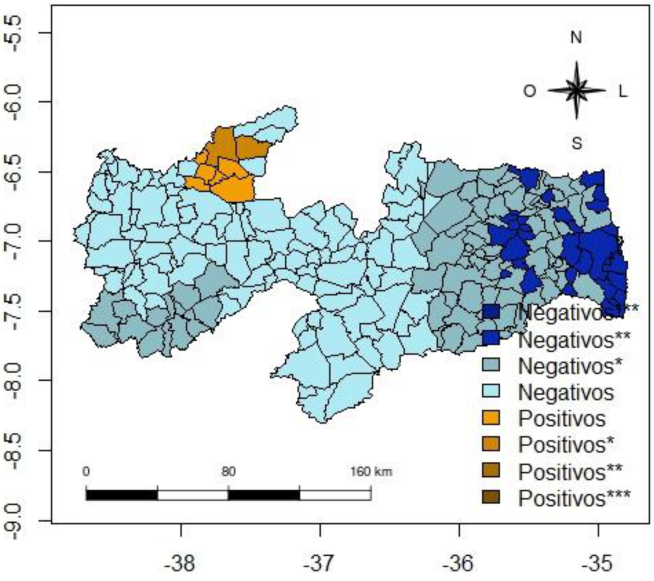

Figure 2. Getis and Ord map of dengue fever for the state of Paraba in 2011.

Comparing the final decision map (Figure 5) with the Besag and Newell map (Figure 3), it was observed that only 7 of them are on the Besag and Newell map. On the other hand, all significant cities on the Besag and Newell map are present on the final decision map.

Comparing the dengue final decision map (Figure 5) with the dengue risk map (Figure 4), it was verified that the cities on the final decision map present risk above 1.25, but not all cites with risk above 1.25 in the relative risk map are present on the final decision map. Finally, the result allows us to state that the methodology identified the cities with high relative risk of dengue fever in the state of Paraba in the year of 2011.

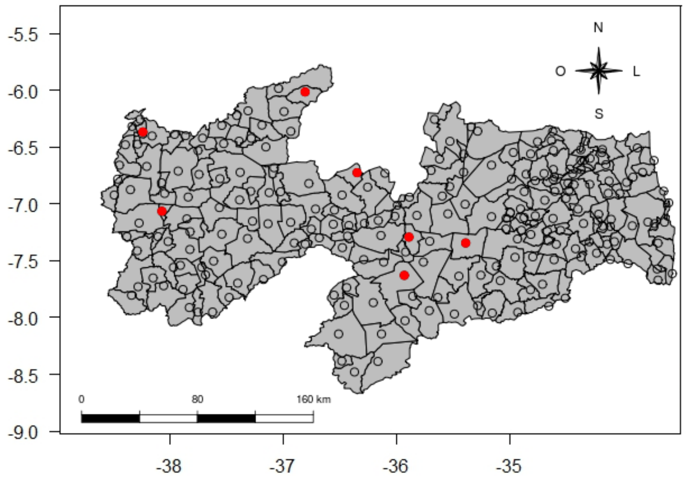

Figure 3. Besag e Newell map of dengue fever for the state of Paraba in 2011.

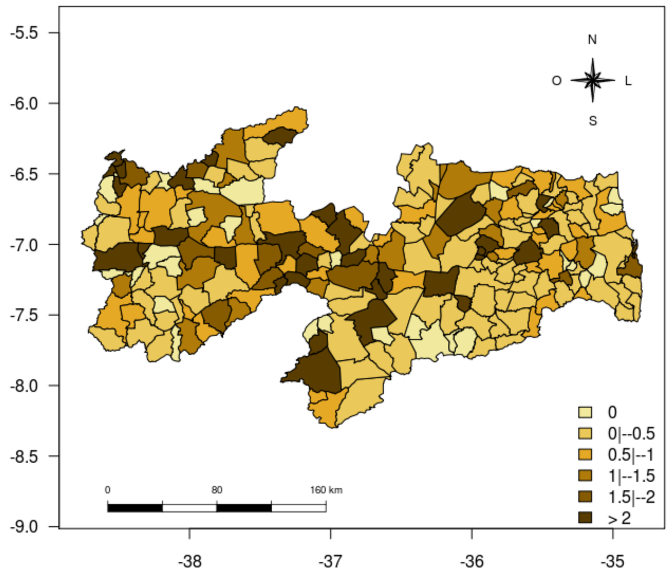

Figure 4. Map of the relative risk of dengue fever for the state of Paraba in 2011.

V. Conclusions

In this paper, we presented a new methodology for the combination of spatial clustering methods. We also presented a rule for building that combination based on majority voting. A study case was presented using epidemiological data of dengue fever from state of Paraba, Brazil, in the year of 2011.

Based on the results achieved, it is possible to affirm that the combination of spatial clustering methods using the majority voting rule presented coherent results and those results were better than each individual classifier. At the end, the methodology identified the priority and non-priority regions for dengue fever in the state of Paraba, Brazil.

VI. Acknowledgments

This project is partially supported by CAPES. It is also partially supported by grant 310470/2012-9 of the National Coun-

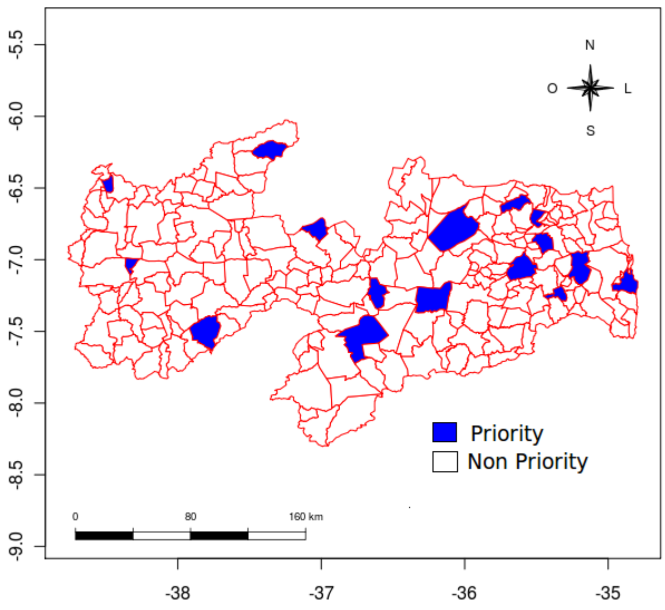

Figure 5. Decision map of the combination of spatial clustering methods for dengue fever in the state of Paraba in 2011.

cil for Scientific and Technological Development (CNPq) and is related to the National Institute of Science and Technology “Medicine Assisted by Scientific Computing”(181813/2010-6) also supported by CNPq.

References

[1] A. Abrams, K. Kleiman, and M. Kulldorff, “Gumbel based p-value approximations for spatial scan statistics”, International Journal of Health Geographics, vol. 9, December 2010, pp. 1-12.

[2] A. Jain, R. Duin, J. Mao, and S. Member, “Statistical Pattern Recognition: A Review”, IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 22, January 2000, pp. 4-37.

[3] L. Anselin, “Spatial data analysis with GIS: an introduction to application in the social sciences,” Fundamental Research in Geographic Information and Analysis, University of California, August 1992.

[4] J. Besag and J. Newell, “The detection of clusters in rare diseases,” Journal of the Royal Statistical Society, vol. 154, June 1991, pp. 143-155.

[5] F. Breve, M. Ponti, and N. Mascarenhas, “Combining Methods to Stabilize and Increase Performance of Neural Network-Based Classifiers,” in Proceding of the XVIII Brazilian Symposium on Computer Graphics and Image Processing, 2005.

[6] M. A. Costa and R. M. Assunção, “A fair comparison between the spatial scan and the Besag-Newell disease clustering test,” Environmental and Ecological Statistics, vol. 12, iss: 3, September 2005, pp. 301-319.

[7] T. Damoulas and A. M. Girolami, “Combining Feature Spaces for Classification,” Pattern Recognition, vol. 42, April 2009, pp. 2671-2683.

[8] R. P. W. Duin and D. M. J. Tax, “Experiments with Classifier Combining Rules,” in MCS ’00 Proceedings of the First International Workshop on Multiple Classifier Systems, vol.1857, June 2000, pp. 16-29.

[9] T. K. Ho, J. J. Hull, and S. N. Srihari, “Combination of Structural Classifiers,” in Proceedings of 1990 IAPR Workshop on Syntactic and Structural Pattern Recognition, 1990.

[10] J. J. Hull, A. Commike, and T. K. Ho, “Multiple Algorithms for Handwritten Character Recognition,” International Workshop on Frontiers in Handwriting Recognition, Montreal, Canada, April 2-3, 1990.

[11] H. Izakian and W. Pedrycz, “A new PSO-optimized geometry of spatial and spatio-temporal scan statistics for disease outbreak detection,”, Swarn and Evolutionary Computation, vol.4, June 2012, pp. 1-11.

[12] G. M. Jacquez, The Handbook of Geographic Information Science. Blackwell Publishing Ltd, 2008.

[13] J. Kersten, “Simulateneous Feature Selection and Gaussian Mixture Model Estimation for Supervised Classification Problems,” Pattern Recognition, vol. 47, August 2014, pp. 2582-2595.

[14] J. Kittler, M. Hatef, R. P.W. Duin, and J. Mata, “On Combining Classifiers”, IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 20, March 1998, pp. 226-239.

[15] M. Kulldorff and N. Nagarwalla, “Spatial Disease Clusters: Detection and Inference,” Statistics in Medicine, vol.14, 1995, pp. 799-810.

[16] L. I. Kuncheva, Combining Pattern Classifiers: Methods and Algorithms. John Wiley & Sons, Inc., Hodoken, New Jersey, 2004.

[17] Y. D. Lan and L. Gao, “ A new model combining multiple classifiers based on neural network,” Fourth International Conference on Emerging Intelligent Data and Web Tecnologies, September 2013, pp. 154-159.

[18] C. M. Bishop, Pattern Recognition and Machine Learning. Spring, London, 2006.

[19] C. Nadal, R. Legault, and C. Suen, “Complementary Algorithms for the Recognition of Totally Unstrained Handwritten Numerals,” in Proceedings of 10th International Conference on Pattern Recognition, vol. 1, June 1990, pp. 443-449.

[20] K. Rothman, Modern Epidemiology. The United States of America, Little, Brown and Company, 1986.

[21] C. Suen, C. Nadal, R. Mai, and L. Lam, “Recognition of Totally Unconstrained Handwritten Numerals Based on the Concept of Multiple Experts,” in Proceedings of International Workshop on Frontiers in Handwriting Recognition, Montreal, Canada, 1990.

[22] X. G. L. Tong, “Learning Opponent’s Indifference Curve in Multi-issue Negotiation Combining Rule Based Reasoning with Bayesian Classifier”, in in Proceedings of 6th International Conference on Natural Computation, vol. 6, August 2010, pp. 2916-2920.

[23] E. G. Knox, “Detection of clusters”. In: Methodology of enquiries into disease clustering. P. Elliot (ed.) Small Area Health Statistics Unit. London, 1989, pp. 17-20.

[24] R. M. Moraes and L. S. Machado, “A New Class of Assessment Methodologies in Medical Training Based on Combining Classifiers”, in Proceedings of XII Safety, Health and Environment World Congress (SHEWC’2012). 22-25 July, 2012, pp. 91-95.

[25] R. M. Moraes, J. A. Nogueira, and A. C. A. Sousa, “A New Architecture for a Spatio-Temporal Decision Support System for Epidemiological Purposes”, in Proceedings of the 11th International FLINS Conference on Decision Making and Soft Computing (FLINS2014), 17-20 Agosto, 2014, Brazil, pp. 17-23.

[26] Z-H. Zhou, “Ensemble Methods: Foundations and Algorithms”. CRC Press, Boca Raton, 2012.