Relevance Analysis based on graph theory and spectral analysis

L.M Morales-Villamil [1] , M.C. García-Murillo, D.H. Peluffo-Ordoñez, C.G. Castellanos-Domínguez

***Abstract*—Category 1. We present a new method for relevance analysis based on spectral information, which is done from a graph theory point of view. This method is carried out by using Gaussian Kernels instead of conventional quadratic forms and then avoiding the need of a linear combination-based representation. For this end, it is implemented an extended approach for relevance analysis using alternative Kernels, in this case, exponential ones. For assessing the proposed method performance, it is applied a clustering algorithm commonly used and recommended by literature: normalized cuts based clustering. Experimental results are obtained from the processing of well known image and toy data bases. Results are comparable with those reported in the literature.**

***Index Terms*— Eigen-solution, exponential kernels, Gaussian kernels, graph theory, spectral clustering.**

I. Introduction

In pattern recognition context, determining the most relevant features is, in almost all cases, a crucial stage to design an automatic classification system. On this regard, many approaches have been proposed that are focused on several problems, such as compactness guarantee, separability, classification performance, dimensionality reduction, feature extraction, among others [1]. For instance, the conventional principal component analysis (PCA) is applied to determine how many features are relevant according to an explained variance criterion [2], [3]. One of the greatest disadvantages of the variance analysis-based methods is that they cannot successfully be applied on problems where features are highly uncorrelated, which for difficult classification data do not guarantee separability, as well as large data.
By other hand, there exist many methods for solving the previous mentioned problems but they are usually computationally expensive [4]. In general, there is no a generic established method for relevance analysis that can be applied to different types of data, achieving a good tradeoff between computational cost and performance in terms of classification. It has been proposed many methods that accomplish an admissible performance but increasing the computational cost. In contrast, there exist other methods that require low computational cost but decreasing classification performance.

In this work, a new method for relevance analysis, applied on difficult data bases employing spectral information about data, is presented that is done from a graph theory point of view. This method is carried out by using Gaussian Kernels instead of conventional quadratic forms and then avoiding the need of a linear combination-based representation. For this end, it is implemented an extended approach for relevance analysis using alternative Kernels, in this case, exponential ones. For assessing the proposed method performance, it is applied a clustering algorithm commonly used and recommended by literature: normalized cuts based clustering. For assessing the validity of the proposed method was tested with experimental toy data base and real data bases, making use of the Fisher criterion and Clustering Index. Performance and effectiveness is evaluated in terms of reduced dimension, which, obtaining a classification of features that represent an ideal way of the databases analyzed and in turn allow the reliability of the proposed system. Results are comparable with those reported in the literature.

[1] lmmoralesv@unal.edu.co, macgarciamu@unal.edu.co, dhpeluffoo@unal.edu.co, cgcastellanosd@unal.edu.co

II. Theoretical background

A. Multi-class Spectral Clustering (MCSC)

A weighted graph can be represented as $\mathbb{G} = (\mathbb{V}, \mathbb{E}, \mathbf{\Omega})$, where $\mathbb{V}$ is the set of either nodes, $\mathbb{E}$ is the edge set, and $\mathbf{\Omega}$ represents the relationship among nodes, that is the kernel matrix or *affinity* matrix as well. Given that, each affinity matrix entry ω_{ij} of $\mathbf{\Omega} \in \mathbb{R}^{N\times N}$ represents the weight of the edge between i-th and j-th element, it must be a non-negative value. Value N is the number of considered samples or nodes. In addition, for a non-directed graph, it holds that $\omega_{ij} = \omega_{ji}$. Therefore, affinity matrix must be chosen as symmetric and positive semi-definite one. In spectral analysis, term $\mathbb{V} = \{1, \ldots, n\}$ represents the indices of data set to be grouped. The aim of spectral clustering is to decompose $\mathbb{V}$ into K disjoint subsets, then, $\mathbb{V} = \cup_{k=1}^{K} \mathbb{V}_k$ and $\mathbb{V}_l \cap \mathbb{V}_k = \emptyset$, $\forall l \neq k$, such decomposition is done, commonly, by using spectral information and orthonormal transformations. Data to be clustered are to be denoted as $\boldsymbol{X} \in \mathbb{R}^{N\times d} = (\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_N^\top)^\top$,where $\boldsymbol{x}_i \in \mathbb{R}^d$ is the i-th sample related to i-th node.

In matrix representation terms, the aim of MCSC is to determine a binary indicator matrix $\boldsymbol{M} = (\boldsymbol{m}^{(1)\top}, \ldots, \boldsymbol{m}^{(K)\top})$, where each vector $\boldsymbol{m}^{(k)}$ is a column vector formed by data point membership regarding cluster k. Each entry ik of matrix $\boldsymbol{M}$ is defined as

$$m_{ik} = \lfloor i \in \mathbb{V}_k \rfloor, \quad i \in \mathbb{V}, \quad k = 1, \ldots, K,$$

where notation $\lfloor \cdot \rfloor$ stands for a binary indicator - it equals to 1 if its argument is true and, otherwise, 0. Also, because each node can only belong into one partition, the condition

$\boldsymbol{M}\mathbf{1}_K = \mathbf{1}_N$ must be satisfied, where $\mathbf{1}_d$ is a d-dimensional all ones vector.

Then, the well-known k-way normalized cuts-based clustering, described in [5], can be written as:

$$\max_{\boldsymbol{M}} \varepsilon_{\boldsymbol{M}} = \frac{1}{K} \frac{\text{tr}(\boldsymbol{M}^\top \boldsymbol{\Omega} \boldsymbol{M})}{\text{tr}(\boldsymbol{M}^\top \boldsymbol{D} \boldsymbol{M})} \tag{1a}$$

$$\text{s. t.} \quad , \boldsymbol{M} \in \{-1, 1\}^{N \times K}, \quad \boldsymbol{M}\mathbf{1}_K = \mathbf{1}_N \tag{1b}$$

where $\boldsymbol{D} \in \mathbb{R}^{N \times N}$ is the degree matrix related to weights or affinity matrix, defined as $\boldsymbol{D} = \text{Diag}(\boldsymbol{\Omega}\mathbf{1}_N)$. Notation $\text{Diag}(\cdot)$ denotes a diagonal matrix formed by its argument vector. Expressions (1a) and (1b) are the formulation of normalized cuts optimization problem, named (*NCPM*).

III. Relevance Analysis-based Data Projection

Let us consider the notation given in Table III.

Term	Notation	Description
Original data matrix	$\boldsymbol{X}$	$\boldsymbol{X} \in \mathbb{R}^{N \times d}$
Rotation matrix	$\boldsymbol{Q}$	$\boldsymbol{Q} \in \mathbb{R}^{d \times d}, \boldsymbol{Q}^\top \boldsymbol{Q} = \boldsymbol{I}_d$
Truncated rotation matrix	$\widehat{\boldsymbol{Q}}$	$\widehat{\boldsymbol{Q}} \in \mathbb{R}^{d \times p}, \widehat{\boldsymbol{Q}}^\top \widehat{\boldsymbol{Q}} = \boldsymbol{I}_d,\ p < d$
Projected data	$\boldsymbol{Y}$	$\boldsymbol{Y} \in \mathbb{R}^{N \times d}, \boldsymbol{Y} = \boldsymbol{X}\boldsymbol{Q}$
Truncated projected data	$\widehat{\boldsymbol{Y}}$	$\widehat{\boldsymbol{Y}} \in \mathbb{R}^{N \times p}, \widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{Q}}$
Reconstructed data	$\widehat{\boldsymbol{X}}$	$\widehat{\boldsymbol{X}} \in \mathbb{R}^{N \times d}, \widehat{\boldsymbol{X}} = \widehat{\boldsymbol{Y}}\widehat{\boldsymbol{Q}}^\top$

For obtaining a rotation matrix $\widehat{\boldsymbol{Q}}$ such that $\widehat{\boldsymbol{Y}}$ contains the projected vectors that most contribute to the explained variance, in [6] is introduced the following optimization problem:

$$\min_{\widehat{\boldsymbol{Q}}} ||\boldsymbol{X} - \widehat{\boldsymbol{X}}||_{\boldsymbol{A}}^2 = \max_{\widehat{\boldsymbol{Q}}} \text{tr}(\widehat{\boldsymbol{Q}}^\top \boldsymbol{X}^\top \boldsymbol{A} \boldsymbol{X} \widehat{\boldsymbol{Q}}) \tag{2}$$

$$\text{s. t.} \quad \widehat{\boldsymbol{Q}}^\top \widehat{\boldsymbol{Q}} = \boldsymbol{I}_d \tag{3}$$

where $\boldsymbol{A}$ is a semidefinite positive matrix.

The eigenvectors associated to the p largest eigenvalues of $\boldsymbol{X}^\top \boldsymbol{A} \boldsymbol{X}$, i.e., $\boldsymbol{Q} = \text{eig}(\boldsymbol{X}^\top \boldsymbol{A} \boldsymbol{X})$ is a feasible solution. The p value is established by means of an accumulated variance criterion. As can be inferred from equation 2, when matrix $\boldsymbol{A}$ is chosen as $\boldsymbol{I}$ we have the standard PCA. When $\boldsymbol{A}$ is chosen as a weighted covariance, WPCA versions are accomplished. Because affinity matrix contains the relationship values of all data points, we propose to select matrix $\boldsymbol{A}$ as the affinity $\boldsymbol{\Omega}$ that can be understood as an estimation of a covariance matrix.

IV. Experimental Setup

Database of public domain numerical matrix format is used for the analysis and study of the proposed method. Table I summarizes the real data used for the analysis showing the variability of selected databases.

Experiments are carried out on two well-known database collections: Firstly, a toy data comprising the following several data sets (4Gaussians, Bulls eye 2 circulos, Bulls eye 3 circulos, dataset1, dataset2, dataset3, dataset 4grupos, dataset 5grupos, HappyFace) as is shown in the Figure 1.

TABLE I
Database Summary

Dataset	Source	n	p	k	Description
80x	[7]	1	8	2	–
Auto mpg	[7]	398	6	2	*Multivariate real attributes on automobile description concerning to city-cycle fuel consumption.*
Biomed	[8]	194	5	2	*Multivariate real attributes on blood measurements for "normal" and "carrier" samples.*
Breast	[7]	683	9	2	*Multivariate integer attributes on breast cancer samples.*
Iris	[7]	150	4	3	*Multivariate real attributes from size measurements on iris plants.*
Malaysia	[7]	1	8	2	*This dataset concerns simple measurements on segments of utility symbols.*

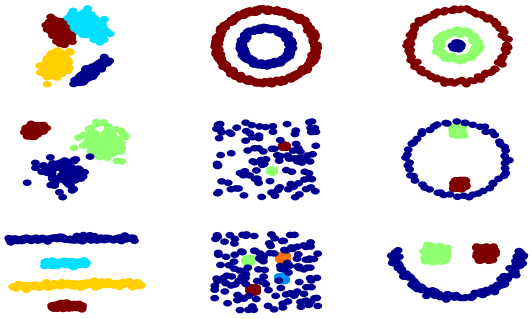

Fig. 1. Considered artificial databases

Secondly, a real databases comprising the following several data sets(iris, 80x, auto_mpg, malaysia, biomed, breast, cbands, chromo, diabetes, ecoli, glass, heart, ionosphere, liver satellite, sonar, spybean1, soybean2, spirals, twonorm, ringnorm, wine, mfeat_fac, mfeat_fou, mfeat_Kar, mfeat_pix, mfeat_zer, mfeat_mor).
Estimation of the group number, k, is based on calculation of the eigenvector set of the affinity matrix [9]. In particular, the scaled exponential affinity matrix $\boldsymbol{\Omega} = \omega_{ij}$ is employed that holds elements defined as follows: [10]

$$\omega_{ij} = \begin{cases} \exp\left(-\frac{d^2(x_i, x_j)}{\sigma_i \sigma_j}\right), & i \neq j \\ 0, & i = j \end{cases} \tag{4}$$

Where $\boldsymbol{X} \in \Re^{n \times p} = (x_1^\top, \ldots, x_n^\top)^\top$ is the data matrix, $x_i \in \mathbb{R}^p$ is its corresponding i-th data point, $\sigma_i = d(x_i, x_N), x_N$ denotes the N-th nearest neighbor, and $d(\cdot, \cdot)$ stands for Euclidian distance. The value of N is experimentally set to be 2.

Table II shows the clustering performance measures with their description.

Lastly, testing within experimental framework is carried out by employing MATLAB Version 7.10.0(R2010a) in a standard PC Acer Nplify TM 802 AMD processor V120, 2.2 GHz and 2 GB RAM memory.

TABLE II
APPLIED PERFORMANCE MEASURES

Measure	Description
Fisher Criterion	Indicates an adequate clustering when the greater is its value.
J	$J = \frac{\sum_{j=1}^{k} \boldsymbol{q}_k - \hat{\boldsymbol{q}}}{\mathrm{tr}(\sum \Sigma_j)}$ where $\boldsymbol{q}_j$ is the mean of i-th cluster, $\hat{\boldsymbol{q}}$ is the mean of whole data $\boldsymbol{X}$ and Σ_j is the covariance matrix associated to cluster j.
Cluster Coherence	It is ranged into $[-1, 1]$ and close to 1 when well clustering.
ε_M	$\varepsilon_M = \frac{1}{k} \sum_{l=1}^{k} \frac{\boldsymbol{M}_l^\intercal \boldsymbol{\Omega} \boldsymbol{M}_l}{\boldsymbol{M}_l^\intercal \boldsymbol{D} \boldsymbol{M}_l}$
Silhouette	It ranges from -1 to 1, being 1 when clustering adequately.
S	$s_i = \frac{\min(\boldsymbol{b}_i - a_i)}{\max(a_i, min(\boldsymbol{b}))}$ where a_i is the average distance from the i-th point to the other points in its cluster, $\boldsymbol{b}_i = (b_1^i, \ldots, b_k^i)$ and b_j^i is the average distance from the i-th point to points from cluster j.

V. RESULTS AND DISCUSSION

Table III shows the numerical results obtained for each of the components of toy data sets, in general we can note that the exponential kernels works significantly better than the conventional method for relevance analysis. This fact can be appreciated in the following components: 4Gaussians, Bulls_eye_3_circulos, dataset2, dataset3, dataset_4grupos and dataset_5grupos.

Table IV shows the numerical results obtained for each of the components of real databases, in general we can note that the exponential kernels works significantly better than the conventional method for relevance analysis. This fact can be apreciated in the following components: iris, 80x, biomed and breast.
Then, be shown the images of relevance analysis of somethings toy and real databases.

TABLE III
J_1, J_2, J_3 TOY DATA

Data Bases	J_1	J_2	J_3
4Gaussians	1.6042	19.7582	29.4229
Bulls eye 2 circulos	0.0001	0.0001	0.0001
bulls eye 3 circulos	0.0845	0.1237	10.6812
dataset1	6.8847	21.7975	15.8113
dataset2	0.2939	0.5037	6.6462
dataset3	1.1296	1.0908	13.7108
dataset4 grupos	0.5379	2.0355	22.9326
dataset5 grupos	0.5237	2.4426	9.1082
Happy Face	1.5271	2.6564	1.4391

- *Toy databases*
- **4 Gaussians:**for this case we can note that the second is the most relevant feature.
- **Bulls eye 2 circulos:**we can note that the first is the most relevant feature.
- **Happy Face:**we can note that the second is the most relevant feature.

Figure 2 show the boxplot for the toy databases; we can note that the three boxplots are symmetric, symmetrical to the right and symmetrical to the right respectively, interquartile range is 1.43, 6.92 and 15.94 respectively. The top quartile has its maximum point in 1.45 , 6.92 and 24.6; we can note that the exponential kernels works significantly better than the conventional method for relevance analysis.

TABLE IV
J_1, J_2, J_3 REAL DATA

Data Bases	J_1	J_2	J_3
Iris	45.9371	64.7114	104.8314
80x	0.6289	0.9808	12.4936
auto_mpg	49.1341	25.9333	26.0328
malaysia	2.1562	2.8964	1.9209
biomed	2.5715	1.5854	2.4243
breast	2.6785	2.6955	4.5962

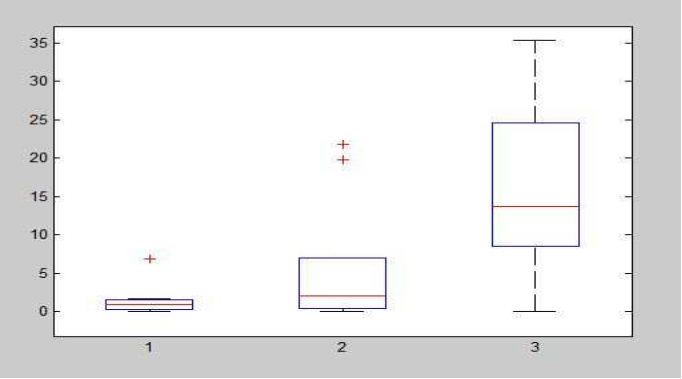

Fig. 2. Boxplot Toy databases

- *Real databases:*

- **iris:**we can note that the first is the most relevant feature.
- **80x:**As show the image the most relevant features are the second and sixth.
- **auto_mpg:**In this case there is a difference between the third feature is the most relevant and all other.
- **malaysia:**In this case there is a difference between the second feature is the most relevant and all other.
- **biomed:**we can note that the fifth is the most relevant feature.
- **breast:**we can note that the first is the most relevant feature.

Figure 3 shows the boxplot for the Real databases; we can note that the three boxplots are symmetrical to the right, interquartile range is 44.36, 25.6 and 26.5 respectively. The top quartile has its maximum point in 45.3, 26 and 26.5.

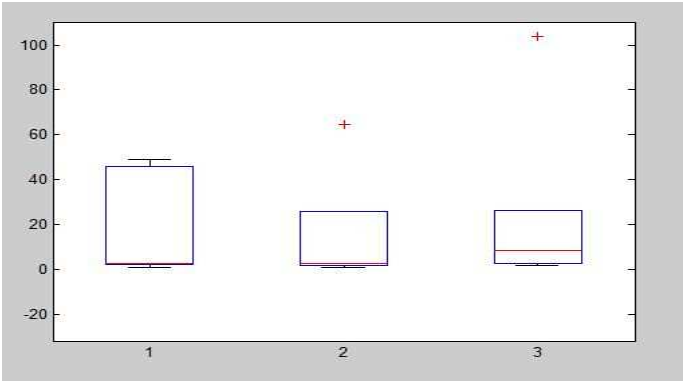

Fig. 3. Boxplot Real databases

A. Clustering Index

- *Toy databases*

It can be noted in Figure 4 that the proposed method to artificial data bases, is satisfactory, because in the Table V shows the corresponding values, which have a value near to one, this indicates the effectiveness of the exponential Kernels, that shows a good measure of clustering. Also we can say that linear projection improves the clustering performance.

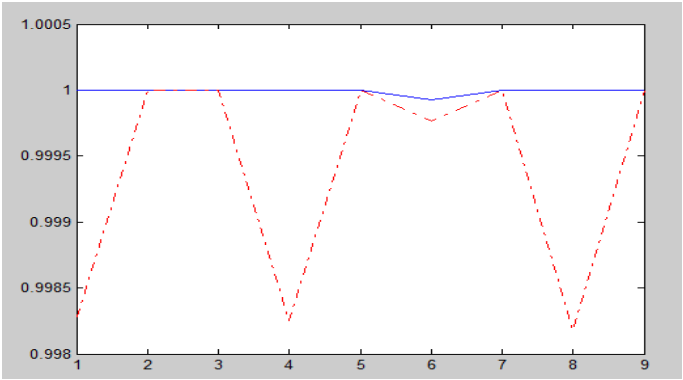

Fig. 4. Clustering coherence of toy data

TABLE V
PLOT FOR CLUSTERING COHERENCE OF TOY DATA

Original	Spectral Clustering
1.0000	0.9983
1.0000	1.0000
1.0000	1.0000
1.0000	0.9983
1.0000	0.9999
0.9999	1.0000
1.0000	0.9998
1.0000	1.0000

As it can be seen in the Figure 5, the distribution obtained from the toy database can be distributed as follows:

– ***Input:*** Presents a non-symmetrical distribution because the median is very close to the upper face, so it is easy to see that approaches the third quartile for this reason the data have a distribution skewed to the left. One can appreciate the outliers of the distribution, these values may represent the effects of extraneous causes that is, measurement error or error in any of the records, these are significant with a red cross in the coordinates (1, 1). Intercualtil range is 0.002, which shows a dispersal in the minimum database, which carries a degree of detachment rate compared to its average value. On the other hand, we see that the top quartile has a peak 1, which indicates that the method is useful, obtaining a satisfactory method.

– ***Output:*** As input, does not present a symmetrical distribution, the median is observed that in this case is the quartile 3 of the picture, it is concluded that in this case presents an asymmetry to the left. Presents a minimum value of 0.998, and in this case not observed outliers which may cause error in the output of the grouping. Intercualtil range is 0.002, this result shows the dispersal in the databases as input, indicating that the degree of detachment is a little higher with respect to its mean value. The proposed method is generally satisfactory, because it is close to unity.

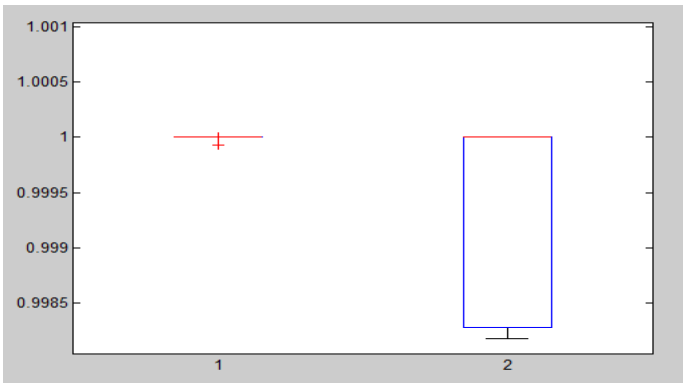

Fig. 5. Boxplot Toy databases

- *Real databases*

By the other hand according to the previews information in Figure 6, gives the results of the Real database. Table VI shows the respective values of both original and obtained by the exponential kernel. We conclude that for this database the method presents a behavior and a high effectiveness, because these measures are near to unity, as the results of the toy databases is possible to say that improves the linear projection clustering performance too.

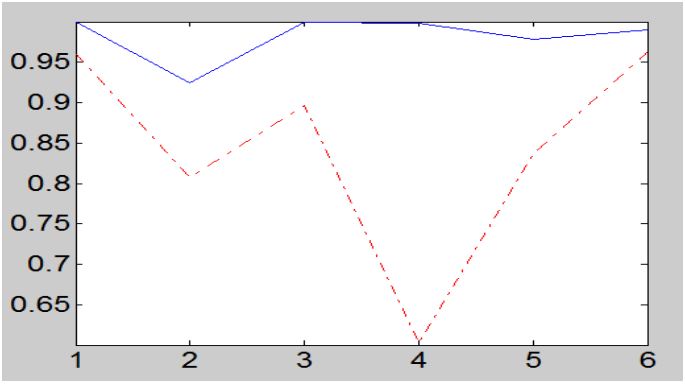

Fig. 6. Clustering coherence of Real data

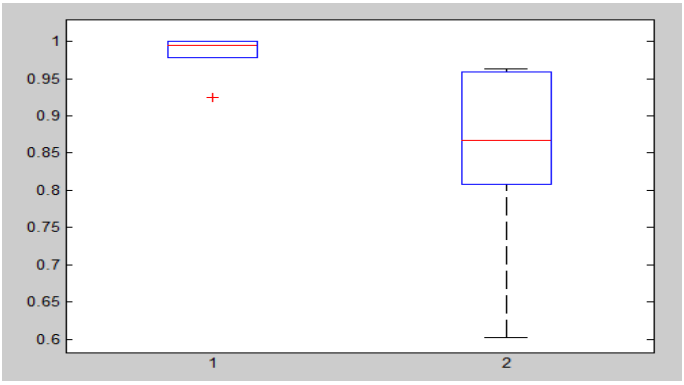

Fig. 7. Boxplot Toy databases

TABLE VI
PLOT FOR CLUSTERING COHERENCE OF REAL DATA

Original	Spectral Clustering
1.0000	0.9592
0.9020	0.8081
1.0000	0.8961
0.9984	0.6025
0.9788	0.8373
0.9906	0.9634

As it can be seen in the Figure 7, the distribution obtained from the real database can be distributed as follows:

–***Input:*** Did not show a symmetrical distribution, also mediated is very close to the upper face, so it is easy to see that approaches the third quartile for this reason the data have a distribution skewed to the left. It can be noted the outliers of the distribution, these values may represent the effects of extraneous causes that is, measurement error or error in any of the registers, these are significant with a red cross in the coordinates (1, 0.923). Intercualtil range is 0.009, which shows dispersal in the minimum database, which carries a degree of distance rate compared to its average value.

–***Output:*** As input, does not present a symmetrical distribution, the median is observed that in this case is near to the underside box, therefore it is concluded that in this case presents an asymmetry to the right. Presents a minimum value of 0.6, and in this case not observed outliers which may cause error in the output of the grouping. Intercualtil range is 0.151, this result shows the dispersal in the databases is a little higher than the input, indicating that the degree of detachment is a little higher with respect to its mean value. The upper level is at 0.963. The proposed method is generally satisfactory, because it is near to unity.

VI. CONCLUSIONS AND FUTURE WORK

We introduced a variant of a linear projection approach for relevance analysis by means of a generalize distance regarding affinity matrix. Projection improves the clustering performance and can be extended as a feature selection method obtaining a relevance vector.

As a future work, more spectral analysis methods and affinity measures will be explored to design an adequate relevance analysis approach keeping a good trade-off between the number of resultant features and performance.

VII. ACKNOWLEDGEMENT

Authors would like to thank Recognition and Digital Signal Processing research group.

REFERENCES

[1] D. H. Peluffo-Ordoñez, J. L. Rodríguez-Sotelo, D. Cuesta-Frau, and C. G. Castellanos-Domínguez, “Estudio comparativo de métodos de selección de caracaterísticas de inferencia supervisada y no supervisada,” *Tecno Lógicas*, no. 23, pp. 149–166, December 2009.

[2] S. Sabha and H. Tamimi, “From appearance matching to feature invariant matching face recognition: Comparisons between pca and sift,” 2012.

[3] V. RADHA and M. PUSHPALATHA, “Comparison of pca based and 2dpca based face recognition systems,” *International Journal of Engineering Science*, vol. 2, 2010.

[4] R. Picard, E. Vyzas, and J. Healey, “Toward machine emotional intelligence: Analysis of affective physiological state,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 10, pp. 1175–1191, 2001.

[5] Y. S. X. and S. Jianbo, “Multiclass spectral clustering,” in *ICCV ’03: Proceedings of the Ninth IEEE International Conference on Computer Vision*. Washington, DC, USA: IEEE Computer Society, 2003, p. 313.

[6] J. C. no Candamil, S. Garcia-Vega, D. P.-O. nez, and C. Castellanos-Domínguez, “A comparative study of weighting factors for wpca based on a generalized distance measure,” *XVII Simposio de tratamiento de señales, imágenes y visión artificial. STSIVA*, 2011.

[7] A. Frank and A. Asuncion, “UCI machine learning repository,” 2010. [Online]. Available: http://archive.ics.uci.edu/ml

[8] R. Duin, P. Juszczak, P. Paclík, E. Pekalska, D. de Ridder, D. Tax, and S. Verzakov, “PR-Tools4.1, a matlab toolbox for pattern recognition,” 2007. [Online]. Available: http://prtools.org

[9] L. Zelnik-manor and P. Perona, “Self-tuning spectral clustering,” in *Advances in Neural Information Processing Systems 17*. MIT Press, 2004, pp. 1601–1608.

[10] D. Peluffo-Ordoñez, C. D. Acosta, and G. Castellanos, “An improved multi-class spectral clustering based on normalized cuts,” *CIARP*, 2011.

Luz Marina Morales-Villamil was born in Manizales, Caldas in 1993. Currently she is student of Electronic and Electric Engineering (fifth semester), at the Universidad Nacional de Colombia - Manizales. Also, she is an active member of Applied Maths Group from the same university. Her main research interest are image processing, telecommunications, biosignals processing and energy quality.

María Camila García-Murillo was born in Manizales, Caldas, in 1993. Currently, she is studying electronic engineering (fifth semester) at Universidad Nacional de Colombia - Manizales. Also, she is an active member of Applied Maths Group from the same university. Her main research interests are image processing, robotics and indutrial automation.

Diego Hernán Peluffo-Ordoñez was born in San Juan de Pasto, Nariño, in 1986. He received his degree in electronic engineering and the M.Eng. degree in industrial automation from the Universidad Nacional de Colombia, Manizales, Colombia, in 2008 and 2010, respectively. Currently, he is PhD student in the same university. His main research interests are applied maths and unsupervised learning and their applications in biosignals analysis.

César Germán Castellanos-Domínguez received his undergraduate degree in radiotechnical systems and his Ph.D. in processing devices and systems from the Moscow Technical University of Communications and Informatics, in 1985 and 1990 respectively. Currently, he is a Full Professor in the Department of Electrical, Electronic and Computer Engineering at Universidad Nacional de Colombia Sede Manizales. He is also the Leader of the Signal Processing and Recognition Group at the same university. His teaching and research interests include information and signal theory, digital signal processing and bioengineering.