K-SEGMENTS CLASSIFIER - A NON-LINEAR APPROACH FOR THE CLASSIFICATION OF SAMPLING DATA

ZAUDIR DAL CORTIVO
Federal University of Paraná
Numerical Methods in Engineering
PPGMNE/UFPR
Centro Politécnico:Jardim das Américas
81531-970, Curitiba (PR)
BRAZIL
zdalcortivo@gmail.com

JAIR MENDES MARQUES
Federal University of Paraná
Numerical Methods in Engineering
PPGMNE/UFPR
Centro Politécnico:Jardim das Américas
81531-970, Curitiba (PR)
BRAZIL
jair.marques@utp.br

Abstract: This paper proposes a method to classify sampling data based on the k-segments algorithm. The data classification efficiency is relevant for this multivariate statistical analysis technique. The method consists of adjusting an a priori defined polygonal line for each class. A new observation is then classified into the class for which polygonal line it has the smallest orthogonal distance. Experimentally, the algorithm is applied to several sets of sampling data and the results are compared with the apparent error rate, which demonstrates the good performance of this methodology.

Key–Words: Discriminant analysis, k-segments algorithm, principal curves

1 Introduction

In many fields of science, the discrimination and classification of sampling data in order to discover relations and patterns in a data set, is a fundamental step of many tasks. For example, understanding and controlling the variation of certain procedures may be important in production, if a product finds itself within the specified tolerances. When more than one machine is being used for production, one can test whether all machines are producing within the same set of specifications [18].

There are several techniques that can be used for the classification of data, such as neural networks, knn (*k*-th nearest neighbor), svm (support vector machine), naive bayes, and Fisher discriminant analysis, among others. The way in which an observation is classified is different for each method. The classification or allocation can be defined as a set of rules that will be used to allocate new objects [14]. For example, for a set with k different classes C_1,C_2...,C_k in which one wants to sort a new individual $\underline{x}^{'}$ = (x_1, ..., x_p) in a space formed by p variables, the knn method identifies the *k*-nearest neighbors of the vector $\underline{x}$ that one wants to classify (the distance between the vectors is calculated), and $\underline{x}$ is classified in the group that resulted in a greater number of neighbors. In Fisher discriminant analysis, on the other hand, the centroids (means for each class) are calculated and a new individual is classified in the group with the smallest distance to the centroid. In all these techniques, the efficiency of the classification of a new individual is always relevant, and a good classification should result in small errors. That is, there should be a small likelihood of misclassification.

There are many situations in which classification and discrimination techniques can be applied. [27] developed an iterative method based on KDE (Kernel Fisher Discriminant Analysis) for pattern classification. [17] applied a new classifier to large data repositories in order to improve the efficiency of classification. These repositories involved large amounts of data collected on people by governments and public and private companies. The information that can be obtained from these repositories are important for the planning of public policies, or to increase and improve the services offered by businesses and governments. For the manipulation of this data, a random data disturbance is performed so as not to violate the privacy of individuals. For this, a new algorithm called NSVDist (Non-homogeneous generalization with Sensitive Value Distributions) was used. In experiments with eight data sets and three additional different classification algorithms, the NSVDist algorithm presented better accuracy than the other three

classifiers used for this type of data. [20] applied discriminant analysis to identify and analyze the performance of strategic classes of Brazilian clothing products, describing their characteristics and comparing them with indicators defined throughout the study. They used data from 510 companies in the sector in the year 2006. The techniques used were data envelopment analysis, cluster analysis to identify groups and discriminant analysis to validate them. [28] combined the technique of *k*-nearest neighbor classification with LAD (Tree Through Stacking) in two different types of data: the macroeconomic and risk parents (these data were collected from 27 countries), with the goal of predicting the risk of economic crisis in a country.

New approaches to classify and discriminate sampling data have appeared in the scientific community in recent decades, among which the discriminant analysis based on non-linear principal curves stands out. Principal Curves (PC), presented by [9][10] is a generalization of linear components and provides a smooth (infinitely distinguishable) one-dimensional (parameterized) approximate curve for a set of data in R^p. Other defintions for PC arose after the work by Hastie and Stuetzle. For [16], the non-linear principal components in NLPCA (Nonlinear Principal Components Analysis), are obtained through autoassociative neural networks. [24] proposed an incremental method to find the principal curve, called *k*-segments. Segments are adjusted and connected to form a polygonal line. New segments are inserted at each iteration until a criterion of optimization is reached.

In the literature, various works can be found that use principal curves for classification. [3] developed a model for the extraction and classification of data for which they proposed an algorithm to improve the performance of the fit of the principal curve. This algorithm is a combination of the original algorithm by [10] and the one by [1]. [7] used the biplot methodology with Fisher discriminant analysis and Principal Curves for the classification of sampling data. The paper shows that the incorporation of Principal Curves provides for the best rating of the data. [22] proposed a classification of ships with principal curves based on the k -segments algorithm. [26] proposed a new classifier for microarray data using principal curves. A principal curve is calculated for each class and a new sample observation is classified in the class with the curve at the smallest distance. Experimental results show that PC performs better when the sample size is small. [19] investigated the efficiency of classification using MPs (Morphological profiles) constructed from the characteristics of the NLPCA.

This work proposes an algorithm for the classification of sampling data. With this algorithm, the principal curve is extracted from the Fisher discriminant spaces matrix of each class. The principal curve used is the *k*-segments from [24]. This algorithm generates a polygonal line and the principal curve. The proposed algorithm uses the polygonal line to classify a new observation sample,which is labeled in the class whose polygonal has the smallest Euclidean distance by orthogonal projection. The performance of the algorithm is measured by the apparent error rate and some data sets from the UCI Machines are used for the simulation.

This paper is organized as follows: Section 2 provides a brief theoretical review of the tools used in this work. Section 3 describes the algorithm and the experimental results.

2 Theoretical review

2.1 Discriminant Analysis

The multivariate technique known as Discriminant Analysis deals with the issues of allocating new objects (or observations) to previously defined sets. One of the goals is to determine the variables that best discriminate between the classes, and to use them to create discriminant functions that will be employed to allocate new individuals, objects or observations in the appropriate class. This discriminant function optimizes allocation [6].

For a sampling data point x and a set with *k* different classes C_1,C_2,,C_k, the posterior probability of $\underline{x}$ being classified as belonging to class C_i, is:

$$P(C_i/\underline{x}) = \frac{p(\underline{x}/C_i)P(C_i)}{\sum_i p(\underline{x}/C_i)P(C_i)} \tag{1}$$

for i = 1, ..., k. In general, the classification may not be based on probability density, since this information in unknown. The classification is therefore formulated in terms of discriminant functions. Fisher gave to the problem of discriminating between k populations the following focus: A matrix of discriminant spaces Y = A^t X is created. This matrix contains the discriminant coefficients that maximize the ratio of the variances between the classes and within the classes [11][29].

$$\frac{w^t S_B w}{w^t S_w w} \tag{2}$$

Where S_w is the covariance matrix within each class, and is defined as:

$$S_W = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\underline{x}_{ij} - \overline{\underline{x}}_i)(\overline{\underline{x}}_{ij} - \overline{\underline{x}})^t \tag{3}$$

here S_B is the covariance matrix between classes, and is defined as:

$$S_B = \sum_{i=1}^{k} (\overline{\underline{x}}_i - \overline{\underline{x}}_i)(\overline{\underline{x}}_{ij} - \overline{\underline{x}})^t \quad (4)$$

And x_{ij} is the j-th sample of the i-th class (represented by a column vector), $\overline{\underline{x}}_i$ is the mean vector of class i, and $\overline{\underline{x}}$ is the mean vector of the data set. If $\underline{w}$ is a vector that maximizes the ratio (2), where $\underline{w}$ is a column of A, then $y = \underline{x}^t\underline{x}$ is the linear Fisher discriminant function. It can be shown that $\underline{w}$ is the eigenvector associated with the largest eigenvalue $S_w^{-1} S_B$ [13][14].

With the function $y = \underline{w}^t x$, a point $\underline{x}$ can be allocated to one of the k populations based on the 'discriminant score' $w^t x$. The sample mean $\overline{\underline{x}}_i$ (centroid) has the score $\underline{w}^t\overline{\underline{x}}_i$ =$\underline{z}_i$ A new observation $\underline{x}$ is classified in the class with the smallest distance of the centroid of each class of $\underline{z}_i$. That is,

$$\left|\underline{w}^t\underline{x} - \underline{w}^t\overline{\underline{x}}_j\right| < \left|\underline{w}^t\underline{x} - \underline{w}^t\overline{\underline{x}}_i\right| \quad (5)$$

every $i \neq j$ The Fisher classification assumes that the variables have a Gaussian distribution, and that there is homogeneity of the covariance matrices: $\Sigma_1 = \Sigma_2 = ... = \Sigma_k$= Σ. The main advantage of this technique is computational simplicity. The centroids, distances and the comparison of these require little calculation. Also, it must be considered that the centroid (which is an average) is a consistent Estimator. As disadvantages of this technique, it must considered that:

I. Using centroids can penalize groups with greater variance. The comments in these groups will have greater chance to be classified to other groups.

II. Each class is represented by just one point (centroid). In groups with greater dispersion or asymmetrical, can increase the incorrect classification of points in other groups.

2.2 Principal Curves

Given the p-dimensional random vector $\underline{x}^t = (x_1, x_2, ..., x_p)$ with a second moment, $f(\lambda)$ is a smooth curve (C^∞) in parameterized $\Re^p$ on a closed interval. For each vector $\underline{x}$, $\lambda_f(\underline{x})$ is defined as the nearest point of the $\underline{x}$ curve. The $f(\lambda)$ curve is therefore called the principal curve for the distribution of the random vector $\underline{x}$ if:

$$E(\underline{x}/\lambda_f(\underline{x}) = \lambda) = f(\lambda) = \begin{bmatrix} f_1(\lambda). \\ . \\ . \\ f_p(\lambda) \end{bmatrix} \quad (6)$$

Where $\lambda_f : R^p \mapsto R$ is the projection index. The function $f(\lambda)$ is the mean of all orthogonally projected points on the curve. This property is known as self-consistency [12]. Finding the projection of the data on the curve is equivalent to finding the value of λ that minimizes the distance between the curve $f(\lambda)$ and the point $\underline{x}$. That is:

$$\lambda = argmin_\lambda \|f(\lambda) - \underline{x}\| \quad (7)$$

The continuous multivariate distributions have infinite principal curves [4].

The procedure for finding the main curve begins with the first linear principal component (PCA) drawn from the data. First, the data are projected (orthogonally) on the PCA, and then all points within an interval are used to calculate the mean (figure 1). The means of each sub-interval are used to find the first approximation for the principal curve (by means of a spline, for example). The process is iterative. That is, the points are projected on the curve, the mean of each sub-interval is calculated, and the new approximation of the curve is obtained. This process is repeated until the desired convergence.

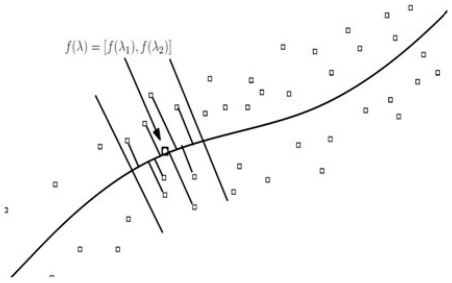

Figure 1: The principal curve of a data set. Each point on the curve is the mean of the projection of all points in the interval.

Principal Points have an important related concept, that shall see later [11]. Which is a set consisting of k classes and for each point x with known distribution. The objective is to determine the class i nearest to the point $\underline{x}$, i = 1, 2, ..., k. This induces a partition of the space of characteristics into so-called Voronoi regions. Given a set S of n points, the target is to determine for each point i of S, which is the V_i region of points that are nearest to $\underline{x}$ than any other point in S. The set of points of each V_i region that minimizes the expected distance of $\underline{x}$, are called principal points. Each principal point is self-consistent, with a mean X equal to the mean of the Voronoi region. For example,

for k = 1, X with a Gaussian distribution, the principal point is the vector of the mean; for k = 2, the principal points are the vectors of the mean of each class. The principal curve can be seen as $k = \propto$ principal points, but restricted to be a principal curve. The principal point is analogous to the centroids obtained by the k-means algorithm. Given that the principal curve points are self-consistent and also are midpoints, these can replace centroids in the classification of new observations sampling. In this way, you don't have just one point to measure the distance the distance to the vector to be classified, which can improve the classification efficiency. The main disadvantage of this technique is that the principal curves generation is a lot more complex if compared to the calculation of centroids.

Definitions and alternative methods for estimating principal curves have been suggested after the pioneering work of Hastie and Stuetzle. [16] used neural networks to obtain non-linear principal components (NLPCA). A more probabilistic approach was given by Tibshirani [13], who proposed that to find the principal curve, a penalized log-likelihood function should be minimized. [15] proposed the polygonal line algorithm, which starts with a single line and adds new vertices at each iteration. The position of the lines is updated for all vertices, such that the expected values of the squares of the distances of the points that are projected on the curve are minimized. [24] proposed the k-segments algorithm for the extraction of principal curves. This method uses a probabilistic definition to find the principal components by maximizing a function in a way similar to the Tibshirani function. Various algorithms are used to find the polygonal line: one starts with the k-means algorithm to find the subset that will contain the segment and the line in V_i is drawn (k-lines algorithm). The k-lines algorithm is then adapted to find the segment. Finally, the connection of segments is made to form the polygonal line (Hamiltonian path algorithm). For each insertion of a new segment, the curve is optimized.

A line is defined as $s(\lambda) = c + u\lambda, \lambda \in R$ and the Euclidean distance of a sample observation $\underline{x}$ to the line is defined by $d(\underline{x}, s) = inf_{t \in R} \|s(\lambda) - \underline{x}\|$. $V_1, V_2, ..., V_k$ are subsets, called Voronoi regions, so that $V_i = \{\underline{x} \in X_n / i = argmin_j d(\underline{x}, s_j)\}$. That is, V_i contains all points nearest to the i-th line. To find s_i, i = 1, ..., k, the total of the square of the distance should be minimized $\sum_{i=1}^{k} \sum_{x \in V_i} d(x, s_i)^2$. The method is incremental. That is, it starts the algorithm with k = 1. The first Voronoi regions is determined and the first line for this region. Then the line is transformed into a segment: The line is 'cut' in $\frac{3\sigma}{2}$ based on the centroid of V_i, where σ^2 is the variance. The number of regions is gradually increased until the maximum number of segments (k) initially set by the user is reached, or until a shutdown criterion is satisfied. The formation of the polygonal line is done through the connection of segments by the HP algorithm (Hamiltonian path).

2.3 Error Rate (ER) and Orthogonal Projection

Since the objective is classification, the performance of a discriminant rule should be evaluated according to its ability to correctly classify, or by its failure to classify. According to [25], the error rate or bad-classification rate is the commonly used criterion to evaluate the performance of a classifier. This rate represents the proportion or percentage of incorrectly classified patterns. Together with the error rate, a confusion matrix or bad-classification matrix is developed. Each element of this matrix represents the number of patterns of class j that were classified as class i.

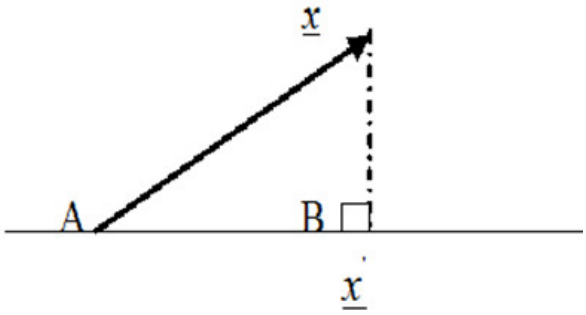

Figure 2: The principal curve of a data set. Each point on the curve is the mean of the projection of all points in the interval.

If $\underline{x}'$ is the orthogonal projection of the vector $\underline{x}$ and $\overrightarrow{Ax'}$ the projection of the vector $\overrightarrow{Ax}$ on the segment AB (figure 2), then the projection $\overrightarrow{Ax}$ is defined by:

$$\overrightarrow{Ax'} = Proj\frac{\overrightarrow{Ax}}{\overrightarrow{AB}} = \frac{\overrightarrow{Ax}\overrightarrow{AB}}{\overrightarrow{AB}\overrightarrow{AB}}\overrightarrow{AB} \tag{8}$$

The coordinates of the projection point of $\underline{x}$.

$$\underline{x}' = A + \frac{\overrightarrow{Ax}\overrightarrow{AB}}{\overrightarrow{AB}\overrightarrow{AB}}\overrightarrow{AB} \tag{9}$$

If $\overrightarrow{u}$ is the versor (unit vector) of $\overrightarrow{AB}$, then:

$$\underline{x}' = A + \frac{\overrightarrow{Ax}\overrightarrow{u}}{\overrightarrow{u}\overrightarrow{u}}\overrightarrow{u} \tag{10}$$

The Euclidean distance is calculated from $\underline{x}$ to $\underline{x}^{'}$. This distance is the shortest distance in the vector $\underline{x}$ at any point of the segment AB.

3 Methods

The k-segments algorithm brings the center of the principal curve to the origin of the coordinate space, since it removes the mean of the set of sampling data. As such, all classes retain the same center, which is the origin of the coordinate system. This fact can result in poor performance of the classifier, since the classifier algorithm uses the distance measurement of the vector $\underline{x}$ to the polygonal line of each class to perform the classification. Given the importance of maintaining the separation between the centers of classes, since the objective is to highlight the differences between the classes, through the extraction of discriminating information, the k-segments algorithm is applied to each class.

Obtaining the polygonal line on the matrix of discriminant spaces $Y = A^tX$ of each class is due to the fact that the Fisher criterion is very interesting for the breakdown of data, since it is easier to distinguish one group from another if the sum of squares between classes is large with respect to the sum of the squares within the classes [11].

The choice of principal curves as an auxiliary tool in the classification of sampling data is based on the definition of principal curves, originally given by [10], as one-dimensional curves that pass through the 'middle' of a data set in a multidimensional space. The two algorithms have similarities because both measure the Euclidean distance of each class. The k-segments algorithm measures the distance of the polygonal line that 'passes' in the middle of each class, determined by the Voronoi regions. The Fisher method measures the distance of the centroids of each class.

The classification of a vector $\underline{x}$ is briefly done as follows: the polygonal line is adjusted for each class of the data set. If $x_{ij}^{'}$ is the orthogonal projection point on the segment i of polygonal line j, then the vector $\underline{x}$ is classified in the class with smallest distance to $x_{ij}^{'}$.

The classifier algorithm can be described by the following steps: Input Data: The data set X with its k classes and the number of segments per class (s).

1. Start with the linear Fisher discriminant analysis. Get the matrix of 'discriminating scores' A. That is, the matrix of eigenvectors that contains the discriminant coefficients that maximize the ratio of the variances between the classes and within them.

2. Perform $Y = A^tX$ (matrix of discriminant functions). Separate Y per class, C_i, i = 1, , k.

3. Apply the k-segments algorithm for each C_i. For each C_i the matrices of nodes (edges) and their coordinates (vertices) are obtained.

4. Determine the orthogonal point of projection $x_j^{'}$, j= 1,..., s, where s is the number of segments per class. Projection of the vector $\underline{x}$ on each segment of C_i. Calculation of the Euclidean distances of $\underline{x}_j^{'}$ to $\underline{x}$ for each segment of class C_i. Perform $d_i = min_{j=1,2,...,s}\|x - \underline{x}_j^{'}\|$.

5. The sample observation $\underline{x}$ is classified in the group that contains the smallest distance. That is, $\underline{x} \in C_i$, is such that d_i ¡ d_j, $i \neq j$, i, j = 1, ..., k.

The advantage of this algorithm is that it not only utilizes a central value, because the principal curves are formed by self-consistent points which are also midpoints. The line generated by the algorithm can determine a non-linear adjustment to the elements of the class, which can improve the classification efficiency. The main disadvantage of this technique is the computational algorithm k-segments utilizes several other algorithms and the development of a software complexity is far superior to the calculation of centroides. The algorithm may not be as efficient in sets in which the transformation effected by discriminant analysis did not effectively separates the classes, that is, the intersection between classes is large or the distribution of discriminant scores has spherical distribution and also when there is the intersection of principal curves.

4 Experimental Results

There are several possible techniques for data classification and the classifier efficiency is relevant, because it minimizes the chance of incorrectly classifying a sample element. But why use the FDA in this work? Because the Fisher's linear discriminant analysis (FDA) is a technique commonly used for data classification, sampling existent in various statistical software. Is a great tool for supervised classification with many applications, due to its simplicity, ruggedness and predictive efficiency [30]. And, according to [31] the FDA, compared with 19 other classification techniques, presented sample classification performance, along with the LOG technique (logistic Union discriminant analysis) and higher including the quadratic discriminant analysis.

In order to study the efficiency of the method, it was experimentally applied to various data sets. Some of the sets were obtained from the UCI repository [23], the data repository of the University of California, Irvine. The datasets used are labeled as follows: Wine, Tiroide, Iris, Glass and Wave. The medical set was obtained from [5], and the alcohol set from the Tanagra data mining tutorials [21]. The Wilk's

lambda statistic is the ratio of the sum of squares within groups and the total sum of squares and assesses the difference between centroids of classes. When the value of Λ is close to 1, meaning that if you have an efficient classifier. To evaluate if the discriminant function is statistically significant, the Bartlett test to investigate the value of p-value, the result of which shall be not less than 0.05 significance level of 5%. Table 1 presents the results of these tests for each set.

Table 1: Statistical tests

Dataset	function	λ of Wilk's	Bartlett χ^2	p-value
Wine	1	0,02	666,79	0
	2	0,19	276,28	0
Thyroid	1	0,12	446,49	0
	2	0,58	115,37	0
Iris	1	0,02	545,58	0
	2	0,78	35,64	0
Alcohol	1	0,16	132,54	0
	2	0,75	20,99	8E-04
Glass	1	0,08	522,92	0
	2	0,43	173,59	0
	3	0,71	71,69	0
	4	0,87	29,72	0,003
	5	0,94	12,15	0,033
Wave	1	0,29	6125,44	0
	2	0,57	2827,72	0
Medical	1	0,13	102,04	0
	2	0,48	36,01	0

In table 1 are presented the results of the test of homogeneity of covariance matrices, obtained in each of the sets worked. For the wine, you gotta $\Lambda = 0.02$ and $\Lambda = 0.02$, for functions 1 and 2 respectively, these being good results for this test (the worst value is 1). Bartlett's test, p-value presented the values $p = 0$ and p = 0 (less than 0.05) for the two discriminant functions, that the level of significance of 5%, reject the null hypothesis that the centers of the groups are significantly different for functions 1 and 2. For the other sets the interpretation is similar to that analysis.

The entire experiment was carried out in the MatLab software. Three subroutines were used: the first, called distance, performs the Fisher discriminant analysis and the calculation of the distances of the vector $\underline{x}$ to the centroid of each class, for the classification according to the Fisher method. The second subroutine, called space, gets the vertice and edge matrices of the k-segments algorithm. The number of segments chosen by class is $k = 3$. The third subroutine, called 'k segments distance', calculates the distances of segments that form the polygonal line, by orthogonal projection.

The hypothesis test that assesses the statistical significance of the discriminatory power of the discriminant function(s) for the classification through Fisher's method is applied to all sets [8], and the first two discriminant functions had a level of confidence of 95%. These tests were performed in the Statgraphics software.

The first procedure is to find the matrix of 'discriminant scores' A for each data set and to separate them by each class. The next procedure was to trace the polygonal lines in each class using the data on the matrices of the discriminating spaces $A'X$. At this moment, the vertice and edge matrices are obtained, which provide the coordinates of the ends of each segment that forms the polygonal line. Then the projection of x on the observation segment is made and Euclidean distances are calculated for each polygon, the vector is classified in the class that contains the polygon with the shortest distance from x. Next, a comment is made on each set and the results are obtained. Tables 2, 3, 4, 5, 6, 7, 8 and 9 show the results obtained in the classification using the two methods through the confusion matrix. This matrix is a very effective way to represent the accuracy of the general classification, and also of individual cases for each class. The main diagonals of each matrix represent the correct classification for the set under analysis (number highlighted in bold in each one of the following tables). The performance of the Fisher method was only better in individual cases in some classes, such as in table 6 where the k-segment method obtained 96% correct classification against 99% for the Fisher method in the first class. In all sets, however, the k-segment method was better in the general classification of data.

4.1 Iris

Thanks to Fisher, this set is quite well-known in discriminant analysis [10]. It is composed of 3 classes, with 50 samples in each class and 4 variables. The confusion matrix for each method is presented in table 2. The results reveal the good performance of both methods, with a small advantage for the k-segments method. This good performance can be explained by the fact that this set does not have a spherical distribution. On verification of figures 3 and 4, it is easy to see that discriminant analysis is effective in the separation of the classes of this set, with small intersection in classes 2 and 3, with the centroids in the center of each class. Princial curves "pass" through the middle of each class.

Table 2: Confusion matrix for the Iris set.

Fisher stat		FDA			FDA k-segmentos		
classe	Tamanho da Classe	Classe Predita			Classe Predita		
		1	2	3	1	2	3
1	50	50	0	0	50	0	0
		100%	0%	0%	100%	0%	0%
2	50	0	48	2	0	49	1
		0%	96%	4%	0%	98%	2%
3	50	0	1	49	0	1	49
		0%	2%	98%	0%	2%	98%
Percentagem de casos classificados corretamente: 98%							98,66%

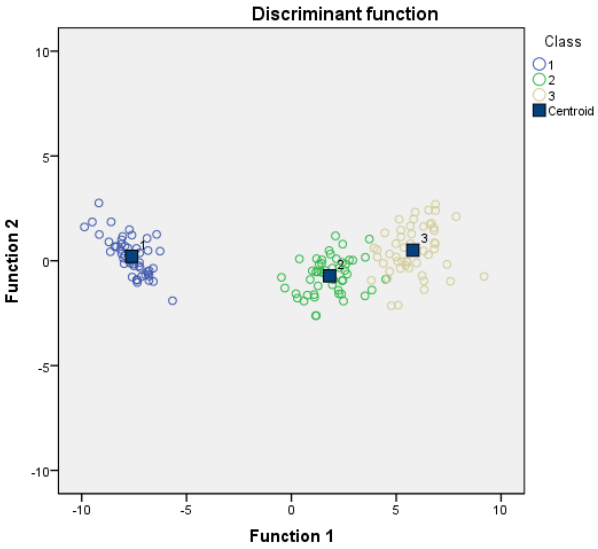

Figure 3: Centroid - Iris dataset.

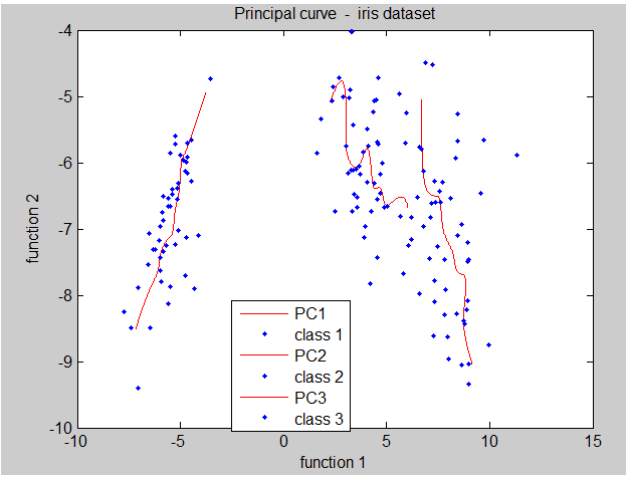

Figure 4: PC - Iris dataset.

4.2 Medical

This set consists of 54 observations, collected in a survey in the field of medicine, referring to the clinical outcomes of a group of 54 participants who answered three tests. Although the Fisher method is more efficient in the classifications in groups 1 and 2, in general the k-segment method performed better with 92.59% according to the results presented in table 3.

Table 3: Confusion matrix for the Medical set

Fisher		FDA			FDA k-segments		
class	Size of the class	Predicted Class			Predicted Class		
		1	2	3	1	2	3
1	26	22	3	1	25	1	0
		84,62%	11,54%	3,85%	96,15%	3,85%	0,00%
2	18	0	17	1	1	15	2
		0,00%	94,44%	5,56%	5,56%	83,33%	11,11%
3	10	0	1	9	0	0	10
		0,00%	10,00%	90,00%	0,00%	0,00%	100,00%
Cases classified correctly: 88,8%							92,50%

With the transformation of the FDA about the together, presented in Figure 5 and 6, the classes are well defined. But the intersection of the classes 1 and 2 and also by greater dispersion of points in class 1 generated the largest number of incorrect classifications by the FDA. The PC was more efficient in classes 1 and 3, but due to the scattering of circular shape in the class 2, this class presented lower performance, although in general classification features superior performance.

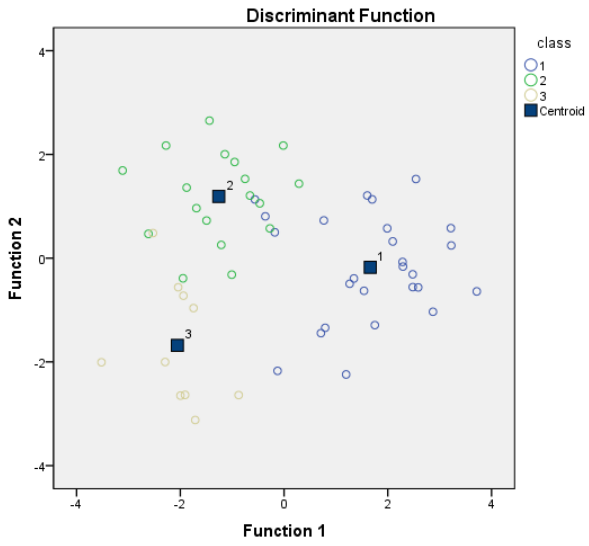

Figure 5: Centroid - Medical dataset.

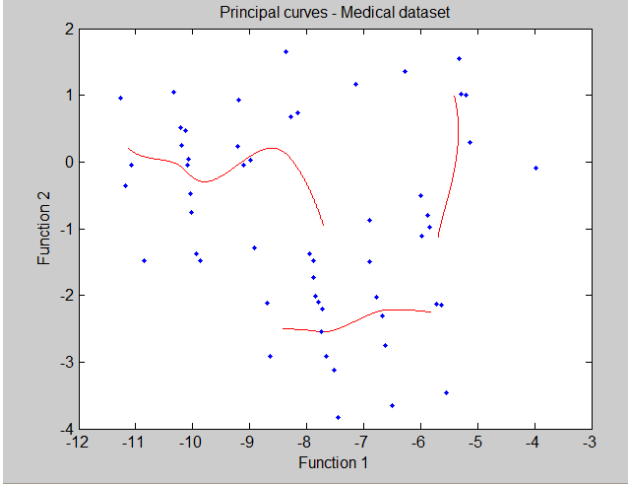

Figure 6: PC - Medical dataset.

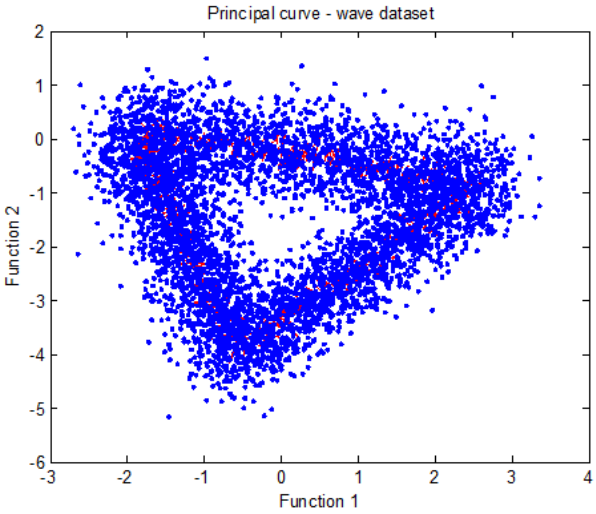

Figure 7: PC - Wave dataset.

4.3 Wave

This is an artificial set of three classes of waveforms, generated by a C-language software, obtained by [2]. Each class consists of a random convex combination of two waveforms sampled with noise added. This set contains 5000 observations and 21 variables.

Table 4: Confusion matrix for the Wave set

Fisher		FDA			FDA k-segments		
class	Size of the class	Predicted Class			Predicted Class		
		1	2	3	1	2	3
1	1657	1327	152	178	1460	82	115
		0%	9%	11%	88%	5%	7%
2	1647	103	1451	93	210	1355	82
		6,25%	88,10%	5,65%	12,75%	82,27%	4,98%
3	1696	75	71	1550	142	98	1456
		4,4%	4,2%	91,4%	8,4%	5,8%	85,8%
Cases classified correctly: 86,56%							86,52%

Even though the confusion matrix reveals a slightly better result for the Fisher method, with accuracy in the distance measurement for $\epsilon < 0.1$ by truncation, the k-segment method has a performance of 91.64% against 88,82% for Fisher (table 4). Figure 7 shows the PC for this set, with the PC at the center of the classes.

4.4 Wine

The data set wine has 3 classes of data . Class 1 has 59 sets of data, class 2 has 71 and class 3 has 48 sets of data. The number of variables is 13. Alcohol, MalicAcid, Ash, AlcalinityOfAsh, magnesium, TotalPhenols, Flavanoids, NonflavanoidPhenols, Proanthocyanins, COLORIN-intensity, hue, OD280/OD315 and proline.

Table 5: Confusion matrix for the Wine set.

Fisher stat		FDA			FDA k-segmentos		
classe	Tamanho da Classe	Classe Predita			Classe Predita		
	Class	1	2	3	1	2	3
1	59	59	0	0	59	0	0
		100%	0%	0%	100%	0%	0%
2	71	0	71	0	0	71	0
		0%	100,00%	0,00%	0%	100,00%	0,00%
3	48	0	0	48	0	0	48
		0%	0%	100%	0%	0%	100%
Cases classified correctly: 100%							100,00%

The k-segment algorithm has performace equal, because the two methods have 100% of correct classification of the data, as presented in table 5. In this set, as shown in figures 8 and 9, the separation of the classes was perfect, no intersection of scores of classes and with the principal curves and centroids Center classes.

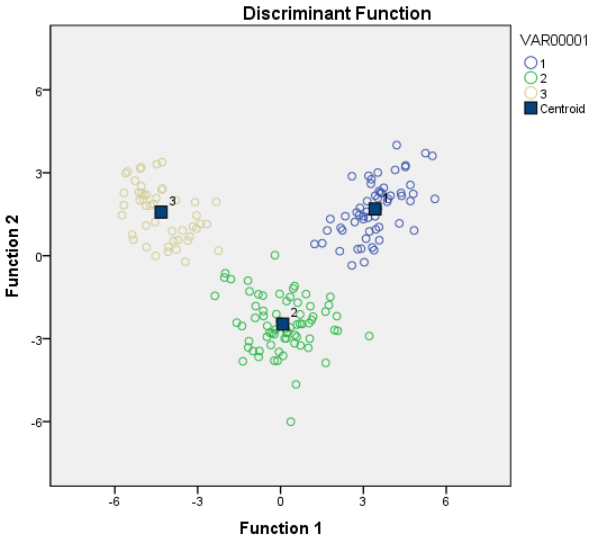

Figure 8: Centroid - Wine dataset.

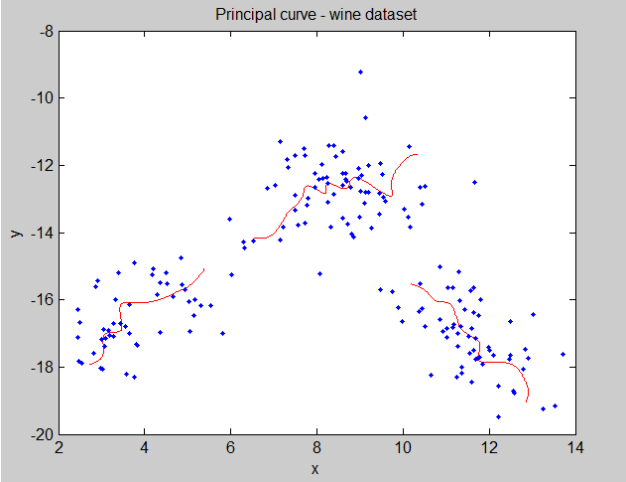

Figure 9: PC - Wine dataset.

4.5 Alcohol

With the data file alcohol, the objective is to predict the type of alcohol (kirsch, mirabelle and poire). This set has six variables (butanol, etc.), with 77 sample observations. In the analysis of the confusion matrix, the result is excellent for both methods in the kirsch class, and the good behavior for the set is mostly based on this class. The descriptive analysis confirms this result in table 6. In figure 10 and 11, scores of class 1 are separated from classes 2 and 3, but the intersection of scores of classes 2 and 3 is great. In these classes the classification was less efficient because of this intersection, also due to the proximity of centroids (FDA) and the intersection of the principal curves.

Table 6: Confusion matrix for the Alcohol set.

Fisher stat		FDA			FDA k-segmentos		
classe	Tamanho da Classe	Classe Predita			Classe Predita		
	Class	Kirsch	Mirab	Poire	Kirsch	Mirab	Poire
Kirsch	18	18	0	0	18	0	0
		100%	0%	0%	100,00%	0,00%	0,00%
Mirab	29	0	23	6	0	22	7
		0%	79,31%	20,69%	0,00%	75,86%	24,14%
Poire	30	0	9	21	0	7	23
		0%	30%	70%	0,00%	23,33%	76,66%
Cases classified correctly: 80,52%							81,81%

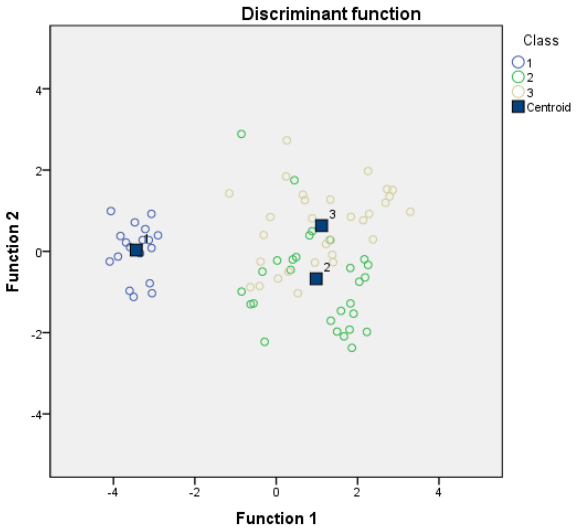

Figure 10: Centroid - Alcohol dataset.

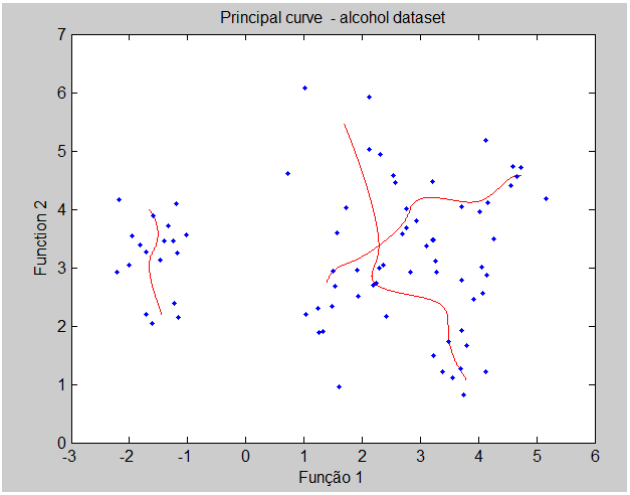

Figure 11: PC - Alcohol dataset.

4.6 Thyroid

This set consists of 3 classes, the problem is to determine whether the thyroid of a patient is in the normal state or not (hyperthyroidism or hypothyroidism). The diagnosis (the class label) was based on a full

medical report, including clinical history, examinations and etc. The good performance of both methods in global classification (94,42% and 97,21%), can be observed in table 7, mainly the results obtained in the second and third class with 100% of correct classification when using the k-segments algorithm.

Table 7: Confusion matrix for the Thyroid set.

Fisher stat		FDA			FDA k-segmentos		
classe	Tamanho da Classe	Classe Predita			Classe Predita		
	Class	1	2	3	1	2	3
1	150	149	1	0	147	2	1
		99%	1%	0%	98%	1%	1%
2	35	5	30	0	3	32	0
			85,71%	0,00%	9%	91%	0%
3	30	6	0	24	0	0	30
		20%	0%	100%	0%	0%	100%
Cases classified correctly: 94,42%							97,21%

In figures 12 and 13, the concentration of class 1 scores around the centroid, affects the performance of the algorithm k-segment, which was exceeding this rating, while for the method of centroides this is the ideal concentration to its efficiency. Already in classes 2 and 3, the scores are scattered longitudinal manner, which favors the k-segments algorithm, with 100% correct classification. For the method of centroids, this form of dispersion is not ideal, because the same having only one point per class.

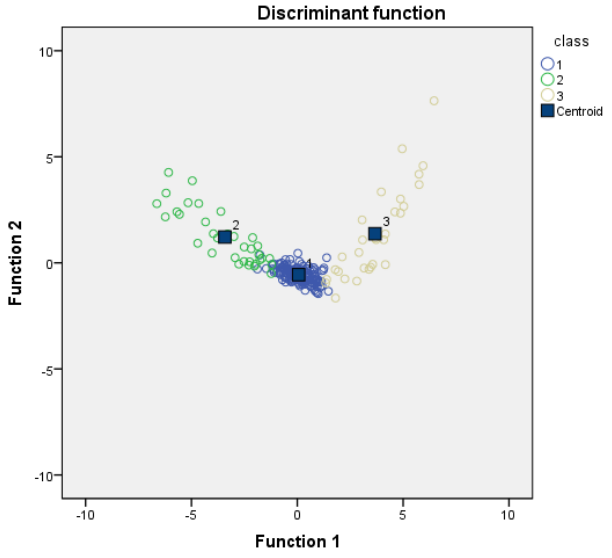

Figure 12: Centroid - Thyroid dataset.

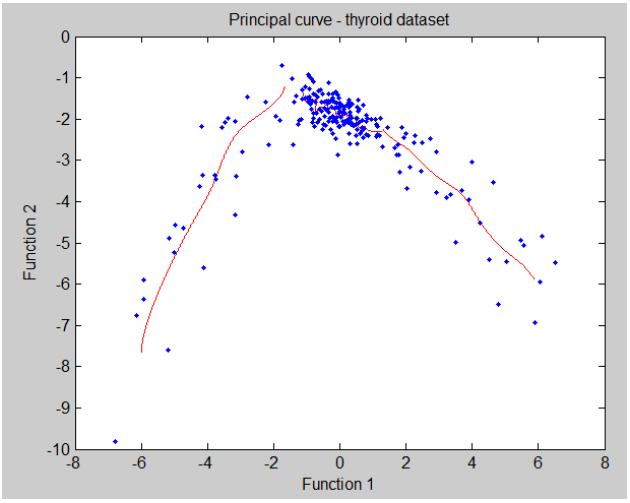

Figure 13: PC - Thyroid dataset.

4.7 Glass

The Glass set is composed of 214 samples with 6 classes. The classification study of types of glass was motivated by criminology. The type of glass collected at a crime scene, can be used as evidence. The k-segments method performed better in the classification, and both classification methods had reasonable results.

Table 8: Confusion matrix (FDA)for the Glass set.

					Fisher
					Predict Class
1	2	3	4	5	6
46	14	10	0	0	0
65,70%	20,00%	14,30%	0,00%	0,00%	0,00%
16	41	12	4	3	0
21,10%	54,00%	15,80%	5,30%	4,00%	0,00%
3	3	11	0	0	0
17,70%	17,70%	64,70%	0,00%	0,00%	0,00%
0	2	0	10	0	1
0,00%	15,40%	0,00%	76,90%	0,00%	7,70%
1	1	0	0	7	0
11,10%	11,10%	0,00%	0,00%	77,80%	0,00%
0	1	1	2	1	24
0,00%	3,50%	3,50%	6,90%	3,50%	82,80%
Cases classified incorrectly : 64,95%					

Table 9: Confusion matrix (K-Segments) for the Glass set.

					k-segments
					Predict Class
1	2	3	4	5	6
55	11	4	0	0	0
78,60%	15,70%	5,70%	0,00%	0,00%	0,00%
23	38	5	3	10	0
30,30%	50,00%	6,60%	3,90%	13,20%	0,00%
11	5	1	0	0	0
64,70%	29,40%	5,90%	0,00%	0,00%	0,00%
0	1	1	9	8	1
0,00%	7,70%	7,70%	69,20%	61,50%	7,70%
0	1	0	0	8	0
0,00%	11,10%	0,00%	0,00%	88,90%	0,00%
1	1	0	0	1	26
3,40%	3,40%	0,00%	0,00%	3,40%	89,70%
Cases classified incorrectly 65,88% com $\epsilon < 0{,}1$					

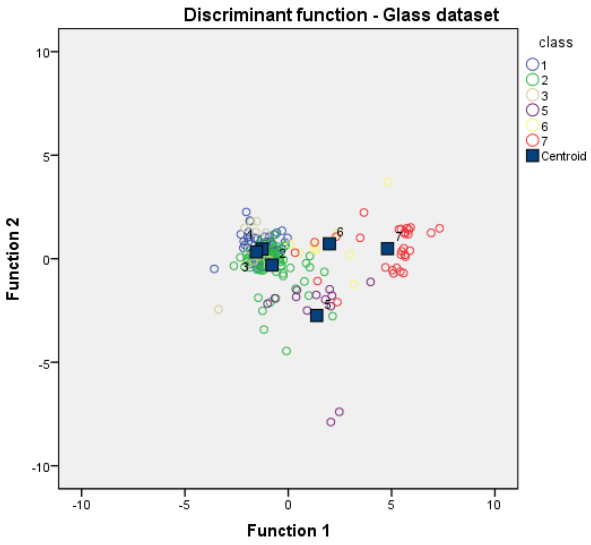

Figure 14: Centroid - Glass dataset.

The analysis of the percentages of incorrect classifications, shown in tables 8 and 9, reveals that the k-segments model has a better predictive ability, although the individual performance in class 2 is worse. In figures 14 and 15, the separation of the classes by the discriminant analysis was not efficient, because the are close and with great centroides intersection between scores of classes. The classification by the two methods used was reasonable.

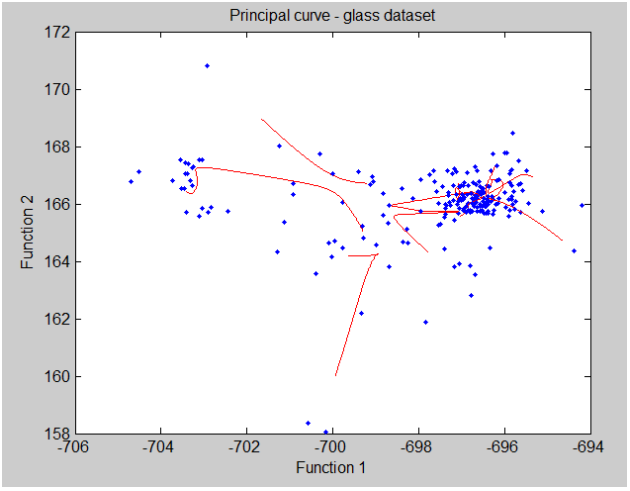

Figure 15: PC - Glass dataset.

5 Conclusion

In this work, an algorithm was developed for the classification of sampling data, called the k-segments classifier. The algorithm was compared with the Fisher method. Experimentally, the proposed algorithm demonstrated good performance compared with the Fisher method, since the probability of misclassification is smaller in all the sets studied. For future searches can study the appropriate number of segments to which the algorithm is more efficient in the classification. Another possibility is the use other algorithms for constructing principal curves as PC Hastie and Stuetzle NLPCA and Kramer.

References:

[1] J. D. Banfield, A. E. Raftery. Ice floe identification in satellite imagens using mathematical morphology and clustering about principal curves. *American Statistical Assoc.*, Vol. 87, 1992, pp. 7-16.

[2] L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone, *Classification and Regression Trees*. Wadsworth International Group, 1984.

[3] K. Chang, J. Ghosh, Principal curve classifer A nonlinear approach to pattern classification. *IEEE Int'l Joint Conf.* Neural Networks, pp. 695-700, 1998.

[4] T. Duchamp, W. Stuetzle, Extremal properties of principal curves in the plane. *Annals of Statistics*, Vol. 24, pp. 1511-1520, 1996.

[5] P. Dugard, J. Todman, H. Staines, *Data sets for approaching multivariante analysis: a pratical introduction*. Routledge, 2009.

[6] F. D. Ferreira, *Análise multivariada*. Lavras: UFLA, 2008.

[7] S. Gardner, N. J. Le Roux, Biplot methodology for discriminate analysis based upon robust methods and principal curves. *Proceedings of the 8th Conference of the International Federation of Classification Societies*, Springer-Verlag, 2002, pp. 169-176.

[8] J. F. Hair, R. E. Anderson, R. L. Taham, W. C. Black, W. C. *Análise multivariada de dados*. Bookman, 2005.

[9] T. Hastie, *Principal curves and Surfaces*. PHD Tesis: Stanford Linear Accelerator Center, 1984.

[10] T. Hastie, W. Stuetzle, Principal curves. *JASA Journal. American Statistic assoc.*, 84, 1989, pp. 502516.

[11] T. Hastie, R. Tibshirani, J. Friedman, *The elements of statistical learning*. Springer Science e Business Media LLC, 2009.

[12] W. W. Hsieh, *Machine learning methods in the environmental sciences*. Cambridge, UK: Cambridge University Press, 2009.

[13] A. J. Izenman, *Modern multivariate statistical techniques*. Springer Science e Business Media LLC, 2008.

[14] R. Johnson, A. D. Wichern, *Applied multivariate statistical analysis*. Upper Saddle River: Prentice Hall, 1998.

[15] B. Kgl, A. Krzyzak, T. Linder, Z. Kenneth, Learning and design of principal curves. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22(3), 2000, pp. 281-297.

[16] M. A. Kramer, Nonlinear principal components analysis using autoassociative Neural Networks. *AIChE Journal*, Vol. 37(2) , 1991, pp. 233 - 243.

[17] M. Last, T. Tassa, A. Zhmudyak, E. Shumeli, Improving Accuracy of classification models induced from anonymized datasets. *Journal Information Sciences: an International Journal*, Vol. 256, 2013, pp. 138-161.

[18] J. M. Lattin, J. D. Carroll, P. E. Green, *Análise de dados multivariados*. Cengage Learning, 2011.

[19] G. Licciardi, P. R. Marpu, J. Chanussot, J. A. Benediktsson, Linear versus nonlinear PCA for the classification of hyperspectral data based on the extended morphological profiles. *IEEE-Geoscience and Remote Sensing Letters*, Vol. 9(3), 2012.

[20] L. M. Santos, M. A. M. Ferreira, B. Tavares, D. R. Dutra, Classes estratégicas e desempenho no setor confeccionista brasileiro. *Gestão e Produção*, 2012.

[21] Tanagra Data Mining Tutorials. Acess in 01 de maio, 2014, http://data-mining-tutorials.blogspot.com.br/2012/11/linear-discriminant-analysis-tools.html.

[22] R. C. Torres, J. M. Seixas, W. Soares Filho, Classificação de sinais de sonar passivo utilizando componentes principais no lineares. *Revista da Sociedade Brasileira de Redes Neurais*, Vol. 2(2), 2004, pp. 60-72.

[23] UCI Machine Learning Repository. Universidade da California de Irvine. Acess in 01/05/2014, http://archive.ics.uci.edu/ml/index.html.

[24] J. J. Verbeek, N. Vlassis, B. Krse, A k-segments algorithm for finding principal curves. *Elsevier: Pattern Recognition Letters*, Vol. 23, 2002, pp. 10091017.

[25] A. Webb, *Statisitical pattern recognition*. John Wiley e Sons, 2002.

[26] L. P. Yunsong, Q. S. Huaijiang, Microarrays data classification basead on principal curves. *Seventh International Conference on Fuzzy Systems and Knowledge Discovery*, 2010, pp. 2199-2202.

[27] G. Minmin, L. Fan. Learning optimal kernel for pattern classification. *Wseas Transactions on Mathematics*, ed. 5, vol.12, pp. 2224-2880, 2013.

[28] N. A. Ramli, M. T. Ismail, H. C. Wooi. An analysis on two different data sets by using ensemble of k-nearest neighbor classifiers. *Wseas Transactions on Mathematics*, vol.13, pp. 2224-2880, 2014.

[29] J. Gao, L. Fan. Kernel-based discriminant analysis with QR decomposition and its application to face recognition. *Wseas Transactions on Mathematics*, ed. 10, vol.10, pp. 1109-2769, 2011.

[30] D. J. Hand. *Classifier technology and the illusion of progress*. Statistical Science, Vol. 21, pp. 1-15, 2006.

[31] T. S. Lim, Y. S. Shih. A Comparison of prediction accuracy, complexity, and training time of thirty-three. *Machine Learning (Kluwer Academic Publishers*, (Boston), vol. 40, pp. 203-229, 2000.