colloquium exactarum

FEATURE SPACE UNIDIMENSIONAL PROJECTIONS FOR SCATTERPLOTS

PROJEÇÕES DE ESPAÇOS DE CARACTERÍSTICAS UNIDIMENSIONAIS PARA GRÁFICOS DE DISPERSÃO

Danilo Medeiros Eler[1]; Alex C. de Almeida[2]; Jaqueline Batista Martins Teixeira[3]; Ives Renê Venturini Pola[4]; Fernanda Paula Barbosa Pola[5]; Mauricio Araujo Dias[6]; Celso Olivete Junior[7]

UNESP - Universidade Estadual Paulista ``Júlio de Mesquita Filho" Faculdade de Ciências e Tecnologia, Departamento da Matemática e Computação

E-mail: [1]daniloeler@fct.unesp.br; [2]alex1almeida@gmail.com; [3]jt.jaque@gmail.com; [4]ivesrene@gmail.com; [5]fernandapaulab@gmail.com; [6]madias@fct.unesp.br; [7]olivete@fct.unesp.br

ABSTRACT – Multidimensional projection techniques are important tools employed in data set exploration and data mining tasks. The data set instances are described in a multidimensional space and projection techniques can be employed to reduce the data set dimensionality and to aid the visualization of instances relations in a computer screen. Usually, the whole multidimensional space is projected, i.e., if it is composed by distinct feature spaces they are handled as a unique feature space. This work proposes an alternative approach dealing with multidimensional spaces as distinct feature spaces, so multidimensional projections can reduce the dimensionality of each feature space into unidimensional spaces and be visualized by a scatter plot -- each unidimensional space will be associated with an axis. Our approach was compared with the traditional way that projects the whole multidimensional space (feature spaces) into the bi-dimensional space. Experiments with different data sets were performed to evaluate which approach better preserves the groups cohesion on the projected space, revealing our approach with good results.
Keywords: multidimensional projection; scatter plot; dimensionality reduction.

RESUMO – Técnicas de projeção multidimensional são ferramentas importantes empregadas na exploração e mineração de conjuntos de dados. Instancias de conjuntos de dados são descritos em um espaço multidimensional e técnicas de projeção podem ser empregadas para reduzir a dimensionalidade dos dados e para auxiliar a visualizar o relacionamento entre instâncias na tela de um computador. Geralmente, todo o espaço multidimensional é projetado, i.e., se ele é composto por espaços de características diferentes eles são manipulados como um único espaço de característica. Este trabalho

propõe uma abordagem alternativa para lidar com espaços multidimensionais como espaços de características distintos, assim, as projeções multidimensionais podem reduzir a dimensionalidade de cada espaço de característica em espaços unidimensionais e visualiza-los com um gráfico de dispersão – cada espaço unidimensional será associado a um eixo. Nossa abordagem foi comparada com a maneira tradicional que projeta todo o espaço multidimensional (espaço de características) em um espaço bidimensional. Experimentos com diferentes conjuntos de dados foram executados para avaliar qual abordagem preserva melhor a coesão dos grupos no espaço projetado, revelando que nossa abordagem alcançou bons resultados.

Recebido em: 13/09/2016
Revisado em: 13/02/2017
Aprovado em: 25/04/2017

Palavras-chave: projeção multidimensional; gráfico de dispersão; redução de dimensionalidade.

1. INTRODUCTION

Multidimensional projection techniques are commonly used to generate scatter plots in 2D space from multidimensional space (TEJADA; MINGHIM; NONATO, 2003). Usually, they use all instance attributes to project a data set into a 2D space. Additionally, when distinct set of attributes (or feature spaces) are employed, they are combined and projected within a single step. For instance, when dealing with image data sets, feature spaces can be computed based on color, texture and shape. Commonly, in most approaches, the multidimensional projection considers the whole multidimensional space to project into 2D space.

In this work, the traditional approach was adapted that projects the whole multidimensional space. For that, we propose an approach that first uses a multidimensional projection technique to project distinct feature spaces (multidimensional space) into unidimensional spaces. Then, it computes a scatter plot with the projected unidimensional spaces. For instance, if two feature spaces are generated from an image data set, a scatter plot will be computed with these two feature spaces, in which each axis corresponds to a different feature space -- one feature space is projected in X axis and other in Y axes. Furthermore, the user can explore the scatter plot relating distinct feature spaces. In order to evaluate our approach, Silhouette Coefficient measure was employed to analyze the scatter plot that present better cohesion and separation among groups.

The remaining of this paper is organized as follows: in Section 2 is presented a conceptual basis of the multidimensional projection techniques employed in this work; in Section 3 is described the proposed approach; in Section 4 the experiments and results are presented; and finally, in Section 5 is presented the conclusion and future works.

2. MULTIDIMENSIONAL PROJECTION

Multidimensional projections have been used to aid the exploration of multidimensional data sets. For example, the dimensionality reduction eases the data processing and allows better data set visualization into the computer screen. Projection techniques transform data sets described in a $\Re^m$ space (with *m* attributes) into a $\Re^n$ space (m > n) (TEJADA; MINGHIM; NONATO, 2003).

The next sub-sections present the multidimensional projection techniques employed in this work. They were selected because they are precise and fast to compute. Also, the Least Square Projection (LSP) is indicated to deal with document

collections, which is one of the data domain used in the experiments.

2.1. Fastmap

The Fastmap technique is a fast multidimensional projection technique with linear time complexity (FALOUTSOS; LIN, 1995). It is based on traditional geometry theorems, such as Cosine Law, enabling to map instances and groups from a multidimensional space to a lower dimensional space (e.g., 2D or 3D).

The technique proceeds as follows. In the first step, two instances called as pivots are chosen, which are used to establish a line crossing the multidimensional space. These pivots must be as far as possible from each other. In order to improve the performance of the pivots choice, the authors proposed an heuristic process, which steps are described as follows:

- Choose randomly an instance O_x;
- Find the most far instance (O_a) from O_x;
- Find the most far instance (O_b) from O_a;
- Pick O_a e O_b as pivots.

After the pivots are chosen, it is necessary to compute distances from each instance O_i to the pivots O_a and O_b before projecting the remaining instances. This step is based on Cosine Laws (see Equation 1). As shown in Figure 1, the x_i indicates the coordinate of O_i instance in the first hyperplane.

$$x_i = \frac{d(O_a,O_i)^2 + d(O_a,O_b)^2 - d(O_b,O_i)^2}{2 * d(O_a,O_b)} \quad (1)$$

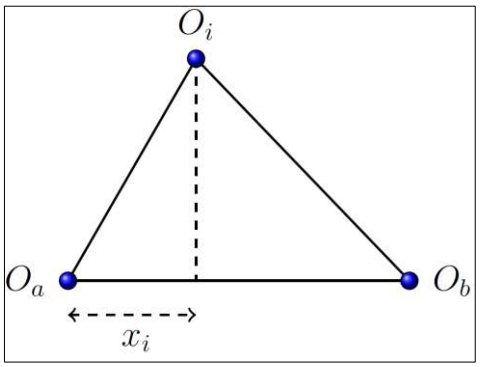


Figure 1. Cosine Laws: Projection on O_aO_b.

Other coordinates can be found by projecting the instances in a hyperplane *H* perpendicular to O_aO_b with dimension *m-1*, as shown in Figure 2. Thus, given two distances to O_i e O_j and its projections (O'_i e O'_j) in the hyperplane *H*, it is possible to obtain the distance function *d'* among the projections, as described in the Equation 2, by updating the distance matrix.

$$d'(O'_i, O'_j)^2 = d(O_i, O_j)^2 - (x_i - x_j)^2 \quad (2)$$

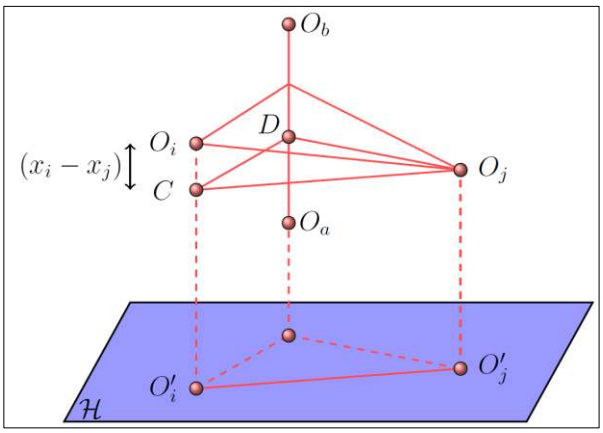


Figure 2. Projection onto hyperplane

Discovering this distance enables the projected instance in a second line on a

hyperplane *H* orthogonal to the first line. Such projection is performed by following the same described steps by getting the instances coordinates from the new dimension. This process is repeated until *p* dimensions are computed. In this paper the process is performed two times (*p=2*).

2.2. Least square projection (LSP)

The Least Square Projection (LSP) technique was proposed by Paulovich et. al. (2008) and employ characteristics from linear and nonlinear methods. Even though LSP uses linear systems, it is defined as a nonlinear technique due to the definition process of control points. LSP was developed to preserve the instances neighborhood relations from multidimensional space onto a lower dimensional space. According to the authors, the LSP time complexity is O(max<$n^{\frac{2}{3}}$, $n\sqrt{k}$>), taking $\sqrt{n}$ as the quantity of control points and the linear system solution with time complexity O($n\sqrt{k}$), considering *k* equals to the number of conditions of A^TA matrix.

The process has the following steps. The first step is to choose the control points to be initially projected in p-dimensional space, by :

1. Executing a cluster algorithms, for instance, the *k-medoids*;
2. Choosing the medoid of each cluster (Medoid is the nearest instance from the cluster centroid).

The control points are used as references to project other instances. They are projected using a fast and precise technique (e.g., fastmap) to get their coordinates in projected space. Following, based on a cluster technique, it is defined a neighborhood (V_i) to each data set instance -- an instance p_i is in the convex hull of V_i. Finally, a system of equations is computed by employing the Equation 3 from which $\tilde{p}_i$ coordinates can be extracted.

$$Lx_1 = 0, Lx_2 = 0, \ldots, Lx_p = 0 \quad (3)$$

Given x_i as the instances coordinates in the projected space, the values of the quadratic matrix *L* can be computed as presented in Equation 4.

$$l_{ij} = \begin{cases} 1, i = j \\ -\alpha_{ij}, p_j \in V_i \\ 0, \text{otherwise} \end{cases} \quad (4)$$

where $\alpha_{ij} = \frac{1}{k_i}$, p_i is the centroid of V_i and *L* matrix is called as Laplacian matrix. The α_{ij} value must be in range [0,1] and $\sum \alpha_{ij} = 1$.

It is necessary to include control points because they have already been projected and give a reference of the projected space coordinates. Thus, the Equation 3 can be rewritten as *Ax = b*. The *A* matrix is composed by two other matrices *L* and *C*, as presented in Equation 5.

$$A = \begin{pmatrix} L \\ C \end{pmatrix} \quad (5)$$

The matrix fields c_{ij} of C are determined by the conditions of Equation 6.

$$c_{ij} = \begin{cases} 1, \text{ if } p_j \text{ is a control point} \\ 0, \text{ otherwise} \end{cases} \quad (6)$$

The vector fields b_i of b is given by the Equation 7, in which $x_{p_{i_c}}$ refers to the control points coordinates in the projected space.

$$b_i = \begin{cases} 0, i \leq n \\ x_{p_{i_c}}, n < i \leq n + nc \end{cases} \quad (7)$$

The least squares approach is employed to solve the resulting linear system, from which is chosen x that minimizes $||Ax-b||^2$. The linear system solution gives the coordinates on the projected space.

The next section describes the proposed approach and discusses how the comparison to traditional approach will be made.

3. THE PROPOSED APPROACH

Usually, feature spaces are acquired from data sets to generate a multidimensional space composed by distinct feature spaces.

In a traditional approach, these feature spaces are aggregated in a unique multidimensional space, which is handled by a multidimensional projection technique, as shown in Figure 3 -- the aggregated feature space is projected wholly.

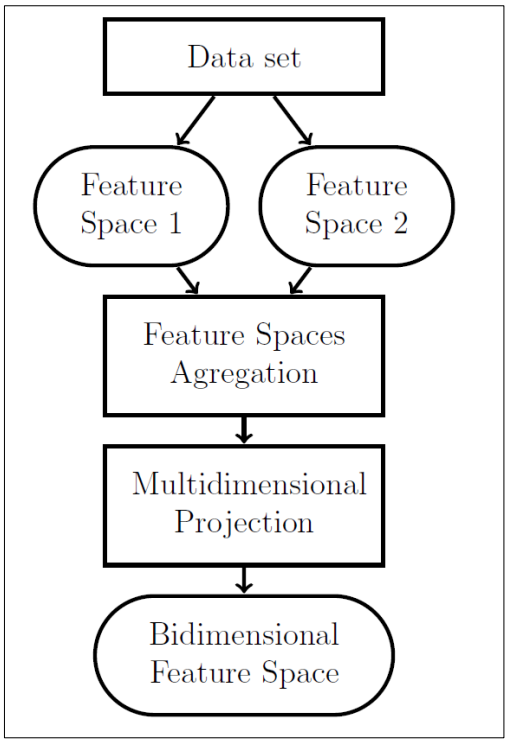


Figure 3. Main steps from traditional approach.

In contrast to traditional approach, this paper proposes a new approach to project each of distinct feature spaces in a unidimensional space, as shown in Figure 4. Following, each unidimensional space is used as a scatter plot axis -- unidimensional feature space 1 as X axis and unidimensional feature space 2 as Y axis.

Figure 5(a) presents a scatter plot generated from a 2D projection computed from a multidimensional space composed by two distinct feature spaces of features acquired from Brodatz image collection (see Section 4 for more details).

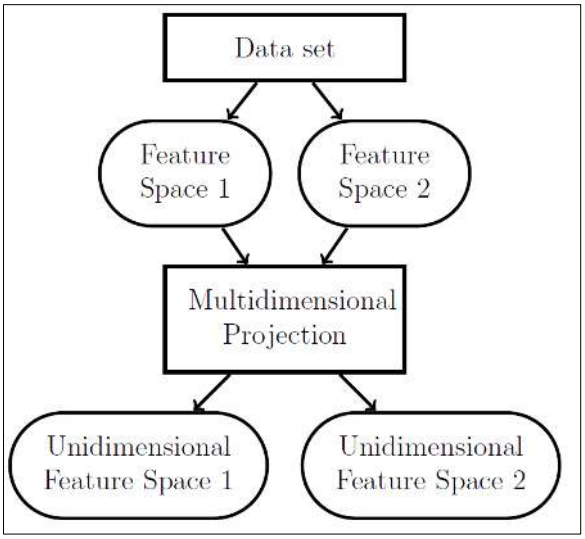


Figure 4. Main steps of proposed approach -- a 2D projection was computed from two aggregated feature spaces.

Figure 5(b) shows a scatter plot generated from two unidimensional projections -- each one was computed from a distinct feature space generated from the Brodatz image collection. In the figure, each circle color indicates the instance label (class).

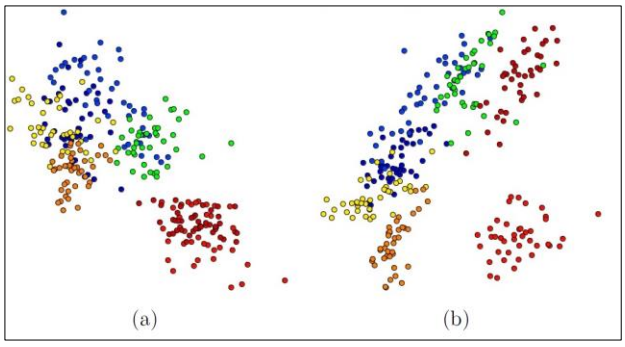


Figure 5. Scatter plots generated from both traditional (a) and proposed (b) approaches for the Brodatz image dataset.

In order to evaluate and compare both approaches, the Silhouette Coefficient can be used to measure the cohesion and separation among instances from clusters (groups) generated by the projections. Given an instance d_i, its cohesion a_i is computed as the mean of distances among d_i and the other instances belonging to d_i's cluster. The separation b_i is the minimum distance computed among d_i and all other instances belonging to other clusters. The Silhouette Coefficient from a projection is given by the mean computed from the coefficients of all instances, as presented in Equation (8), and is valued in the interval [-1,1]. Higher values indicate a better cohesion and separation among the clusters generated by the multidimensional projection technique.

$$S = \frac{1}{n}\sum_{i=1}^{n}\frac{(b_i - a_i)}{max(a_i, b_i)} \qquad (8)$$

The next section presents the experiments and results of employing both approaches to compute scatter plots from data sets.

4. EXPERIMENTS

This section details the performed experiments and shows the results for the proposed approach. Four real data sets were used in experiments, detailed as follows:

- **BRODATZ**: Composed by 280 texture images divided in seven classes. This data set was adapted from Brodatz data set \cite{Brodatz};
- **FIBERS**: Composed by 19000 brain fibers tracks divided in eight classes. This data set was obtained from the 2009 Pittsburgh Brain Competition (PBC) -- Brain Connectivity Challenge (*http://pbc.lrdc.pitt.edu*);
- **CBR-ILP-IR**: Composed by 574 scientific papers divided in three classes from Artificial Intelligence -- Case-Based Reasoning (CBR), Inductive Logic Programming (ILP) and Information Retrieval (IR);
- **NEWS-10**: Composed by 1000 messages from news groups divided in ten classes.

In the experiments, we used two feature spaces computed for each data set. In our proposed approach, each feature space was individually projected, and in the traditional approach the feature spaces were aggregated as a unique feature space.

For the BRODATZ data set, two feature spaces were computed using co-occurrence matrices from Haralick and Gabor filters (BIANCONI; FERNANDÉZ, 2007). The feature spaces were computed as described by Brandoli et al. (2010).

For FIBERS data set, two feature spaces were computed using spatial and curvature features as described by Poco et al. (2012).

For CBR-ILP-IR data set, two feature spaces were computed using terms of high and low frequencies from the document collection -- low frequencies in range [100,1000]; and high frequencies in range [1001,2745].

For the NEWS-10 data set, two feature spaces were computed using high and low frequencies, respectively, term frequencies in range [10,49] and [50,166].

After computing all feature spaces, the approaches were employed to generate the scatter plots. The traditional approach computes 2D projections from the whole multidimensional space composed by aggregation of two distinct feature spaces; and the proposed approach generates scatter plots by projecting distinct feature spaces into a unidimensional space -- each one for each scatter plot axis. Following, Silhouette Coefficients were computed for each resulting scatter plot to evaluate the best approach. The obtained results are shown in Figure 6. LSP results are not reported because this technique could not deal with more than 10,000 instance.

From the results it can be seen that our approach achieved better results in the last two data sets, while the traditional

approach achieved better results for other two data sets.

Additionally, note that the Fastmap technique did not achieve good results when projecting document collection features spaces in 2D space, but our approach improved greatly the Fastmap results.

	Traditional	Proposed	Traditional	Proposed
Data sets	*Fastmap 2D*	*Fastmap 1D1D*	LSP 2D	*LSP 1D1D*
BRODATZ	0.2753	0.1180	0.4679	0.1431
FIBERS	0.5260	0.3964	–	–
CBR-ILP-IR	−0.0269	0.3366	0.3491	0.4442
NEWS-10	−0.0715	0.0630	0.2686	0.4096

Figure 6. Silhouette Coefficients from the scatter plots generated for traditional and proposed approach.

Furthermore, we compute Neighborhood Hit (The Neighborhood Hit computes the percentage of nearest neighbors belonging to the same class of certain instance in the projected space.) plots for those data sets to confirm our evaluation. As shown in Figure 7, the traditional approach reaches better results than the proposed approach. However, as shown in Figure 8, the proposed approach achieved a better precision for document collections.

4. CONCLUSIONS

Multidimensional projection techniques are valuable tools for dealing with multidimensional spaces, aiding in exploration and better understanding of data sets instances relationship. Traditionally, techniques are employed to project a unique multidimensional space in a lower dimensional space (e.g., 2D space). In this work we proposed a new approach for dealing with multidimensional spaces to generate scatter plots from multidimensional projection techniques. Our approach generates 2D scatter plots from two unidimensional spaces, which one computed from a distinct feature space. In order to evaluate our approach, the silhouette coefficient measure was used to evaluate the scatter plot quality regarding the clusters cohesion and separation.

The experiments showed that our approach reached better results depending of the data set used, but greatly on document collections, from which two feature spaces were computed using high and low term frequencies. However, better results for 2D projections for CBR-ILP-IR data set is presented in Eler et al. (2013) -- Silhouette Coefficient equals to 0.66 -- in

which is described an approach to find a threshold to eliminate low frequencies from multidimensional spaces computed from document collections.

This work presented an initial approach of generating scatterplots, even though we reached good results for specific datasets, there is no indication about which approach is better for a unknown dataset. Thus, it is necessary to execute both approaches to decide which is the better. With more experiments and datasets, we will try to analyze the dataset features behavior to stablish a correlation with the best approach.

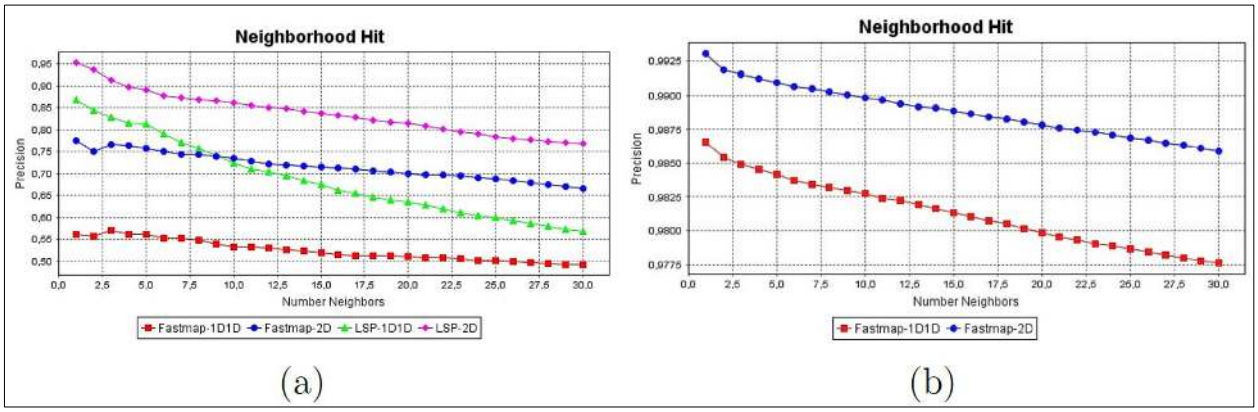


Figure 7. Neighborhood Hit plot for BRODATZ (a) and FIBERS (b) data sets.

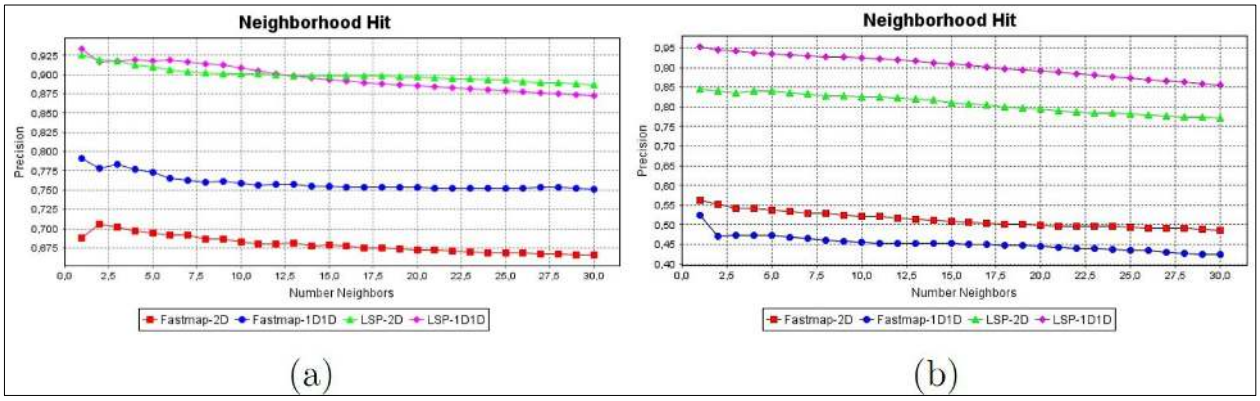


Figure 8. Neighborhood Hit plot for CBR-ILP-IR and NEWS-10 data sets.

In this paper, we could use the indicated threshold to compute the multidimensional space, but we firstly need to establish a method to divide the multidimensional space in two distinct spaces. Therefore, in further works, we will use the thresholds employed in Eler and Garcia

(2013) and propose a way to divide multidimensional spaces computed from

document collections to perform new evaluations with our proposed approach. Furthermore, in future works, more multidimensional projection and evaluation techniques could also be employed to verify the best projection technique to be employed in our approach, confronting with other types of data sets.

All experiments were based on at least two groups of features spaces or they were generated based on the original space. In future works, we intent to deal with multidimensional spaces composed by more than two groups of features, in this case, a new approach to separate the feature spaces have to be developed and evaluated.

REFERENCES

BIANCONI, F.; FERNANDÉZ, A. Evaluation of the effects of gabor filter parameters on texture classification. **Pattern Recognition**, v. 40, n. 12, p. 3325–3335, 2007. https://doi.org/10.1016/j.patcog.2007.04.023

BRANDOLI, B. et al. Visual data exploration to feature space definition. In: THE 2010 SIBGRAPI CONFERENCE ON GRAPHICS, PATTERNS AND IMAGES, 23. SIBGRAPI '10, IEEE COMPUTER SOCIETY. **Proceedings…** Washington, DC, USA, 2010. p.32-39. ttps://doi.org/10.1109/SIBGRAPI.2010.13

BRODATZ, P. **Textures:** a photographic album for artists and designers. New York: Dover, 1966.

ELER, D. M.; GARCIA, R. E., **Using Otsu's Threshold Selection Method for Eliminating Terms in Vector Space Model Computation**, 17th International Conference Information Visualisation, p. 220--226, 2013. https://doi.org/10.1109/IV.2013.29

FALOUTSOS, C.; LIN, K. Fastmap: A fast algorithm for indexing, data mining and visualization of traditional and multimedia databases. In: the International Conference on Management of Data (SIGMOD ’95), San Jose-CA, USA. **Proceedings…** New York: ACM Press, 1995. p. 163–174. https://doi.org/10.1145/223784.223812

PAULOVICH, F.V. et al. Least square projection: a fast high precision multidimensional projection technique and its application to document mapping. **IEEE Transactions on Visualization and Computer Graphics**, v. 14, n. 3, p. 564–575, 2008.

POCO, J. et al. Employing 2D projections for fast visual exploration of large fiber tracking data. **Comp. Graph. Forum**, v. 31, n.3, pt.2, p. 1075-1084, jun. 2012.

TEJADA, E.; MINGHIM, R.; NONATO, L. G. **On improved projection techniques to support visual exploration of multidimensional data sets**. Information Visualization, v. 2, n. 4, p. 218–231, 2003. https://doi.org/10.1057/palgrave.ivs.9500054