GRAPH-BASED FEATURE REPRESENTATION FOR MULTI-CLASS CLASSIFICATIONS USING THE JEFFRIES-MATUSITA DISTANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Visualization of the feature space has gained attention in recent years, especially with the growing importance of explainability in AI. When processing high-dimensional datasets, a common pre-processing step is feature selection. Filter-based feature selection algorithms are not tailored to a specific classification method, but rather rank the relevance of each feature with respect to the target and the task. The Jeffries-Matusita (JM) distance is a measure for separability between two distributions that has been used for filter based feature selection. It calculates for each feature the separability strength between pairs of classes. In this work, a low-dimensional representation of the JM measures for the feature space is formed by diffusion map. Diffusion maps organizes the features by their class separation abilities. Moreover, feature elimination is performed based on the distribution of the points in the low-dimensional space. Experimental results are provided for 3 public datasets and compared with known filter-based feature selection techniques.

1 INTRODUCTION

Geometric based representation and manifold learning techniques have been shown to be a valuable tool for modeling and visualization of complex data (Meyer & Barth, 2016; Moon & Krishnaswamy., 2017; Wang & Lin, 2020). Commonly, manifolds learning techniques are applied to the data of the sample space for revealing its underlying latent factors (Singer & Coifman, 2009; Comeau & Majda, 2017). Nevertheless, graph-based approached for feature visualization and selection have also been proposed and applied in differed domains (Das & Chakraborty, 2017; Roffo & Vinciarelli, 2017; Hashemi & Nezamabadi-pour, 2020). Two notable domains of interest, which typically process datasets of high-dimension for supervised multi-class learning problems, are remote sensing (Georganos & Wolff, 2018; Taşkin & Bruzzone, 2017) and gene expression data (Dutta & Gulati, 2019; Swarnkar & Mitra, 2015).

In this work, we focus on filter feature selection methods using graph-based techniques. In filter methods the selection of features is independent of the classifier used. Other known feature selection techniques are wrapper methods that perform an iterative search for finding the best subset of features (Jović & Bogunović, 2015) and embedded methods, in which the feature selection procedure is an integrated into the learning algorithm (Monteiro & Murphy). However wrapper methods are less suitable for processing high-dimensional datasets because of their high computational complexity (Ghosh & Abraham, 2019).

The Jeffries-Matusita (JM) distance, which improves the Bhattacharya distance (Guorong & Minhui, 1996), is a measure for separability between two distributions. It has been successfully tested for feature selection in remote sensing and hyperspectral datasets (Wei & Gao, 2017; Jung & Ehlers, 2016). Given a multi-class problem, the JM based feature separation is coded by symmetric matrix that holds the class separability strength of each feature. The separation coefficients between two classes take values between 0 and 2, where 2 stands for two classes that are highly separated by the inspected feature. One way to choose the best features is to calculate the average value of the JM matrix for each feature, and leave the features with the highest averaged separation scores (Khosravi & Homayouni, 2018). However, this elimination approach may ignore some features that do not have the highest averaged separation but contribute to specific class-separation pairs. In order to study the JM based feature space, a manifold learning algorithm, diffusion maps (Coifman & Lafon, 2006), is utilized.

Diffusion maps is well known manifold learning technique. It starts by representing the data points as a weighted graph. The spectral decomposition of the graph is used to embed the data into a low-dimensional space. The associate diffusion distance metric ensures that the geometric structure of the data is kept in the low-dimensional space, where Euclidean distances correspond to the diffusion distance in the original space. Application of diffusion maps to the

feature space, where each feature is represented by a JM distance matrix, provides a visual description of the feature distribution with respect to class separability. Furthermore, feature with similar JM distance matrices lie close to one another in the low-dimensional space, thus, feature elimination can be determined based on Euclidean distance in the compact representation.

Graph-based feature visualization and selection is proposed by combining the JM distance and diffusion maps. This follows ideas from (Zeng & Wu, 2018) and (Khosravi & Homayouni, 2018) that propose improved versions of the JM feature selection techniques by considering both separability and redundancy between the features. In (Zeng & Wu, 2018), this redundancy is calculated using the Pearson correlation. In (Khosravi & Homayouni, 2018) the features come from two different modalities, radar and optic instruments. Features from the first modality that have an averaged JM value that is larger than a pre-defined threshold are kept. Then, features from the second modality that have a large correlation coefficient with the features selects from the first modality, are eliminated. Here, we propose to eliminate features that are represented by JM distance matrices based on their distribution in the low-dimensional space. Results are demonstrated on three high-dimensional datasets and are compared with known filter methods for feature selection, the fisher score (Duda & Stork, 2001), RelifF (Robnik-Sikonja & Kononenko, 2003) and Correlation-based feature selection (CFS , Hall (1999)). We show that the proposed geometric-based elimination often results in higher classification accuracy than the other techniques, while also providing a visual representation of the feature space.

2 METHODS

This section reviews the Jeffries-Matusita (JM) distance measure and diffusion maps. Denote the learned dataset by $X = \{x_1, \ldots, x_N\}$, where N is the number of samples, and $x_i \in \mathbf{R}^D$, where D is the dimension of the feature space.

2.1 JEFFRIES-MATUSITA DISTANCE FOR FEATURE SELECTION

Given a feature $x_i \in X$ and a set of classes that are associated with the classification task, $c = 1, 2, \ldots, C$, the JM distance computations result a in matrix of size $C \times C$, in which the separability between two classes c and $\tilde{c}$ is defined by

$$JM_i(c, \tilde{c}) = 2\left(1 - e^{-B_i(c,\tilde{c})}\right), \tag{1}$$

where

$$B_i(c, \tilde{c}) = \frac{1}{8}\left(\mu_{c,i} - \mu_{\tilde{c},i}\right)^2 \frac{2}{\sigma_{c,i}^2 + \sigma_{\tilde{c},i}^2} + \frac{1}{2}\ln\left(\frac{\sigma_{c,i}^2 + \sigma_{\tilde{c},i}^2}{2\sigma_{c,i}\sigma_{\tilde{c},i}}\right) \tag{2}$$

is the Bhattacharyya distance. The values $\mu_{c,i}, \mu_{\tilde{c},i}$ and $\sigma_{c,i}, \sigma_{\tilde{c},i}$ are the mean and variance values of two given classes c and $\tilde{c}$ from the feature x_i.

For the propose of this work, each feature $x_i \in \mathbf{R}^D$ is replaced by its associated JM matrix of size $C \times C$, denoted by JM_i. Thus the input set for the diffusion maps algorithm may be denoted by $\mathbf{X} = \{JM_1, \ldots, JM_N\}$. The size of $\mathbf{X}$ is $N \times C^2$.

2.2 DIFFUSION MAPS

Diffusion maps (DM) allows to model high-dimensional that lie on a non-linear manifold. Here, the input data is comprised of symmetric matrices, that are flattened into vectors of size $1 \times C^2$. This set is denoted by $\mathbf{X}$, and it is the input for the DM algorithm.

A graph $G = (\mathbf{X}, \mathbf{W})$ is constructed from $\mathbf{X}$, where the points in $\mathbf{X}$ are the vertices of G and a kernel matrix $\mathbf{W} \triangleq \mathbf{w}\left(\mathbf{JM_i}, \mathbf{JM_j}\right)$, of size $N \times N$, holds the graph's weighted edges. $\mathbf{W}$ should satisfy the following properties. Symmetric, positive-preserving, and positive semi-definite (see Coifman & Lafon (2006) for details).

The Gaussian kernel

$$\mathbf{W} = \mathbf{w}\left(\mathbf{JM_i}, \mathbf{JM_j}\right) = \mathbf{e}^{\frac{-\|\mathbf{JM_i} - \mathbf{JM_j}\|^2}{2\epsilon}} \tag{3}$$

is a common choice for the weight matrix of G. The scale of the kernel ϵ defines the local neighborhood around each data point in the original space. The value of ϵ should be adapted to the density distribution of the data.

By introducing a scale parameter α, the effect of the non-uniformed data distribution may be controlled. A general normalized form of the kernel is given by

$$\mathbf{W}_\alpha = \mathbf{w}_\alpha(\mathbf{JM_i}, \mathbf{JM_j}) = \frac{\mathbf{w}(\mathbf{JM_i}, \mathbf{JM_j})}{\mathbf{q}^\alpha(\mathbf{JM_i})\mathbf{q}^\alpha(\mathbf{JM_j})}, \quad \mathbf{q}\left(\mathbf{JM_i}\right) = \sum_{\mathbf{JM_j} \in \mathbf{X}} \mathbf{w}\left(\mathbf{JM_i}, \mathbf{JM_j}\right). \tag{4}$$

Here, we set $\alpha = 1$, which results in an approximation of $\mathbf{K}$ to the Laplace-Beltrami operator (Coifman & Lafon, 2006), and it allows to recover the geometry of the data points, regardless of their distribution. A second normalization is performed for generating a Markov transition matrix $\mathbf{K}$ from W^α. This results in

$$\mathbf{K} = \mathbf{D^{-1}}\mathbf{W}_\alpha, \quad \mathbf{D} = \mathbf{d}(\mathbf{JM_i}, \mathbf{JM_i}) = \sum_{\mathbf{JM_j} \in \mathbf{X}} \mathbf{w}_\alpha\left(\mathbf{JM_i}, \mathbf{JM_j}\right). \tag{5}$$

Computing the spectral decomposition of $\mathbf{K}$ results in embedding coordinates for the dataset $\mathbf{X}$. Denote the eigenvalues of $\mathbf{K}$ by $\{\lambda_l\}_{l=0}^{N-1}$ and the left and right eigenvectors by $\{\phi_l\}_{l=0}^{N-1}$ and $\{\psi_l\}_{l=0}^{N-1}$, respectively. Although $\mathbf{K}$ is not a symmetric matrix, it is conjugate to a symmetric matrix, thus the two sets of eigenvectors $\{\phi_l\}_{l=0}^{N-1}$ and $\{\psi_l\}_{l=0}^{N-1}$, are biorthonormal $\langle \phi_m, \psi_l \rangle = \delta_{l,m}$. Therefor, each element in $\mathbf{K}$ can be computed by

$$k(JM_i, JM_j) = \sum_{l=0}^{N-1} \lambda_l \psi_l(JM_i)\phi_l(JM_j). \tag{6}$$

The matrix $\mathbf{K}$ has a decaying spectrum, $\lambda_l \to 0$ as l grows. This allows to approximate the entries of $\mathbf{K}$ in equation 6 by considering a small number of terms d in the sum. Finally, the diffusion maps coordinates are defined by

$$\mathbf{\Psi}(\mathbf{JM_i}) = (\lambda_\mathbf{1}\psi_\mathbf{1}(\mathbf{JM_i}), \lambda_\mathbf{2}\psi_\mathbf{2}(\mathbf{JM_i}), \lambda_\mathbf{3}\psi_\mathbf{3}(\mathbf{JM_i}), \cdots). \tag{7}$$

The first $d \leq C^2$ diffusion maps coordinates are considered, thus the space is reduced.

2.3 Feature clustering and eliminating in the embedded space

The diffusion maps coordinates provide a reduced and compact space that organizes the features by their underlying class separability properties. When using the Gaussian kernel (see equation 3), features with high separability values will be close to one another. Thus, a simple k-means clustering algorithm may be applied in the reduced space and create clusters of features with similar averaged JM separability. Clusters that include features with low averaged separability may then be eliminated.

Another elimination approach is based on the density distribution of the JM matrices in the diffusion maps space. Two points that lie close to each other in the embedded space correspond to two features with similar JM matrices. By setting a scale parameter $\bar{\varepsilon}$ and a multiplication factor a, one of the two points may be eliminated. In particular, we start with the first embedded point $\Psi(JM_1)$ and eliminate all other points that satisfy

$$\|\Psi(JM_1) - \Psi(JM_j)\| \leq a \cdot \bar{\varepsilon}, \qquad \text{for all} \qquad \{\Psi(JM_j)\}_{j=1}^{N}. \tag{8}$$

The distances are the Euclidean distances in the low-dimensional space. The value for $\bar{\varepsilon}$ is set as the average of the minimal distances between each point and its closest neighbor in the embedding space. Then, the elimination process continues by setting the next point $\Psi(JM_2)$ in equation 8 (assuming that it was not eliminated in the previous step). The process continues by scanning all of the embedded features.

3 Datasets

Three public datasets (UCI) belonging to multi-class classification tasks are considered.

1. **Isolet:** The dataset holds 617 features extracted from voice recordings of 150 subjects, who spoke out each of the English alphabets. The task is to classify the correct letter, hence there are 26 classes.
2. **Crop:** The data holds bi-temporal optical-radar data for cropland agricultural classification. The full dataset includes 174 features and seven crop type classes.
3. **Obesity:** The data is used for the estimation of obesity levels in individuals from the countries of Mexico, Peru and Colombia, based on their eating habits and physical condition. It includes 16 features and seven obesity type classes.

4 RESULTS

The datasets were first split into train and test sets. In Obesity and Isolet the train-test split was 70% - 30%. In the Crop dataset, 3361 samples were randomly chosen for training and 1441 for test. Classification on the test datasets was performed by application of multi-class SVM, KNN and random forests. Experiments were repeated three times. The features from the train sets of the three datasets were first replaced by their associated JM matrices, as explained in Section 2.1. This resulted in a dataset denoted by $\mathbf{X}$. Then, diffusion maps was applied to $\mathbf{X}$ and feature elimination was performed. The eliminated feature set from the train data was fed into the classifiers, and the class of the test samples was predicted based on these features.

Table 1 presents the results for the Isolet dataset. The first row considers all of the features. The second and third rows, named *high-cluster* and *high + medium cluster* consider features that were selected after application of k-means, with $k = 3$ in the low-dimensional space. Each of the resulted clusters was associated with the mean average JM separation strength of its embedded points. Next, points were eliminated by application of equation 8, once with $a = 1$ and once with $a = 2$ in the 2-dimensional space. The last rows plot the classification results of fisher score and reliefF features selections, once by keeping 50% of the features and once by keeping 30% of the features. The last row shows the classification results after application of CFS. The columns hold the averaged accuracy from 3 test runs for each type of classifier. It can be seen that DM based eliminate with $a = 2$ mostly achieved higher accuracy than the fisher score and reliefF with 30%, these all hold approximately the same number of features. In addition, DM based eliminate with $a = 2$ performs better than reliefF with 50% of the features and it is quite comparable to the fisher score with 50% of the features. Note the in these two cases the number of features for DM eliminate with $a = 2$ is much smaller than fisher and reliefF. Visualization of the feature space is plotted in Figure 1 in the Appendix.

Table 1: Classification results for the Isolet dataset

Method	**Num. of features**	**SVM**	**KNN**	**Random Forests**
All features	617	0.962	0.87	0.899814736
High cluster (DM)	103	0.656	0.588	0.601
High+Med. cluster (DM)	235	0.896	0.813	0.822
Eliminate (DM, a=1)	367	0.963	0.89	0.895
Eliminate (DM, a=2)	190	0.952	0.891	0.875
Fisher-score (50%)	309	0.955	0.903	0.886
Fisher-score (30%)	185	0.933	0.884	0.877
reliefF (50%)	309	0.949	0.86	0.856
reliefF (30%)	185	0.936	0.837	0.826
CFS	27	0.688	0.638	0.692

The results and low-dimensional plots for the Crop and Obesity dataset are detailed in the Appendix. In both of these cases the DM elimination approach resulted in high accuracy rates and a relatively small number of features.

5 CONCLUSIONS

In multi-class classification tasks, the JM index provides an informative measure about the class separation strength of a feature. However, it is not intuitive to analyze the $C \times C$ matrices that are calculated, especially when the number of features and C are large. Embedding of the JM matrices by diffusion maps in a low-dimensional space provides a visualization of the feature space. In addition, it allows to perform a geometric-based elimination approach, which results in competitive classification performance when compared to known feature selection techniques. In future work we plan to test different kernel types and to improve the elimination procedure by choosing the best embedding dimension.

REFERENCES

Ronald R. Coifman and Stéphane Lafon. Diffusion maps. *Applied and computational harmonic analysis*, 21(1):5–30, 2006.

Zhizhen Zhao Dimitrios Giannakis Comeau, Darin and Andrew J. Majda. Data-driven prediction strategies for low-frequency patterns of north pacific climate variability. *Climate Dynamics*, 48(5-6):1855–1872, 2017.

Saptarsi Goswami Amlan Chakrabarti Das, Amit Kumar and Basabi Chakraborty. A new hybrid feature selection approach using feature association map for supervised and unsupervised classification. *Expert Systems with Applications*, 88:81–94, 2017.

Peter E. Hart Duda, Richard O. and David G. Stork. *Pattern Classification*. Wiley Interscience, 2001.

Sriparna Saha Dutta, Pratik and Saurabh Gulati. Graph-based hub gene selection technique using protein interaction information: Application to sample classification. *IEEE journal of biomedical and health informatics*, 23(6):2670–2676, 2019.

Tais Grippa Sabine Vanhuysse Moritz Lennert Michal Shimoni-Stamatis Kalogirou Georganos, Stefanos and Eleonore Wolff. Less is more: Optimizing classification performance through feature selection in a very-high-resolution remote sensing object-based urban application. *GIScience & remote sensing*, 55:221–242, 2018.

Ritam Guha Ram Sarkar Ghosh, Manosij and Ajith Abraham. A wrapper-filter feature selection technique based on ant colony optimization. *Neural Computing and Applications*, 32:1–19, 2019.

Chai Peiqi Guorong, Xuan and Wu Minhui. Bhattacharyya distance feature selection. In *In Proceedings of 13th International Conference on Pattern Recognition*, volume 2, pp. 195–199. MIT Press, 1996.

Mark Andrew Hall. *Correlation-based feature selection for machine learning*. 1999.

Mohammad Bagher Dowlatshahi Hashemi, Amin and Hossein Nezamabadi-pour. Mgfs: A multi-label graph-based feature selection algorithm via pagerank centrality. *Expert Systems with Applications*, 142:113024, 2020.

Karla Brkić Jović, Alan and Nikola Bogunović. A review of feature selection methods with applications. In *In 2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO)*, pp. 1200–1205. IEEE, 2015.

Richard Jung and Manfred Ehlers. Comparison of two feature selection methods for the separability analysis of intertidal sediments with spectrometric datasets in the german wadden sea. *International journal of applied earth observation and geoinformation*, 52:1527–1554, 2016.

Abdolreza Safari Khosravi, Iman and Saeid Homayouni. Msmd: maximum separability and minimum dependency feature selection for cropland classification from optical and radar data. *International Journal of Remote Sensing*, 39(8):2159–2176, 2018.

Alexander M. Benison Zachariah Smith Meyer, François G. and Daniel S. Barth. Decoding epileptogenesis in a reduced state space. In *In 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 152–157. IEEE, 2016.

Sildomar T. Monteiro and Richard J. Murphy. Embedded feature selection of hyperspectral bands with boosted decision trees. In *In 2011 IEEE International Geoscience and Remote Sensing Symposium*, pp. 2361–2361. IEEE.

David van Dijk Zheng Wang William Chen-Matthew J. Hirn-Ronald R. Coifman Natalia B. Ivanova Guy Wolf Moon, Kevin R. and Smita Krishnaswamy. Phate: a dimensionality reduction method for visualizing trajectory structures in high-dimensional biological data. In *BioRxiv, 120378*. 2017.

M. Robnik-Sikonja and I. Kononenko. Theoretical and empirical analysis of relieff and rrelieff. *Machine Learning*, 51(1-2):23–69, 2003.

Simone Melzi Umberto Castellani Roffo, Giorgio and Alessandro Vinciarelli. Infinite latent feature selection: A probabilistic latent graph-based ranking approach. In *In Proceedings of the IEEE International Conference on Computer Vision*, pp. 1398–1406. IEEE, 2017.

Radek Erban Ioannis G. Kevrekidis Singer, Amit and Ronald R. Coifman. Detecting intrinsic slow variables in stochastic dynamical systems by anisotropic diffusion maps. *Proceedings of the National Academy of Sciences*, 106(38): 16090–16095, 2009.

Tripti Swarnkar and Pabitra Mitra. Graph-based unsupervised feature selection and multiview clustering for microarray data. *Journal of biosciences*, 40(4):755–767, 2015.

Hüseyin Kaya Taşkin, Gülşen and Lorenzo Bruzzone. Feature selection based on high dimensional model representation for hyperspectral images. *IEEE Transactions on Image Processing*, 26(6):2918–2928, 2017.

UCI. In *https://archive.ics.uci.edu/ml/datasets.php*.

Hau-Tieng Wu Po-Hsun Huang Cheng-Hsi Chang Chien-Kun Ting Wang, Shen-Chih and Yu-Ting Lin. Novel imaging revealing inner dynamics for cardiovascular waveform analysis via unsupervised manifold learning. *Anesthesia & Analgesia*, 130(5):1244–1254, 2020.

Honglin He-Haijian Hao Wei, Zhanyu and Wei Gao. Automated mapping of landforms through the application of supervised classification to lidar-derived dems and the identification of earthquake ruptures. *International Journal of Remote Sensing*, 38(23):7196–7219, 2017.

Hui Lin-Enping Yan Qian Jiang Hongwang Lu Zeng, Wen and Simin Wu. Optimal selection of remote sensing feature variables for land cover classification. In *In 2018 Fifth International Workshop on Earth Observation and Remote Sensing Applications (EORSA)*, pp. 1–5. IEEE, 2018.

A Appendix

Figure 1 plots the two-dimensional JM feature representation from the Isolet dataset. The points are colored by their average JM values. It can be seen that the first diffusion coordinate captures this property, however, there are other factor that differentiate between the JM matrices, which are captured in the second diffusion coordinate.

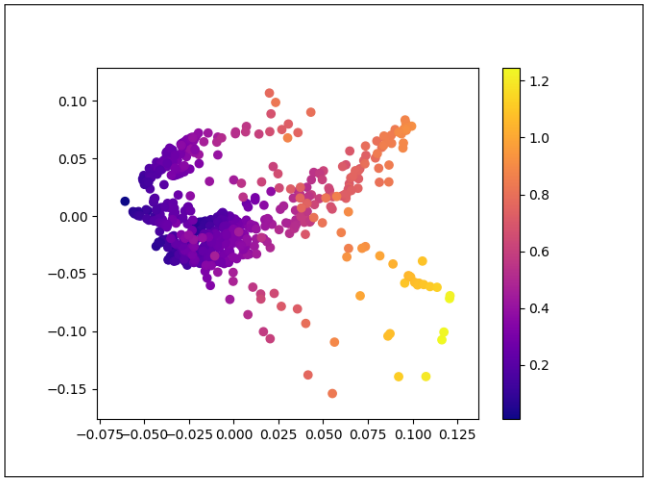

Figure 1: Embedding of the feature space of the Isolet dataset by the first two diffusion maps coordinates.

The 174 computed and embedded JM features from the Crop dataset are presented in Figure 2 (left). The features of the Obesity dataset, represented by the first two diffusion maps coordinates are presented in Figure 2 (right). In both sub-figures the x-axis corresponds to $\lambda_1\psi_1$ and the y-axis to $\lambda_2\psi_2$.

Table 2 plots the classification results for the Crop dataset. Like in the Isolet example, it can be seen that DM-eliminate with $a = 2$ leaves approximately a third of the original features, performs equal or better than fisher and reliefF with

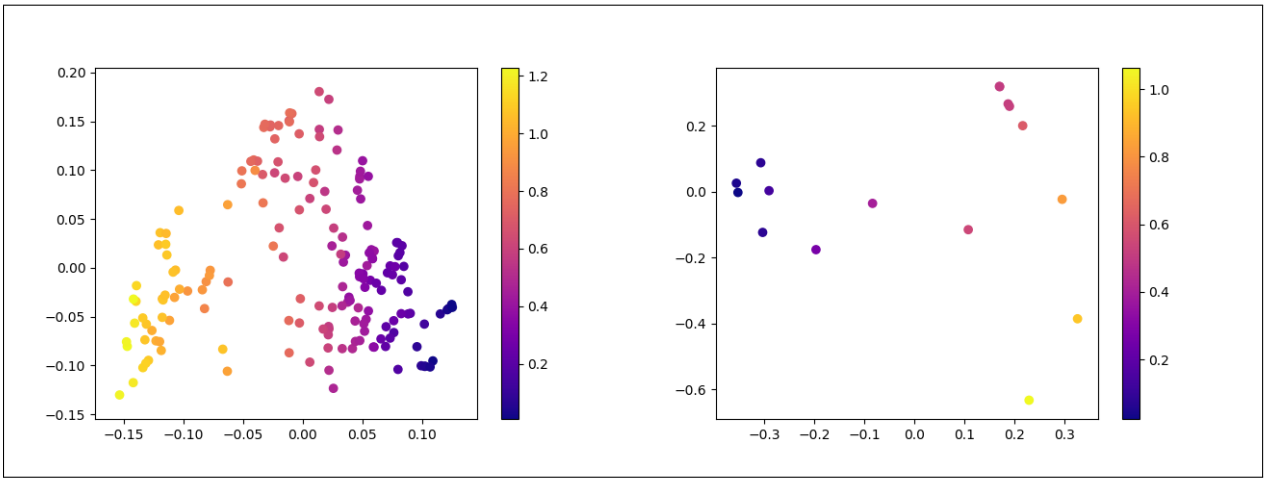

Figure 2: Embedding of the feature space of the Crop dataset (left) and the Obesity dataset (right) by the first two diffusion maps coordinates.

30% and quite similar to fisher and reliefF with 50%. A similar advantage of the DM-eliminate method with $a = 2$ is seen for the Obesity dataset, in Table 3.

Table 2: Classification results for the Crop dataset

Method	**Num. of features**	**SVM**	**KNN**	**Random Forests**
All features	174	0.985	0.981	0.978
High cluster (DM)	41	0.938	0.956	0.954
High+Med. cluster (DM)	84	0.976	0.982	0.974
Eliminate (DM, a=1)	117	0.984	0.983	0.974
Eliminate (DM, a=2)	59	0.983	0.979	0.974
Fisher-score (50%)	87	0.984	0.983	0.973
Fisher-score (30%)	52	0.978	0.978	0.969
reliefF (50%)	87	0.986	0.981	0.974
reliefF (30%)	52	0.983	0.975	0.97
CFS	12	0.93	0.935	0.951

Table 3: Classification results for the Obesity dataset

Method	**Num. of features**	**SVM**	**KNN**	**Random Forests**
All features	16	0.87	0.806	0.922
High cluster (DM)	5	0.689	0.731	0.754
High+Med. cluster (DM)	9	0.804	0.788	0.867
Eliminate (DM, a=1)	10	0.826	0.806	0.904
Eliminate (DM, a=2)	6	0.905	0.896	0.937
Fisher-score (50%)	8	0.874	0.853	0.928
Fisher-score (30%)	5	0.74	0.799	0.834
reliefF (50%)	8	0.896	0.793	0.911
reliefF (30%)	5	0.906	0.847	0.934
CFS	11	0.876	0.801	0.927