COMPUTATIONAL SCIENCE AND TECHNIQUES
Volume 4, Number 1, 2016, 487-499
© Klaipėda University, 2016

Publisher: Klaipėda University
http://journals.ku.lt/index.php/CST
Online ISSN: 2029-9966
DOI: 10.15181/csat.v4i1.1095

CONTROL POINT SELECTION FOR DIMENSIONALITY REDUCTION BY RADIAL BASIS FUNCTION

Kotryna Paulauskienė, Olga Kurasova
Vilnius University, Institute of Mathematics and Informatics
kotryna.paulauskiene@mii.vu.lt, olga.kurasova@mii.vu.lt

Abstract. The paper presents the results on the dimensionality reduction technique which is based on radial basis function (RBF) theory. The technique uses RBF for mapping multidimensional data points into a low-dimensional space by interpolating the previously calculated position of so-called control points. This paper analyses various ways of selection of control points (*regularized orthogonal least squares* method, *random* and *stratified* selections). The experiments have been carried out with 8 real and artificial data sets. Positions of the control points in a low-dimensional space are found by principal component analysis. Combinations of RBF technique with *random* and *stratified* selections outperformed RBF with *regularized orthogonal least squares* algorithm regarding to computation time analysing all data sets. We demonstrate that *random* and *stratified* selections of control points are efficient and acceptable in terms of balance between projection error (*stress*) and time-consumption.

Keywords: dimensionality reduction, radial basis function, selection of control points, large data set.

Introduction

With fast evolution of technology and science the amount of the data has been growing up in the last years. Various domains as science, engineering, telecommunications, finances are facing with the big data. Regardless of the data volume, the data is high-dimensional, i.e., each data point is characterized by many features (variables). One of the problems with high-dimensional data is that, in many cases, not all the measured features are important for understanding the underlying phenomena of interest (Fodor, 2002). Dimensionality reduction approaches extract low dimensional data from the high-dimensional input data. Dimensionality reduction (projection) techniques map data points from m-dimensional space to a smaller d-dimensional space $(d < m)$. Among classical dimensionality reduction methods we may mention principal component analysis (PCA) and multidimensional scaling (MDS) as ones of the best known methods. The main idea of PCA is to reduce the dimensionality of data by performing a linear transformation and rejecting a part of the components, variances of which are the smallest ones (Sorzano, et al., 2014). The goal of MDS is to find low-dimensional points, such that the distances between the points in the low-dimensional space were as close to the data proximities as possible (Borg & Groenen, 2005). Dealing with large data sets, MDS suffers from drawback of computer memory resources. The problem of projection process arises, because huge distance matrixes are used and they require large memory resources and computation time. One of the solutions is to use parallel and distributed computing or cloud computing. Another solution is, at first, to project a subset of data samples, called control points, latter, to project remaining points taking into account positions of control points. The control points are a subset of data samples which are projected in a low-dimensional space, the information got from the control points is used to project the rest part of data samples. Recently the dimensionality reduction methods based on manipulation of control points have been proposed: a part-linear multidimensional projection (PLMP) (Paulovich, et al., 2010), a local affine multidimensional projection (LAMP) (Joia, et al., 2011), a projection with a radial basis function (RBF) (Amorim, et al., 2014). When the number of data points is very large the relative MDS (Naud & Duch, 2000) and landmark MDS (de Silva & Tenenbaum, 2004) methods might be used. The mentioned methods are attractive as they avoid calculations with huge distance matrixes (if they are used in the algorithm) and needs less computer memory and computation time. However, the projection quality depends on the number of control points and the manner of selection of control points, thus, the selection should be done properly. The set of control points has a direct impact in the quality of the final projection results (Amorim, et al., 2014). The goal of the paper is to estimate various ways of selection of control points including the usage of radial basis function technique in order to determine which way is more effective.

The paper is organized as follows. Section 1 presents the related works. Section 2 reviews dimensionality reduction techniques based on radial basis function. Section 3 introduces the ways of selection of control points. Section 4 shows some experimental results. Finally, conclusions are drawn.

1. Related works

The section reviews some dimensionality reduction methods based on selection of control points. Paulovich et al. have proposed a multidimensional projection technique called part-linear multidimensional projection (PLMP) (Paulovich, et al., 2010). PLMP is a linear mapping, which uses a subset of data samples to define a global linear map. This method enables the embedding of high-dimensional data points in a visual space while avoiding extensive computation of distances between data instances. It requires only distance information between pairs of representative samples.

A local affine multidimensional projection (LAMP) method also uses a subset of data samples and their location in the visual space (Joia, et al., 2011). LAMP relies on a mathematical formulation derived from orthogonal mapping theory, what ensures robustness and accuracy to the process. Dimensionality reduction by PLMP and LAMP using data sets of various volumes has been presented in (Paulauskienė & Kurasova, 2014).

A piecewise Laplacian-based projection (PLP) method uses a force-based scheme to place the subset of data samples in the visual space. The remaining data instances are projected using several local Laplacian-like operators, which are built from disjoint local neighbourhood graphs (Paulovich, et al., 2011).

A landmark MDS method runs the classical MDS to embed a chosen subset of the data in a low dimensional space. Each remaining data point is located within this space given knowledge of its distances to the subset points (de Silva & Tenenbaum, 2004).

A relative MDS method is proposed in (Naud & Duch, 2000). This method also uses a subset of initial data set and then maps this subset using the MDS algorithm. The remaining points of initial data are added to the mapped points using the relative mapping. The papers (Bernatavičienė, et al., 2006), (Bernatavičienė, et al., 2007) focus on a strategies of selecting a subset in relative MDS, too.

The quality of projections by the before mentioned methods depends on the number of control points and the way of selection of control points, thus, the selection should be done properly. Usually a *random* selection of the control points is used. Recently the selection based on forward-selection and *orthogonal least squares* techniques have been proposed (Amorim, et al., 2014). This selection is grounded on a deterministic algorithm that selects those data instances that better explain the entire data set. A *stratified* selection of control points might be considered as an alternative to a *random* selection and the selection based on *orthogonal least squares*.

2. Dimensionality reduction with radial basis function

A novel multidimensional projection technique based on radial basis function theory has been introduced by Amorim et al. (Amorim, et al., 2014). Consider a data set $X \subset R^m$ with n points. Let $X_S = \{x_1, \dots, x_k\} \subset X, k \ll n,$ be a set of control points, for which the set of corresponding low-dimensional points $Y_S = \{y_1, \dots, y_k\} \subset R^d$, $d < m$ is calculated in advance using any dimensionality reduction method ($d = 2$). RBF projection finds the function $s\colon R^m \to R^d$ of the form:

$$s(x) = \sum_{x_i \in X_S} \lambda_i \phi(\|x - x_i\|), \qquad (1)$$

in such a way that the function s interpolates the position of each control point, i.e., $s(x_i) = y_i, i = 1,..,k$. The function $\phi\colon R_+ \to R$ is called RBF kernel. There are numerous functions that can be used as a kernel, more information about them can be found in (Amorim, et al., 2014).

The real-value coefficients λ_i have to be calculated to satisfy the interpolation condition. Thus, a linear system with k equations $s(x_i) = y_i, i = 1,..,k$ has to be solved. The system can be written in matrix form as

$$\Phi\lambda = y, \qquad (2)$$

where Φ is an interpolation matrix with dimensions $k \times k$, with $\Phi_{ij} = (\|x_i - x_j\|)$; y and λ are 2-columned vectors, each column is accounted for one of the final dimension of the output. Let $\phi_{ij} = (\|x_i - x_j\|)$, $\lambda_i = (\lambda_i^1, \lambda_i^2), y_i = (y_i^1, y_i^2)$, then Equation (2) can be written as

$$\begin{bmatrix} \phi_{11} & \cdots & \phi_{1k} \\ \vdots & \ddots & \vdots \\ \phi_{k1} & \cdots & \phi_{kk} \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_k \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_k \end{bmatrix}. \qquad (3)$$

When the coefficients λ are calculated, the function s is fully determined and it can be used to approximate the remaining instances of the data set.

3. Selection of control points

3.1. Selection of control points using method based on orthogonal least squares

Amorim et al. have proposed to use orthogonal least squares (OLS) method for selection of control points (Amorim, et al., 2014). Originally OLS method is used for selection of centres in RBF (Chen, et al., 1991). It is important to note RBF is a linear regression model. Assume we have N control points candidates $\{x_i, y_i\}_{i=1}^{N}$, where y_i is the output corresponding to control point x_i. So, the first step is randomly select N candidates for control points and to calculate their low-dimensional projections. If all x_i are used as control points, Equation (1) can be rewritten as:

$$s(x_t) = \sum_{i=1}^{N} \lambda_i \phi(\|x_t - x_i\|)\,, 1 \leq t \leq N, \tag{4}$$

Let $\phi_i(t) = \phi(\|x_t - x_i\|$, the desired output y_t can be expressed as

$$y_t = \sum_{i=1}^{N} \lambda_i \, \phi_i(t) + e_t, 1 \leq t \leq N, \tag{5}$$

where $e(t) = y_t - s(x_t)$ is an error between the desired output y_t and the approximated output $s(x_t)$. $e(t)$ will be zero when all candidates are used as control points, but the goal of the method is to reduce the set of control points. Equation (5) can be written in the matrix form as

$$y = \Phi\lambda \, + e, \tag{6}$$

where $y = [\, y_1 \dots y_N]^T$, $\Phi = [\phi_1 \dots \phi_N]$, ϕ_i=$[\phi_i(1) \dots \phi_i(N)]^T$, $\lambda = [\lambda_1 \dots \lambda_N]$ and $e = [e_1 \dots e_N]$.

Equation (6) has the form of a linear regression model and the vectors ϕ_i should be referred to as regressors (Amorim, et al., 2014).

The selection of control points starts with an empty set of regressors (control points) and one regressor from the set of candidates is selected at a time. Each selection is made in such a way to maximally decrease the squared error $e^T e$. Applying the concept of the OLS method, which transforms the set of ϕ_i into a set of orthogonal basis vectors, it is possible to calculate regressors individual contributions.

The regression matrix Φ can be decomposed as (Amorim, et al., 2014)

$$\Phi = WA, \tag{7}$$

where A is an upper-triangular matrix with diagonal 1 and $W = [w_1 \dots w_N]$ with orthogonal columns that satisfy condition, that $w_i^T w_j = 0$, if $i \neq j$. The model (6) can be rewritten as

$$y = Wg + e, \tag{8}$$

with $A\lambda = g$.

To prevent overfitting regularization technique is applied. The error to be minimized can be expressed as:

$$e^T e + \beta g^T g, \tag{9}$$

where $\beta \geq 0$ is a regularization parameter. This error formulation renames the technique to Regularized Orthogonal Least Squares (ROLS). Equation (9) can be rewritten as

$$e^T e + \beta g^T g = y^T y - \sum_{i=1}^{N} (w_i^T w_i + \beta) g_i^2. \tag{10}$$

Dividing (10) by $y^T y$ yields

$$\frac{e^T e + \beta g^T g}{y^T y} = 1 - \frac{\sum_{i=1}^{N} (w_i^T w_i + \beta) g_i^2}{y^T y} \tag{11}$$

and the regularized error reduction ration due to w_i is defined as

$$error_i = \frac{\sum_{i=1}^{N} (w_i^T w_i + \beta) g_i^2}{y^T y}. \tag{12}$$

At each step of the selection, the control point x_i associated with vector w_i and maximum value of $error$ is included in the set of control points.

The *stress*, given by Equation (13) (Borg & Groenen, 2005), of the remaining point candidates is also calculated.

$$stress = \frac{\sum_{ij} \left(d(x_i, x_j) - d(y_i, y_j) \right)^2}{\sum_{ij} \left(d(x_i, x_j) \right)^2}, \tag{13}$$

where $d(x_i, x_j)$ and $d(y_i, y_j)$ are distances between instances (points) in the initial (m-dimensional) and the reduced dimensionality (d-dimensional) spaces. Using all candidates as control points the *error* (12) is reduced at most, but not necessarily the *stress* (13). The goal is to select a limited amount of points that better explains the data set and potentially reduces the *stress*. The detailed ROLS algorithm for selection of control points can be found in

(Amorim, et al., 2014). The authors of ROLS algorithm have proposed to stop selecting control points when the maximum number of control points is reached and then the iteration with minimum *stress* is found.

3.2. Proposed strategies of selection of control points

We have observed that for each vector y $(d = 2)$ two different sets of control points are obtained and this special case is not discussed in (Amorim, et al., 2014), therefore we propose to join those control point sets in one set by selecting either the *unique* data points or *matching* data points. Here *unique* means that all the different points without repetitions are selected from two sets of control points, while *matching* means that we select only the control points which can be found in both data sets. Thus, we have a set of control points which will be used in RBF technique.

In this research, we also investigate *random* and *stratified* selection of control points. *Stratified* selection means that the same proportion of points as is in a full data set was taken from each class. The advantage of *random* and *stratified* selection ways is that these algorithms are simple, therefore the calculation time is short.

In this paper, the RBF technique with various strategies of selection of control points have been analysed:

1. RBF technique with the ROLS algorithm for *unique* control points selection (*RBF, ROLS, unique*). The ROLS algorithm stops, when it selects the maximum number of control points.
2. RBF technique with ROLS algorithm for *matching* control points selection (*RBF, ROLS, matching*). The ROLS algorithm stops, when it selects the maximum number of control points.
3. RBF technique with ROLS algorithm for selection of control points (*RBF, ROLS, min stress*). The ROLS algorithm stops, when the maximum number of control points is reached and when the iteration with minimum *stress* is found. This iteration indicates the number of control points which will be used in further calculations.
4. RBF technique with *random* selection of control points (*RBF, random*).
5. RBF technique with *stratified* selection of control points (*RBF, stratified*).

4. Experimental results

8 data sets are used in the experimental investigations. The *Yeast*, *Image segmentation*, *Waveform*, *Page blocks*, *MAGIC gamma telescope*, *Letter* recognition data sets are taken from ”UCI Repository of Machine Learning Databases” (http://archive.ics.uci.edu/ml/), *Helix* and *Swiss rolls* data sets are generated by us. The data sets vary in size (number of instances), data dimensionality (number of features) and number of classes. The short descriptions of the data sets are presented in Table 1.

A personal computer (Intel i5-3317U CPU 1,7 GHz (Max Turbo 2.6 GHz), with 2 cores and 12 GB of RAM memory) is used in experimental investigation. The ROLS algorithm for selection of control points and RBF technique for dimensionality reduction are implemented in *MATLAB R2012b*.

Table 1. Data sets

Name	Number of instances (n)	Number of features (m)	Number of classes (l)
Yeast	1453	8	10
Image segmentation	2 086	19	7
Waveform	5 000	21	3
Helix	5 000	3	2
Page blocks	5406	10	5
Letter recognition	18 668	16	26
MAGIC gamma telescope	18 905	10	2
Swiss roll	30 000	3	2

When using a computer with other characteristics, absolute values of the results would change, but the same ratio value between different ways would remain.

In this work, we apply the widely used multiquadratics RBF kernel: $\phi(r) = \sqrt{c^2+(\varepsilon r)^2}$, $c = \varepsilon = 1$. Two number of candidates and sizes of control point sets have been investigated. For the first selection of control points the following parameters are used: number of candidates $N = 200$, maximum number of control points is equal to 30. For the second one: number of candidates is $N = 300$, maximum number of control points is 100.

Positions of control points in a low dimensional space ($d = 2$) are found using the PCA method. In order to estimate that RBF technique with various ways of selection of control points gives appropriate results we compare RBF with the linear projection method – PCA. RBF could be compared with the non-linear method – multidimensional scaling (MDS), but this method suffers from drawback of computer memory then large data sets are analysed.

The quality of techniques has been evaluated according to a popular quality metric called *stress* function given by Equation (13) and the execution time in seconds. For each data set 100 experiments with different number of candidates of control points or *random* (or *stratified*) set of control points are executed.

Table 2 shows the *stress* values, obtained by various techniques, the best values are presented in bold. In addition the min/max and variance (Var) values are given. The smallest *stress* values are obtained using RBF technique with ROLS algorithm for five (from eight) data sets (*Yeast, Letter recognition, Waveform, Magic gamma telescope, Swiss roll*). The results show that it is unimportant how the join of two control point sets was made (*matching* or *unique*). Although the RBF technique with ROLS algorithm gives better *stress* values, the difference of *stress* values between combination of RBF with ROLS algorithm and RBF with *stratified* (or *random*) selection is not significant. The variance values indicate good results i.e., that the spread of the values from the mean is small for all techniques. The experimental results show that RBF technique combined with ROLS algorithm and *stratified* selection of control points gives *stress* values which differ insufficiently from *stress* obtained by PCA.

Table 2. Projection error (*stress*) values for eight data sets.

Name		**RBF, ROLS, matching (~30 cp)**	**RBF, ROLS, unique (~30 cp)**	**RBF, ROLS, min stress (~30 cp)**	**RBF, strata (~30 cp)**	**RBF, random (~30 cp)**	**PCA**
Yeast	Stress	**0.221**	0.224	0.228	0.225	0.244	0.153
	Min	0.156	0.156	0.164	0.183	0.180	
	Max	0.383	0.395	0.401	0.281	0.360	
	Var	0.002	0.002	0.003	4.85E–04	0.001	
Page blocks	Stress	0.090	0.083	0.104	0.068	**0.063**	6.95E–05
	Min	0.014	0.015	0.016	0.004	0.003	
	Max	0.593	0.436	0.498	0.284	0.251	
	Var	0.010	0.008	0.009	0.003	0.003	
Letter recognition	Stress	0.259	**0.258**	0.274	0.279	0.289	0.190
	Min	0.220	0.217	0.224	0.236	0.242	
	Max	0.327	0.343	0.438	0.327	0.339	
	Var	5.33E–04	5.92E–04	0.0016	3.71E–04	5.04E–04	
Image segmentation	Stress	0.191	0.195	0.197	**0.160**	0.166	0.175
	Min	0.121	0.136	0.142	0.129	0.137	
	Max	0.606	0.557	0.571	0.209	0.293	
	Var	0.005	0.006	0.005	1.91E-04	4.12E-04	
Waveform	Stress	0.119	**0.119**	0.128	0.133	0.136	0.085
	Min	0.107	0.107	0.108	0.117	0.116	
	Max	0.136	0.145	0.159	0.170	0.160	
	Var	3.45E–05	4.17E-04	1.51E–04	8.83E–04	1.0622–04	
MAGIC gamma telescope	Stress	0.159	**0.140**	0.159	0.164	0.165	0.067
	Min	0.083	0.075	0.084	0.105	0.097	
	Max	0.287	0.296	0.343	0.317	0.267	
	Var	0.003	0.002	0.003	0.001	0.002	
Helix	Stress	0.043	0.044	0.044	0.029	**0.028**	0.012
	Min	0.015	0.016	0.015	0.015	0.015	
	Max	0.173	0.145	0.193	0.062	0.054	
	Var	7.98E–04	6.54E–04	8.92E–04	7.56E–05	7.59E–05	
Swiss roll	Stress	**0.088**	0.089	0.095	0.095	0.093	0.056
	Min	0.061	0.064	0.058	0.063	0.064	
	Max	0.180	0.151	0.194	0.135	0.147	
	Var	6.45E–04	4.36E–04	7.37E–04	2.26E–04	2.74E–04	

Figure 1 shows the comparison of projection quality combining the RBF technique with various ways of selection of control points and the projection error (*stress*) values obtained by PCA is also included. It is evident that the smallest *stress* values are for the projection which was obtained by PCA as the whole data set is projected at ones unlike methods which use the position of control points. The comparison shows that the mean *stress* values are very similar and differ insignificant for combinations of RBF technique with various selection ways with ~ 30 control points analysing all data sets.

Though the *stress* values are similar but the computation time varies. Figure 2 presents mean time for eight different data sets. The results show that the mean computational time of projection using RBF technique with ROLS algorithm is 2.4 times greater than using RBF method with *stratified* (or *random*) selection.

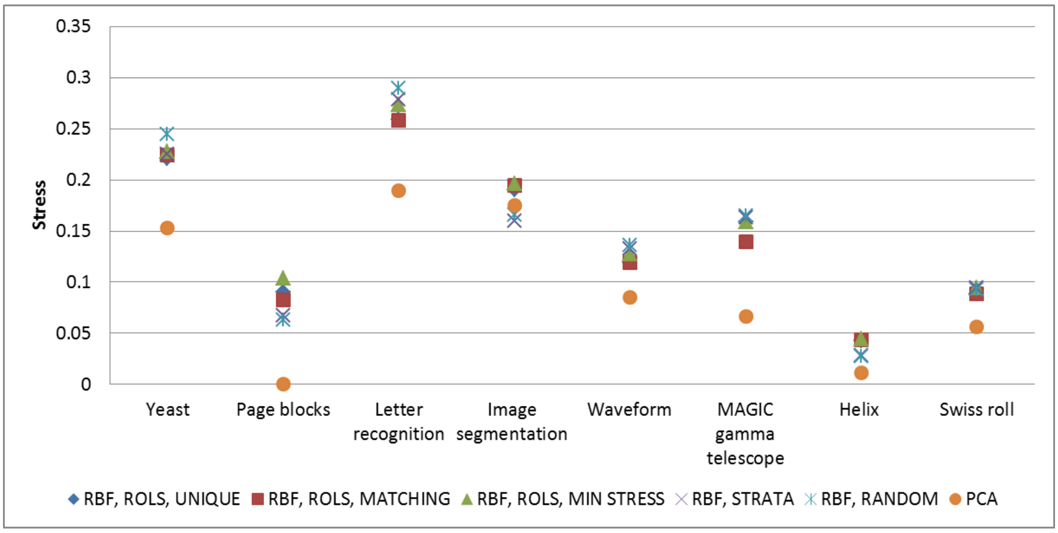

Fig. 1. **Mean stress values using different techniques for eight data sets.**

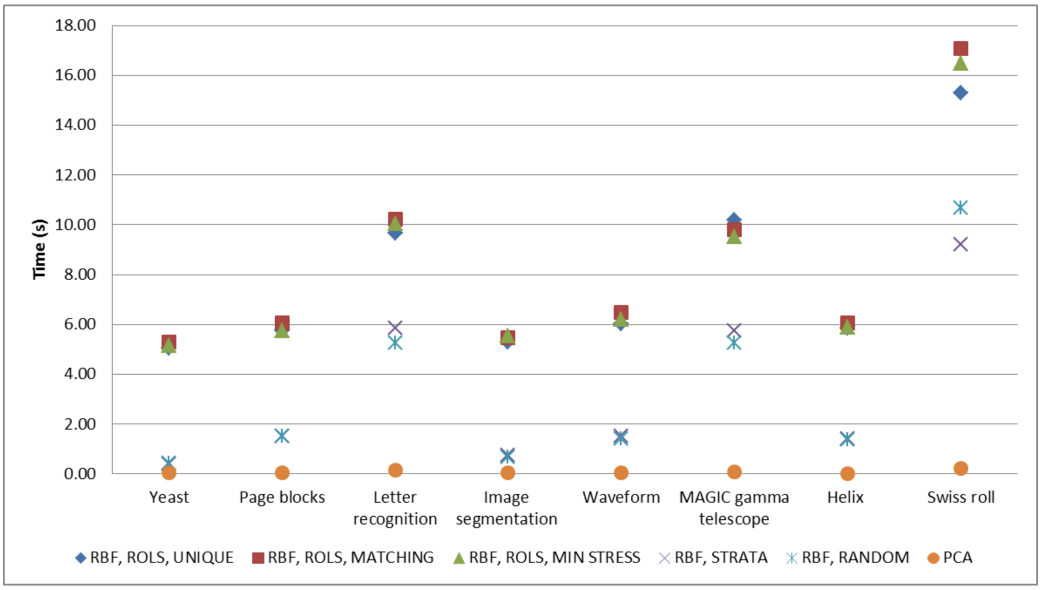

Fig. 2. **Mean computing time using different techniques for eight data sets.**

For example, the projection of the *Waveform* data set is found in about 6 seconds using RBF with ROLS algorithm and in about 1.65 seconds using RBF with *stratified* (or *random*) selection. Fast computational time is caused by simple selection algorithms.

The visualization of the *Waveform, Helix, MAGIC gamma telescope, Swiss roll* data sets when their dimensionality is reduced to two by the RBF technique combined with ROLS and *stratified* selection is presented in Fig. 3. Labels and units for both axes are not presented, because we are interested in observing the interlocation of points on a plane only. The images show that the positions of points, obtained by RBF combined with ROLS algorithm and *stratified* selection are similar to positions obtained by PCA.

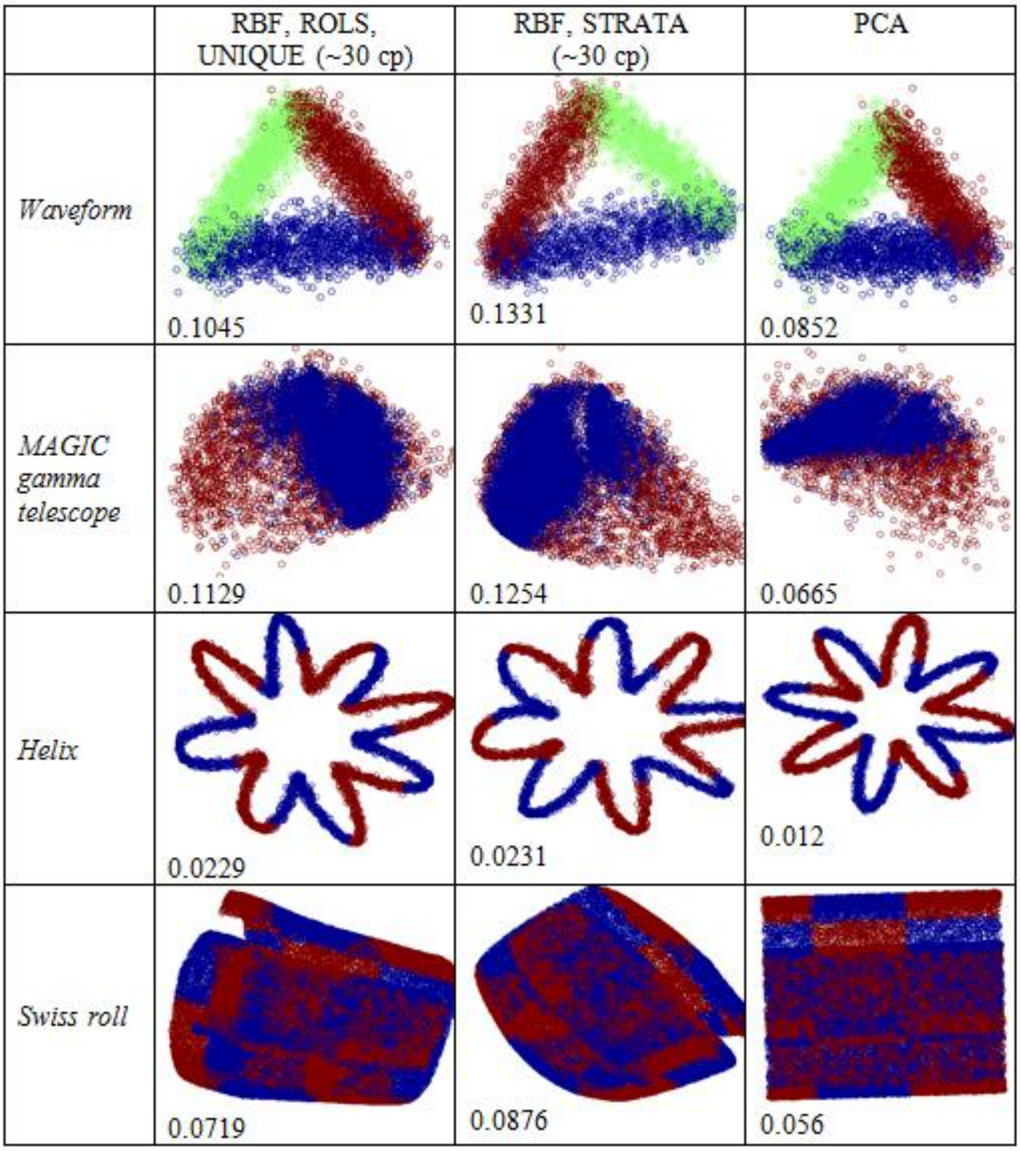

Fig. 3. **Visualization of four data sets using ROLS algorithm and *stratified* selection. Projection error values are shown in the bottom left.**

The experiments with bigger sets of candidates (300 points) and control points (100 points) are also carried out. The results obtained are presented in Table 3. In this experiment, two ways are investigated: RBF technique with ROLS algorithm; RBF technique with *stratified* selection. Better *stress* values are presented in bold. It is evident that when the size of control points set is increasing, the *stress* values become smaller for both ways. Table 3 shows that RBF technique with ROLS algorithm gives smaller stress values than RBF technique with *stratified* selection analysing six from eight data sets (*Yeast, Letter recognition, Waveform, Magic Gama telescope, Helix, Swiss roll*). The mean stress values in these ways are 0.107 and 0.111, respectively, thus the difference is not essential.

However the computation time vary significantly for those two ways and the mean computational time differs 8.6 times. For example, the projection of the *Letter recognition* data set is found in about 35.2 seconds using RBF with ROLS algorithm and in about

5.6 seconds using RBF with *stratified* selection, and the stress values are 0.21 and 0.22, respectively. Although the RBF with ROLS algorithm with increasing number of candidates and control points gives better stress values, but there is a trade-off between projection quality and computation time. It can be noted that increasing number of control points does not much influence the computation time of combination of RBF technique with *stratified* selection. It can be emphasized that RBF technique with *stratified* selection saves the computation time and the loss of *stress* accuracy is not significant comparing to RBF with ROLS algorithm. The analysis of *Image segmentation* data set has shown that *stress* is smaller obtained by RBF technique with *stratified* selection than obtained by PCA.

Table 3. Projection error (*stress*) and time values for eight data sets.

		RBF, ROLS, UNIQUE (~100 cp)	RBF, STRATA (~100 cp)
Yeast	Stress	**0.178**	0.189
	Time	29.729	0.571
Page blocks	Stress	0.032	**0.019**
	Time	29.729	1.615
Letter recognition	Stress	**0.207**	0.221
	Time	35.219	5.574
Image segmentation	Stress	0.161	**0.144**
	Time	29.280	0.737
Waveform	Stress	**0.0950**	0.103
	Time	30.225	1.565
MAGIC gamma telescope	Stress	**0.0972**	0.121
	Time	31.923	6.428
Helix	Stress	**0.0176**	0.0179
	Time	29.762	1.4928
Swiss roll	Stress	**0.065**	0.071
	Time	39.163	11.7306
Mean	Stress	0.107	0.111
	Time	31.879	3.714

Conclusions

In the paper, we have analysed dimensionality reduction by using RBF technique with various selection ways of control points. The combinations of RBF technique with ROLS algorithm, *random* and *stratified* selections are shown to be effective in terms of projection error (*stress*). The results have shown that combinations of RBF with *random* and *stratified* selections outperformed RBF with ROLS algorithm regarding to computation time analysing all data sets. The mean computational time differs 2.4 times for size of 30 control points and 8.6 times for size of 100 control points. Though the *stress* values for RBF with ROLS algorithm outperformed the RBF with *stratified* (or *random*) selection for the most cases but the difference was not essential. The results have shown that the execution of RBF technique with *stratified* or *random* selection is fast and not limited essentially in size of control points.

It can be emphasized that the loss of projection error accuracy is not significant compared to the computation time we save using RBF technique with *stratified* or *random* selections. We conclude that the *random* and *stratified* selection ways are attractive for their simplicity and they can be used instead of ROLS algorithm when dimensionality reduction task is solved.

References

Amorim, E., Brazil, E., Nonato, L. & Samavati, F., 2014. *Multidimensional projection with radial basis function and control points selection.* Yokohama, s.n., pp. 209-216.

Bernatavičienė, J., Dzemyda, G., Kurasova, O. & Marcinkevičius, V., 2006. Strategies of selecting the basis vector set in the relative MDS. *Technological and Economical Development of Economy,* 12(4), pp. 283-288.

Bernatavičienė, J., Dzemyda, G. & Marcinkevičius, V., 2007. Conditions for optimal efficiency of relative MDS. *Informatica,* 18(2), pp. 187-202.

Borg, I. & Groenen, P., 2005. *Modern Multidimensional Scaling: Theory and Applications.* 2 ed. New York: Springer.

Chen, S., Cowan, C. & Grant, P., 1991. Orthogonal least squares learning algorithm for radial basis function networks.. *IEEE Transactions on Neural Networks,* 2(2), pp. 302-309.

de Silva, V. & Tenenbaum, J., 2004. *Sparse multidimensional scaling using.* [Online] Available at: http://pages.pomona.edu/~vds04747/public/papers/landmarks.pdf [Accessed 03 09 2015].

Fodor, I. K., 2002. *A survey of dimension reduction techniques.* [Online] Available at: https://e-reports-ext.llnl.gov/pdf/240921.pdf [Accessed 03 09 2015].

Joia, P. et al., 2011. Local affine multidimensional projection. *IEEE Transactions on Visualization and Computer Graphics,* 17(12), pp. 2563-2571.

Naud, A. & Duch, W., 2000. Interactive data exploration using MDS. *Proceedings of the Fifth Conference: Neural Networks and Soft Computing,* pp. 255-260.

Paulauskienė, K. & Kurasova, O., 2014. Analysis of dimensionality reduction methods for various volume data (in Lithuanian). *Information Technology. 19th Interuniversity Conference on Information Society and University Studies (IVUS 2014),* pp. 114-121.

Paulovich, F. et al., 2011. Piecewise laplacian-based projection for interactive data exploration. *Computer Graphics Forum,* 30(3), p. 1091–1100.

Paulovich, F. V., Silva, C. T. & Nonato, L. G., 2010. Two-phase mapping for projecting massive data sets. *IEEE Transactions on Visualization and Computer Graphics,* 16(6), pp. 1281-1290.

Sorzano, C., Vargas, J. & Pascual-Monato, A., 2014. A survey of dimensionality reduction techniques. *CoRR abs/1403.2877.*

K. Paulauskienė is a doctoral student at Vilnius University Institute of Mathematics and Informatics. She obtained BSc degree in 2003 and MCs degree in 2005, both in the field of Statistics in Vilnius University. Her main research interests are dimensionality reduction methods for large data sets.

O. Kurasova received the Ph.D. in computer science from Institute of Mathematics and Informatics jointly with Vytautas Magnus University in 2005, Lithuania. Recent employment is at Institute of Mathematics and Informatics of Vilnius University, as senior researcher, and at the Informatics Department of Lithuanian University of Educational Sciences as associate professor. Her research interests include data mining methods, optimization theory and applications, artificial intelligence, neural networks, visualization of multidimensional data, multiple criteria decision making, multi-objective evolutionary algorithms parallel computing.

KONTROLINIŲ TAŠKŲ PARINKIMAS DIMENSIJAI MAŽINTI NAUDOJANT RADIALINĘ BAZINĘ FUNKCIJĄ

Kotryna Paulauskienė, Olga Kurasova

Santrauka

Šiame darbe nagrinėjamas dimensijos mažinimo metodas, kuris remiasi radialinių bazinių funkcijų (RBF) teorija. Pradžioje randamos tik dalies duomenų aibės taškų, vadinamų kontroliniais taškais, koordinatės sumažintos dimensijos erdvėje, pagal kurias naudojant RBF randamos likusiųjų duomenų aibės taškų projekcijos. Tyrime nagrinėjami įvairūs kontrolinių taškų parinkimo būdai (*ortogonaliųjų mažiausių kvadratų* metodas, *atsitiktinis* ir *stratifikuotas* parinkimai). Tyrimas atliktas naudojant 8 duomenų aibes. Kontrolinių taškų koordinatės sumažintos dimensijos erdvėje randamos pagrindinių komponenčių analizės metodu. Tyrimo rezultatai parodė, kad *atsitiktinis* ir *stratifikuotas* kontrolinių taškų parinkimas yra efektyvūs išlaikant kompromisą tarp projekcijos paklaidos ir skaičiavimo laiko.

Pagrindiniai žodžiai: dimensijos mažinimas, radialinės bazinės funkcijos, kontroliniai taškai, didelės apimties duomenų aibės.