VILNIUS UNIVERSITY

KOTRYNA PAULAUSKIENĖ

MASSIVE DATA VISUALIZATION BASED ON DIMENSIONALITY REDUCTION AND PROJECTION ERROR EVALUATION

Summary of Doctoral Dissertation
Physical Sciences, Informatics (09P)

Vilnius, 2018

The dissertation work was carried out at Vilnius University from 2011 to 2017.

Scientific Supervisor

Prof. Dr. Olga Kurasova (Vilnius University, Physical Sciences, Informatics – 09P).

The dissertation is defended at the Dissertation Defence Counsil:

Chairman

Prof. Dr. Romas Baronas (Vilnius University, Physical Sciences, Informatics – 09P).

Members:

Prof. Dr. Stefano Bonnini (University of Ferrara, Italy, Physical Sciences, Informatics – 09P),
Prof. Dr. Rimatas Butleris (Kaunas University of Technology, Technological Sciences, Informatics Engineering – 07T),
Dr. Remigijus Paulavičius (Vilnius University, Physical Sciences, Informatics – 09P),
Prof. Dr. Artūras Serackis (Vilnius Gediminas Technical University, Technological Sciences, Informatics Engineering – 07T).

The dissertation will be defended at the public meeting of the Council in the auditorium 203 of the Institute of Data Science and Digital Technologies of Vilnius University on the 24th of September, 2018 at 12:00.
Address: Akademijos str. 4, LT-04812 Vilnius, Lithuania.
The summary of the doctoral dissertation was distributed on the 23rd of August, 2018.
The dissertation is available at the library of Vilnius University.

VILNIAUS UNIVERSITETAS

KOTRYNA PAULAUSKIENĖ

DIMENSIJŲ MAŽINIMU PAGRĮSTAS DIDELĖS APIMTIES DUOMENŲ VIZUALIZAVIMAS IR PROJEKCIJOS PAKLAIDOS VERTINIMAS

Daktaro disertacijos santrauka
Fiziniai mokslai, informatika (09P)

Vilnius, 2018

Disertacija rengta 2011–2017 metais Vilniaus universitete.

Mokslinė vadovė

prof. dr. Olga Kurasova (Vilniaus universitetas, fiziniai mokslai, informatika – 09 P).

Disertacija ginama viešame disertacijos Gynimo tarybos posėdyje:

Pirmininkas

prof. dr. Romas Baronas (Vilniaus universitetas, fiziniai mokslai, informatika – 09P).

Nariai:

prof. dr. Stefano Bonnini (Feraros universitetas, Italija, fiziniai mokslai, informatika – 09P),
prof. dr. Rimatas Butleris (Kauno technologijos universitetas, technologijos mokslai, informatikos inžinerija – 07T),
dr. Remigijus Paulavičius (Vilniaus universitetas, fiziniai mokslai, informatika – 09P),
prof. dr. Artūras Serackis (Vilniaus Gedimino technikos universitetas, technologijos mokslai, informatikos inžinerija – 07T).

Disertacija bus ginama viešame disertacijos Gynimo tarybos posėdyje 2018 m. rugsėjo 24 d. 12 val. Vilniaus universiteto Duomenų mokslo ir skaitmeninių technologijų instituto 203 auditorijoje.
Adresas: Akademijos g. 4, LT-04812 Vilnius, Lietuva.
Disertacijos santrauka išsiuntinėta 2018 m. rugpjūčio 23 d.
Disertaciją galima peržiūrėti Vilniaus universiteto bibliotekoje ir VU interneto svetainėje adresu: *www.vu.lt/lt/naujienos/ivykiu-kalendorius*.

1 Introduction

1.1 Research area and relevance of the problem

With fast evolution of technology and science the amount of the data has been growing up in the last years. Various domains, such as science, engineering, telecommunications, finances are facing the massive 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. Dimensionality reduction approaches allow us to understand better the high-dimensional data. Dimensionality reduction (projection) techniques map data points from high-dimensional space to the projection space so that data structure is preserved as much as possible. Data visualization approaches present data in a graphical form and enable people to understand data better and deeper. Visualization of millions of points in a scatter plot does not make sense, because they might concentrate and overlap when massive data is analysed. There are various visualization systems, but most of them support visualization of aggregated data or the visualization can be performed by few features. If we aim to take into account all the data features and to visualize not aggregated data, but rather to identify the position of each point, we need to employ visualization approach that helps us to comprehend a large amount of data and to identify each data point location among the data. Furthermore, when one applies visualization techniques which are based on dimensionality reduction, the projection quality needs to be evaluated. Most dimensionality reduction methods involve the optimization of a certain criterion. One way to assess the projection quality is to evaluate the value of this criterion. The problem arises when projections obtained by different methods need to be compared. Usually the projection error is used in order to evaluate the results of dimensionality reduction. When calculating projection error, often the distances between points are taken into account. When we deal with large data sets, the problem of estimating the projection error arises, huge distance matrices (or vectors) are used and they require large memory resources. Dimensionality reduction by some dimensionality reduction methods is very quick even for rather large data sets, but the evaluation of the projection error still remains a complicated problem.

Thus, in this dissertation the following issues are being solved:

1. Projection error calculation for massive data sets.
2. Massive data sets visualization without point overlapping in projection space.

1.2 The object of research

The object of research:

- massive multidimensional data sets;
- dimensionality reduction techniques for massive data visualization and projection error evaluation.

1.3 The Aim and Tasks of the Research

The goal of this research is to develop projection error evaluation approaches for massive data as well as to propose visualization approach for massive data.

To achieve this goal, it is necessary to solve the following tasks:

- to perform analytical review of dimensionality reduction methods which are used for massive data visualization and to analyse the projection error quality evaluation techniques;
- to propose projection error calculation approaches for multidimensional massive data;
- to do a comparative analysis of proposed ways to calculate projection error and existing ways;
- to propose and investigate dimensionality reduction-based visualization approach for massive data which would allow us to avoid visual data points overlapping and yet preserve data structure;
- to do application analysis of proposed approaches while dealing with real data.

1.4 Research methods

To analyse the scientific, experimental and practical achievements in the fields of dimensionality reduction and data visualization, information retrieval, organization, analysis, benchmarking and aggregation methods were used. Based on experimental

research method, the analysis of proposed approaches was performed and comparative and generalization methods were used to evaluate the outcome.

1.5 Scientific novelty

1. Proposed ways to evaluate projection error are suitable for massive data sets. One of them is based on building the sample of data set, the second one on dividing the data set into the smaller data sets.
2. Proposed new approach of massive data visualization lets us to visualize data without points overlapping and keeps the structure of the data.
3. Comprehensive analysis of various dimensionality reduction techniques was performed while solving the dimensionality reduction problem.

1.6 Statements to be defended

1. Proposed ways to evaluate projection error can be applied to calculate the projection error for massive data.
2. Proposed visualization approach allows us to visualize massive data sets by avoiding points overlapping but yet preserving data structure.

1.7 Approbation of the research

The main results of the dissertation were published in 6 research papers: 3 papers are published in periodicals, reviewed scientific journals, one of them in journal indexed in Clarivate Analytics Web of Science; 3 papers are published in conference proceedings. The main results have been presented and discussed at 3 national and 3 international conferences.

1.8 Outline of the dissertation

The dissertation consists of 6 chapters and the list of references. The chapters of the dissertation are as follows: Introduction, Review of dimensionality reduction techniques, Projection error calculation and data visualization approach for massive data, Experimental results, Applications of proposed solutions for *Weather* data analysis, Conclusions. The dissertation also includes the list of notation and abbreviations. The scope of the work is 119 pages including 25 figures and 19 tables. The list of references consists of 89 sources.

2 Review of dimensionality reduction and visualization techniques

Firstly, in this chapter the definitions of big data are discussed. With fast evolution of technology and science, the amount of the data has been growing in the last years. International Data Corporation's Digital Universe study predicts that the world's data will amount to 44 zettabytes by 2020 [1]. The meaning of the term "big data" is still the subject of some disagreement, but it generally refers to data that is too big or too complex to process on a single machine [2]. Usually, big data is characterized by three main components: volume, velocity, and variety [3]. Later two more components were introduced: value and veracity [4]. In this thesis, high-dimensional and large volume data is considered to be massive data.

Secondly, in this chapter the dimensionality reduction methods (Multidimensional scaling [5], Principal component analysis [6], Independent component analysis [7], Random projection [8]) are reviewed. The dimensionality reduction methods based on control points selection (Part-linear multidimensional projection [9], Local affine multidimensional projection [10], Multidimensional projection with radial basis function and control points selection [11]) are introduced as well.

Thirdly, the overview of projection error evaluation measures such as stress function [5], Spearman's rho [12], Konig's topology measure [12], silhouette [13], Renyi entropy [14] is provided.

In addition, the analysis of data mining tools (WEKA [15], KNIME [16], ORANGE [17]) is presented. Finally, the technologies usually used for big data analysis and data visualization tools are reviewed. Problem of overlapping data points is raised and existing approaches solving the overlapping points issue are discussed.

3 Projection error calculation and dimensionality reduction-based visualization approach for massive data

In this chapter two ways to calculate projection error are proposed when massive data sets are analysed using a personal computer without any particular programming and technologies for high performance computing. A new approach for massive data

visualization without visual point overlapping and preserving data structure is proposed as well.

In this work all proposed approaches are implemented in MATLAB, however, the proposed and explored ways could be applied for other programming environments by using analogical functions. MATLAB was chosen due to several reasons. MATLAB provides a range of numerical computation methods for analysing data, developing algorithms, and creating models. Furthermore, one of the key features of MATLAB is that it uses processor-optimized libraries for fast execution of matrix and vector computations. This feature is important when working with dimensionality reduction techniques in which distance matrices are used.

3.1 Projection error calculation

Usually, dimensionality reduction can be evaluated using the projection error given in the following formula [5]:

$$E_{\text{Stress}} = \frac{\sum_{ij}\left(d(X_i,X_j)-d(Y_i,Y_j)\right)^2}{\sum_{ij}\left(d(X_i,X_j)\right)^2}, \qquad (1)$$

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, respectively.

There are several well-known basic ways for projection error calculation using MATLAB:

- to calculate the projection error using the loop for each data point (usually *for*) by the formula (1);
- to use MATLAB function *pdist* to compute distances for the high dimensional data set and for the data set of the reduced dimensionality, then to apply the formula (1).

An advantage of the first way is that huge distance matrices are not used. In such a way, the computer memory resources are saved. Instead of distance matrices, the projection error is calculated using the loop for each data point summing the numerator and the denominator of the formula (1). The second way uses MATLAB function *pdist* which computes the Euclidean distance between pairs of objects in m by n data matrix X and the Euclidean distance between pairs of objects in m by d reduced dimensionality matrix Y. To save memory space and computation time, pairwise distances are formed not

as a matrix, but as a vector. The vector contains only unique distances between the data points. The analysis of large data sets has shown that, in the first case (when the loop is used), the computation time with 250000 instances is about 2 hours, in the second case (when the function *pdist* issued), the computation time is very fast, but with more than 30000 instances the computer runs out of memory (12 GB) [18]. Hence, it is necessary to search ways to reduce memory usage and computation time needed for calculation of distances.

In this work, two efficient solutions of projection error evaluation for large data sets are proposed:

- to calculate the projection error not for the full data set, but only for the data sample;
- to calculate the projection error for the full data set, but dividing the data set into the smaller data sets.

3.1.1 Obtaining the data sample

In the first way, the projection error evaluation is based on the data sample. Usually in statistics, the data population is too large for the researcher to attempt to analyse all of its members. A small, but carefully chosen sample can be used to uncover the population. The data sample reflects the characteristics of the population from which it is drawn. So, in order to save the computation time and to reduce the usage of operating memory, the projection error could be evaluated only for a data sample. Data samples can be found by these sampling methods: random sampling and stratified sampling. Having the data sample, the projection error is calculated by the formula (1) not for full data set, but only for the data sample. MATLAB function *pdist* is used for projection error calculation.

3.1.2 Dividing the data set into the smaller data sets

The second proposed way calculates the projection error for divided data set. The algorithm for calculating the projection error when the data set is divided into the smaller data sets can be summarized as follows:

Step 1: the initial data set and the data set of the reduced dimensionality are divided into the smaller data sets;

Step 2: Euclidean distances between pairs of the instances for each smaller data set in the high-dimensional and the reduced dimensionality spaces are calculated. The distances are calculated using MATLAB function *pdist*.

Step 3: for each smaller data set the numerator and the denominator of the formula (1) are calculated.

Step 4: Euclidean distances between the instances of each of two smaller data sets in the high-dimensional and the reduced dimensionality spaces are calculated using MATLAB function *pdist2* (this function computes pairwise distances between two sets of instances).

Step 5: for each possible pair of smaller data set the numerator and the denominator of the formula (1) are calculated.

Step 6: the projection error is calculated dividing the sum of numerators by the sum of denominators obtained in *steps* 3 and 5.

Dividing the data set into the smaller data sets allows us to avoid running out of computer memory. As MATLAB functions *pdist* and *pdist2* are used, the distances between points are found very fast. It is necessary to emphasize that this way to evaluate projection error does not influence the projection error value i.e., it remains the same as it would be if calculated for non divided data set. Pseudo-code for projection error calculation dividing the data set into the smaller data sets is provided in figure 1.

```
Input: data – multidimensional points;
       proj – points of reduced dimensionality;
       A – two column matrix (the elements of the first (second) column
       indicate the data index, corresponding to the beginning (end) of the
       smaller data sets);
       groups – the number of the smaller data sets.
Output: Stress – projection error.
BEGIN
//For each smaller data set
FOR i=1:groups
      data_temp=pdist(data(A(i,1):A(i,2),:))
      proj_temp=pdist(proj(A(i,1):A(i,2),:))
      nomin_temp(i)=sum((data_temp-proj_temp).^2)
      denom_temp(i)=sum(data_temp.^2)
END
//For the instances of each of two smaller data sets
numerator=0; denominator=0
FOR i=1:groups
FOR j=i+1:groups
      data=(pdist2(data(A(i,1):A(i,2),:),data(A(j,1):A(j,2),:)))
      proj=(pdist2(proj(A(i,1):A(i,2),:),proj(A(j,1):A(j,2),:)))
      numerator=numerator+sum(sum((data-proj).^2))
      denominator=denominator+sum(sum(data.^2))
END
END
//Projection error is calculated
Stress=(numerator+sum(nomin_temp))/(denominator+sum(denom_temp))
END
```

Figure 1. Pseudo-code for projection error calculation, dividing the data set into the smaller data sets

3.2 Dimensionality reduction based visualization approach for massive data

Visualization approach is needed to present the main meaningful information when massive data sets are analysed. Visualization of a large amount of data points in a scatter plot in most cases will end up taking a certain shape, fully filled with data points, and it will not be an informative representation of the data. Thus it does not make sense to visualize all the data points. A conventional way is to cluster the data and then select representatives from each cluster, but in this case, the representatives of sparse clusters or outliers can be lost. Representatives of sparse clusters or outliers can carry important information about the data under investigation. Another problem we face is that some dimensionality reduction techniques cannot handle a large amount of data. MDS is usually

unable to deal with millions of points and requires a lot of computational time [18], [19]. There is no need to visualize all the data points if it can be done by visualizing a data sample, which will uncover characteristics of the data set.

Thus, in this work it is proposed to visualize not the whole data set, but the data sample. If one has the data sample the projection can be found quickly and then can be visualized.

The proposed visualization approach consists of two main stages:

Stage 1: proper selection of a data subset and calculation of the data subset projection;

Stage 2: visualization of the projection of the data subset without point overlapping.

The proposed visualization approach is provided in figure 2.

Consider a data set $X \in R^m$ with n points. Let $X_S = \{x_1, \dots, x_s\} \subset X, s < n$, be a subset of the data set, for which a set of corresponding low-dimensional points $Y_S = \{y_1, \dots, y_s\} \subset R^d,\ s < n,\ d = 2$ is computed using any dimensionality reduction method. The final data subset for visualization is $Y_L = \{y_1, \dots, y_l\} \subset Y_S, l \leq s < n$. Final subset for visualization Y_L contains less data points than Y_S, since we eliminate the overlapping of points in *stage 2*.

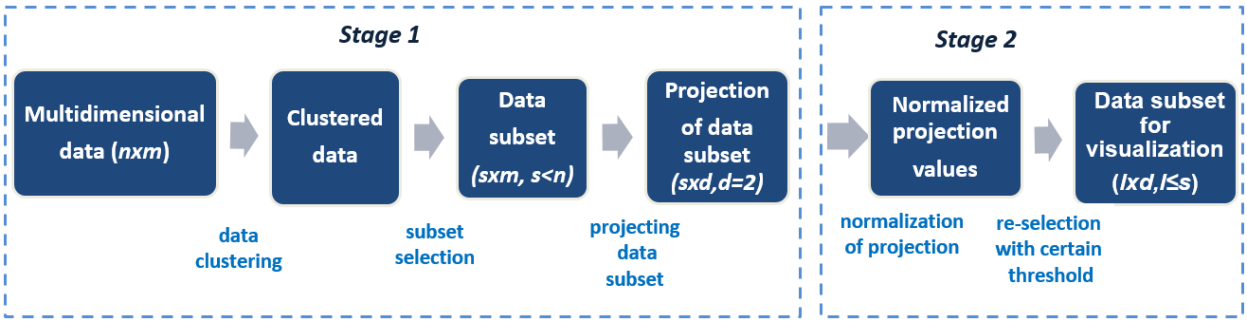


Figure 2. The proposed visualization approach

3.2.1 Selection of data subset (stage 1)

An important task of the proposed approach is a proper selection of data subset. As it was mentioned before a conventional way is to cluster the data and then select representatives from each cluster, but in this case, we could lose the outliers. Data clustering is one of the most popular techniques in data mining. It is a process of partitioning an unlabelled data set into clusters, where each cluster contains data points that are similar to one to another and different from that of other clusters with respect to a

certain similarity measure [20]. For data subset selection, the following methods can be used: simple random sampling, systematic sampling, stratified sampling, cluster sampling, etc. [21]. But in this case, only few representatives from sparse clusters and not all outlying observations, that can be significant in data analysis and knowledge discovery will be selected. An outlier can be defined as an observation that is far distant from the rest of observation [22]. Outliers can carry important information about the data under investigation. Rarely distributed points that can make a separate cluster are important as well, especially when talking about medical data or fraud analysis [20]. Therefore, it is important to select (not to lose) those points (representatives from sparse clusters or outliers) which would be excluded if the standard sampling methods were used. Thus, in this thesis, new approach for data subset selection and for massive data visualization is proposed. The main idea of the proposed selection is to take into account the density of points. This is implemented via data clustering, i.e. the sum of distances from each data point to the centre per cluster and the number of points per cluster are estimated.

The proposed data subset selection can be summarized as follows:

Step 1: data clustering is performed to divide the high-dimensional points of the $n \times m$ data matrix X into τ clusters (known classes can be considered as clusters as well, in such a case data clustering can be not applied);

Step 2: for each cluster i, the sum of distances from each point X_j^i to cluster centre M_i is calculated by formula $D_i = \sum_{j=1}^{N_i} d(X_j^i, M_i)$, where N_i, $i = 1, \dots, \tau$ is the number of points in cluster i;

Step 3: the size s of data subset is determined, i.e. the number of points that will be selected as candidates for visualization is defined;

Step 4: the ratio $r_i = D_i/N_i$ is calculated, which indicates how many points from each cluster should be selected. If $r_i = 0$, then any one point from i cluster will be selected into the data sample. The number of points to be selected from each cluster into the data subset is calculated by the formula $N_i' = \frac{r_i \times s}{\omega}$, where $\omega = \sum_i^\tau r_i$, and s – the size of data subset;

Step 5: the data subset X_S of size $s \times m$ is selected;

Step 6: dimensionality of points of the data subset X_S is reduced by projection technique and the matrix Y_S is obtained. The size of matrix Y_S is $s \times d$.

How to select the size s of data sample remains an open question. Since the points will be visualized in the computer display, it makes sense to tie-up this measure to the resolution of monitor. Suppose we have a common display with resolution 1280×1024 pixels and we aim for the picture to take a fifth of the computer monitor. If one point has 256 pixels, then we will be able to visualize 1024 points. So it is recommended to select $s = 1024$ points. This fact is illustrated in figure 3. Here two-dimensional points are evenly distributed on the square. It can be seen that the points are not overlapping. When visualizing real-world data, the points will not be so evenly distributed and they might concentrate in groups, consequently not scattering so widely and sparsely; we can expect a part of them to overlap. Such visual overlapping will be eliminated in the next stage of the proposed visualization approach (see subsection 3.2.2).

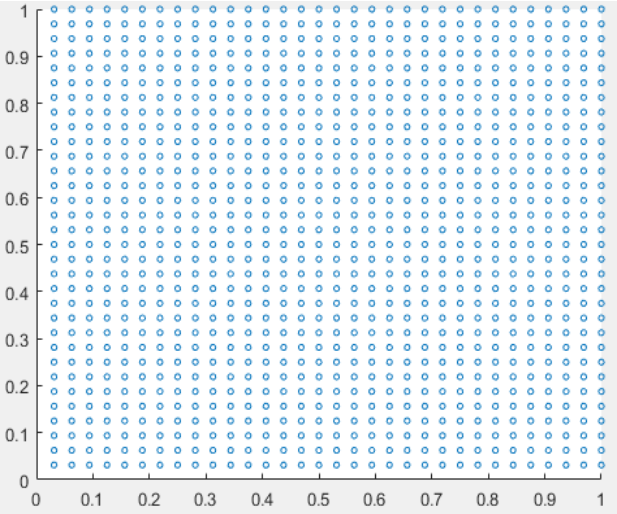


Figure 3. Visualization of the points in the interval $[0,1]$, the size of data subset: $s = 1024$

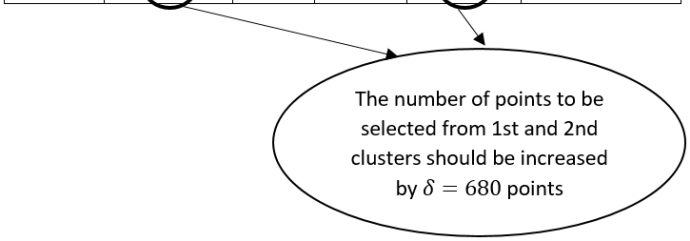
There may be cases when the number of points N_i' to be selected from a cluster, calculated in *Step 4*, is larger compared to the actual number of points N_i in a cluster i. In such cases, we suggest to select all the points from the cluster and increase the number of points to be selected from the remaining clusters according to their ratio r_i. The value of increase is $\delta = N_i' - N_i$. Table 1 provides an example where the number of points to be selected from the first and second clusters is increased with respect to their ratio r_i when the *Random3* data set is analysed ($n = 2515, m = 10, \tau = 3$, the size of data subset is $s = 1024$). It can be seen that according to the ratio $r_3 = 4$ of the third cluster, $N_3' = 695$ points should be selected to the data subset, but the third cluster contains only $N_3 = 15$

points. Thus, we select all 15 points from the third cluster to the data subset and increase the number of points to be selected from the remaining clusters by $\delta = 680$ points with respect to their ratios r_1 and r_2.

Projection of the data subset X_S in *Step 6* can be found by various dimensionality reduction techniques.

Table 1. Example of redistribution of the number of points to be selected from the clusters

Cluster *(i)*	Number of points per cluster (N_i)	D_i	Ratio (r_i)	Number of points from cluster (N_i')	Number of points from cluster (N_i') (increased)
1	500	461	0.922	156	492
2	2000	1941	0.971	165	517
3	15	60	4	695	15



3.2.2 Data visualization without overlapping (stage 2)

Usually when huge amount of points is visualized, points overlap in a scatter plot. If it is necessary to identify each point (or position of the point) we need to eliminate the points which visually overlap and cover each other. Let us define that the subset of points of reduced dimensionality is Y_S and subset to be visualized is Y_L , $l \leq s$. The proposed data subset visualization without overlapping can be summarized as follows:

Step 1: the values of features of the initial subset of reduced dimensionality Y_S should be normalized in the range $(0, 1)$, so that the minimal value of each feature is equal to 0, and the maximal one are equal to 1. The normalization is performed in order to set the same value of the *threshold* (see *Step 2*) for all data sets and to be able to compare obtained results;

Step 2: the normalized data subset points of the reduced dimensionality Y_S, are re-selected with a certain *threshold* t. The *threshold* controls the density of the points. The re-selection is performed in the following way: the distance matrix Δ for the points of reduced dimensionality is calculated; if the distance from one point to another is less than

t, then the point is eliminated from the initial normalized data subset $Y_{S'}$. The size of the final data subset Y_L which will be visualised is $l \times m$ ($l \leq s$, $d < m$);

Step 3: the data subset Y_L is visualized in a scatter plot.

4 Experimental results

In this section, experimental investigations are provided. Firstly, all the data sets used for the experimental analysis are described. 21 real and artificial data sets were used in the experimental investigations. The data sets varied in size (number of instances) and data dimensionality (number of features). Number of instances varied from 150 to 1000000. Number of features varied from 3 to 166. A personal computer (MS Windows 8 operating system, Intel i5-3317U CPU 1.7 GHz (Max Turbo 2.6 GHz), with 2 cores and 12 GB of RAM memory) was used for the experimental investigation. Due to the limited size of summary of doctoral thesis, detailed results cannot be provided.

Secondly, the comparative analysis of dimensionality reduction techniques is provided. The comparative analysis of MDS, PCA, ICA, RP, LAMP, PLMP, RBF methods showed that it takes up to one minute to find the projection in visual space while analysing various sizes (150–1000000) of data sets, except MDS and LAMP methods. However, the smallest projection error values are obtained by MDS and LAMP methods. Dimensionality reduction methods based on control points selection require a reduced amount of distance information to carry out the embedding into a visual space, speeding up projection calculation process. RP method is attractive because of its simplicity and fast execution time, but it is not suitable for data visualization. Experimental investigation of projection quality evaluation methods showed that in order to compare projections obtained by different methods several quality evaluation measures need to be used.

Thirdly, the experiments have been carried out in order to show the performance of proposed ways to calculate projection error. The experimental investigation with different data sets has shown that in order to decrease computation time, the projection error can be evaluated precisely using just data sample and not full data set. In all data sets analysis, the differences between the projection error values of the data samples and the full data are not significant. Furthermore the projection error calculation for the data sample significantly saves computation time. The results have shown that dividing data set into the smaller data sets allows us to calculate the projection error for data of 450000

instances in an appropriate time (1 h 14 min). Analysis results of the data set of 450000 instances indicated that the computational time differs almost 6 times calculating projection error by dividing data set into the smaller data sets compared to calculating projection error using the loop for each data point. Dividing data set into the smaller data sets allows us to calculate projection error for 15 times $\left(\frac{450000 \text{ instances}}{30000 \text{ instances}}\right)$ bigger data set compared to calculating projection error calculation way using non divided data set and using the special function for fast execution of matrices and vectors.

Finally, the proposed visualization approach was experimentally investigated. The proposed visualization approach consist of two stages: selection of data subset; visualization of the projection of the data subset. It has been shown that the subset up to 1000 points is enough to visualize in order to reveal the structure of multidimensional data. Experimental investigation with different size of data subsets has shown that proposed data subset selection way preserves the structure both of sparse clusters and the data. The investigation results have shown that overlapping of data subset points can be eliminated by applying a reselection of points with a certain threshold. Two threshold values were proposed. The position of certain point that is not a member of the data subset can be found by the nearest neighbour in a scatter plot.

5 Application of proposed approaches for *Weather* data analysis

In this section, applications of proposed approaches (projection error calculation for massive data (section 3.1) and visualization approach (section 3.2)) were applied for *Weather* data analysis and experimental results are provided.

5.1 Data description

To show the performance of the proposed projection error calculation and visualization approaches, some experimental investigations are carried out with real world data. In this section *Weather* data was analysed, based on Weather Underground data (https://www.wunderground.com/). Weather Underground provides local and long range weather forecasts, weather reports, maps and tropical weather conditions for locations worldwide. The measurements from three weather stations were selected for experimental

analysis (Vilnius (EYVI), Singapore (WSAP), Yakutsk (UEEE)), which were collected during the years 2016–2017. Weather data is described by seven features:

- temperature (C),
- dew point (C),
- humidity (%),
- wind-chill (C),
- wind speed (km/h),
- pressure (pHa),
- visibility (km).

40939 measurements were from Vilnius (EYVI) weather station, 15680 measurements were from Singapore (WSAP) weather station and 2922 – form Yakutsk (UEEE) weather station. Overall *Weather* data set had 59641 observations. Since this data set is high dimensional and the number of observations is quite big, it is suitable to demonstrate the proposed approaches and to reveal their advantages. In order to have a more realistic interpretation *Weather* data was split into three classes, which corresponded to three weather stations, or into twelve classes, which corresponded to the months of the year. Even though the investigations showed that the most accurate and the smallest projection error is obtained by MDS, but the personal computer runs out of memory while analysing data set of 59541 instances. Therefore MDS was not chosen for further analysis. Dimensionality reduction methods based on control points (LAMP, PLMP, RBF) did not show obvious advantage in terms of projection error and calculation time. Thus, PCA was used to reduce the dimensionality of the *Weather* data ($d = 2$). This method was chosen due to fast calculations when massive data is analysed. In figure 4 visualization of multidimensional points by weather station is provided (yellow points corresponds to Yakutsk, blue points – Vilnius, turquoise points – Singapore). In figure 5 visualization of multidimensional points by months is provided. In figure 4 three different climate zones can be seen clearly. In figure 5, we can see that points corresponding to spring, summer and autumn months in Vilnius are closer to points which correspond to Singapore, while points corresponding to cold months in Vilnius are closer to points which correspond to Yakutsk.

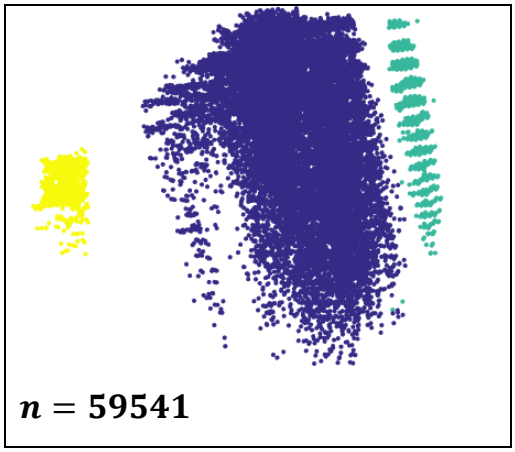


Figure 4. Visualization of weather data set by weather station

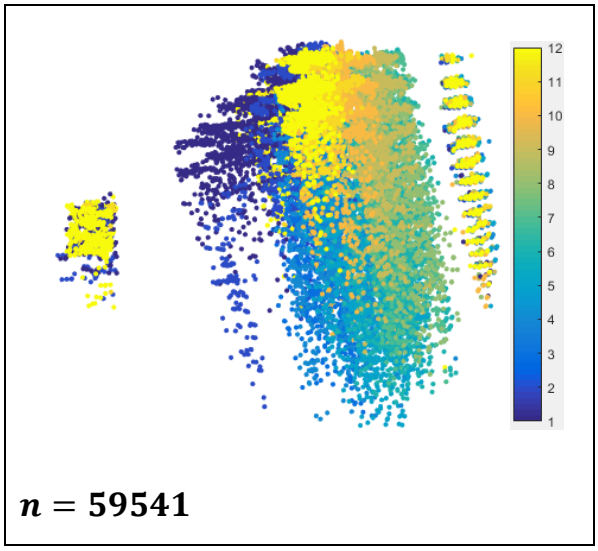


Figure 5. Visualization of weather data set by month (the months by colours are provided on the scale)

5.2 Projection error calculation

In this section the performance of the proposed solutions of projection error calculation is shown when weather data is analysed.

First way – to calculate the projection error not for the full data set, but for the data sample. *Second way* – to calculate the projection error for the full data set, dividing the data set into the smaller data sets.

When applying the *First way*, three sizes of data sample were investigated: n''=10000, 5000, 1000. Data samples were selected by applying the random sampling.

Table 2 shows the projection error values and the calculation time for different data samples of *Weather* data. It has to be mentioned that clustering time is not involved in calculation time. The projection error values and calculation time for the full data set were also obtained and presented in bold. Since the *Weather* data set contained more than 30000 points the projection error for the full *Weather* data set was calculated using the proposed *Second way*. While calculating the projection error using the proposed *Second way* the *Weather* data set was divided into smaller sets of the size of 10000 points. Projection error values for *Weather* data subsets were calculated by applying MATLAB function *pdist.* Table 2 shows that the differences between projection error values for the data samples and the full *Weather* data set are minor. However, the computation time obviously differs a lot.

Table 2. Projection error and computation time values in seconds (s) for different size of *Weather* data subsets (n'' – sample size)

(n'')	**Projection error**	**Time (s)**
59541	**0.0130**	**61.18**
10000	0.0133	1.32
5000	0.0134	0.14
1000	0.0132	0.04

The visualization of the *Weather* data set and its data samples, when dimensionality was reduced by PCA, is presented in figure 6. The images show that the distribution of points of data samples is similar to that of the full data set. It can be emphasized that the loss of projection error accuracy is not significant compared to the computation time we save. The experimental results show that the projection error can be evaluated for the data sample when massive data sets are analysed.

Applying the *Second way* the projection error of the full *Weather* data set is calculated in about 61.18 s. The most time consuming way for projection error calculation is to use the loop for each data point: it takes about 289.95 s. Using the *pdist* function calculating the projection error for non divided *Weather* data set, causes the computer to run out of memory resources.

Experimental results prove that both proposed ways of projection error calculation can be applied for real world massive data analysis.

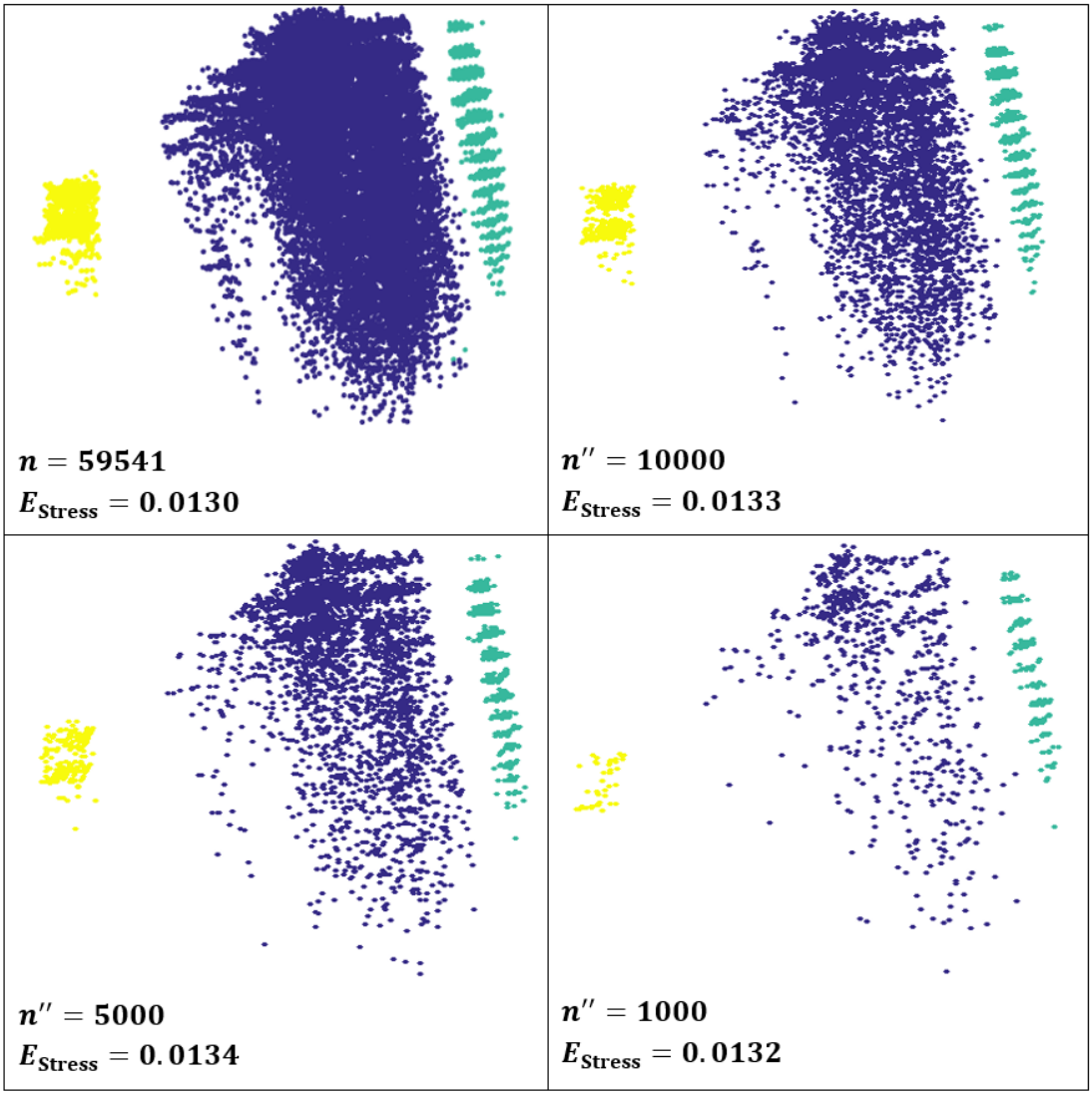


Figure 6. Visualization of *Weather* data set and the data samples. Sample size (n'') and projection error values (E_{Stress}) and are shown at the bottom left

5.3 Weather data visualization

In this section the performance of the proposed visualization approach is shown when *Weather* data is analysed. The *proposed approach* consists of two stages:

Stage 1 – data subset selection in which:

- multidimensional data has been clustered by a *k-medoids* method [20];
- the number of clusters τ has been determined by the Calinski-Harabasz clustering evaluation criterion [23];
- after the density of data sample has been evaluated, the data sample was selected.

Stage 2 – visualization of the data subset without visual points overlapping.

To evaluate the *proposed approach* of the data subset selection, it was compared to the *stratified* sampling. In the stratified random sampling, a population is divided into smaller sub-groups called strata. The random sampling is applied within each subgroup or stratum [24]. To make it easier the size of data subset is set to be $s = 1000$.

Table 3 shows the comparison of two subset selection methods (*stratified* sampling and the *proposed approach* of subset selection). *Weather* data set was divided into three clusters. It can be seen that the number of selected points from each cluster differs depending on the ways of selection. The highest difference among selection ways can be seen for the second cluster. The visualization results of the *Weather* data set and its subsets, selected by different methods, i.e. the *stratified* sampling and the *proposed approach* of subset selection, are presented in figure 7. Data points corresponding to January and February months in Vilnius are clustered with points corresponding to Yakutsk. Data points which correspond to June–August months in Vilnius are clustered with data points which correspond to Singapore. The rest of Vilnius data points makes the separate cluster. The images show that the distribution of the points of the *Weather* data subset, obtained by the *proposed approach*, is similar to that of the full data set, while the *stratified* subset does not retain the structure of the second cluster (the blue points), i.e. only few points from the second cluster corresponding to Vilnius are selected to the data subset.

Table 3. Selection of data subset by two different methods

Cluster (i)	**Full data set**	**Subset by the *proposed approach***	***Stratified* subset**
1	34640	397	582
2	3180	261	53
3	21721	342	365

The second stage is visualization of subset. In this stage the overlapping of points is eliminated. First of all the initial data subset of reduced dimensionality (obtained in *Stage 2*) was normalized in the range $(0; 1)$. Afterwards the normalized data subset points of the reduced dimensionality were re-selected with a certain *threshold* values $t = 0.01$, $t = 0.0075$.

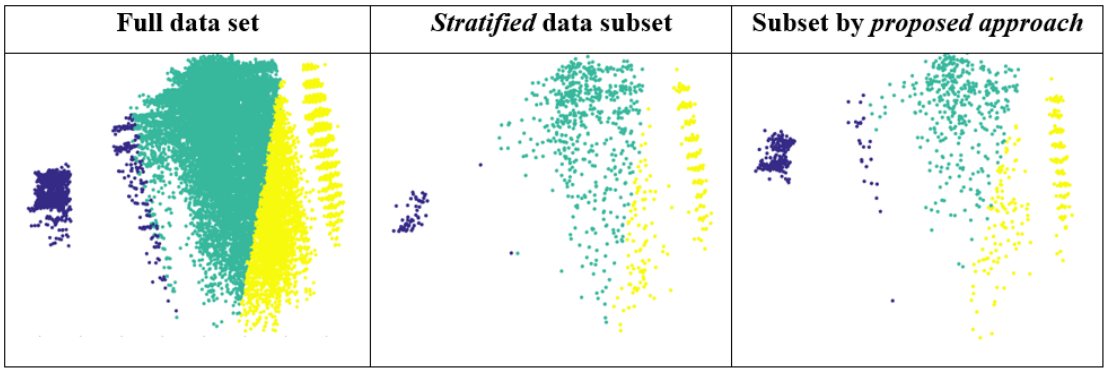


Figure 7. Comparison of visualization results using different ways of data subset selection

Figure 8 shows the visualization of the *Weather* data subset points after the re-selection of points. The results have shown that the overlapping of points is eliminated with $t = 0.01$ and allow us to identify each data points in a scatter plot.

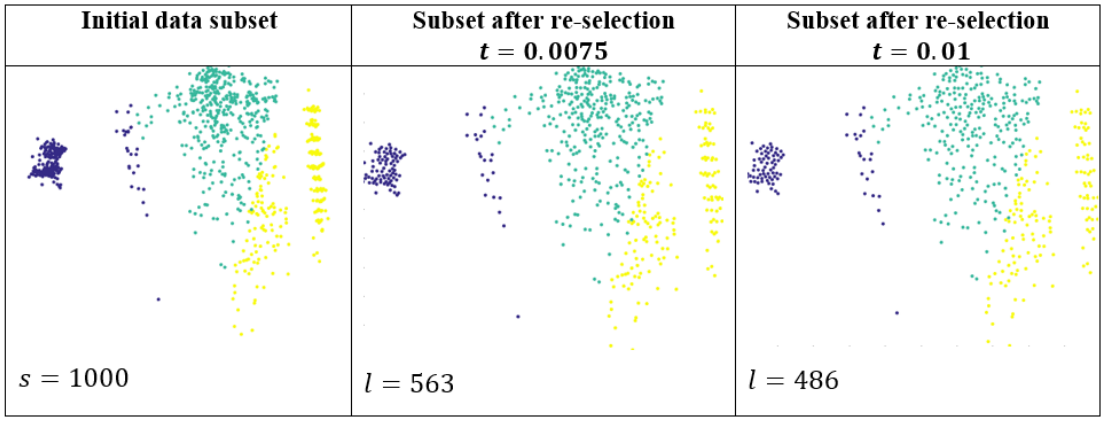


Figure 8. Visualization results after the *Weather* data subset was re-selected with different threshold values. The size of data subset is shown at the bottom left

5.4 Generalization

The experimental investigation with *Weather* data set has shown that in order to decrease computation time, the projection error can be evaluated precisely using just a data sample, and not a full data set. The differences between the projection error values of the data samples and full data are not significant. The results have shown that dividing data set into the smaller data sets allows us to calculate the projection error for massive data set and not to run out of computer memory. Visualization of *Weather* data by proposed visualization approach has proved that selecting the data sample using the

proposed way and visualizing its two-dimensional points, the structure of *Weather* data is preserved.

Weather data analysis has shown that the approaches proposed in this work can be applied to solve real world data tasks.

6 Conclusions

In this work, the following results were achieved: various classic dimensionality methods and methods which are based on control point's selection were investigated; various projection evaluation measures were investigated; ways to calculate projection error for massive data sets were proposed and explored; a new dimensionality reduction-based visualization approach for massive data was proposed and explored. The performance of the proposed approaches on the real world data analysis was shown.

Experimental investigation revealed the advantages of proposed projection error calculation and visualization approaches for massive data sets. Experimental investigation and results point to these conclusions:

1. The projection error can be evaluated precisely using just data sample when massive data sets are analysed. The projection error calculation for the data sample significantly saves computation time. For some of investigated data sets the calculation time is $2-9$ times or even few hundred times shorter.
2. Dividing data set into the smaller data sets allows us to shorten the projection error calculation time 6 times compared to calculating projection error using the loop for each data point. Projection error calculation when data set is divided allows to avoid running out of computer memory and to calculate the projection error for the data set which is 15 times bigger than it would be possible to calculate using non divided data set and using special function for fast execution of matrices and vectors.
3. The proposed data subset selection way is efficient in terms of preserving the data structure while analysing various test data sets. It has been shown that the subset up to 1000 points is enough to visualize in order to reveal the structure of multidimensional massive data.

7 List of publications

The articles published in the peer-reviewed periodical publications:

- Paulauskienė, Kotryna; Kurasova, Olga. Control Point Selection For Dimensionality Reduction By Radial Basis Function. *Computational Science and Techniques*. ISSN 2029-9966. 2016, vol. 4(1), pp. 487–499.

- Paulauskienė, Kotryna; Kurasova, Olga. Projection error evaluation for large data sets. *Nonlinear Analysis: Modelling and Control*. ISSN 1392-5113. 2016, vol. 21(1), pp. 92–102 (Clarivate Analytics Web of Science, Impact Factor 2017: 0.896.

- Paulauskienė, Kotryna; Kurasova, Olga. Investigating abilities of data mining systems to analyse various volume data sets. Information sciences. *Information sciences.* Vilnius: Vilnius university publishing house. ISSN 1392-0561. 2013, vol. 65, pp. 85–95 (in Lithuanian).

The articles published in the conference proceedings:

- Paulauskienė, Kotryna; Kurasova, Olga. Analysis of dimensionality reduction methods for various volume data. *Information Technology: 19th Interuniversity Conference on Information Society and University Studies (IVUS 2014).* Kaunas: Technologija. ISSN 2029-4832. 2014, pp. 114–121 (in Lithuanian).
- Paulauskienė, Kotryna; Kurasova, Olga. Evaluation of projections obtained by dimmensionality reduction techniques. *Proc. of the Lithuanian Mathematical Society*. ISSN 0132-2818. 2014, vol. 55, ser. B, pp. 137–142 (in Lithuanian).
- Paulauskienė, Kotryna; Kurasova, Olga. A new dimensionality reduction-based visualization approach for massive data. *WSCG 2017 Posters Proceedings*: *Computer Science Research Notes*. ISSN 2464-4617. 2017, pp. 19–24.

Abstracts in the conference proceedings:

- Paulauskienė, Kotryna; Kurasova, Olga. Massive data visualization via selecting a data subset. 8-th International workshop on Data analysis methods for software systems: Abstracts book, Druskininkai, 1–3 December, 2016, pp. 48.

- Paulauskienė, Kotryna; Kurasova, Olga. Improvement of projection error evaluation for massive data sets. 6-th International Conference on Advanced Technology and Sciences: Abstract book, Riga, 12–15 September, 2017, pp. 70.

About the author

Kotryna Paulauskienė was a PhD student at Vilnius University Institute of Mathematics and Informatics from 2012 to 2017. She obtained BSc degree in 2003 and MCs degree in 2005, both in the field of Statistics in Vilnius University. From 2003 to 2006 she worked at Lithuanian Health Information Centre as senior statistician, from 2006 she continued as head of causes of death register. From 2009 to 2017 she worked at Institute of Hygiene as head of causes of death register. From 2017 Kotryna Paulauskienė works in Danske Bank Global Services Lithuania Group Risk Management as senior analyst.

List of references used for summary of dissertation

[1] International Data Corporation, The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things, 2014. [Online]. Available:https://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm. [Accessed 30 7 2018].

[2] S. Landset, T. Khoshgoftaar, A. N. Richter and T. Hasanin, "A survey of open source tools for machine learning with big data in the Hadoop ecosystem," *Journal of big data,* vol. 2, no. 24, pp. 1–36, 2015.

[3] D. Laney, "3D data management: controlling data volume, velocity and variety," *Meta group,* 2001.

[4] Y. Demchenko, P. Grosso, C. de Laat and P. Membrey, "Addressing big data issues in scientific data infrastructure," in *International Conference on Collaboration Technologies and Systems (CTS)*, San Diego, 2012.

[5] I. Borg and P. Groenen, Modern Multidimensional Scaling: Theory and Applications, 2 ed., New York: Springer, 2005.

[6] K. Pearson, "On Lines and Planes of Closest Fit to Systems of Points in Space," *Philosophical Magazine,* vol. 2, no. 11, pp. 559–572, 1901.

[7] A. Hyvarinen, "Independent component analysis: recent advances," *Philosophical Transactions of the Royal Society A,* vol. 371, no. 1984, 2013.

[8] E. Bingham and H. Mannila, "Random projection in dimensionality reduction: applications to image and text data," in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, 2001.

[9] F. V. Paulovich, C. T. Silva and L. G. Nonato, "Two-phase mapping for projecting massive data sets," *IEEE Transactions on Visualization and Computer Graphics,* vol. 16, no. 6, pp. 1281–1290, 2010.

[10] P. Joia, F. V. Paulovich, D. Coimbra, J. A. Cuminato and L. G. Nonato, "Local affine multidimensional projection," *IEEE Transactions on Visualization and Computer Graphics,* vol. 17, no. 12, pp. 2563–2571, 2011.

[11] E. Amorim, E. Brazil, L. Nonato and F. Samavati, "Multidimensional projection with radial basis function and control points selection," in *IEEE Pacific Visualization Symposium*, Yokohama, 2014.

[12] O. Kurasova and A. Molytė, "Quality of quantization and visualization of vectors obtained by neural gas and self-organizing map," *Informatica,* vol. 22, no. 1, pp. 115–134, 2011.

[13] P. Tan, M. Steinbach and V. Kaumar, Introduction to data mining, Boston: Addison-Wesley, 2005.

[14] A. Gupta and R. Bowden, "Evaluating dimensionality reduction techniques for visual category recognition using Renyi entropy," in *19th European Signal Processing Conference (EUSIPCO 2011)*, Barcelona, 2011.

[15] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten, "The WEKA Data Mining Software: An Update," *SIGKDD Explorations,* vol. 11, no. 1, pp. 10–18, 2009.

[16] M. Berthold, N. Cebron, F. DILL, T. Gabriel, T. Kotter and T. Meinl, "KNIME: The Konstanz Information Miner," *Data Analysis, Machine Learning and Applications, Studies in Classification, Data Analysis, and Knowledge Organization,* pp. 319–326, 2008.

[17] T. Curk, J. Demšar, Q. Xu, G. Leban, U. Petrovič and I. Bratko, "Microarray data mining with visual programming," *Bioinformatics,* vol. 21, no. 3, pp. 396–398, 2005.

[18] K. Paulauskienė and O. Kurasova, "Analysis of dimensionality reduction methods for various volume data," in *Information Technology. 19th Interuniversity Coneference on Information Society and University Studies (IVUS 2014)*, Kaunas, 2014.

[19] P. Pawliczek and W. Dzwinel, "Interactive Data Mining by Using Multidimensional Scaling," *Procedia Computer science,* vol. 18, pp. 40–49, 2013.

[20] J. Han and M. Kamber, Data Mining Concepts and Techniques, 2 ed., San Francisco: Morgan Kaufman Publishers, 2006, p. 743.

[21] S. L. Lohr, Sampling: Design and Analysis, 2nd edition, Boston: Brooks/Cole, 2010.

[22] G. S. Maddala, Introduction to Econometrics 2nd ed., New York: MacMilan, 1992, p. 89.

[23] U. Maulik and S. Bandyopadhyay, "Performance evaluation of some clustering algorithms and validity indices," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 12, pp. 1650–1654, 2002.

[24] Y. Ye, Q. Wu, J. Z. Huang and L. X. Li, "Stratified sampling for feature subspace selection in random forests for high dimensional data," *Pattern Recognition*, vol. 46, no. 3, pp. 769–787, 2013.

DIMENSIJŲ MAŽINIMU PAGRĮSTAS DIDELĖS APIMTIES DUOMENŲ VIZUALIZAVIMAS IR PROJEKCIJOS PAKLAIDOS VERTINIMAS

1 Tyrimo sritis ir problemos aktualumas

Mokslo, inžinerijos, telekomunikacijų, finansų, medicinoje ir kitose srityse nuolat susiduriama su didelės apimties duomenų aibėmis. Vystantys technologijoms kaupiamų duomenų apimtys sparčiai didėja. Nors objektų skaičius didelis, kiekvieną iš jų nusako ir daug požymių, tačiau analizuojant daugiamačius duomenis dažnai ne visi būna svarbūs. Dimensijos mažinimo metodai leidžia mums geriau suprasti daugiamačius duomenis. Metodų tikslas – pateikti objektus, apibūdinančius duomenis mažesnės dimensijos erdvėje (projekcijos erdvėje), taip, kad būtų kiek galima tiksliau išlaikyti tam tikrą duomenų struktūrą ir būtų lengviau apdoroti ir vizualizuoti didelės dimensijos duomenis. Duomenų vizualizavimas leidžia geriau suprasti turimus duomenis, pastebėti išskirtinumus, grupavimosi tendencijas, tarpusavio ryšius, t. y. atskleisti duomenų struktūrą. Analizuojant daugiamačius didelės apimties duomenis nėra tikslinga vizualizuoti keliasdešimt tūkstančių ar milijoną taškų sklaidos diagramoje, nes jie gali susitelkti ir vienas kitą perdengti. Yra sukurta įvairių duomenų vizualizavimo sistemų, tačiau dauguma iš jų duomenis leidžia vizualizuoti tik agreguotus arba juos vizualizuoti pagal keletą daugiamačių duomenų požymių. Vis dėlto norint atsižvelgti į visus duomenų aibės požymius ir matyti ne agreguotus duomenis, o identifikuoti kiekvieno taško poziciją, trūksta vizualizavimo būdų. Be to, taikant vizualizavimo metodus, pagrįstus duomenų dimensijos mažinimu, reikia įvertinti gautos projekcijos kokybę. Dažniausiai dimensijų mažinimo metodas turi savo kriterijų, pagal kurį ieškoma optimali projekcija. Gauta projekcija gali būti vertinama taikant tą patį kriterijų. Tačiau norint įvertinti keliais metodais gautas projekcijas, naudojami kiti nuo metodo nepriklausantys matai, atspindintys įvairias duomenų ypatybes. Dažniausiai dimensijų mažinimo metodų rezultatams tyrimuose vertinti naudojama projekcijos paklaida. Projekcijos paklaidos reikšmėms skaičiuoti dažnai naudojami atstumai tarp taškų. Nagrinėjant didelės apimties duomenų aibes kyla projekcijos paklaidos įvertinimo problema, kadangi apskaičiuoti ją naudojamos didelės apimties atstumų matricos, o šioms apskaičiuoti gali pritrūkti personalinio kompiuterio operatyviosios atminties. Nors ir analizuojant didelės apimties

duomenų aibes dimensija tam tikrais dimensijos mažinimo metodais gali būti sumažinama labai greitai, tačiau dėl anksčiau minėtos priežasties projekcijos paklaidos įvertinimas trunka labai ilgai arba reikalauja daug kompiuterio operatyvios atminties.

Taigi šioje disertacijoje sprendžiamos šios pagrindinės problemos:

1. Projekcijos paklaidos apskaičiavimas analizuojant didelės apimties duomenų aibes.
2. Didelės apimties duomenų aibės vizualizavimas išvengiant duomenų aibės taškų persidengimo projekcijos erdvėje.

2 Tyrimo objektas

Disertacijos tyrimo objektas:

- didelės apimties daugiamačiai duomenys;
- dimensijų mažinimo metodai didelės apimties daugiamačiams duomenis vizualizuoti ir projekcijos paklaidų įvertinimas.

3 Darbo tikslas ir uždaviniai

Darbo tikslas – sukurti didelės apimties duomenų projekcijos paklaidos apskaičiavimo būdus ir pasiūlyti duomenų vizualizavimo strategiją didelės apimties duomenims vizualizuoti.

Siekiant tikslo būtina spręsti šiuos uždavinius:

- atlikti dimensijų mažinimo metodų, skirtų daugiamačiams duomenims vizualizuoti, ir projekcijos kokybės įvertinimo būdų analitinę apžvalgą;
- pasiūlyti daugiamačių duomenų projekcijos į mažesnio matmenų skaičiaus erdvę apskaičiavimo būdus, leidžiančius projekcijos paklaidą vertinti didelės apimties duomenims;
- eksperimentiškai palyginti pasiūlytus didelės apimties duomenų projekcijos apskaičiavimo būdus su jau žinomais būdais;
- pasiūlyti ir ištirti didelės apimties duomenų aibės vizualizavimo strategiją, leidžiančią išvengti duomenų aibės taškų persidengimo ir išlaikyti bendrą duomenų struktūrą;
- pasiūlytus sprendimus pritaikyti realių duomenų vizualiosios analizės uždaviniui.

4 Tyrimo metodai

Analizuojant dimensijos mažinimo ir duomenų vizualizavimo srities mokslinius ir eksperimentinius pasiekimus naudoti informacijos paieškos, sisteminimo, analizės, lyginamosios analizės ir apibendrinimo metodai. Remiantis eksperimentinio tyrimo metodu, atlikta statistinė duomenų ir tyrimų rezultatų analizė, o šios rezultatams įvertinti naudotas apibendrinimo metodas.

5 Darbo mokslinis naujumas

1. Pasiūlyti du projekcijos paklaidos apskaičiavimo būdai, tinkami didelės apimties duomenų aibėms. Vienas iš jų grindžiamas duomenų aibės imties sudarymu, antrasis – duomenų aibės dalijimu į dalis.
2. Pasiūlyta nauja vizualizavimo strategija, leidžianti vizualizuoti didelės apimties duomenų aibes, išvengti duomenų aibės taškų persidengimo ir išlaikyti bendrą duomenų struktūrą.
3. Atlikta išsami įvairių dimensijos mažinimo metodų, sprendžiant projekcijos paieškos uždavinį, lyginamoji analizė.

6 Ginamieji teiginiai

1. Pasiūlyti projekcijos paklaidos apskaičiavimo būdai yra tinkami apskaičiuoti projekcijos paklaidą didelės apimties duomenų aibėms.
2. Pasiūlyta nauja vizualizavimo strategija yra tinkama didelės apimties duomenų aibėms vizualizuoti, išvengti duomenų aibės taškų persidengimo ir išlaikyti bendrą duomenų struktūrą.

7 Darbo rezultatų praktinė reikšmė

Pasiūlyti projekcijos paklaidos apskaičiavimo būdai leidžia sutaupyti skaičiavimo laiką ir kompiuterio operatyviąją atmintį bei leidžia projekcijos paklaidą apskaičiuoti didelės apimties duomenų aibėms. Pasiūlyta didelės apimties duomenų aibių vizualizavimo strategija leidžia vizualizuoti didelės apimties duomenų aibes, išlaikyti duomenų struktūrą ir išvengti taškų persidengimo. Pasiūlytas duomenų aibės imties sudarymo būdas gali būti naudojamas ne tik didelės apimties duomenims vizualizuoti, bet

ir sprendžiant duomenų analizės uždavinius įvairiose srityse. Visi disertacijoje pasiūlyti būdai gali būti taikomi sprendžiant realius duomenų analizės uždavinius.

8 Disertacijos struktūra

Disertaciją sudaro 6 skyriai ir literatūros sąrašas. Disertacijos skyriai: Įvadas, Dimensijos mažinimo ir vizualizavimo metodų apžvalga, Projekcijos paklaidos apskaičiavimas ir strategija didelės apimties duomenims vizualizuoti, Eksperimentinių tyrimų rezultatai, Pasiūlytų sprendimų taikymas meteorologinių duomenų aibės analizei, Bendrosios išvados. Be to, disertacijoje pateiktas naudotų žymėjimų ir santrumpų sąrašai. Visa disertacijos apimtis – 119 puslapių, juose pateikti 25 paveikslai ir 19 lentelių. Disertacijoje remtasi 89 literatūros šaltiniais.

9 Išvados

Tiriant dimensijų mažinimo metodus darbe gauti šie rezultatai: ištirti įvairūs dimensijos mažinimo metodai, iš jų klasikiniai gerai žinomi metodai ir metodai, kuriuose projekcija randama remiantis valdymo taškais; ištirti įvairūs projekcijos kokybės įvertinimo matai; pasiūlyti ir ištirti projekcijos paklaidos apskaičiavimo būdai didelės apimties duomenų aibėms; pasiūlyta ir ištirta didelės apimties duomenų aibių vizualizavimo strategija, leidžianti neprarasti retų klasterių ir bendros duomenų struktūros ir vizualizuoti duomenų aibės taškus be persidengimo; pademonstruotas disertacijoje pasiūlytų sprendimų pritaikymas sprendžiant realų uždavinį, kai analizuojama meteorologinių duomenų aibė.

Atlikti tyrimai atskleidė darbe pasiūlytų projekcijos paklaidos apskaičiavimo būdų ir duomenų aibių vizualizavimo strategijos naudą tirti didelės apimties duomenų aibėms. Eksperimentinių tyrimų rezultatai leidžia daryti šias išvadas:

1. Projekcijos paklaida gali būti vertinama pagal duomenų aibės imtį analizuojant didelės apimties duomenų aibes. Projekcijos paklaidos duomenų aibės imčiai skaičiavimo laikas yra trumpesnis nei skaičiuojant visai duomenų aibei. Vienoms nagrinėtoms duomenų aibėms laikas sutrumpėja 2–9 kartus, tačiau yra duomenų aibių, kurioms skaičiavimo laikas sutrumpėja šimtais kartų.
2. Projekcijos paklaidos skaičiavimas dalijant pradinę duomenų aibę į dalis leidžia sutrumpinti skaičiavimo laiką beveik 6 kartus lyginant su projekcijos paklaidos

skaičiavimo būdu, taikančių ciklą, kuriame paklaida skaičiuojama panariui kiekvienam duomenų aibės taškui. Projekcijos paklaidos skaičiavimui, kai pradinė duomenų aibė dalijama į dalis, pakanka personalinio kompiuterio operatyviosios atminties 15 kartų didesnei duomenų aibei lyginant su projekcijos paklaidos skaičiavimo būdu, kai naudojama nedalyta duomenų aibė ir speciali funkcija, pritaikyta greitam atstumų skaičiavimui.

3. Duomenų aibės vizualizavimo strategijoje pasiūlytas duomenų aibės imties sudarymo būdas išlaiko duomenų struktūrą įvairioms testinėms duomenų aibėms. Tiriant duomenų vizualizavimą be persidengimo, parodyta, kad daugiamačiams taškams vizualizuoti sklaidos diagramoje pakanka iki 1000 taškų, kad būtų atskleista bendra duomenų struktūra.

Trumpai apie autorę

Kotryna Paulauskienė Vilniaus universitete įgijo statistikos bakalauro (2003 m.) ir magistro (2005 m.) laipsnius. 2012–2017 m. studijavo informatikos krypties doktorantūrą Vilniaus universiteto Matematikos ir informatikos institute. 2003–2006 m. dirbo Lietuvos sveikatos informacijos centre vyr. statistike, nuo 2006 m. pradėjo vadovauti Mirties priežasčių registrui. 2009 – 2017 m. dirbo Higienos institute Mirties priežasčių registro vadove. Nuo 2017 m. Kotryna Paulauskienė dirba vyr. analitike Danske Bank Globalių paslaugų centre Rizikos valdymo grupėje.

Kotryna Paulauskienė

MASSIVE DATA VISUALIZATION BASED ON DIMENSIONALITY REDUCTION AND PROJECTION ERROR EVALUATION

Summary of Doctoral Dissertation
Physical Sciences
Informatics (09P)
Editor Jolita Pons

Kotryna Paulauskienė

DIMENSIJŲ MAŽINIMU PAGRĮSTAS DIDELĖS APIMTIES DUOMENŲ VIZUALIZAVIMAS IR PROJEKCIJOS PAKLAIDOS VERTINIMAS

Daktaro disertacijos santrauka
Fiziniai mokslai
Informatika (09P)
Redaktorė Jorūnė Rimeisytė