ISSN 2464-4617(print) ISSN 2464-4625(CD) CSRN 2703 Computer Science Research Notes
http://www.WSCG.eu

A new dimensionality reduction-based visualization approach for massive data

Kotryna Paulauskienė
Institute of Mathematics and Informatics,
Vilnius University
Akademijos str. 4
LT-08663 Vilnius, Lithuania
kotryna.paulauskiene@mii.vu.lt

Olga Kurasova
Institute of Mathematics and Informatics,
Vilnius University
Akademijos str. 4
LT-08663 Vilnius, Lithuania
olga.kurasova@mii.vu.lt

ABSTRACT

We live in a big data and data analytics era. The volume, velocity, and variety of data generated today require special methods and techniques for data analysis and inferencing. Data visualization tools allow us to understand the data deeper. One of the straightforward ways of multidimensional data visualization is based on dimensionality reduction and illustrated by a scatter plot. However, visualization of millions of points in a scatter plot does not make a sense. Usually, data sampling or clustering is performed before visualization to reduce the amount of the visualized points, but in such a case, meaningful outliers can be rejected and will not be visualized. In this paper, a new approach for massive data visualization without point overlapping is proposed and investigated. The approach consists of two main stages: selection of a data subset and its visualization without overlapping. The experiments have been carried out with ten data sets. The efficiency of subset selection and visualization of data subset projection is confirmed by a comprehensive set of comparisons.

Keywords

Massive data, dimensionality reduction, data visualization, data subset, visualization without overlapping.

1. INTRODUCTION

Today, big data handling is still challenging. The big data analysis helps us to do insights that lead to better decisions and strategic moves. The analysis of large amounts of data helps us to reveal hidden patterns, correlations, and other insights. Usually, the analysed real-world data is of high-dimensionality. In general, each data instance (point) is characterized by many features. In order to visualize data on a 2D or 3D space, the dimensionality needs to be reduced to two or three. Dimensionality reduction (projection) techniques extract lower-dimensional data from the high-dimensional input data [1]. The techniques map the data points from m-dimensional space to a smaller d-dimensional one ($d < m$). 2D and 3D spaces are used for data visualization. Data visualization approaches present data in a graphical form and enable people to understand the data better and deeper. Good data visualization yields better models and predictions and allows discovering the unexpected information [2], [3]. The size of data, obtained and generated, at present is huge and continues to increase every day [4]. Usually, big data is characterized by three main components: volume, velocity, and variety [5], [6]. In this paper, high-dimensional and large volume data is considered to be massive data. Visualization of a large amount of data points in a scatter plot in most cases will end with some shape, fully filled with data points, and it won't be an informative representation of the data. Human eyes have a difficulty in extracting meaningful information when the data becomes extremely large. Not many existing visualization systems are designed to present meaningful and high-quality information for human perception of big data [7]. Another problem we face is that some dimensionality reduction techniques cannot handle a large amount of data. The multidimensional scaling algorithm is usually unable to deal with large amount of points and requires much computational time [8], [9]. Thus, in this research, we suggest to visualize not a full data set but only a subset of the data. Having only a subset of high-dimensional points, the projection can be found fast and then visualized in a scatter plot [10]. Also, there are many open source and commercial data visualization tools, but in most of them dimensionality reduction methods are not implemented, visualization can be performed by two or three features or visualization is applied to aggregated (counted, averaged, min/max values) data [11], [12], [13] The following most common types of graphs can be mentioned: scatter plot, line chart, bar

chart, pie chart, histogram, infographic, and others [14]. Thus, we need a visualization approach that helps us to comprehend a large amount of data and to identify each data point location among the data.

2. VISUALIZATION APPROACH

The proposed visualization approach consists of two main stages: (1) proper selection of a data subset and calculation of the data subset projection; (2) visualization of projection of the data subset without point overlapping.

Consider a data set $X \subset R^m$ 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 $Y_L = \{y'_1, \dots, y'_l\} \subset Y_S, l \leq s < n$. The final subset for visualizing Y_L contains less data points than Y_S, since we eliminate the overlapping points. The input in the proposed visualization approach is high-dimensional points of the given data and the obtained output is the data subset points of reduced dimensionality to be visualized.

1.1. Selection of a data subset

An important task of the proposed strategy is a proper selection a data subset. A conventional way is to cluster the data and then select representatives from each cluster, but in this case, we can 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 with respect to a certain similarity measure and different from that of other clusters [15], [16]. For data subset selection the following methods can be used: simple random sampling, systematic sampling, stratified sampling, cluster sampling, etc. [17]. But in this case, not all outlying observations that can be significant in data analysis and knowledge discovery are selected. For example, only one or two outliers will be selected to the data subset because they usually are far from other observations and their density is not high. An outlier can be defined as an observation that is far distant from the rest of observations [18]. An outlier may be due to variability in the measurement or it can indicate an experimental error, but also it can be due to a chance or some natural process of the construct that is being measured. Outliers are often considered as an error or noise, however, sometimes they can carry important information about the data under investigation [19]. The aim is to select (not to lose) those points (outliers) which would be excluded if the standard sampling methods were used.

Thus, in this research, we propose a new approach for data subset selection and for massive data visualization. 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 k clusters;

Step 2: for each cluster, the sum of distances ($sumD_i$) from the cluster centre to each point is calculated; the number of points N_i, $i = 1, \dots, k$ per cluster is calculated;

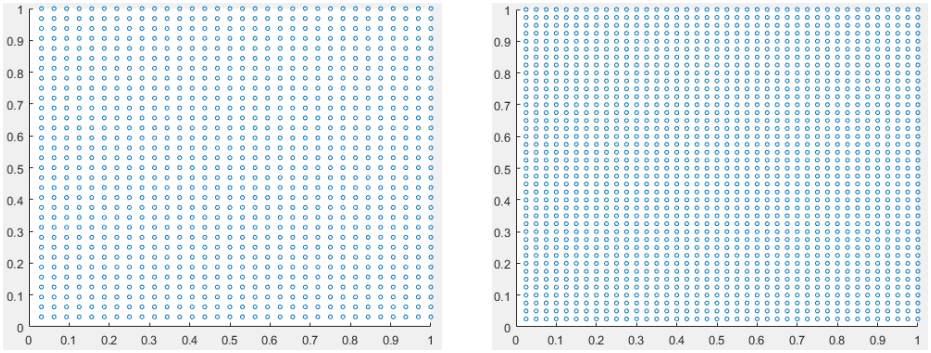
Step 3: the size s of data subset is determined, i.e. the number of points that will be selected as candidates for visualization is defined;

Step 4: the ratio ($r_i = sumD_i/N_i$) is calculated. The number of points to be selected from each cluster into the data subset is calculated by the formula $N'_i = \frac{r_i \times s}{v}$, where $v = \sum_i^k r_i$, and s – the size of data subset;

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

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

The size s of the data subset (*Step 3*) is the number of instances (points) that will be candidates for visualization. We recommend selecting $s = 1000$ as the maximal number of points. 1000 points yield a good balance between data representation and quality of the final mapping. This fact is illustrated in Figure 1. Here two-dimensional points are evenly distributed on the 1×1 rectangle. We see that the points are not overlapping. However, when the number of points is equal to 1600 (Figure 1b), they are very close to each other. When visualizing real-world data, the points are not so evenly distributed and they can concentrate in groups, and consequently do not scatter so widely and sparsely. If 1000 points of real-world data are taken for visualization, they will be more concentrated than that in Figure 1a, and it is maybe that a part of them will be overlapping. Such visual overlapping will be eliminated in the next step of the proposed visualization approach (see subsection 2.2).



a) b)

Figure 1. Visualization of points in the interval [0,1]. The size of data subset: a) $s = 1024$, b) $s = 1600$.

There may be cases that the number of points N_i' to be selected from a cluster, calculated in *Step 4*, is larger compared with the actual number of points N_i in a cluster i. In those 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 *Random* data set is analysed ($n = 2\,515, m = 10, k = 3$, the size of data subset is $s = 1\,000$). It can be seen that according to the ratio $r_3 = 4$ of the third cluster, $N_3' = 679$ points should be selected to the data subset, but the third cluster contains only $N_3 = 15$ points. Thus, we select all the 15 points from the third cluster to the data subset and increase the number of points to be selected from the remaining clusters by $\delta = 664$ points with respect to their ratios r_1 and r_2.

Cluster (i)	Number of points per cluster (N_i)	$sumD_i$	Ratio (r_i)	Number of points from cluster (N_i')	Number of points from cluster (N_i') (increased)
1	500	461	0.922	156	479
2	2 000	1 941	0.971	165	506
3	15	60	4	679	15

The number of points to be selected from 1st and 2nd clusters should be increased by $\delta = 664$ points

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

Projection of the initial data subset X_S in *Step 6* can be found by various dimensionality reduction techniques. In this paper, well known projection technique – multidimensional scaling is used to reduce the dimensionality of the initial data subset [20]. In this paper, the dimensionality of a data subset is reduced to two ($d = 2$) due to the visualization purpose.

1.2. Data visualization without overlapping

As usual, points are overlapping in a scatter plot, when their huge amount is visualized. 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 the input be a subset of points of reduced dimensionality Y_S and the output be a subset to be visualized $Y_L, s \leq l$. 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 were equal to 0, and the maximal one were 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 the results obtained;

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 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 is $l \times m$ ($l \leq s, d < m$);

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

3. EXPERIMENTAL RESULTS

To show the performance of the proposed visualization approach, some experimental investigations are carried out. 10 benchmark test data sets are visualized by the proposed approach. *Magic gamma telescope*, *Waveform*, *Wine quality*, *Letter recognition*, *Musk*, *Animals*, *Skin segmentation*, *Image segmentation*, *Yeast* data sets are taken from UCI Machine Learning Repository [21], a *Random*

data set is generated by us, where the numbers are uniformly distributed in the intervals [0, 1.0] – 1st cluster; [1.5, 2.5] – 2nd cluster; [6.0, 11.0] – 3rd cluster. Each data set analysed has some specific characteristics. Short descriptions of the data sets are presented in Table 2.

Multidimensional data sets have been clustered by a *k-medoids* method [15]. The number of clusters *k* has been determined by the Calinski-Harabasz clustering evaluation criterion [22]. To evaluate the *proposed approach* of the data subset selection, we compare it 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 [23]. In this paper, the relevant strata are identified using the *k-medoids* method.

Name	n	m	k
Random	2 515; 15 020	10	3
Magic gamma telescope	18 905	10	3
Waveform	5 000	21	3
Wine quality	3 961	11	2
Letter recognition	18 668	16	2
Musk	6 581	166	3
Animals	16 384	72	4
Skin segmentation	51 444	3	2
Image segmentation	2 086	19	4
Yeast	1 453	8	2

Table 2. Data sets (n – number of instances, m – number of features, k – number of clusters).

Table 3 shows the comparison of two subset selection methods (*stratified* sampling and the *proposed approach* of subset selection). Considering paper size limit only three data sets (*Random, Magic gamma telescope, Musk*) are analysed deeply. The columns "*Subset by the proposed approach*" and "*Stratified subset*" provide the information how many points (N_i') should be selected from each cluster to a data subset. The comparison shows that it is important how the subset is obtained, i.e. the numbers of points in clusters of subsets differ especially when *Random* and *Magic gamma telescope* data sets are analysed.

The visualization results of the *Random* 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 2. The images show that the distribution of the points of the *Random* 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 third cluster (the red points), i.e. only one point from the third cluster is selected to the data subset. 526 points of the *Magic gamma telescope* data set are selected from the second cluster (the green points) by the *proposed approach* and 194 points by the *stratified* method. The opposite situation is with the third cluster (the red points) 195 and 424 points are selected, respectively. Our *proposed approach* takes into account the density of points, the higher the ratio r_i, the more points from the cluster are selected to the data subset and vice–versa. It is obvious from Table 3 and Figure 2 that the structure of the *Musk* data subsets, obtained by the compared methods, differs insignificantly ([323; 335; 342] and [383; 354; 263]). That is why the ratio and the number of points per cluster is similar for all clusters.

The next stage of massive data visualization is elimination of overlapping points. In this paper, we have determined the *threshold* values t according to which overlapping of points can be eliminated. Figure 3 shows the comparison of two *threshold* values with ten different data sets. The size of all data subsets is $s = 1000$, and all the data subsets have been normalized. Figure 3 illustrates the number of points in the subset that finally will be visualized without overlapping with various values of the threshold. It is obvious that a smaller value allows us to obtain a smaller number of points. However, this fact depends on the particular data set. Figure 2 shows some examples of visualization of the subset points without overlapping. We recommend a *threshold* value which varies in the interval $[0.0075, 0.01]$, furthermore, the re-selection should be applied to normalized data subsets. The results have shown that the point overlapping is eliminated with $t = 0.01$ for all data sets.

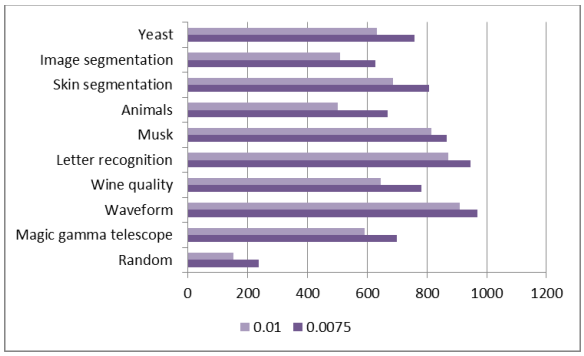


Figure 3. Comparison of two *threshold* values ($t = 0.01, t = 0.0075$) .

It should be noted that, if we are interested in a particular data set point that has not been selected as a member of the data subset and not visualized in the scatter plot, we can find its nearest neighbour in a high dimensional space and its projection in the 2D space. The position of projection of the nearest neighbour approximately shows where a certain point should be.

Data set	Cluster (i)	Number of points per cluster (N_i)	$sumD_i$	Ratio (r_i)	Subset by the *proposed approach*	Subset by the *proposed approach* (re-distributed)	*Stratified* subset
Random	1	14 000	13 410	0.96	125	468	932
	2	1 000	1 048	1.05	136	512	67
	3	20	114	5.7	739	20	1
Magic gamma	1	7 218	533 352	13 969	279	X	382
	2	3 672	511 559	26 342	526	X	194
	3	8 015	413 549	9 754	195	X	424
Musk	1	2 518	2 259 955	898	323	X	383
	2	2 332	2 172 782	932	335	X	354
	3	1 731	1 644 631	950	342	X	263

X – re-distribution of the number of points to be selected is not applied

Table 3. Selection of a data subset by two different methods.

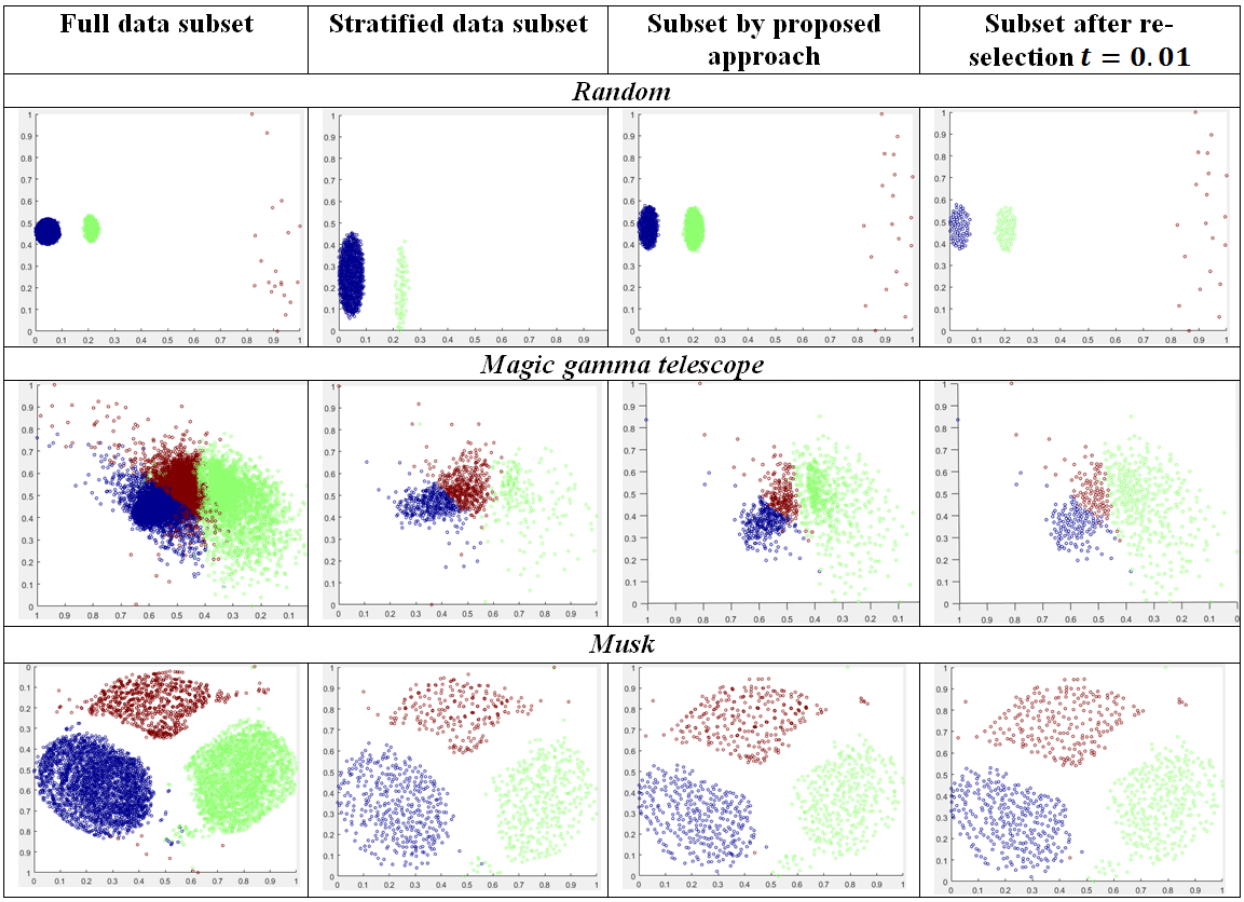


Figure 2. Comparison of scatterplot using various selections of a data subset.

4. CONCLUSIONS

In this paper, we have proposed a new strategy of massive data visualization. The *proposed approach* consists of two stages – data subset selection and visualization of the data subset without visual overlapping. The proposed data subset selection is efficient in terms of preserving outliers in a data subset. We have shown that the precisely selected subset can represent the massive data set well. The investigation results have shown that overlapping of subset points can be eliminated by applying a re-selection of points with a certain *threshold*. We have proposed two *threshold* values. Using them, the algorithm eliminates the overlapping of points. The position of a certain point that is not a member of a subset can be found by the nearest neighbour in a scatter plot. We conclude that the *proposed approach* can be applied as a new way of visualizing massive data sets.

5. REFERENCES

[1] L. P. J. van der Maaten, E. O. Postma and H. J. van den Herik, "Dimensionality reduction: a comparative review," 2009.

[2] D. Cook, E. K. Lee and M. Majumder, "Data Visualization and Statistical Graphics in Big Data Analysis," *Annual Review of Statistics and Its Application,* vol. 3, pp. 133-159, 2016.

[3] J. Bernatavičienė, D. Dzemyda, O. Kurasova, V. Marcinkevičius, V. Medvedev and P. Treigys, "Cloud Computing Approach for Intelligent Visualization of Multidimensional Data," *Advances in Stochastic and Deterministic Global Optimization. Optimization and Its Applications,* vol. 107, pp. 73-85, 2016.

[4] I. A. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani and S. U. Khan, "The rise of "big data" on cluod computing: Review and open research issues," *Information systems,* vol. 47, pp. 98-115, 2015.

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

[6] S. Sagiroglu and D. Sinanc, "Big data: A review," in *2013 International Conference on Collaboration Technologies and Systems (CTS)*, San Diego, 2013.

[7] R. Agrawal, A. Kadadi, X. Dai and F. Andres, "Challenges and Opportunities with Big Data Visualization," in *7th International Conference on Management of computational and collective intElligence in Digital EcoSystems*, Caraguatatuba, 2015.

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

[9] 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.

[10] K. Paulauskienė and O. Kurasova, "Projection error evaluation for large data sets," *Nonlinear Analysis: Modeling and Control,* vol. 21, no. 1, pp. 92-102, 2016.

[11] M. Gounder, V. Iyer and A. Mazyad, "A survey on business intelligence tools for university dashboard development," in *2016 3rd MEC International Conference on Big Data and Smart City (ICBDSC)*, Sultanate of Oman, 2016.

[12] A. Shukla and S. Dhir, "Tools for Data Visualization in Business Intelligence: Case Study Using the Tool Qlikview," *Information Systems Design and Intelligent Applications,* vol. 434, pp. 319-326, 2016.

[13] J. Miller, Big Data Visualization, Birmingham: Packt Publishing, 2017.

[14] S. G. Archambault, J. Helouvry, B. Strohl and G. Williams, "Data visualization as a communication tool," *Library Hi Tech News,* vol. 32, no. 2, pp. 1-9, 2015.

[15] J. Han and M. Kamber, Data Mining Concepts and Techniques, San Francisco: Morgan Kaufman Publishers, 2006.

[16] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognition Letters,* vol. 31, no. 8, pp. 651-666, 2010.

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

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

[19] O. Maimon and L. Rockah, Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers, Kulwer Academic Publishers, 2005.

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

[21] M. Lichman, "UCI Machine Learning Repository," University of California, Irvine, School of Information and Computer Sciences, 2013. [Online]. Available: http://archive.ics.uci.edu/ml. [Accessed 1 1 2017].

[22] 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.

[23] 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.