Enhanced Visualization of Customized Manufacturing Data

Olga Kurasova
Mykolas Romeris University
Ateities str. 20
LT-08303 Vilnius, Lithuania
olga.kurasova@mif.vu.lt

Virginijus Marcinkevičius
Mykolas Romeris University
Ateities str. 20
LT-08303 Vilnius, Lithuania
virginijus.marcinkevicius@mif.vu.lt

Birutė Mikulskienė
Mykolas Romeris University
Ateities str. 20
LT-08303 Vilnius, Lithuania
birute.mikulskiene@mruni.eu

ABSTRACT

Recently, customized manufacturing is gaining much momentum. Consumers do not want mass-produced products but are looking for unique and exclusive ones. It is especially evident in the furniture industry. As it is necessary to set an individual price for each individually manufactured product, companies face the need to quickly estimate a preliminary cost and price as soon as an order is received. The task of estimating costs as precise and timely as possible has become critical in customized manufacturing. The cost estimation problem can be solved as a prediction problem using various machine learning (ML) techniques. In order to obtain more accurate price prediction, it is necessary to delve deeper into the data. Data visualization methods are excellent for this purpose. Moreover, it is necessary to consider that the managers who set the price of the product are not ML experts. Thus, data visualization methods should be integrated into the decision support system. On the one hand, these methods should be simple, easily understandable and interpretable. On the other hand, the methods should include more sophisticated approaches that allowed reveal hidden data structure. Here, dimensionality-reduction methods can be employed. In this paper, we propose a data visualization process that can be useful for data analysis in customized furniture manufacturing to get to know the data better, allowing us to develop enhanced price prediction models.

Keywords

Data visualization, dimensionality reduction, machine learning, cost/price estimation and prediction, customized furniture manufacturing.

1. INTRODUCTION

A visualization is a powerful tool in data exploration. A human being is able to understand visually presented information much better than that shown in other forms. Visualization allows data analysts to delve deeper into the data. Good data knowledge enables developing or selecting methods for further data analysis to solve specific tasks, such as classification, prediction, etc. When solving real-world problems, usually, data are of specific nature and complex structure, thus, sophisticated methods for data visualization are needed. Commonly, the real-world data are multidimensional. So, data dimensionality reduction-based visualization methods are widely employed in order to see the general structure of the data. On the other hand, if the developed methods are integrated into a decision support system, and its users are not familiar with data mining and machine learning, visualization tools should be simple, easily understandable and interpretable. The challenge is to reconcile these two aspects of visualization methods. Moreover, the methods must best reveal the characteristics of the data being analyzed, considering the specificities of the domain. In this paper, we propose a data visualization process in order to adapt it to a specialized decision support system for an early price estimation in customized furniture manufacturing when data visualization is used by untrained experts for rapid prediction experiments.

2. CUSTOMIZED FURNITURE MANUFACTURING

Recently, customized furniture manufacturing is gaining much momentum. Consumers do not want mass-produced products but are looking for unique and exclusive furniture. Furniture manufacturing company faced with custom furniture pricing, which must be done quickly, but with sufficient accuracy. Here, machine learning techniques can be employed. The price estimation problem can be solved as a prediction problem using machine learning techniques [Kur21].

Usually, the furniture price consists of two components: the total cost of production C_p and the profit P_p. The price P_w is a sum of these components, $P_w = C_p + P_p$. The cost C_p consists of direct costs and overhead costs. Direct cost includes direct material costs, direct labor costs, and other costs. Production costs, administrative costs, the cost of disposing belong to overhead costs. Usually, the early cost estimation is the most relevant problem [Nia06]. It is a complicated and time-consuming process due to the need to evaluate many components in the early design stage when information is most limited. When the cost is estimated, the profit is added as a certain percentage of the cost.

If machine learning methods are to be used, it is necessary to have data from the domain in question. When analyzing customized furniture manufacturing data, one data item is a product (furniture), characterized by a set of various features: item measurement (length, height, width, weight, the volume of the bounding box); material data (the materials used for production and their costs); operational data (operation list and the time required to complete the operation process); labor data (much manual work, expensive machine tools are used); production time (the customer's requirement for production time); batch size (more the same products reduce the cost of one product); manufacturing complexity (a qualitative parameter indicative of the uniqueness of the item and complexity of the work). Let's denote these features by $x_1, x_2, \ldots, x_m$. They are used as independent variables if regression-based methods are used for the cost prediction. In this case, the cost is a dependent variable. Let's denote it by y.

The cost estimation by machine learning is challenging due to the very wide variety of custom furniture. The prediction accuracy is not as high as one would like. Thus, it is purposeful to apply visualization methods to delve deeper into the data using visualization output as an additional source of information for external experts. The visual analysis will help select appropriate data subsets and data features that should be used in machine learning to improve cost prediction results.

3. DATA VISUALIZATION METHODS

Data visualization is crucial in the big data era. The data visualization helps to notice the totality of the analyzed data, allows for better knowledge of the data, resulting in easier decision making [Med17]. The data visualization can be used for various purposes: initial data cognition, exploratory data analysis [Bat17], interpretation of machine learning [Cha20a], etc. This is especially important if the person working with the data is not an expert in data mining and machine learning. Data should be presented in such a form that they would easily understandable and interpreted. Many methods for data visualization are developed [Sor14], [Wan15], [Dze13]. The methods can be divided into two large groups: direct visualization and dimensionality reduction-based visualization [Dze13]. When using direct visualization methods, the data are presented in a visual form acceptable to a human being. Scatterplots, bar plots, histograms, and others are assigned to this group. Dimensionality reduction is one of the major data abstraction techniques in visual analytics. It aims to represent multidimensional data in low-dimensional spaces while preserving most of its relevant structure, such as outliers, clusters, or underlying manifolds [Sac16].

Principal component analysis (PCA) is the best known and most popular approach for dimensionality reduction. It finds a linear subspace that aims to maintain most of the variability of the data [Wan16]. PCA is a linear projection method, while multidimensional scaling (MDS) is a nonlinear one. MDS aims to find low-dimensional points such that the distances between the points in the low-dimensional space were as close to the proximities of multidimensional points as possible [Dze13]. The t-distributed stochastic neighbor embedding (t-SNE) based on non-convex optimization has become the *de facto* standard for visualization in a wide range of applications [Aro18]. The state-of-the-art t-SNE algorithm manages to create low-dimensional representations that accurately capture complex patterns from the high-dimensional space, showing them as well-separated clusters of points [Cha20b]. Recently, one more group of dimensionality reduction approaches becomes popular—autoencoder neural networks [Wan16]. They can cope with large amounts of data because deep learning strategies are applied to their training.

The problem arises how to select more appropriate visualization methods and how to organize the whole visualization process. In Fig. 1, the proposed data visualization process is depicted. In the beginning, a data set should be formed. After that, feature selection can be helpful. It can be performed manually or applying some feature selection techniques. The next step is data visualization. At first, simple direct visualization methods can be used (scatterplots, histograms, etc.). After that, more sophisticated methods should be employed. As the data to be analyzed usually are multidimensional, dimensionality reduction-based visualization methods can help dive into the hidden structure of the data. As a result, a set of visual presentations is obtained. A data analyst/decision maker should review and interpret the results obtained, i.e., to look for data properties and interrelationships that would help solve the main task of the data analysis—a classification, prediction, etc. If the obtained results are satisfactory, the final decision making can be performed, and the

process is completed. Otherwise, the process is continued selecting a subset of the data set if it turns out that the data items form groups and it is appropriate to analyze individual groups. Then the visual analysis is performed for data subsets. The steps are repeated until the final decision is made.

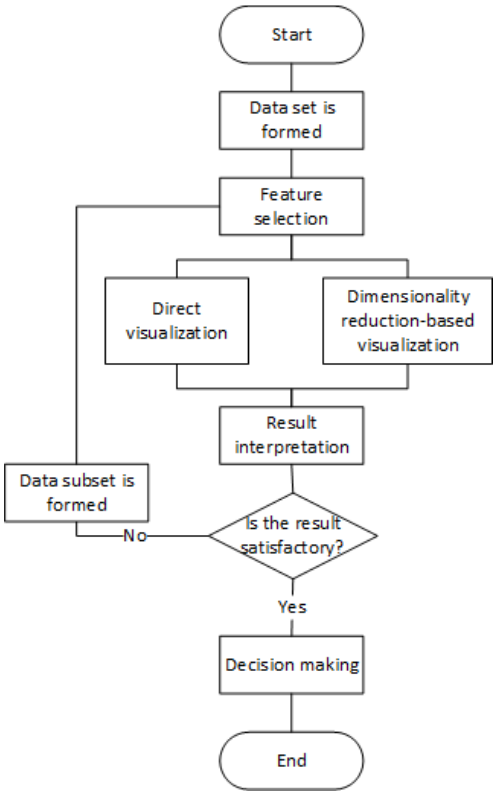

Figure 1. Data visualization process

The proposed visualization processes should be adapted to a domain to be investigated. As mentioned before, our domain is related to customized furniture manufacturing. Thus, historical data on already manufactured products should be collected. The data can be described by metal and wood processing times, various material features, number of different components, etc. A subset of these features is selected. After that, some direct visualization techniques can be used for initial explorative analysis. Scatterplots and histograms are most suitable here. In order to see the general structure of the data, the well-known principal component analysis and multidimensional scaling are irreplaceable. However, more modern methods, t-SNE and autoencoder neural networks, must be used, too. Moreover, integrating cluster analysis into visualization allows us to dive much deeper into the data and notice important insights. Here, we suggest using a specific clustering technique—Louvain algorithm [Tra19]. Suppose the results of the performed analysis on the whole data set do not satisfy the decision maker. In that case, a data subset should be selected, and the visual analysis is repeated for this data subset.

4. VISUAL ANALYSIS

The real manufacturing data for 1007 products provided by a Lithuanian furniture manufacturing company are used in the visual analysis to demonstrate the proposed visualization process. The data gathered over the last five years include the real prices of these products. A set of products includes items of various sizes and complexity (from small pieces of furniture to large furniture kits). Each product is characterized by some features $x_1, x_2, \dots, x_m$. The selected features are described in Table 1, $m = 17$. Here, the price is denoted by y. All the feature values are numerical. Thus, a data matrix can be formed. Let's denote the i^{th} product by X_i. Then we have a $d \times m$ matrix $\mathbf{X} = (x_{ij})$, $i = 1, \dots, d$, $j = 1, \dots, m$. The i^{th} row of this matrix corresponds to the vector $X_i = (x_{ij})$, $j = 1, \dots, m$, $i \in \{1, \dots, d\}$, which elements are the feature values of the i^{th} product, $d = 1007, m = 17$. The data mining software *Orange* is used for data visualization [Dem13] (https://orangedatamining.com). For training autoencoder neural network and visualizing the results obtained, the following *python* libraries are used: *tensorflow*, *keras*, *scikit-learn*, *numpy*, *pandas*, *seaborn*, *matplotlib*.

x_i	**Notation**	**Short descriptions**
$x_1 \dots x_5$	m10-m50	metal processing times
$x_6 \dots x_9$	w60-w90	wood processing times
x_{10}	m	total meters of materials
x_{11}	m^2	total square meters of materials
x_{12}	kg	total weight of materials
x_{13}	qty	total amount of materials
x_{14}	qty_ parts	number of parts
x_{15}	qty_ diff_parts	number of different parts
x_{16}	qty_ materials	number of different materials
x_{17}	qty_ order	quantities (order size)
y	price	price

Table 1. Manufacturing data features

At first, initial data exploration should be performed. Here, we confine ourselves to visual analysis. It is purposefully to see a general structure of data. Histograms or other simple direct visualization techniques are usable for this purpose. In Fig. 2, a histogram of the price is presented. We can see that the data set includes a lot of cheap products. About 700 items are cheaper than 200 Eur. Only a few expensive products are costing more than 1000 Eur. Knowing this information, the data analyst can divide

the data into several groups according to the size of the prices.

The histograms of other data features can be explored to see data similarities and dissimilarities. Due to the limited size of the paper, here, we present only a feature histogram. In Fig. 3, we see a frequency of the number of different parts of furniture. In many cases, the number of different parts is not greater than 20. Only a few pieces of furniture consist of many components.

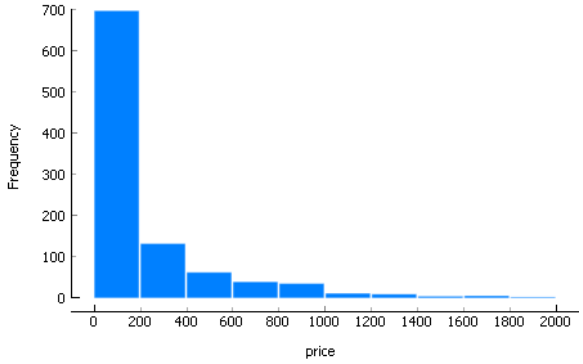

Figure 2. Price distribution

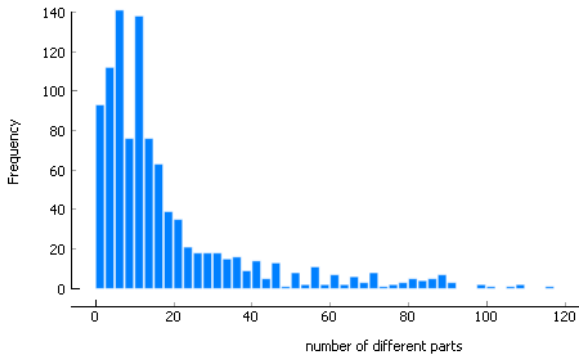

Figure 3. Distribution of different parts of furniture

As the data are multidimensional, dimensionality reduction methods can be employed for data visualization. Various approaches can be used. At first, the data dimensionality is reduced by PCA. If we want to represent the data of low-dimensionality in a 2D Cartesian coordinate system, only two first principal components (PC1 and PC2) can be used (Fig. 4). Here, a point corresponds to a product (a piece of furniture). The points are colored according to the price size. We can see that the points corresponding to cheaper products are huddled in one place. Meanwhile, the points corresponding to expensive furniture are distributed farther apart. It should be noted that two principal components explain only 53% of the variance. In order to get 80%, even seven principal components need. However, in that case, the data cannot be represented visually.

The well-known fact that PCA is a linear dimensionality reduction method. If the data are of nonlinear nature, nonlinear dimensionality reduction methods can be more suitable. Thus, the data are visualized by MDS (Fig. 5). Here the Euclidean distance is used as a data proximity measure. We can see not only accumulations of points but also data outliers. Using PCA, two yellow points were distant from other points. In the case of MDS, two yellow points remain outliers; additionally, two blue points are far away from the majority of the other points.

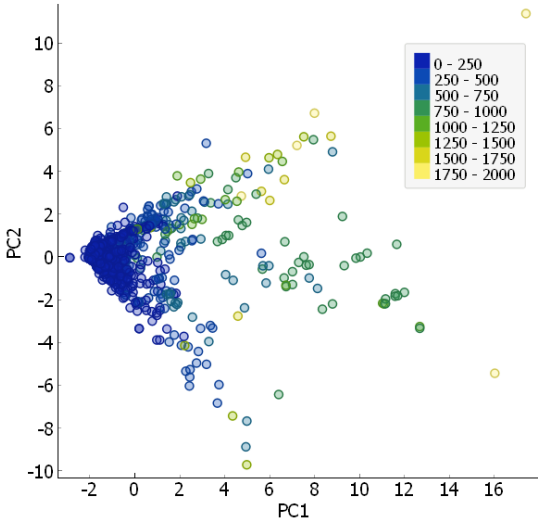

Figure 4. The data visualized by two principal components

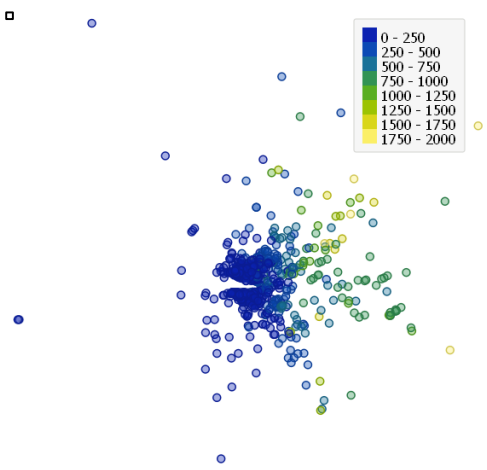

Figure 5. The data visualized by MDS

It is interesting to see how the data are distributed if the state-of-the-art t-SNE is used. The result is presented in Fig. 6. We can see that the data representation differs from Fig. 4 and 5. Here, some data clusters can be observed. One large group of the points is monitored at the bottom of the image. A smaller cluster is obtained at the top of the image. It should be noted that the points corresponding to more expensive products (green and yellow points) form

some clusters. Meanwhile, in the cases of PCA and MDS, no such clusters were obtained. It is worth noting that no clustering method was used here. We interpret cluster formation only by observing the image.

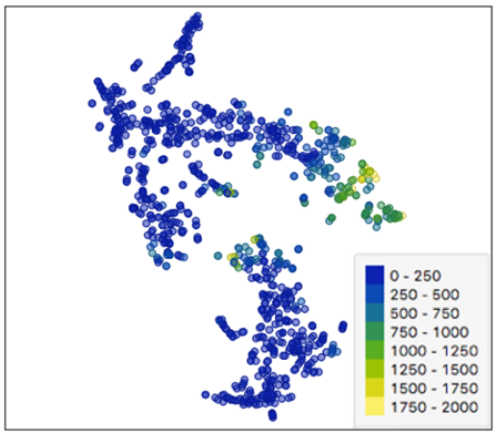

Figure 6. The data visualized by t-SNE

However, clustering methods can also be employed. Data clustering is commonly used in visual analysis. In this investigation, we use a specific clustering technique—Louvain algorithm [Tra19]. It is a hierarchical clustering algorithm that recursively merges communities into a single node and executes the modularity clustering on the condensed graphs. Originally it is proposed for community detection, but it can be used to solve other clustering problems. The furniture data also be clustered by the Louvain algorithm, and after that, they are visualized by t-SNE (Fig. 7). Here points are colored according to the clusters to which they are assigned. The Louvain algorithm automatically identifies six clusters. It is purposeful to see how clusters are related to the price size. Let's take the points of cluster C6 (magenta) and represent them in a scatter plot (Fig. 8). The points are colored according to the price size. We can see that cluster C6 consists of the points corresponding to the furniture cheaper than 250 Eur. In Fig. 8, a few points fell outside of their cluster. It is necessary to review these products and look for the reasons for this. The reasons can be different: some products can be priced in an unusual way, there are inaccuracies in fixing the feature values, etc.

In Fig. 9, the data are visualized by an autoencoder neural network. We can see that the points (blue, orange, and green) corresponding to cheaper products are huddled close together; meanwhile, the points corresponding to more expensive products (red and magenta) are spread far apart.

It should be noted that the set of the analyzed products includes items of various sizes and complexity. There are a lot of small pieces of furniture, but large furniture kits are also included. It is purposeful to perform a visual analysis of the data subsets. Let's take a subset of the products, the price of which is between 100 and 500 Eur. The data of these products are visualized by t-SNE (Fig. 10). Here the clusters are obtained by the Louvain algorithm. We can see three large clusters.

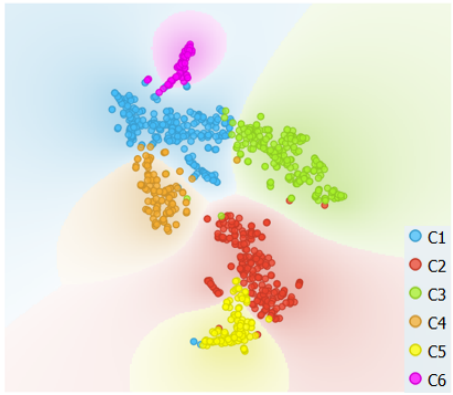

Figure 7. The data clustered by Louvain algorithm and visualized by t-SNE

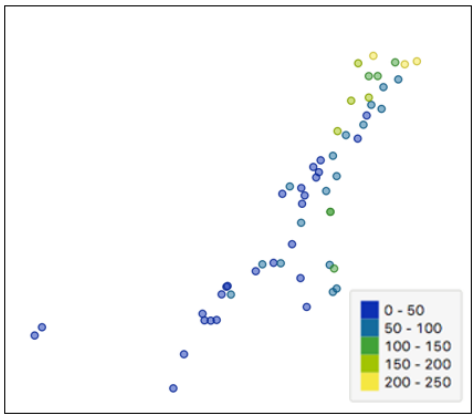

Figure 8. Representation of the cluster C6

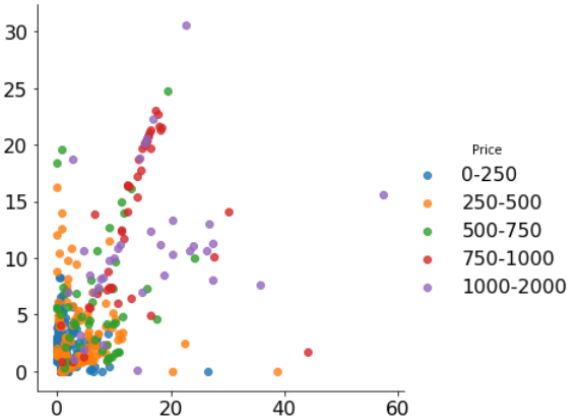

Figure 9. The data visualized by autoencoder neural network

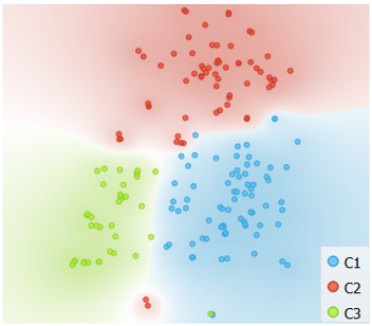

Figure 10. The data subset clustered by Louvain algorithm and visualized by t-SNE

5. CONCLUSIONS

The paper aims to highlight the benefits of visualization in solving data mining tasks. Visual analysis is performed in a very specific domain—customized furniture manufacturing. Here, the main task is to predict the prices of the products before a designed phase. We have proposed a data visualization process that included various visualization approaches that would be useful in explorative data analysis. The approaches will allow decision makers to know the data better and, as a result, will be able to estimate/predict more accurate product prices. Moreover, the proposed process will be implemented into a decision support system. This system is designed for managers and constructors of furniture companies who want to quickly and conveniently evaluate the received individual furniture order and provide the preliminary price of the products. In the system, a machine learning-based module is integrated for price prediction, too. The visual analysis module will enhance this decision support system and ensure its usability and efficiency.

6. ACKNOWLEDGMENTS

This project has received funding from European Regional Development Fund (project No 01.2.2-LMT-K-718-01-0076) under grant agreement with the Research Council of Lithuania (LMTLT).

7. REFERENCES

[Aro18] Arora, S., Hu, W., Kothari, P. K. An analysis of the t-SNE algorithm for data visualization. In Proceedings of the 31st Conference on Learning Theory, PMLR, vol. 75, pp. 1455-1462, 2018.

[Bat17] Batch, A., Elmqvist, N. The interactive visualization gap in initial exploratory data analysis. IEEE Transactions on Visualization and Computer graphics, vo. 24, no. 1, pp. 278-287, 2017.

[Cha20a] Chatzimparmpas, A., Martins, R. M., Jusufi, I., Kerren, A. (2020). A survey of surveys on the use of visualization for interpreting machine learning models. Information Visualization, vol. 19, no. 3, pp. 207-233, 2020.

[Cha20b] Chatzimparmpas, A., Martins, R. M., Kerren, A. t-viSNE: interactive assessment and interpretation of t-SNE projections. IEEE Transactions on Visualization and Computer Graphics, val. 26, no. 8, pp. 2696-2714, 2020.

[Dem13] Demsar, J., Curk, T., Erjavec, A., Gorup, C., Hocevar, T., Milutinovic, M., Mozina, M., Polajnar, M., Toplak, M., Staric, A., Stajdohar, M., Umek, L., Zagar, L., Zbontar, J., Zitnik, M., Zupan, B. Orange: Data Mining Toolbox in Python, Journal of Machine Learning Research, vol. 14, pp. 2349−2353, 2013.

[Dze13] Dzemyda, G., Kurasova, O., Žilinskas, J. Multidimensional data visualization: methods and applications. New York: Springer, 2013.

[Kur21] Kurasova, O., Marcinkevičius, V., Medvedev, V., Mikulskienė, B. Early cost estimation in customized furniture manufacturing using machine learning. International Journal of Machine Learning and Computing, vol. 11, no. 1, pp. 28-33, 2021.

[Nia06] Niazi, A., Dai, J. S., Balabani, S., Seneviratne, L. Product cost estimation: Technique classification and methodology review. Journal of Manufacturing Science and Engineering, vol. 128, no. 2, pp. 563-575, 2006.

[Med17] Medvedev, V., Kurasova, O., Bernatavičienė, J., Treigys, P., Marcinkevičius, V., Dzemyda, G. A new web-based solution for modelling data mining processes. Simulation Modelling Practice and Theory, vol. 76, pp. 34-46, 2017.

[Sac16] Sacha, D., Zhang, L., Sedlmair, M., Lee, J. A., Peltonen, J., Weiskopf, D., Keim, D. A. Visual interaction with dimensionality reduction: A structured literature analysis. IEEE Transactions on Visualization and Computer Graphics, vol. 23, no. 1, pp. 241-250, 2016.

[Sor14] Sorzano, C. O. S., Vargas, J., Montano, A. P. A survey of dimensionality reduction techniques, 2014. arXiv preprint arXiv:1403.2877.

[Tra19] Traag, V. A., Waltman, L., van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. Scientific Reports, vol. 9, no. 5233, 2019.

[Wan15] Wang, F., Sun, J. Survey on distance metric learning and dimensionality reduction in data mining. Data mining and knowledge discovery, 29(2), 534-564, 2015.

[Wan16] Wang, Y., Yao, H., Zhao, S. Auto-encoder based dimensionality reduction. Neurocomputing, vol. 184, pp. 232-242, 2016