How to evaluate a subspace visual projection in interactive visual systems? A position paper

Lydia Boudjeloud-Assala*

Université de Lorraine
CNRS, LORIA, F-57000 Metz, France

Abstract

This paper presents a position paper on subspace projection evaluation methods in interactive visual systems. We focus on how to evaluate real information rendered through the visual data projection for the mining of high dimensional data sets. To do this, we investigate automatic techniques that select the best visual projection and we discuss how they evaluate the projections to help the user before interactivity. When we deal with high dimensional data sets, the number of potential projections exceeds the limit of human interpretation. To find the optimal subspace representation, there are two possibilities, the first one is to find the optimal subspace which reproduces what really exists in the original data: getting the existing clusters and/or outliers in the projection. The second possibility consists in researching subspaces according to the knowledge discovery process: discovering novel, but meaningful information, such as clusters and/or outliers from the projection. The problem is that visual projection cannot be in adequation with the subspaces. In some cases, the visual projection can show some things that do not really exist in the original data space (which can be considered as an artifact). The mapping between the visual structure and the real data structure is as important as the efficiency and accuracy of the visualization. We examine and discuss the literature of Information visualization, Visual analytic, High dimensional data visualization, and interactive data mining and machine learning communities, on how to evaluate the faithfulness of the visual projection information.

Index Terms: Human-centered computing—Visualization—Visualization application domains—Visual analytics; Human-centered computing—Visualization—Empirical studies in visualization

1 Introduction

The curse of dimensionality phenomena as defined by Belmann [3] appears when analyzing and organizing data in high-dimensional spaces but it deso not occur in low-dimensional settings. When the dimensionality increases, the volume of the space increases so fast that the available data becomes sparse. The human ability to model the visual space is limited to just three dimensions. To overcome the curse of dimensionality, the common approach is to apply dimension reduction methods such as PCA [22] or feature selection methods (dimension, attribute selection). When data is projected in the reduced space, generally it is difficult to interpret the visualization and to evaluate the truthfulness of the visual information. Data dimensionality is a major limiting factor. Finding relations, patterns, and trends over numerous dimensions is, in fact difficult, because the projection of n-dimensional objects over two dimensional spaces carries necessarily some form of information loss.

Techniques like principal component analysis (PCA) [22], multi dimensional scaling (MDS) [23] and t-Distributed Stochastic Neighbor Embedding (T-SNE) [45] offer traditional solutions by creating data embedding that try to preserve distances as much as possible in the original multi dimensional space in the two dimensional space projection. However, in terms of interpretation, these techniques assessed several drawbacks. Indeed, it is difficult to interpret the observed patterns in terms of the original data space. The dimensionality reduction or feature selection methods do not solve some others problems, that we can expect in high dimensional data. For instance, there can be different views of the data set where the same data points might be grouped differently in different subspace perspectives [40]. Indeed, some data points can belong to one cluster in one subspace and belong to some other clusters in other subspaces, or they can be considered as outliers in a different subspace. In such scenarios, the clustering algorithms as well as the visual analysis based on the whole data space may fail. If we introduce interactivity during this step, how can the user interpret correctly what he or she visualizes?

In this paper, we systematically examine and discuss the literature of Information visualization, Visual analytic, High dimensional data visualization, Interactive Data mining and machine learning communities, and we classify them according to the whole data space exploration approach (about 50 papers). Indeed, there are two possibilities to investigate the whole data space. The first possibility is to find the reduced space which reproduces what really exists in the original data. In this case, the knowledge of the whole data set structure guides the search of the subspace projection or the feature selection. For example, this can be carried out by searching the subspace representation for the existing clusters or outliers. The second possibility applies when we do not know what exists in the original data space. In this case, the approach is to find some interesting reduced spaces according to some optimal subspace structures, where we can extract important information (discover new clusters, outliers, . . .) In this case, the data structure in the subspace guides the search, and it can be useful for the multi-view clustering (When multiple sets of features are available for each individual object), or subspace clustering, for instance. These possibilities can reproduce spaces that show things that do not really exist in the original data space, which can be considered as an artifact.

The mapping between the visual structure and the real data structure is as important as the efficiency and the accuracy of the visualization, and these two criteria (efficiency and accuracy) can be evaluated on the whole data set space and on the subspace data set. One promising approach is the work of Tatu et al. [39], where they see if there is a correlation between what the human perceives and what the machine detects. In fact, there are no user studies able to inspect the relationship between what the human can detect and what the machine can detect as data patterns. In the related works, visual quality metrics have recently been introduced to automatically extract interesting visual projections out of a large number of available candidates to explore high dimensional data sets. For instance,

*e-mail: lydia.boudjeloud@loria.fr

theses metrics permit the user to search within a large set of scatter plots (in a scatter plot matrix) and select the views that contain the best separation between clusters. We think that this kind of metrics can help the user to trust the visualization, and help him interact correctly with the process. We describe the related work in the fields of information visualization, visual quality metrics, high-dimensional data and interactive machine learning and then we propose a discussion before concluding. We try to follow a framework to understand the given hypothesis of evaluation: getting the same information (clusters, outliers, . . .) in the projection and/or getting novel, but meaningful, information from the projection in each field.

2 Information visualization

Information visualization research can be divided into three categories: basic or foundational work, transitional approaches to create and refine techniques, and application-driven efforts [21]. We focus on how to measure and evaluate the effectiveness of various proposed approaches in the information visualization community, according to the framework that we propose. Indeed, we propose a classification for different methods through the information that they get, how to evaluate it and if they use or not the whole date space.

Based on the whole data information

In this proposal the abstracted data set and projection is based on the knowledge of whole data set and the known repartition of data points. However, authors did not consider the perceptual issues, which are very dependent on the used visualization tool. Three *measures of representativeness* when using filtering, sampling, clustering, and summarizing, are developed in 2006 by Cui et al [11], to reduce the data point number and to be the most accurate, in visualization. These measures are based on histogram comparisons, nearest neighbor computations, and statistical data properties. The authors define two measures, *Data Abstraction Level* [11] and *Data Abstraction Quality* [11]. To evaluate and validate these measures, the authors propose an interactive tool where each view of the data generates quality measures, they use bar charts to display the data abstraction level, nearest neighbor measure and normalized histograms. Therefore, analysts can control the abstraction quality, with a compromise between relative data density, the degree to which outliers are preserved, response time, display clutter, and information loss.

One of the obvious techniques to measure information loss is to use the *Entropy* during the visualization process rather than in the total information content of the data set. Moreover, there are several techniques commonly use for data transformation in visualization that provide an implicit measure of information loss, as Purchase et al. [31] pointed in 2008, where they try to present some theoretical foundations from an information visualization point of view.

For example, Multi Dimensional Scaling (MDS) [23], a process commonly used for dimensionality reduction, provides a *measure of Stress*, which captures the difference between the distances between points in the original dimensioned space and the corresponding distances in the subspace. When using Principal Component Analysis (PCA) [22] to perform this reduction, the loss can be measured from the dropped components and the part of the *restituted data Inertia*. Similarly, Stochastic Neighbor Embedding (SNE) [46] is equivalent to minimizing the *mismatch between Squared Distances* in the two spaces, and the loss of information can be computed based on this difference. It is interesting to study the proposed measures and see how efficiently they can be used to evaluate subspace data projection without prior knowledge of the data repartition and if they can be evaluated.

Discovering meaningful information

The second classification concerns measures that evaluate or conduct the subspace data projection without prior knowledge of the data repartition. Generally, the proposed methods are based on *Stress measures* and focus on the embedding, which coincide with mapping that minimizes the error in target space. These methods are generally based on iterative optimization where they minimize the *Stress error* and then propose the optimal projection. Some methods propose to use *Proximity Relationships* on a low dimension space such as those based on graph [15].

The *Local graph modeling* idea is to divide the data into small subspaces and to propose a local optimal projection of the data. For example, *Locally Linear Embedding* (LLE) [34] models the data by extracting local patterns with *intrinsic geometry*. The *Local Intrinsic Geometry* has a property according to which it stays unchanged under transformations like translation, rotation or scaling. Hence, the local linear relationships of points in data space can be used to evaluate target subspace projection. Similar to LLE, Piecewise Laplacian-based Projection (PLP) [28] makes the assumption that every data point can be approximated through a convex combination of its neighbors. To evaluate the local solution and the attachment of the different local solutions, the PLP method use the *Stress-based Force Scheme* [42].

To evaluate and validate the proposed approach, the authors propose an interactive tool where the user can interact with the projected data set through its representation as a k-nearest neighbor graph and adjust neighborhoods or samples by simply moving data points within the embedding. Due to the local subspaces optimization and the local patches between different spaces being randomly chosen, there is no guaranty that global features can be preserved, the authors try to make up for this by introducing user interaction.

These two approaches are based on minimizing a neighborhood function which can find optimal local solution but can also produce artifact in the projection. This artifact cannot be detected or evaluated unless we use the interactive tool and the user's knowledge.

Discussion

Although, several methods are proposed in the field information visualization, it is difficult to define effective visualization and how to measure it. There is no universal definition of visualization effectiveness. Most of the existing definitions are incomplete and only focus on one aspect of effectiveness. The existing research suffers from the lack of a theoretical framework and they have deeply affected the design and evaluation of visualization.

There is another major problem facing user studies today which is the lack of standard benchmark databases, benchmark tasks, and benchmark measures [50]. The user study procedures have not been standardized. There has been some progress in this area, for example, the Information Visualization Benchmarks Repository [30] has been established but limited to IEEE challenges from 2003 at 2006. However, we can refer to associated web sites of the different IEEE Vast challenges. More importantly, fully annotated benchmark databases for major application areas of information visualization, such as computer security and bioinformatics (biovis conference challenge), are needed. In addition, benchmark task specifications, standardized user study procedures, as well as baseline measures need to be developed.

According to the retrospective study of the evaluation practices of Isenberg et al. [19], it is certain that, there is a lack of mathematical measures to evaluate methods in the area of visual information. Generally, authors focus on the user or participant tests, and evaluate the methods by asking the viewer of the resulting visualization questions. As mentioned by Isenberg et al. [19], only 46% of all papers were evaluated by the authors [19] using the qualitative result inspection scenario. It consists in asking the user to agree to a proposed model or visualization tool results by inspecting a proposed visualization. It was followed by 35% of papers using Algorithm performance to evaluate their proposition. Less than 5% of the reviewed papers used User performance, Understanding environment practices, Visual data analysis, Evaluating Communication through visualization and Evaluating collaborative data analysis scenarios, which concern data analysis process.

We understand that, generally, the community of information visualization gives great importance to the visual evaluation and do not go any further in assessing the faithfulness of the visual information restituted. The risk of visual evaluation is to obtain information that cannot exist in the whole data space which can be considered as an artifact.

3 VISUAL QUALITY METRICS

The visual quality measures are developed to help the user in the interactive methods, visual clutter and data abstraction in ergonomic graphic [4, 5, 29, 38, 44].

In 2011, Bertini et al. [6] provide an overview of techniques that use quality metrics to help and find meaningful patterns in high-dimensional data according to the visual exploration. In their survey, they focus on the different metrics and how they conduct different steps of the information visualization process. It seems that these measures are also based on the pixels and colors of the image space obtained by the different visualization methods, either according the initial information in the whole data set or according the image visual space.

Based on the whole data information

Among measures that use the whole data set information, we can mention the *Lie Factor* introduced by Tufte [44] in 1982, and the *Visual Clutter Measures* for parallel coordinates, scatter plot matrices, star glyphs and dimensional stacking proposed by Peng et al. [29] in 2004. They use these measures to provide the best dimension orders with low visual clutter. These measures are based on the total number of outliers between neighboring dimensions for the parallel coordinates technique. For the scatter plot matrices the proposed measure focuses on finding structure in plots rather than outliers, and is based on the correlation between two dimensions. To reduce the clutter using star glyphs, the proposed measure is based on minimizing the total occurrence of unstructured rays in glyphs. Finally, the clutter measure for dimensional stacking is the proportion of occupied bins aggregated with each other versus small isolated bins. These measures and the proposed reordering dimension algorithms require a high computational time for small data sets with fewer than 10 dimensions.

With the same idea to measure the clutter of data abstraction, and get the model to measure it with visual density in two dimensional scatter plots, Bertini et. al. [4] in 2004 develop a *Clutter Measure* that represents the percentage of colliding pixels of all possible permutations. This measure is similar to the *Histogram Difference Measure* developed in 2006, by Cui et al. [11], who developed measures of representativeness in the visualization when using different visual operations to reduce the data point number. In addition to *Histogram Density Measure* [11], Tatu et al. [38] in 2009, define *Class Density Measure*, *Similarity Measure* and *Overlap Measure* on classified data, based on the pixels and colors of the image obtained by the visualization. Tatu et al. [38] propose to rank visualizations based on features, according to a specified user task.

Discovering meaningful information

In the visual quality metrics fields, it is difficult to find measures that are not based on the whole data information. It can be explained by the need to evaluate the concordance with some existing and known information. Very few measures to our knowledge are developed to discover new information without being based on the whole data information. We can cite *Rotating Variance Measure* and *Hough Space Measure* [38], developed for unclassified data, without any information on the data. It is defined to find linear or non-linear correlations and clusters in the data sets, respectively. These measures are based on the pixels and colors of the image space obtained by the visualization. This approach provides a number of potentially useful candidate visualizations, which can be used as a starting point for interactive data analysis.

Discussion

Among the most promising work in this area is the work of Tatu et al. [38] in 2009, where they propose automatic analysis methods to extract potentially relevant visual structures from a set of candidate visualizations. They present measures for Scatter Plots and Parallel Coordinates visualization methods, for unclassified data, without any information on the data, as well as classified data information. To evaluate how the visual cluster detection of the user is correlated with series of selected metrics, Tatu et al. [39] in 2010 propose a user's evaluation. The authors evaluate the correlation between the scores of the selected visualization with the score obtained by the selected quality measures. This approach may provide an answer to the questions we ask. It can be used to evaluate what is restituted by the visualization according to what the measure provides. Different quality metrics are proposed to automate the demanding search through large spaces of alternative visualizations, allowing the user to concentrate on the most promising visualizations suggested by the quality metrics.

Bertini et al. [6] provide a good state of the art of quality metrics used in high dimensional data analysis, and try to show how the different proposed metrics are applied on the different steps of the data analysis and visualization process, including interactivity with the process. A principal focus application of visual metrics presented above is to apprehend the high dimensional data, using classical visualization methods such as parallel coordinates and scatter plot matrices, and many papers focus on this problematic [5, 6, 38, 39].

4 HIGH DIMENSIONAL VISUALIZATION

High dimensional data sets contain hundreds of variables (attributes, dimensions), that are difficult to explore. One of the consequences is that the traditional visualization methods cannot represent effectively this kind of data. One solution consists in employing dimensionality reduction prior to visualization. Numerous dimensionality reduction methods are available, and many approaches are introduced to evaluate the projected data.

High dimensional data spaces analysis consists in combined features measured with different properties. In some cases, the relationships between the different properties may not be clear to the user, but these properties can be revealed in the appropriate dimension projection or combination. It is often not sufficient to see different data properties when we take only one subspace.

However, different subspaces may show complementary, conjointly, or contradicting relations between data items and data properties. The whole data set information may remain embedded in sets of subspaces. For a large number of candidate subspaces, they apply hierarchical grouping and filtering to obtain a smaller set of interesting groups of subspaces for interactive analysis.

Based on the whole data information

In 2007, Aupetit [2] proposes to visualize any measure associated to a reference projected plan or to a set of projected data, by coloring the corresponding Voronoï cell in the projection space, in order to evaluate the faithfulness of the visualization of continuous multi-dimensional data, based on their projection into a two dimensional space. The author tries to say where the high-dimensional manifolds have been modified through a reduction or a projection and tries to evaluate how faithful the projection is to the original data. The proposed approach is specific to one type of the projection (SOM: self organizing map method). It is difficult and costly to apply this approach to other projection methods (PCA, MDS, T-SNE) or visualization methods generally applied to high dimensional data sets (parallel coordinates or scatter plot matrices).

Some measures are based on a *Similarity Function* defined on subspace pairs according to two main criteria that are the *Overlap* of the sets of dimensions that constitute the respective subspaces, and *Resemblance* in the data topology given in the respective subspaces. As is presented in Sedlmair et al. [37] with the visual interactive system for subspace based analysis in high dimensional data. They use the *Tanimoto Similarity* [33] on the contained dimensions in the respective subspace. They also, compare subspaces with regard to the distribution of their data using *Similarity Measure*, which is very close to the *Clusters Stability* concepts [47], to evaluate clustering in data mining.

Discovering meaningful information Tatu et al. [40], propose in 2012 another method for the visual analysis of high dimensional data in which they employ an interestingness guided subspace search algorithm to detect a candidate set of subspaces. They introduce the *Subspace Similarity Function*, they visualize the subspaces and provide navigation facilities tools to explore interactively large sets of subspaces. We can compare and relay subspaces respecting the involved dimensions and clusters with this approach.

Few reduction approaches take the importance of several structures into account and few provide an overview of the structures existing in the high dimensional data set. For an exploratory analysis, as well as for many other tasks, several structures may be interesting. Exploration of the whole high dimensional data set without reduction may also be desirable [16].

Measuring and evaluating subspace clustering results is not trivial due to the different information contained in subspace clustering results such as subspaces, number of objects in cluster, and overlapping between subspaces and/or clusters. In this topic, an interactive data analysis and visualization tool for subspace clustering, ClustNails [41], is introduced according to the subspace clustering tasks to deal with high dimensional data sets, using these different measures. They use the *Tanimoto Similarity* [33] on the contained dimensions in a respective subspaces. They also, compare subspaces with regard to their data distribution using the *Similarity measure* to compare between them.

Discussion Automated methods are employed to analyze dimensions, using a range of quality metrics, providing one or more measures of interestingness for individual dimensions. Through ranking, a single interestingness value is obtained, based on several quality metrics, these measures provided by statistical data exploration, and industrial data analysts, such as *entropy*, *correlation*, *variance*, *skewness*,....

Generally, the methods propose an interactive environment where the user is provided with many possibilities to explore and gain understanding of the high dimensional data set. Guided by this, the user can explore the high dimensional data set and interactively select a subset of the potential most interesting variables.

To guide the exploration, these approaches use the real information existing in the whole data set to find subspace data projection and exploration. Very few methods take into account the existing information in the subspace data projection and evaluate it, the evaluation is generally left to the user, visually.

5 Interactive data mining and machine learning

Our objective is to help determine whether what the eyes can see really exists, and whether it requires user intention or if it is just a visual artifact. In the data mining literature, there are measures that are used to evaluate the subset of dimensions according to what really exists in the original dimensional space, according to the classification results both in supervised and unsupervised classification (clustering) [17,27]. But, they do not specify whether what exists in the subspace is the reflection of the structure that really exists in the original space.

We want to present and discuss measures that inform signification, if we lose information in the subspace or generate structures that can be considered as artifact. We cannot do this survey without addressing the issue of measures that are introduced in the data analysis, statistics, data mining and machine learning communities, to evaluate their own supervised or unsupervised classification problems, particularly those dedicated to interactivity.

Based on the whole data information For supervised classification, the evaluation is done on classification data sets where the class labels are known, and this information is considered as the truth against which the different methods are compared and evaluated. There are also many *Internal* and *External* measures generally known as *clustering validity indices* to evaluate clustering results. The *internal clustering quality measures* are based generally on sums of square distances to cluster centers or ratio of between-cluster to within-cluster similarities. The *internal clustering quality measures* are based on comparison and evaluation with classified data sets with known class labels.

Desgraupes [13] developed an *R package* that includes all recent internal and external measures. For the outlier detection problem, Aggarwal [1] provide a large overview of the literature, such as linear methods, proximity-based methods, subspace methods, and supervised methods; with data domains, such as, text, categorical, mixed-attribute or time-series. Schubert et al. [36] propose a measure for comparing and ranking *outlier scores* and discuss about the relationship and differences to typical ranking evaluation measures.

Is generally considered as unsupervised problem, the proposed methods are evaluated on similarity and redundancy of existing outlier in the whole data. In particular, this measure provides for the first time the means to select members of an ensemble for outlier detection. But it does not indicate the variability of the threshold that help us to declare a data point as an outlier neither the difference between group outliers and cluster.

Discovering meaningful information For the unsupervised classification (clustering), several methods are proposed. For instance, to evaluate the *stability* of the clustering algorithm, the same clustering algorithm is applied repeatedly to perturbed versions of the original data. Then a *stability* score is computed to evaluate if the results of the algorithm are stable or unstable. If the results are unstable, the algorithm is considered as unsuitable to use. U. von Luxburg et al. [48] provide a large overview of the literature on clustering stability.

Nevertheless, most of these evaluation measures evaluate the complete clustering result and not each cluster separately. Only two criteria, the *Wemmert-Gançarski measures* [12] provide such evaluation, for each cluster separately. However, this measure is based on the distances to other cluster centers. The *intra-class inertia* also provides an individual cluster evaluation, but is biased by the cluster's size and the variance of the dimensions.

To avoid these different problems, some answers were given by introducing measures to evaluate an individual cluster [7, 14]. These measures concern only clustering and cluster evaluations, they do not take into account the overlapping clusters, the subspace projection of cluster reliability or existing outliers. we can mention for example, the work of [8] that use this kind of measures.

In 2016, [8] propose a semi-interactive system for visual data exploration using an iterative clustering that combines an automatic approach with an interactive one. They propose a framework to improve the interactivity between the user and the data analysis process, allowing him or her to participate actively in the iterative clustering tasks using a two-dimensional projection. Defining a cluster by its seed (center) and its limit, the proposed approach allows the user to modify the automated values or to define new seeds and the associated cluster limit himself or herself. The user can evaluate the obtained cluster based on the evaluation measure and can also choose to let the automated approach find optimal seeds and then interact with the process to iterate the clustering process according to his or her visual perception and domain knowledge.

In 2015, Bruneau et al. [10] describe Cluster Sculptor, a novel interactive clustering system that allows a user to iteratively update the cluster labels of a data set, and an associated low-dimensional projection. The system is fed by clustering results computed in a high-dimensional space, and uses a 2D projection, both as support for overlaying the cluster labels, and to make for engaging user interaction. The user can inject his or her domain of knowledge progressively, crafting an updated 2D projection and the associated clustering structure that combines his or her preferences and the manifolds underlying the data. Via interactive controls, the distribution of the data in the 2D space can be used to amend the cluster labels, or reciprocally, the 2D projection can be updated so as to emphasize the current clusters.

In order to assist the user in better understanding and utilizing PCA, the iPCA system [20], developed in 2009, visualizes the results of principal component analysis using multiple coordinated views and a rich set of user interactions. The iPCA system [20] allows the analyst to re-position a point in the 2D projection, and see how other values change. These interactions can be useful for revealing relationships in the data that might otherwise not be recognized. The proposed approach is based on the Correlation view, on coefficient and on the relationships (scatter plot) between each pair of variables that are presented to the user.

Brown et al. [9] propose an interactive system, *Dis-Function*, that allows a data expert to interact directly with a visual data representation to define an appropriate distance function. This system allows the user to move incorrectly-positioned data points to locations that reflect his or her understanding of the similarity of those data points relative to the other data points.

The EvoGraphDice [43] presents a prototype to modify the characteristics that are not visible in views based on the primary set of dimensions using an interactive evolutionary algorithm. The fitness evaluation of evolutionary algorithms evaluates the suggested visualization taking into account user interactions and internal metrics. The user interaction criterion tries to adapt user preferences in the fitness function while the internal metrics evaluate the relations between variables.

Discussion Finally, depending on the context, and on what the user or the data specialist expects, several approaches can be used. Often, it is very hard to quantify mathematically the faithfulness of visual projection. It is important to take into account the users constraints to evaluate and model the problem with user (analyst) centric perspective in order to get meaningful truthfulness evaluation.

Consequently, an interesting future work is to discover or set up descriptions of key indicators, or summaries of key milestone results in the model structure and thus allow users to play an important role in the improvement of the performance at each iteration of the IML system.

Another interesting challenge for future work is to better illustrate the logic of the model and the decisions made. Recently, some proposals have been made for this purpose [24, 25]. We can cite, for example, the work of [24] who developed a probabilistic program induction algorithm. They propose simple stochastic models to represent concepts, related to each other by parts, or subparts and by spatial relations. The authors demonstrated that their algorithm achieves reasonable human (user) performance on a point classification task, while outperforming recent deep learning approaches. However, for tasks where learning data is abundant, such as speech recognition, less explainable deep learning approaches always outperform the algorithm. There is still a long way to go to develop more explainable models for this type of task [26].

6 Conclusion

We start out with the assumption that there are two possibilities to search the subspaces and how they are evaluated. Our selection of evaluations in this paper is not exhaustive, and is limited by our own work, our knowledge of the field and our personal experience in result evaluation. We review the literature on different communities: information visualization, visual analytic, high dimensional visualization and interactive data mining and machine learning systems to try answering the question: how to evaluate the truth in the subspace projection for interactive systems?.

Generally they compare the known information in the whole data set to find the optimal subspace projection, which is not the absolute truth. There may be an optimal subspace in another structure different from the one that exists in the original space. We believe in the complementarity of the mathematical criterion and visual evaluation, that allows the user or the data specialist to evaluate the truth of what he sees.

In the literature, it seems the user constraints of the user are not taken into account enough in the learning process of evaluating the truthfulness of the visualization. We believe only the cooperation of multiple notions from the different research fields cited before can be used in the exploration of big and massive data sets without any knowledge of the data.

We want to consider only the evaluation aspect of the results that can be done automatically through a mathematical criteria or a comparison. But as we have seen it in von Luxburg et al. [49], this problem exists also in clustering evaluation. This discussion can serve also as regards multi-view clustering problems [18], and eventually for dynamic clustering to deal with data stream.

Finally, visual analysis is very useful for the experts in order to correctly configure the operating mechanisms of the models and to design ways to improve the performance of the model. However, experts can also introduce some bias into the analysis process. this bias can come from each step of the process through interactivity, or with the interaction process. An interesting direction is to develop visual analysis techniques that measure and quantify this bias in data processing, model building and visualization. These types of approaches allow experts to quickly identify potential problems in the analysis or interaction model and allow it to react accordingly [35]. It is difficult to model the different types of uncertainties as well as the interactions. During the analysis process, there are uncertainties due the machine (imperfect learning patterns) and uncertainties that come from the human user (an incorrect expert feedback). These two types of uncertainties interact and influence each other. If the system presents misleading information to the experts, the experts may interact in incorrect ways causing the model to change. Allowing the experts to overfit the results can also bias the results of the model [32].

Therefore, it would be interesting to propose approaches where the interactions and uncertainties induced by these methods are taken into account in the model to a lesser extent [8]. In the longer term, an ideal semi-interactive visual system would allow the user to choose to interact or not in all stages of the process thus integrating its constraints without questioning human interaction or the machine in the treatment process.

References

[1] C. C. Aggarwal. *Outlier Analysis*. Springer New York, 2013.

[2] M. Aupetit. Visualizing distortions and recovering topology in continuous projection techniques. *Advances in Computational Intelligence and Learning, Neurocomputing*, 70(7-9):1304–1330, March 2007.

[3] R. Bellman. *Adaptive Control Processes*. Princeton University Press, Princeton, NJ, 1961.

[4] E. Bertini and G. Santucci. By chance is not enough: preserving relative density through nonuniform sampling. In *Proceedings of the Eighth International Conference on Information Visualisation*, p. 622–629, 2004.

[5] E. Bertini and G. Santucci. Quality metrics for 2d scatterplot graphics: automatically reducing visual clutter. In *Proceedings of the 4th International Symposium on SmartGraphics*, p. 77–89, 2004.

[6] E. Bertini, A. Tatu, and D. Keim. Quality metrics in high-dimensional data visualization: An overview and systematization. In *Proceedings of the IEEE Transaction on Visualization and Computer Graphics*, vol. 17, p. 2203–2212, 2011.

[7] L. Boudjeloud-Assala and A. Blansché. Iterative evolutionary subspace clustering. In *International Conference on Neural Information Processing*, vol. 1, pp. 424–431, 2012.

[8] L. Boudjeloud-Assala, P. Pinheiro, A. Blansché, T. Tamisier, and B. Otjacques. Interactive and iterative visual clustering. *Information Visualization*, 15(3):181–197, 2016.

[9] E. T. Brown, J. Liu, C. E. Brodley, and R. Chang. Dis-function: Learning distance functions interactively. In *2012 IEEE Conference on Visual Analytics Science and Technology, VAST 2012, Seattle, WA, USA, October 14-19, 2012*, pp. 83–92, 2012.

[10] P. Bruneau, P. Pinheiro, B. Broeksema, and B. Otjacques. Cluster sculptor, an interactive visual clustering system. *Neurocomputing*, 150:627–644, 2015.

[11] Q. Cui, M. Ward, E. Rundensteiner, and J. Yang. Measuring data abstraction quality in multiresolution visualization. In *Proceedings of the IEEE Symposium on Information Visualization*, pp. 709–716, 2006.

[12] C.Wemmert, P. Gançarski, and J. Korczak. A collaborative approach to combine multiple learning methods. *International Journal on Artificial Intelligence Tools*, 9(1):59–78, 2000.

[13] B. Desgraupes. Clustecrit: An r package for computing clustering quality indices. R package, 2013. http://cran.r-project.org/web/packages/clusterCrit/.

[14] S. Dormieu and N. Labroche. Snow, un algorithme exploratoire pour le subspace clustering. In *Extraction et Gestion des Connaissances*, pp. 79–84, 2013.

[15] D. Engel, L. Hüttenberger, and B. Hamann. A survey of dimension reduction methods for high-dimensional data analysis and visualization. In *Visualization of Large and Unstructured Data Sets: Applications in Geospatial Planning, Modeling and Engineering - Proceedings of IRTG 1131 Workshop 2011, VLUDS 2011, June 10-11, 2011, Kaiserslautern, Germany*, pp. 135–149, 2011.

[16] S. J. Fernstad, J. Shaw, and J. Johansso. Quality based guidance for exploratory dimensionality reduction. *Information Visualization Journal*, p. 24, 2013.

[17] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *of Machine Learning Research*, 3:1157–1182, 2003.

[18] B. Hanczar and M. Nadif. Precision-recall space to correct external indices for biclustering. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, pp. 136–144, 2013.

[19] T. Isenberg, P. Isenberg, J. Chen, M. Sedlmair, and T. Möller. A Systematic Review on the Practice of Evaluating Visualization. In *Proceedings of the IEEE Transaction on Visualization and Computer Graphics*, vol. 19, p. 10, 2013.

[20] D. H. Jeong, C. Ziemkiewicz, B. D. Fisher, W. Ribarsky, and R. Chang. ipca: An interactive system for pca-based visual analytics. *Comput. Graph. Forum*, 28(3):767–774, 2009.

[21] C. Johnson, R. Moorhead, T. Munzner, H. Pfister, P. Rheingans, and T. Yoo. *NIH/NSF Visualization Research Challenges Report*. IEEE Press, 2006.

[22] I. Jolliffe. *Principal Component Analysis*, vol. 2nd ed. Springer, NY, 2002.

[23] J. B. Kruskal and M. Wish. *Multidimensional Scaling*. Beverly Hills and London: Sage Publications, 1978.

[24] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350:1332–1338, 2015.

[25] B. Letham, C. Rudin, T. H. McCormick, and D. Madigan. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *Annals of Applied Statistics*, 9(3):1350–1371, 2015.

[26] S. Liu, X. Wang, M. Liu, and J. Zhu. Towards better analysis of machine learning models: A visual analytics perspective. *Visual Informatics*, 1(1):48–56, 2017.

[27] L. C. Molina, L. Belanche, and . Nebot. Feature selection algorithms: A survey and experimental evaluation. In *International Conference on Data Mining, ICDM'2003*, pp. 306–313, 2002.

[28] F. V. Paulovich, D. M. Eler, J. Poco, C. P. Botha, R. Minghim, and L. G. Nonato. Piecewise laplacian-based projection for interactive data exploration and organization. In *Proceedings of the 13th Eurographics / IEEE - VGTC Conference on Visualization*, EuroVis'11, pp. 1091–1100, 2011.

[29] W. Peng, M. Ward, and E. Rundensteiner. Clutter reduction in multidimensional data visualization using dimension reordering. In *Proceedings of the IEEE Symposium on Information Visualization*, p. 89–96, 2004.

[30] C. Plaisant. Information visualization repository. http://www.cs.umd.edu/hcil/InfovisRepository/ acceded on november 2013, 2007.

[31] H. Purchase, N. Andrienko, T. Jankun-Kelly, and M. Ward. *Information Visualization: Human-Centered Issues and Perspectives*, chap. Theoretical foundations of information visualization, pp. 46–64. Number 4950 in Lecture notes in computer science. A. Kerren and J.T. Statsko and J.D. Fekete and C. North, 2008.

[32] D. Ren, S. Amershi, B. Lee, J. Suh, and J. D. Williams. Squares: Supporting interactive performance analysis for multiclass classifiers. *IEEE Trans. Vis. Comput. Graph.*, 23(1):61–70, 2017.

[33] D. J. Rogers and T. T. Tanimoto. A computer program for classifying plants. *Science*, 132(3434):1115–1118, 1960.

[34] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *SCIENCE*, 290:2323–2326, 2000.

[35] D. Sacha, H. Senaratne, B. C. Kwon, G. P. Ellis, and D. A. Keim. The role of uncertainty, awareness, and trust in visual analytics. *IEEE Trans. Vis. Comput. Graph.*, 22(1):240–249, 2016.

[36] E. Schubert, R. Wojdanowski, A. Zimek, and H.-P. Kriegel. On evaluation of outlier rankings and outlier scores. In *SDM*, pp. 1047–1058, 2012.

[37] M. Sedlmair, A. Tatu, T. Munzner, and M. Tory. A taxonomy of visual cluster separation factors. *Computer Graphics Forum (Proc. EuroVis 2012)*, 31(3):1335–1344, 2012.

[38] A. Tatu, G. Albuquerque, M. Eisemann, J. Schneidewind, H. Theisel, M. Magnor, and D. Keim. Combining automated analysis and visualization techniques for effective exploration of high dimensional data. In *IEEE Symposium on Visual Analytics Science and Technology*, pp. 59–66, 2009.

[39] A. Tatu, P. Bak, E. Bertini, D. A. Keim, and J. Schneidewind. Visual quality metrics and human perception: an initial study on 2D projections of large multidimensional data. In *Proceedings of the Working Conference on Advanced Visual Interfaces (AVI '10)*, pp. 49–56, 2010.

[40] A. Tatu, F. Maaß, I. Färber, E. Bertini, T. Schreck, T. Seidl, and D. A. Keim. Subspace search and visualization to make sense of alternative clusterings in high-dimensional data. In *Procedings of IEEE Symposium on Visual Analytics Science and Technology (VAST)*, pp. 63–72. IEEE CS Press, 2012.

[41] A. Tatu, L. Zhang, E. Bertini, T. Schreck, D. A. Keim, S. Bremm, and T. von Landesberger. ClustNails: Visual Analysis of Subspace Clusters. *Tsinghua Science and Technology, Special Issue on Visualization and Computer Graphics*, 17(4):419–428, Aug. 2012.

[42] E. Tejada, R. Minghim, and L. G. Nonato. On improved projection techniques to support visual exploration of multidimensional data sets. *Information Visualization*, 2(4):218–231, 2003.

[43] W. C. Ticona, N. Boukhelifa, and E. Lutton. Evographdice: Interactive evolution for visual analytics. In *Proceedings of the IEEE Congress on Evolutionary Computation, CEC 2012, Brisbane, Australia, June 10-15, 2012*, pp. 1–8, 2012.

[44] E. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT, 1982.

[45] L. van der Maaten and G. Hinton. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, Nov 2008.

[46] L. van der Maaten and G. Hinton. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[47] U. von Luxburg. Clustering stability: An overview. *Foundations and Trends in Machine Learning*, 2(3):235–274, 2009.

[48] U. von Luxburg. Clustering stability: an overview. In *Foundations and Trends in Machine Learning*, vol. 2, pp. 235–274, 2010.

[49] U. von Luxburg, R. Williamson, and I. Guyon. Clustering: Science or art? In *Workshop on Unsupervised Learning and Transfer Learning, JMLR Workshop and Conference Proceedings*, vol. 27, pp. 65–79, 2012.

[50] Y. Zhu. Measuring effective data visualization. In *Proceedings of the 3rd international conference on Advances in visual computing*, vol. 2 of *ISVC'07*, pp. 652–661. Springer-Verlag, 2007.