How to evaluate a subspace visual projection? A tentative survey

Lydia Boudjeloud-Assala*
Université de Lorraine
CNRS, LORIA, F-57000 Metz, France

ABSTRACT

This paper contributes a survey of subspace projection evaluation methods in visualization. We focus on how to evaluate real information rendered in the visual data projection for the mining high dimensional data sets. For this, we investigate automatic techniques that select the best visual projection and discus on how they evaluate the projections. When we deal with high dimensional data sets, the number of potential projections exceeds the limit of human interpretation. To find optimal subspace representation, there are two possibilities, the first one consists to search the optimal subspace which reproduce what really exist in the original data (getting the existing clusters and/or outliers in the projection). The second possibility consists to search subspaces according to the knowledge discovery process (discover novel, but meaningful, clusters, outliers, ..., from the projection). The problem is that visual projection cannot be in adequation with the subspaces. In some cases the visual projection can shown some things that are not really exist in the original data space (which can be considered as an artifact). The mapping between visual structure and real data structure is as important as the efficiency and accuracy of visualization. We examine and discuss the literature of Information visualization, Data mining, Visual analytic, High dimensional data visualization communities; on how to evaluate the meaningfulness 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

Curse of dimensionality phenomena as defined by Belmann [3] appears when analyzing and organizing data in high-dimensional spaces that do 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 are limited to just three dimensions. To overcome the curse of dimensionality, the common approach is to apply dimension reduction methods such as PCA [18] or feature selection methods (dimension, attribute selection). When data are projected in the reduced space, generally it is difficult to interpret the visualization and 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 loss information. Techniques like principal component analysis (PCA) [18], multi dimensional scaling (MDS) [19] and Stochastic Neighbor Embedding (T-SNE) [36] offer traditional solutions by creating data embedding that try to preserve as much as possible distances in the original multi dimensional space in the two dimensional space projection. However, in terms of interpretation, these techniques have problems, 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 [32]. Indeed, some data points can belonging one cluster in one subspace and belonging some other clusters in other subspace or can be considered as outliers in other one. For such scenarios, the clustering algorithms as well as visual analysis based on the whole data space may fail. If we introduce the interactivity in this step, how can the user interpret correctly what he visualise? There are two possibilities to investigate the whole data space, the first possibility is to find the reduced space which reproduce what really exist in the original data. In this case, the knowledge of the whole data set structure guide the search of the subspace projection or the feature selection. For example, search the subspace representation for the existing clusters or outliers. The second possibility is that we don't know what exist in the original data space and find some interesting reduced space according optimal subspace structure, where we can extract important information (discover new clusters, outliers, ...). In this case, the data structure in the subspace guide the search, it can be useful for the multi-view clustering, or subspace clustering, for instance. These possibilities can reproduce spaces that shown some thing that are not really exist in the original data space, which can be considered as an artifact. The mapping between visual structure and real data structure is as important as the efficiency and accuracy of visualization, and these two tasks can be evaluated in the whole data set space and in the subspace data set. One promising approach is the work of Tatu et al. [31], 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 human detected and what machine detected as data patterns. However, it is obvious how these measures can help users to deal with large data sets. We focus on how to evaluate some consequences on these possibilities for the visual data mining in high dimensional data sets and how evaluate information existing in the data. In the related works, visual quality metrics have been recently introduced to automatically extract interesting visual projections out of a large number of available candidates to explore high dimensional data sets. Theses metrics permit, for instance, 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 to interact correctly with the process. We describe the related work on fields of information visualization, visual quality metrics, high-dimensional data and data mining and then we propose the discussion before the conclusion. We try to follow a framework to understand the given hypothesis of evaluating : getting the same information (clusters, outliers, ...) in the projection and/or getting novel, but meaningful, information from the projection in each field. According to the different fields, the framework can be divided on methods that are based or that use the information in the whole data set or not.

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

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 [17]. 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 of different methods through the information that they get, how evaluate it and if they use or not the whole date space.

Getting the same information of whole data space In this proposition 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, to reduce the data point number and to be the most accurate, in the visualization are developed in 2006 by Cui et al. [8]. These measures are based on histogram comparisons, nearest neighbor computations, and statistical data properties. The authors define two measures, *Data Abstraction Level* [8] and *Data Abstraction Quality* [8]. 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 the relative data density, the degree to which outliers are preserved, response time, display clutter, and information loss. The *Data Abstraction Level* measure represents the ratio between the size of the abstracted data set and the original data set; and *Data Abstraction Quality* measure represents the degree to which the abstracted data set represents the original data set [8]. At a given data abstraction level, the *data abstraction quality* will vary based on the different abstraction methods used, and the *data abstraction level* can be considered as a very coarse *data abstraction quality* measure. To measure the *data abstraction quality*, authors use the difference between the normalized histograms of the original data set and the abstracted data set. For the *Nearest Neighbor Measure* [8], authors defined it as the normalized average of distances between every data points in the original data set to its representative from a cluster. They consider that each data point in the original data set has a nearest neighbor in the abstracted data set, designed as its representative. They suppose that the data points in the original data set, are represented by the same representative in the abstracted data set form a cluster. One of the obvious techniques to measure loss of information 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 in commonly use of for data transformation in visualization that provide an implicit measure of information loss, as Purchase et al. [25] pointed in 2008, where they try to present some theoretical foundations from an information visualization point of view. For example, Multi Dimensional Scaling (MDS) [19], 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) [18] for performing this reduction, the loss can be measured from the dropped components and the part of the *restituted data Inertia*. Similarly, Stochastic Neighbor Embedding (SNE) [37] is equivalent to minimizing the *mismatch between Squared Distances* in the two spaces, and the loss of information can be computed by this difference. It is interesting to study the proposed measures and see how well they can be used to evaluate subspace data projection without prior knowledge of the data repartition and if they can be.

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 focuses 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 dimensional as those based on graph [12]. Local graph modeling idea is to divide the data into small subspaces and to propose local optimal projection of the data. For example, Locally Linear Embedding (LLE) [27] models the data by extracting local schemas with *intrinsic geometry*. The idea is based on the linear approximation of data points by a convex linear combination of its neighborhood. The *Local Intrinsic Geometry* has a property that 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 the target subspace projection. Similar to LLE, Piecewise Laplacian-based Projection (PLP) [22] makes the assumption that every data point can be approximated by a convex combination of its neighbors. The proposed approach divides the data in smaller subsets, each is involved separately to propose a global projection preserving a global relationships among subsets. To evaluate the local and the attachment of the different local solution, the PLP method use the *Stress-based Force Scheme* [34]. 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 randomly chosen of the local patches between different spaces, there is no guaranty that global features can be preserved, the authors try to make up for this with introducing the user interaction. These two approaches are based on minimizing a neighborhood function which can find optimal local solution but can also produce artefact in the projection. This artefact can not be detected or evaluated only if we use the interactive tool and the user knowledge.

Discussion Although, several methods are proposed in information visualization field, it is difficult to define effective visualization and how to measure it. Some studies are proposed trying to find a response on how to define effective visualization. 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 [41]. The user study procedures have not been standardized. As a result, the user study data are not generally comparable with each other. In 2007, Zhu [41] points out a problem that has deeply affected the design and evaluation of visualization, that is really a lack of standards for measuring the effectiveness of visualization as well as a lack of standardized procedures. There has been some progress in this area, for example, the Information Visualization Benchmarks Repository [24] has been established but limited to IEEE challenges from 2003 at 2006, however, we can refer to associated web site 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. In 2013, Isenberg et al. [16] present a retrospective study of the evaluation practices in papers published at the IEEE Visualization conference. They try to reflect in the information visualization community on evaluation through an understanding of

the characteristics and goals of different evaluations. The authors extend a coding scheme previously established by Lam et al. [20] in 2011, and they enumerate and study eight scenarios to evaluate how different papers conduct their visual evaluation according to these scenarios. Four scenarios focused on data analysis process, that are, Understanding environment practices, Visual data analysis, Evaluating communication through visualisation and Evaluating collaborative data analysis. And four scenarios focused on the user evaluation and algorithms performance, that are, User performance, User experience, Algorithm performance, Qualitative result inspection [16]. It is certain that, there is a lack of mathematical measures to evaluate methods in visual information area. Generally authors focus on the users or participants tests, and evaluate with asking the viewer of the resulting visualization. As mentioned by Isenberg et al. [16], it was only 46% of all papers that were evaluated by the authors [16] that use the qualitative result inspection scenario. It consists in asking the user to agree a proposed model or visualization tool results by inspecting a proposed visualization. Followed by 35% of papers used Algorithm performance to evaluate their proposition. Less than 5% of reviewed papers used User performance, Understanding environment practices, Visual data analysis, Evaluating Communication through visualisation 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 not going further in assessing the faithfulness of the visual information restituted. The visual evaluation risk is to obtain some information that cannot exist in whole data space which can be considered as artefact.

3 Visual quality metrics

The visual quality measures are developed to control interactive analysis on the visualization in terms of visual quality, visual clutter and data abstraction in ergonomic graphic [4, 5, 23, 30, 35]. In this section, it is difficult to decompose the methods on them getting the same information or novel information because they are generally based on the information that exist in the original data. Indeed, some of these measures use the whole data set information. Therefor, the classification in this section is based on if these measures are based on the existing information on whole data sets or not. 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 the visual exploration. In their survey, they focus on the different metrics and how they drive 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 site the *Lie Factor* introduced by Tufte [35] in 1982, represents the ratio between size of the effect shown in the graphic and size of effect in data and the data density which represents the ratio between drawn data entries and the graph area. We can cite also the *Visual Clutter Measures* for parallel coordinates, scatter plots matrices, star glyphs and dimensional stacking proposed by Peng et al. [23] 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 plots matrices the proposed measure focus 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 dimensions algorithms have a high computational time for a small data sets with less than 10 dimensions. With the same idea to measure the clutter 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. The idea is to partition two dimensional scatter plots into blocks, compare data densities of the original data set and the data densities in each block, and this measure represents the percentage of matching blocks. This *Clutter Measure* is performed to find an optimal sampling level [5]. This measure is similar to the *Histogram Difference Measure* developed in 2006, by Cui et al. [8], 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* [8], Tatu et al. [30] 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. [30] 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 discovering new information and are not based on the whole data information. We can site *Rotating Variance Measure* and *Hough Space Measure* [30], developed for the unclassified data, without any information of the data, 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. [30] 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 of 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. [31] in 2010 propose a user 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 which is restituted by the visualization according to what the measure provide. 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. A principal focus application of visual metrics presented above is to apprehend the high dimensional data, using classical visualization methods such as parallels coordinates and scatter plots matrices, and many papers focus on this problematic [5, 6, 30, 31]. This is why we introduce the next section, focusing on the high dimensional data sets and how to evaluate different projections and visualization in a smaller subset of dimensions. We try to understand, in the corresponding state of the art, how they evaluate the reproduced information in the data projections.

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 can not represent effectively this kind of data. A 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. The high dimensional data spaces analysis consists to 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 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 pair 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 to a two dimensional space. The author defines specific measures for a self organizing map method (SOM) and the corresponding Voronoï cells, and show how they allow estimating visually whether some part of the projection is or is not a reliable image of the original manifolds. The author tries to say where the high-dimensional manifolds have been modified by reduction or the 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). It is difficult and costly to apply this approach to other projection methods (ACP, 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. [29] with the visual interactive system for subspace based analysis in high dimensional data. They use the *Tanimoto Similarity* [26] on the contained dimensions in a respective subspace. They also, compare subspaces with regard to their data distribution using *Similarity Measure*, which is very close to the *Clusters Stability* concepts [38] to evaluate clustering in data mining.

Discovering meaningful information Tatu et al. [32], 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 *Subspace Similarity Function*, they visualize the subspaces and provide navigation facilities to explore interactively large sets of subspaces. We can compare and relay subspaces respecting involved dimensions and clusters with this approach. Few reduction approaches take the importance of several structures into account and few provide an overview of structures existing in the high dimensional data set. For 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 [13]. 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 [33], is introduced according to the subspace clustering tasks to deal with high dimensional data sets, using these different measures. They use the *Tanimoto Similarity* [26] on the contained dimensions in a respective subspace. They also, compare subspaces with regard to their data distribution using *Similarity measure* to compare between them.

Discussion Automated methods are employed to analyse the dimensions, using a range of quality metrics, providing one or more measures of interestingness for individual dimensions. Through ranking, a single value of interestingness is obtained, based on several quality metrics, these measures provided from 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 potentially most interesting variables. In the related work on high dimensional data visualization, the approaches are generally proposed to guide the high dimensional data exploration, but they have a high computational time. They use measures provided from statistical data exploration, data analysis, and image segmentation. 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 Data mining

Our objective is to help determine whether what the eyes see really exists, and requires user intention or 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) [14, 21]. But not in the reverse direction, not they tells whether what exist in the subspace is the reflection of the structure that really exists in the original space. We want to persent and discuss measures that inform, if we loose information in the subspace or generate structures that can be considered as artefact. We can not 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.

Based on the whole data information For the 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 sum 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 [10] 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 on methods and techniques commonly used in outlier analysis. Such as linear methods, proximity-based methods, subspace methods, and supervised methods; with data domains, such as, text, categorical, mixed-attribute or time-series. We note that, there is a lack in the evaluation area of outliers detection. Schubert et al. [28] try to fill this gap. They propose a measure for comparing and ranking *outlier scores* and discuss about the relationship and differences to typical ranking evaluation measures. For a problem that 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 doesn't 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, 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. [39] 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* [9] provide such evaluation, for each cluster separately. However, this measure are based on the distances to other cluster centers. The intra-class inertia also provides an individual cluster evaluation, but is biased by the cluster size and the dimensions variance. To avoid these different problems, some answers were given by introducing measures to evaluate an individual cluster [7, 11]. These measures concern only clustering and clusters evaluations, they do not take into account the clusters overlapping, the subspaces projection of clusters reliability or existing outliers. In biclustering tasks, generally a large part of the points do not belong to any biclusters and some biclusters may overlap. For these reasons, several classic performance measures of clustering can not be used in biclustering. The evaluation of biclustering algorithm results is based on same performance indices that are used in the clustering evaluation *external* and *internal* indices. The *external indices* estimate the similarity between a biclustering solution and a priori knowledge. The *internal indices* compare intrinsic information about data with the biclustering solution produced by an algorithm. Hanczar and Nadif [15] propose a reliable evaluation procedure of a biclustering algorithm. They present and analyse the main external indices in the precision-recall space with a theoretical correction for each measure to remove the size bias that advantages the largest biclusters.

Discussion Finally, depending on the context, and 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 devise meaningful truthfulness evaluation. In this approach, the user is involved in making decisions, which is hardly the case, in data mining tasks.

6 Conclusion

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

Generally they compare with 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 with another structure different from the one exist in the original space. In this case, how we can evaluate this information if the new structure do not is available?. We believe on 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 exist a lack of taking into account constraints of the user in the learning process to evaluate the truth of the visualization. We believe that only through the cooperation of multiple notions from the different research fields cited before, can methods for projection evaluation be used in the exploration big and massive data set without any knowledge of the data. We want to consider only the evaluation aspect of the result that can be done automatically by the mathematical criteria or by the comparison. But as we have seen it in von Luxburg et al. [40], this problem exist also in the clustering evaluation. This discussion can be served also to related multi-view clustering problem [15], and eventually for the dynamic clustering to deal with data stream. For these possibilities a reliable evaluation procedure of a visual projection should include several steps. The first step consists in testing the algorithm on artificial data sets where the true results are known. The second one consists in testing the algorithm on the same artificial data sets in different subspaces (dimensions combination). The third step consist in computing the coordinates in the projected space. The fourth step consist in using the external indices to measure the performance of the tested algorithm and analyze its behavior in different subspaces (dimensions combination) and projection. If the results on subspaces (dimensions combination) and projection go in the same direction with in whole data set where the truth results are known, we can draw reliable conclusions. We believe that different point of views are important to evaluate the truthless of visualization, they must be contributed together with the visual interactive contribution, for all data mining and visual analysis preoccupations, as for clustering as outliers detection problems. At first sight, this framework seems to be exactly what we are looking for. At second sight, one realizes that it is not so obvious how to implement it in practice. Finally, one real answer to the question is to step back from a purely mathematical criteria and algorithmic point of view or only information visualization point of view. What is missing is to address the user and the data specialist on the focus center, for this one proposition is to use interactivity for a "users' problem" centric perspective, in order to propose a precise and meaningful evaluation 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] 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.

[9] 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.

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

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

[12] 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.

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

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

[15] 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.

[16] 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.

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

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

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

[20] H. Lam, E. Bertini, P. Isemberg, C. Plaisant, and S. Carpendale. Empirical studies in information visualization: Seven scenarios. In *Proceedings of the IEEE Transaction on Visualization and Computer Graphics*, vol. 18, pp. 1520–1536, 2011.

[21] 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.

[22] 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.

[23] 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.

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

[25] 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.

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

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

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

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

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

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

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

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

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

[40] 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.

[41] 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.