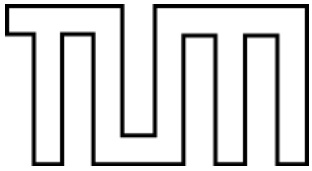


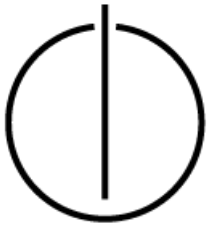
DEPARTMENT OF INFORMATICS
TECHNISCHE UNIVERSITÄT MÜNCHEN

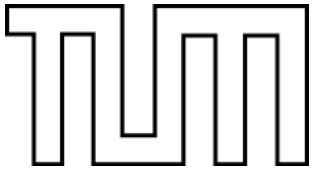
Bachelor's Thesis in Informatics

An Empirical Study in the Quality of Algorithms for Dimensionality Reduction and Visualisation of High-Dimensional Data

---

Lyubomir Stoykov





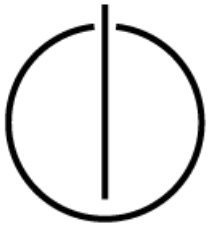
DEPARTMENT OF INFORMATICS
TECHNISCHE UNIVERSITÄT MÜNCHEN

Bachelor's Thesis in Informatics

An Empirical Study in the Quality of Algorithms for Dimensionality Reduction and Visualisation of High-Dimensional Data

Eine empirische Untersuchung zur Güte von Algorithmen zur Dimensionsreduktion und Visualisierung von hochdimensionalen Daten

Author:	Lyubomir Stoykov
Supervisor:	Prof. Dr. Florian Matthes
Advisor:	Marin Zec, M. Sc.
Submission Date:	10. February 2016



Declaration

I confirm that this bachelor's thesis is my own work and I have documented all sources and material used.

Munich, 10. February 2016

Lyubomir Stoykov

Abstract

The following bachelor thesis explores and evaluates dimensionality reduction and data visualization algorithms. Their objective is to find low-dimensional, compressed representation of high-dimensional data sets with minimum information loss, where analysis of raw data is beyond the capabilities of current software technologies.

As analysis of big data opens up new possibilities and challenges this leads to very concentrated research efforts and a lot of innovation in the field recently. Therefore there is a research gap for a very much needed, up-to-date comprehensive overview of unsupervised dimensionality reduction techniques, which this papers fills.

Evaluation of suchlike techniques is very challenging task since this is an ill-posed problem and there aren't currently any good mathematical approaches. However, humans' visual system is extremely advanced and sophisticated, much more than any existing algorithm, which is proven by the fact that identifying faces is something that we do on daily basis, yet no algorithm can nearly come close to such accuracy. This is why heuristic approach by visual analysis is generally taken for quality evaluation.

Important to note is that not only metric data has been tested, but a novel attempt to visualize categorical data with dimensionality reduction techniques has been successfully made where the user defines mapping function $f : \mathbb{S}tring \rightarrow \mathbb{N}umber$.

Last but not least, a state-of-the-art web application has been conceptualized and fully implemented where enduser without any technical knowledge is able to apply dimensionality reduction and cluster analysis on his own data sets in a very simple, intuitive way.

Acknowledgement

Firstly, I would like to thank my advisor Marin Zec for actively supporting me with my thesis and being reachable even at unusual times. General acknowledgements goes to the open source community, without which I wouldn't be able to implement the required application for this thesis. At last, I would like to thank my supervisor Prof. Dr. Florian Matthes for giving me the chance to write my thesis at his chair.

Contents

1 Introduction

1.1 Motivation

Advances in computer technologies and digital devices has recently increased data storage to a whole new level. Since real world data is most likely not random and there are certain underlying patterns, it is believed that there is a process that approximately describes the data. This has lead to very concentrated research effort in fields of computer science that analyze big data sets to make use of it, for instance, to make predictions.

At this point some very important questions arise such as is it able to predict weather based on data in the past? Is it able to predict human behavior? Is it able to construct self-driving cars? Well, the answer is most probably – yes, it is. Finally there's a chance to reveal insights into the way humans behave. However, in this context, scientists face a very, very challenging problems – real world data is enormous in both volume and variety. Therefore analysis of raw information is far beyond current human and software capabilities. In this train of thought, means for exponentially faster data processing with less resources and visual data analysis are very much needed.

Subject of the following bachelor thesis is the empirical evaluation and study of suchlike means – namely, techniques for dimensionality reduction and visualization of high-dimensional data sets. Their objective is to find low-dimensional, compressed representation of high-dimensional data sets with minimum information loss. Because they not only improve execution time, but also are able to enhance performance by removing noise, they find application in the core of rapidly developing disciplines that are shaping our future such as artificial intelligence, big data, machine learning, biotechnology and many more. Moreover, the boom of those disciplines encouraged a lot of scientists to further develop new algorithms for high-dimensional data sets.

It is, however, inherently very difficult to evaluate the quality of dimensionality reduction techniques because of the incapability of current research state to reveal the structure of complex high-dimensional structures. This leads to inability to compare the initial data set and its low-dimensional representation generated by dimensionality reduction techniques. Scientists in the community are consistent with

the statement that there isn't currently a reliable mathematical approach to evaluate the quality of suchlike algorithms. This is why, in the following thesis, techniques for dimensionality reduction and data visualization are evaluated by heuristic, empirical approach.

1.2 Research Questions

In the scope of this bachelor thesis following questions are to be answered. Given that no prior knowledge of the initial data is available...

1. What kind of dimensionality reduction algorithms exist?

2. What are some other means to explore patterns in high-dimensional data sets and visualize them?

3. In what extent do dimensionality reduction techniques and means for pattern exploration in high-dimensional data set reveal information about the underlying data?

4. What results do they yield if used to visualize categorical data after a user defined mapping $f : \mathbb{S}tring \rightarrow \mathbb{N}umber$ has been applied?

1.3 Thesis Structure

The structure of the work is as follows. At **Chapter 2** emphasis is put on the contribution of this thesis. Then at **Chapter 3** preliminaries are defined that set the basis for further exploration of the topic. At **Chapter 4** related work is discussed to put the thesis in perspective and elaborate on its importance. **Chapter 5** explores various approaches for dimensionality reduction and data visualization and defines the scope of this paper. At **Chapter 6** a comprehensive overview of unsupervised dimensionality reduction and cluster analysis techniques is proposed and **Chapter 7** evaluates five of them. Moreover, the latter presents the setting of the evaluation, the developed evaluation tool and the chosen datasets. **Chapter 8** proposes discussion and **Chapter 9** elaborates on possibilities for future work. **Chapter 10** draws a conclusion.

2 Contribution

Most importantly, a novel attempt to apply dimensionality reduction and data visualization techniques on non-metric data sets with user-defined function $f : \mathbb{S}tring \rightarrow \mathbb{N}umber$ has been successfully made. This is a step in the future which could lead to new revelations in the field. For example, one could apply dimensionality reduction on morphological matrix to find clustered solutions of a wicked problem and thus decrease the potential candidates for correct answer.

Moreover, to author's best knowledge the last paper to provide a comprehensive overview of dimensionality reduction techniques was written about 10 years ago by L.J.P. van der Maaten and E.O. Postma, H.J. van den Herik. Having in mind the concentrated research efforts recently in this subfield of computer science, this bachelor thesis fills a research gap for a very much needed, up-to-date overview of algorithms for dimensionality reduction and visualization of high-dimensional data. Some other means for high-dimensional data visualization enhancement are also explored.

Furthermore, most heuristic assessment of such techniques take into consideration only a very small number of datasets, which is not necessarily the most objective approach to tackle this problem. Evaluation here is proposed on a much bigger number of data sets for more representative results. Investigation is performed by a careful analysis of the empirical results on specifically designed artificial and real-world datasets.

Last but not least, a state-of-the-art visualization tool is conceptualized and fully implemented where enduser without any technical knowledge is able to apply five dimensionality reduction and data visualization techniques in a very friendly, simple way.

3 Preliminaries

This chapter introduces terminology that would be the basis for further elaboration on technical details of the thesis. Since absolutely correct and formal mathematical definition would be far beyond the scope of this paper, only a broad, intuitive explanation of the terms is proposed so that later concepts in the paper are easily understood.

1. *Intrinsic Dimensionality* – represents the number of features or dimensions that are needed to completely describe the initial data. For example, a n dimensional vector that has two redundant features has intrinsic dimensionality of $n - 2$.

2. *Information Loss* – the term refers to data loss in which information is destroyed or falsified.

3. *Dimensionality Reduction* - the process of applying a technique on a dataset X with dimensionality D to generate a new $d < D$ dimensional representation with minimum information loss.

4. *Subspace* – a space which is contained in another space.

5. *Space* – a set with some underlying patterns or structure.

6. *Set* - a collection of distinct objects

7. *Subset* - a set which is contained in another one.

8. *Empty Set* - the set containing no elements, denoted with $\emptyset$.

9. *Topological space* - a set of X together with a collection of open subsets T that satisfies: The empty set is in T, X is in T, the intersection of a finite number of sets in T is also in T, the union of an arbitrary number of sets in T is also in T. [@LW16]

10. *Euclidean Space* - is the space of all $n - tuples$ of real numbers, $(x_1, x_2, ..., x_n)$ [@SW16]

11. *Manifold* – topological space that locally resembles euclidean space, but globally it may not.

12. *Interior Angle* – an angle inside a shape.

13. *Convex Manifold* – intuitively speaking, a convex manifold is one that has no curves pointing inwards. It means that no interior angle is more than 180 degrees.

14. *Non-convex Manifold* – a manifold that is not convex, therefore it has at least one interior angle with more than 180 degree.

15. *Regression* – refers to the process of estimating the results of a real-valued function based on a finite set of noisy samples as samples. [CM98]

16. *Classification* – refers to the process of identifying to which category an object belongs to. [Ped+11]

17. *Clustering* – refers to the process of automatic grouping of similar objects into sets. [Ped+11]

18. *Feature* – an attribute of an object or vector.

19. *Feature Extraction* – creating a set of new features which describe the initial vector.

20. *Feature Selection* – selecting a subset of feature from the initial vector or set of features.

21. *Visual Analysis* – the term describes the process of drawing conclusion about data set properties based on it's visual representation.

22. *Density* – describes mass of substance per unit volume.

23. *Distribution* – describes how objets are spread over an area and thus their relations.

24. *Prediction Error* – the mean of squared errors of prediction, where error is the difference between true output and its predicted value. [Ize08]

25. *Generalization Error (Infinite Test Error)* – expected prediction error over an independent test set. [Ize08]

26. *Scalability* – the capability of an algorithm to remain efficient and accurate the complexity of the problem increases. [Ize08]

27. *Artificial Intelligence* – a subfield of computer science which focuses on development of machines that could think, solve problems and act rationally like humans do.

28. *Machine Learning* – a subfield of computer science which focuses on development of machines that can learn from past experiences.

29. *Supervised Learning* – problem where an algorithm receives input and its correct output variable before applying it on test sets.

30. *Unsupervised Learning* – problem in which there is no information available to define an appropriate output variable, often referred to as scientific discovery. [Ize08]

31. *Cluster* – a group of similar objects.

32. *A Learning (Training) Set* – data set with correct input and its corresponding output variable used in supervised algorithms for learning purposes.

33. *A Validation Set* – a data set used to adapt the algorithm performance and tune parameters of the classifier.

34. *A Test Set* – a data set to be used for assessing the performance of a completely specified and finally tuned classifier.

35. *Learning Error* – average of generalization error over learning dataset.

36. *Overfitting* – this phenomenon occurs when the model is too complicated. It usually results in a very small learning error and a large generalization (test) error. [Ize08]

37. *Graph* – a non-empty finite set V of elements called vertices and set E of pairs of vertices called edges.

38. *Adjacent Vertices* – pairs of vertices which belong to E.

39. *Graph Neighborhood* – set of all the vertices adjacent to a certain vertex.

40. *Geodesic distance* – the shortest path between two vertices in a graph is called geodesic distance.

41. *Projection* – function that maps a set into a subset.

42. *Map* - a map is a way of associating unique objects to every element in a given set. [@Wei16]

43. *Euclidean distance* – the length of the line connecting two points in the space is defined as their euclidean distance.

44. *Hyperplane* – a hyperplane of a space is a subspace with one dimension less than the original space.

45. *Tangent space* – intuitively speaking, a tangent space at a point of a manifold is the hyperplane that best approximates the manifold at this point.

46. *Curse of Dimensionality* – refers to the phenomenon where the sample size required to estimate a function of several variables grows exponentially with the increasing number of variables. [Wan08]

47. *Semidefinite Programming* – field which is concerned with optimizing linear functions.

48. *Pattern recognition* – a subfield of computer science which identifies regularities in data sets.

49. *Data mining* – a subfield of computer science which focuses on searching massive datasets for structures, relationships, trends, clusters and outliers. It also build models and algorithms for regression, classification, pattern recognition and other machine learning tasks. [Ize08]

50. *Outlier* – an observation that doesn't comply with the usual patterns in the data.

51. *Noisy Data* – meaningless data.

52. *Kernel Function* – a similarity function that takes inputs and yields how similar they are.

53. *Numerical Data* – numerical data is data that is measurable. For example, a person's age.

54. *Ordinal Data* – ordinal data is data which defines ranking order and particular value has no meaning beyond its ability to establish order.

55. *Categorical Data* – categorical data defines categories without any relationships between them. Colors such as blue, yellow and red are an example for categorical data.

4 Related Work

To put the thesis in perspective and emphasize on its importance, the following section elaborates on related work.

Dimensionality reduction and high-dimensional data visualization finds application in the core of scientific fields that recently earn big attention such as machine learning, data mining, artificial intelligence, biotechnology. In this sense, there is plenty of literature about those disciplines of computer science that touch on dimensionality reduction techniques. However, techniques for data reduction there are used just "as is", without any assessment of their quality. At most the underlying mathematical basis is explained. Therefore this papers concentrates on evaluation of algorithms for dimensionality reduction and data visualization.

Moreover, the last paper that proposes a comprehensive overview of dimensionality reduction techniques, to my best knowledge, has been published by L.J.P. van der Maaten and E.O. Postma, H.J. van den Herik [Maa08c] about ten years ago. Considering recent innovations in this field, it has already been a long time since then.

As for evaluation of dimensionality reduction techniques there exist two main approaches in the community – mathematical, based on cost function, and heuristic, based on visual analysis [Ven07].

Mathematical approach doesn't quite suffice to evaluate data reduction quality, because dimensionality reduction is an ill-posed problem cite [Maa]. The high-dimensional structure of the manifold is usually unknown and thus assumptions on the data has to be made that in most cases falsify the results. It is though the minority of papers that tackle evaluation by mathematical means [Ven07]. This fact implies that there aren't currently good theoretical approaches to address the problem. More on why mathematical evaluation doesn't quite suffice is proposed on the section in section 7.

On the other hand, there are quite some papers that focus on visualization to estimate quality of dimensionality reduction [ZZb], [Li04], [GR], [MGMa], [Maa08a]. The problem is that at most 5 datasets are tested against a technique in each paper, which is not sufficient to draw general conclusions. Another problems is that usually the

author of an algorithm assesses its quality and it is human nature to favor our own solutions.

Furthermore, there is lack of research that explores application of dimensionality reduction techniques for visualizing categorical data after a user-defined mapping function is applied, which could find a very broad practical use. For instance, consider application of dimensionality reduction on categorical companies data in order to find clusters of similar ones.

Other than dimensionality reduction techniques not much has been done to empirically evaluate data visualization techniques.

This paper addresses all these issues by proposing a very much needed, up-to-date comprehensive overview of dimensionality reduction techniques and evaluation on its quality on more than ten datasets by a more objective, author-detached approach. A novel attempt to test those techniques against categorical data with a user-defined mapping function has been made. Moreover, new techniques for recognizing data patterns and visualization are taken into consideration. Finally, this work comes with an application for enduser without any technical knowledge to apply state-of-the-art visualization techniques on his own data sets, something that is really missing, because existing frameworks require technical knowledge.

5 Scope

A very important part of this bachelor thesis is to research other techniques than dimensionality reduction to visualize high-dimensional data. In an effort to better understand the numerous methods which have been proposed and select the most suitable ones, this section defines the scope of the paper. Moreover, together with Chapter 6, it gives an answer to the first two research questions.

First and foremost, in regard to metric data, the main focus here is to explore application of algorithms without any user supervision. Therefore only unsupervised techniques for data visualization and dimensionality reduction are taken into consideration, which means that classification and regression methods are not explored. Moreover, there exist very little research on visualizing and reducing dimensions of categorical data, therefore solely algorithms for metric data are taken into account. However, an attempt to map categorical to metric data with a user-defined function and then apply dimensionality reduction techniques for metric data on the result for visualization purposes has been made. The latter approach has been evaluated. In other words, the thesis mainly focuses on unsupervised techniques for metric data.

Nevertheless, it is still rather challenging task to fully define its scope in such an enormous scientific field, therefore Figure 5.1 shows that data visualization can be obtained in mainly two different ways – by adjusting the data or by adjusting the visual representation. Adjusting the visual representation – modern techniques include parallel coordinates and scatter plot matrices. Parallel coordinate visualize each feature of a vector on different axis. On the other hand, scatterplot matrix visualizes only a subset of features at a time on different plots as shown on Figure 5.1. The drawback of those techniques is that they don't really scale. As large data sets become ubiquitous but the screen space for displaying is limited, the size of the data sets exceeds the number of pixels on the screen. Hence, we cannot display all data values simultaneously [Lon09] This limits the applicability of these techniques to real-world data sets that contain thousands of dimensions. As the main focus here is visualization of high-dimensional data, the first major group for data visualization, namely adjusting the visual representation is out of scope.

Adjustment of the data – in contrast to adjustment of visual representation, adjusting data could result in a two or three dimensional vectors which would allow for visual

representation that is easily interpreted and analyzed by human eye. There we define three subfields – feature selection, feature extraction and cluster analysis.

Feature selection refers to the process of reducing dimensionality by selecting a subset of features from the initial vector or set of features. This groups is represented by techniques such as “Low variance filter”, “Forward feature construction”, “Backward feature elimination”, “Low variance filter”. Even if one would assume that most features are redundant, it is still hardly imaginable that in a high-dimensional data sets only two or three features carry most of the information. This is why feature selection techniques are not suitable for high-dimensional data and thus not in scope, since they are not really applicable to reduce the dimension to two or three and thus adjust data for visualization and human perception.

The thesis focuses therefore mainly on cluster analysis and feature extraction algorithms as means to enhance visualization of high-dimensional data. Clustering on the one hand utilizes visual analysis by grouping similar objects together and feature extraction algorithms help to reduce dimensionality of data by deriving new features from the initial data set. From now feature extraction is referred to as dimensionality reduction.

To sum up, the papers mainly focuses on unsupervised dimensionality reduction techniques, but also touches on visualization of high-dimensional data with other means such as cluster analysis.

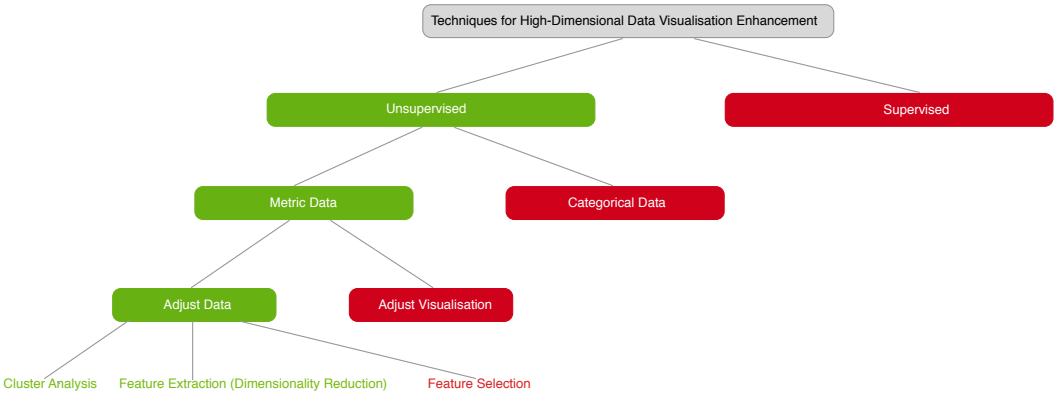


Fig. 5.1: Scope Definition

6 Dimensionality Reduction and Clustering Techniques – an Overview

The following section provides a comprehensive overview of unsupervised techniques for dimensionality reduction and of cluster analysis. This is not meant to be a mathematical guide through all existing algorithms, but rather an overview which helps to understand the basic idea between each technique. Moreover, the information in this section would serve as basis for the decision which techniques are going to be evaluated.

Every technique is briefly explained and a source for more information is given, then we select the techniques that we'll research more in depth and evaluate.

It is also good to know that the main distinction between techniques for dimensionality reduction is the distinction between linear and nonlinear. Linear assume that data lie or near a linear subspace of high-dimensional space. Nonlinear techniques for dimensionality reduction do not rely on the linearity assumption as a result of which more complex embeddings of the data in the high-dimensional space can be identified. [Maa08c]

6.1 Unsupervised Dimensionality Reduction Techniques

1. *Principle Component Analysis* [I.T86] - The basic idea behind PCA is to find vectors, derived from linear combination of the initial set of variables, that retain as much of the variance present in the dataset as possible. These vectors are called principle components. The principle components are mutually orthogonal and account for the variance present in all of the original variables in a descending order. For a lower, d-dimensional representation of the original data, the first d principle components are chosen. After that, principle scores of the variables are computed and plotted in the low-dimensional space represented by the chosen principle components.
   There are two main approaches to solve PCA – by singular value decomposition or by correlation matrix. Figure 6.1 shows the first principle component on two

dimensional data set. The idea can be extended for much higher dimensional data sets.

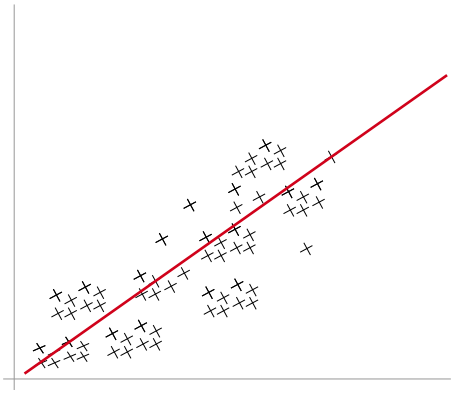


Fig. 6.1: First Principle Component in direction of maximum variance

2. *Factor Analysis* - Factor Analysis tries to find lower dimensional representation of the initial dataset in terms of a new set unobserved variables called factors. While often incorrectly used in exchange with PCA, both are not similar. Main differences include - principal components account for maximal amount of variance of observed variables, while factors account for common variance in the data, component scores are a linear combination of the observed variables weighted by eigenvectors, while factors are linear combinations of the underlying and unique factors. [Suh]

3. *Non-linear PCA* - A few dimensionality reduction techniques extend the idea behind PCA to capture maximum variance of the data, but improve on that and propose non-linear solution. While PCA identifies only linear correlations between variables, non-linear PCA uncovers both linear and non-linear correlations, without restriction on the character of the nonlinearities present in the data. [Kra91]

    a) *Principle Curves* [HS89] - While PCA draws a straight line through the data set in attempt to minimize the sum of squared deviations and maximize the variance as shown on Figure 6.2, principle curves uses a smooth curve to summarize the data. Principle curves are subject to the constraint that any point the curve is an average of all the data points that are projected onto it. It means that at any point a principle curve best summarizes the data. The algorithm starts with a simple curve and iteratively checks if it is consistent with the constraints. If not, the curve is adjusted and the process repeats until convergence.
    Since the concept was proposed by Hastie and Stuetzle in 1989, a considerable number of refinements and further developments have been reported. [Non]

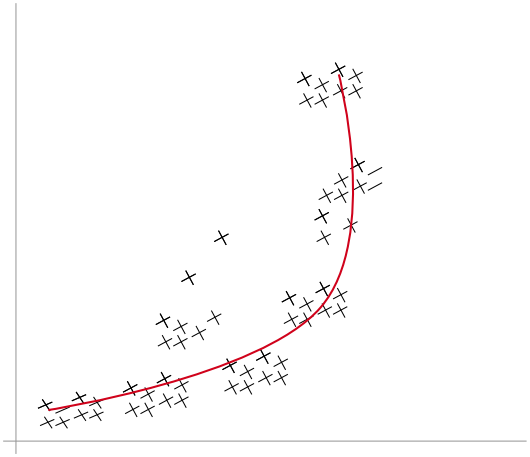


Fig. 6.2: First Principle Curve in direction of maximum variance

b) *Kernel PCA* [BS98] - is another nonlinear variation of PCA with increasing popularity. Instead of using covariance matrix, kernel PCA derives its principle components from a kernel matrix. The idea is to apply a kernel function which generates new data points that are non-linearly correlated to the initial data set. Figure 6.3: represents the basic idea behind kernel PCA. Then kernel matrix is constructed as a dot product of the new high-dimensional data points (interesting are not principal components in input space but principal components of variables, or features, which are nonlinearly related to the input variables). Since the new variables are nonlinearly related to input space, the principle components become nonlinear in input space.
Crucial to kernel PCA is the fact that all necessary computations are carried out by the use of a kernel function in input space. [Non]

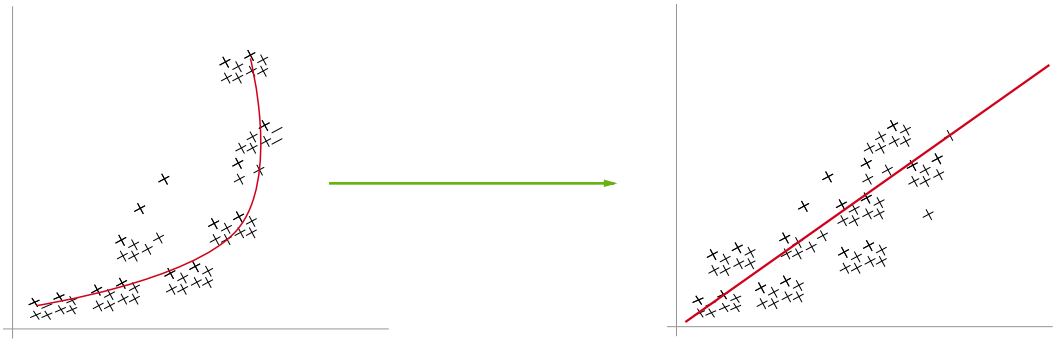


Fig. 6.3: Nonlinear to linear mapping with kernel function

4. *Classical Multidimensional Scaling* [Tor52] - CMDS was introduced was first introduced by Torgerson. Firstly, the algorithm computes pairwise distances between all points in high-dimensional space. Then it maps those points into low-dimensional representation and tries to preserve distances as much as

possible. Classical multidimensional scaling is therefore minimizes following stress function: $f : (X, Y) = \sum_{i,j}(\|x_i - x_j\| - \|y_i - y_j\|)^2$, where x_i are high-dimensional and y_i low-dimensional datapoints.

5. *Sammon's Mapping* [Sam69] - as classical multidimensional scaling has a big shortcoming – it emphasizes much more on retaining the distance between point which are far away from each other in the input space, sammon mapping improves on the raw stress function introduced by CMDS and defines a new one as $f : (X, Y) = \frac{1}{\sum_{i,j}\|x_i - x_j\|} \frac{\sum_{i \neq j}(\|x_i - x_j\| - \|y_i - y_j\|)^2}{\|x_i - x_j\|}$, where x_i are high-dimensional and y_i low-dimensional datapoints.
   Sammon's stress function penalizes points which are initially close in the input space, but widely separated in the output space much more than points which were far away, but in are mapped in the same neighborhood. This, of course, could lead to false neighborhoods. [SLF07]

6. *Kruskal's Stress Function* - there exist one more variation of classical multi-dimensional scaling. The Kruskal stress function which has to minimized is defined as $\sqrt{(\sum_{i<j} \frac{(d_{ij} - d_{ij}')^2}{d_{ij}'}}$), where $d_i j$ is the distance between points in the initial and $d_i j'$ the distance between points in the reduced data set. This approach, however, is also flawed and encourages tears in the manifold, because the bigger the $d_i j$, the smaller kruskal's function.

7. *Data-Driven High-Dimensional Scaling* [SLF07] - DDHDS is a novel techniques firstly introduced by Sylvain Lespinats, Michel Verleysen, Senior Member, IEEE, Alain Giron, and Bernard Fertil. It addresses and improves on all the issues of Sammon, Kruskal and MDS in two ways. Firsly, a It introduces 1) a specific weighting of distances between data taking into account the concentration of measure phenomenon - distances between high-dimensionalobjects are usually very concentrated around their average. [Don]
   Secondly, it approaches handling of short distances in the originaland output spaces in a symmetric way, avoiding false neighbor representations while still allowing some necessary tears in the original distribution. More precisely, the weighting is set according to the distribution of distances in the data set.

8. *RankVisu* [SL09] - RankVisu can be thought of as another extensino of MDS and is designed to preserve rank of neighborhood rather than distance. RankVisu is especially useful on difficult tasks (when the preservation of distance cannot be achieved satisfyingly). Indeed, the rank of neighborhood is less informative than distance (ranks can be deduced from distances but distances cannot be deduced from ranks) and its preservation is thus easier. Simple, but good because most MDS focus on mapping function which does not consider the preservation of very small distances as important, since the cost is small with

respect to the preservation of large dissimilarities, therefore neighborhoods may be distorted for the benefit of the overall mapping. This is why ranking approach makes sense, it assigns as much importance on preserving small distances as on preserving big ones

9. *Isomap* - Isometric maps method improves on multidimensional scaling and seeks to preserve geodesic instead of euclidean distances between points while mapping the observed data into fewer dimensions. Isomap has the shortcoming that it doesn't perform well on non-convex manifold or one which consists of many sub-manifolds, which is inherent problem on all distance-based algorithms.

10. *Maximum Variance Unfolding* - similarly to MDS and Isomap, MVU defines a neighborhood graph on the data. The difference is that MVU explicitly unfolds the manifold – maximizes euclidean distances between data points, but keeps the local structure.This is done with means of simidefinite programming.
Maximum variance unfolding is basically multidimensional scaling applied on an unfolded manifold. The unfolding is needed because MDS will incorrectly consider two distanced points as neighbors based on euclidean distances, when the manifold is curved. Unfolding seeks to solve this problem. [Wan08]

11. *Locally Linear Embedding* [LKS] - LLE concentrates on preserving the local structure of manifold, which allows for successful embedding of non-convex manifolds. Firstly, LLE defines every datapoint in terms of its neighbors as follows $f : X_i = \sum_j X_j W_j$. Once the algorithm has found W, it optimizes the following function $f : Y = \sum_i \|(Y_i - \sum_j W_{ij} Y_j)^2\|$ in order to find low-dimensional datapoints Y. In other words, it seeks to keep relations between neighbors in the low and high-dimensional space.

12. *Modified Locally-Linear Embedding* [ZZb] - MLLE is very similar to LLE. The difference is that instead of a single weight-vector W, MLLE uses multiple weights to represent a high-dimensional datapoint in terms of its neighbors.

13. *Hessian Locally-Linear Embedding* [DG] - HLLE, initially introduced by David L. Donoho and Carrie Grimes, is an extension of LLE, but actually conceptually very similar to Laplacian Eigenmaps. It tends to produce higher quality visualisation than LLE, but is computationally very expensive and even the basic idea behind is mathematically too involved for this paper. Intuitively speaking HLLE attempts to minimize the curviness of the high-dimensional manifold when embedding it into a low-dimensional space by finding the eigenvectors of a matrix H that describes the curviness of the manifold around data points. A Hessian matrix is used to measure the curviness at every datapoints.

14. *Relational Perspective Map* [Li04] - the key idea of RPM algorithm is its exploitation of the properties of closed manifold (the torus) to keep the configuration in balance as shown on Figure 6.4. The RPM algorithm keeps the image points together with torus as a force-directed multiparticle system: the image points are considered as particles that can move freely on the surface of the torus, but can not escape the surface. The particles exert repulsive forces on each other so that, guided by the forces, they themselves to a configuration that visualizes the relational distances.

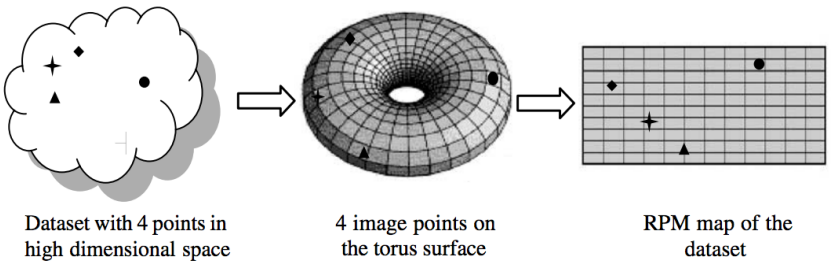


Fig. 6.4: Model of the RPM method. Adapted from [Li04]

15. *Laplacian Eigenmaps* [BN] - LEIGS is very similar to LLE in the sense that both preserve local properties of the manifold. However, in Laplacian Eigenmaps the local properties are based on pairwise distances between neighbors. Therefore a distance-based function is defined which has to be optimized for its smallest value.
    The cost function is defined as $f : Y = \sum_{i,j}(y_i - y_j)^2 w_{ij}$, where y_i and y_j are data points from the output space and large weights w_{ij} correspond to small distances between data points x_i and x_j in the input space. [Maa08c]

16. *Diffusion Maps* [RRC06] - DMaps was first introduced by Ronald R. Coifman and Stéphane Lafon and uses Markov matrix to define a random walk over the data. With its help and specially defined functions, "diffusion maps", it obtains new description of the input space which enables embedding the manifold into euclidean space, where the distance between points describes their relationships. These relationships are then preserved as well as possible in the low-dimensional representation. Please refer to the paper where the algorithm is firstly introduced, because more mathematical description is too heavy for this thesis.

17. *Diffeomorphic Dimensionality Reduction* [WS] - the algorithm slightly resembles diffusion maps and constructs mappings called "diffeomorphic" which maps the data set near to a subspace of lower dimension. Then the output space is projected onto the lower-dimensional space.

18. *Manifold Sculpting* [MGMb] - this techniques takes a little different approach through a "process of progressive refinement" - dimensions are iteratively removed and data points projected incrementally to new, lower-dimensional space. The algorithm first fixes a point, finds its k neighbors and then defines relationships between the point itself and its neighbors in terms of euclidean distances. This is done for all data points. After that, iterative process begins where all the values in the set of dimensions that will be eliminated by the projection are scaled by a factor < 1. After that the values in the set of dimensions that will remain are adjusted so that the relationships defined between neighbors are preserved. When the dimensions that are going to be removed contain only values that are null, they are dropped and data is projected.

19. *Stochastic Neighborhood Embedding* [HR] - SNE is a probabilistic approach for dimensionality reduction. SNE defines similarities between neighbors in input space in terms of probabilities. The similarity between x_i and x_j is therefore defined in proportion to their probability density under a Gaussian centered at x_i as $f : p_{ij} = \frac{\exp \frac{-\|x_i - x_j\|^2}{\mathcal{N}^2}}{\sum_{i \neq k} \exp \frac{-\|x_i - x_k\|^2}{\mathcal{N}^2}}$ .Those probabilities are approximated as much as possible in low-dimensional representation by optimizing a cost function.

20. *t-Distributed Stochastic Neighborhood Embedding* [Maa08b] - T-SNE is an extension of SNE proposed by Laurens van der Maaten and Geoffrey Hinton. It has two improvements over SNE

    a) The cost function is symmetric.

    b) t-student distribution is used instead of Gaussian.

21. *Topologically Constrained Isometric Embedding* [BK] - TCIE is relatively new algorithm for nonlinear dimensionality reduction that uses global information to reduce dimensionality. The main contribution is that it detects and ignores geodesic distances which may be inconsistent because of potential non-convexity of the manifold. It basically resembles Isomap, but ignores geodesic distances that are falsified by holes or non-convex structures in the manifold.

22. *Local Tangen Space Alignment* [ZZa] - LTSA, as the name of the algorithm implies, is a technique that explores local properties of high-dimensional manifold using the local tangent space of each point. LTSA first assumes local linearity of the manifold and defines mapping from high-dimensional datapoint to its local tangent space. Since the local assumptions would hold also for

low-dimensional representation, LTSA assumes the existence of linear mapping from the output space to the same local tangent. Finally, attempt to align both mapping function is made.
In other words, LTSA simultaneously searches for the coordinates of the low-dimensional data representations, and for the linear mappings of the low-dimensional data points to the local tangent space of the high-dimensional data. Based on the intuition that when a manifold is correctly unfolded, all of the tangent hyperplanes to the manifold will become aligned. [Wan08]

6.2 Cluster Analysis Techniques

1. *Hierarchical clustering* - distance-based clustering technique, where similarity of objects is directly related to the distance between them.
   The algorithm for identifying clusters can be described as follows

   a) Assign each item to different cluster.

   b) Find the most similar pair of clusters and merge them into a single one.

   c) Compute similarity between the new cluster an each other one.

   d) If only one cluster left, stop

   e) Otherwise repeat 1b and 1c

   Similarity can be computed in three different ways:

   a) *Single-Link Similarity* - the distance shortest between any point in the first cluster to any point in the second cluster is defined as their similarity.

   b) *Complete-Link Similarity* - the longest shortest between any point in the first cluster to any point in the second cluster is defined as their similarity.

   c) *Average-Link Similarity* - the average shortest between any point in the first cluster to any point in the second cluster is defined as their similarity.

2. *Centroid Based Clustering* - CBC techniques first partition the space by k mean vectors. Then they form groups of items together which are closest to one of the k mean vectors. A very big shortcoming of those such techniques is that they rely on predefined k. $K - means$ clustering is probably the most popular centroid based clustering.

3. *Distribution Based Clustering* - DBC defines member of the same cluster as object which belong to the same distribution. This is done by sampling random objects from a distribution. One of the biggest advantages of those type of clustering techniques, similar to density based clustering, is that very little input is required by the user.
   Moreover, they enable discovering of clusters with arbitrary shape. [XX]

4. *Density Based Clustering* - the intuition behind density based clustering algorithms is that they define clusters as groups of different density, in other words clusters are dense regions, separated by regions of lower density as shown on.

7 Dimensionality Reduction and Cluster analysis for Data Visualization – Evaluation

This Chapter answers the last two research questions.

Humans' visual system is extremely advanced and sophisticated, much more than any existing algorithm, which is proven by the fact that identifying faces is something that we do on daily basis, yet no algorithm can nearly come close to such accuracy. Therefore visualizing high-dimensional data in low-dimensional space leverages our unique ability to recognize non-trivial, complex patterns and relationships in the data.
This is why a well-designed visual representations can replace cognitive calculations with simple perceptual inferences and improve comprehension, memory, and decision making. [@JHO16]

Unfortunately, reducing dimensionality to enhance data visualization is inherently linked to information loss, thus the challenge is not only to create effective and engaging visualizations, but also to be able to tell how representative they are. What is the quality of the low-dimensional representation is in this sense a very important question.

Now that is clear how significant the topic of visual representation actually is, it is time to evaluate the quality of such techniques.

7.1 Dimensionality Reduction - Background and Context

Assessment of dimensionality reduction techniques has proven to be a rather challenging task because of the lack of suitable quality measures in the information visualization field. [Ven07]
Nevertheless, the scientific community in this field has proposed some good means which are explored in the next subsection.

7.1.1 Metric Data

The problem of evaluating quality of low-dimensional representations has been mainly tackled in two ways – mathematically and heuristically.

1. Mathematical Approach - about 10% of the evaluations include mathematical approach by defining a cost function. [Ven07] There exist Hn types of cost function, defined as follows

   a) The raw cost function: $f : (X, Y) = \sum_{i,j}(\|x_i - x_j\| - \|y_i - y_j\|)^2$

   b) The Sammon's cost function: $f : (X, Y) = \frac{1}{\sum_{i,j}\|x_i - x_j\|} \frac{\sum_{i \neq j}(\|x_i - x_j\| - \|y_i - y_j\|)^2}{\|x_i - x_j\|}$

   c) The Kruskal cost function: $\sqrt{(}\sum_{i<j} \frac{(d_{ij} - d_{ij}{}')^2}{d_{ij}{}'})$

   where x_i are high-dimensional, y_i low-dimensional datapoints, $d_i j$ the distance between points in the initial and $d_i j'$the distance between points in the reduced data set.

   This type of solutions have two big shortcomings. First and foremost, the general idea behind cost functions is to measure distance between points in high-dimensional space and penalize pairs in low-dimensional representation, which are not compliant with their initial measurements. Ideally, the distance between each pair A and B in high-dimensional space is equal to the distance between their equivalents in low-dimensional space.
   The first short coming is that the measurements are distance based. Since the manifold's structure is unknown, certain assumptions has to be made. Euclidean based distances assume that the distance between A and B is a straight line. Geodesic distances assume that the distance between A and B is the shortest path in the graph of all points. Both euclidean and geodesic distances don't follow the exact structure of the manifold, because it may be curved, it may contain holes or consist of more sub-manifolds. It means that the very basis for defining a cost function is flawed and could already falsify results.
   Furthermore, the raw cost function emphasizes on retaining distance between points that are further away and doesn't really penalize the case where certain neighborhood is distorted. The Sammon cost function improves on this problem and puts more emphasize on preserving distances that were originally small. This doesn't come without cost, since it could lead to false neighborhoods of points that were originally far away from each other. Such a poor behavior was recognized by the author himself, who mentions that separated classes in high-dimensional space may not be separated in the low-dimensional

representation [Sam69]. On the other hand, the Kruskal cost function could lead to tears in the manifold [SL] because of the likelihood that a small distance in the original space is associated with a large distance in the output space.

Another mathematical approach, ranking-based was proposed by Jarkko Venna in his dissertation. The author defines two criteria. The lesser their sum, the better is the visualization.

a) *Trustworthiness - Let N be the number of data samples and $r(i,j)$ be the rank of the data sample j in the ordering according to the distance from i in the original data space. Denote by $U_k(i)$ the set of those data samples that are in the neighborhood of size k of the sample i in the visualization display but not in the original data space. The measure of trustworthiness of the visualization is $M_{trust}(k) = 1 - A(k)\sum_{i=1}^{N}\sum_{j\in U_k(i)}(r(i,j)-k)$ where $A(k)$ scales between 0 and 1 [Ven07].*
   Note that the error gets its maximum value when the ranks in the input and output space are reversed.

b) *Continuity - Let $V_k(i)$ be the set of those data samples that are in the neighborhood of the data sample i in the original space but not in the visualization, and let $l_{(i,j)}$ be the rank of the data sample j in the ordering according to the distance from i in the visualization display. The effects of the projection are measured by $M_{cont}(k) = 1 - A(k)\sum_{i=1}^{N}\sum_{j\in V_k(i)}(l(i,j)-k)$ [Ven07].*

The idea behind trustworthiness is to penalize points that weren't neighbors in the input space, but in the output space are in the same neighborhood. The idea behind continuity is to penalize points that are originally neighbors, but in the input space are far from each other. Although the idea is novel and yields good results, it is still flawed. Trustworthiness and continuity don't take into consideration the relationships or ranking of points that are both neighbors in the input and output space. If their ranks are reversed, no penalty is applied.

Third mathematical approach is based on classification. Firstly, dimension reduction algorithm is applied. Then classifier is trained in the low-dimensional result. Quality of dimensional reduction is based then on the classifier's performance, because if dimensionality reduction was good, then it should recognize different classes and thus train the classifier very well. The drawback is that classes may not be well differentiated in the input space itself, which could lead to poor performance of the classifier, although dimensionality was reduced without much information loss. Therefore this approach fails to

estimate the quality of dimensionality reduction techniques, but could be used for comparison between algorithms as proposed in [Maa08c].

All these facts lead to falsification of results, that is why mathematical approach is not further explored in the scope of this thesis. One more proof that theoretical approach doesn't quite suit dimensionality reduction evaluation is that only about 15% of the papers that asses such techniques used it. [Ven07]

2. *Heuristic Approach* - for evaluation of dimensionality reduction techniques a heuristic approach is taken. Ten datasets are selected where dimensionality reduction is applied on each and every one of them. Results are assessed by visual analysis, because, as already mentioned, humans' visual system is much more advanced than any existing algorithm in recognizing sophisticated data patterns, for example – face recognition. Direct visual analysis and comparison between input and output space for metric data is possible, because sets with known high-dimensional representation has been specifically selected. To evaluate quality of dimensionality reduction on categorical data a survey has been done.

7.1.2 Categorical Data

Due to lack of results in the area of direct non-metric data reducement, in this paper a novel attempt has been made to visualize categorical data, indirectly, after appliance of dimensionality reduction techniques on non-metric data with a user-defined function $f : \mathbb{S}tring \rightarrow \mathbb{N}umber$.
To my best knowledge this is the first paper to apply and evaluate such approach. The results, however, are very important for science and have a lot of implications. For instance, if dimensionality reduction could be applied on categorical data and yield good results, this would be a step in the future. One use case would be appliance on a morphological matrix to find clustered solutions of wicked problem and thus exponentially reduce the number of possible answers.
Moreover, since data set with unknown high-dimensional structure has been used, indirect approach for visual analysis is taken to evaluate quality of the low-dimensional representation. After dimension of the initial data is reduced, people have been asked to elaborate on their interpretation of the visual representation, to find patterns, structures, similarities or inconsistencies. The answers are then explored in depth for compliance with the initial high-dimensional data. Unfortunately, only one data set has been evaluated, because of time constraints.

7.2 Cluster Analysis - Background and Context

7.2.1 Metric Data

Evaluation of cluster analysis is mathematically much easier problem to solve, since this is a classification problem. The question that needs to be addressed is – are the points in the low-dimensional representation classified in the same cluster as in the high-dimensional space? In contrast to dimensionality reduction here manifold structure and relationships between clusters are not important.

There exist two main approaches – internal and external.

1. *Internal Evaluation* – a function of the given data is defined as quality measurement, called objective function. This method is used rather as comparison between algorithms, not as quality measurement, since objective function by itself provides no indication of "how much better" is one partitioning than the other.
   Furthermore, estimates of objective functions in the literature are far from being accurate. [Kog07]

2. *External Evaluation* – quality is being evaluated using external information that has not been used in the algorithm itself, such as class labels. After cluster analysis has been applied and every input value assigned to cluster, those are compared with the initial class labels. For instance, let's say there are 4 data points with values a, b, c, d and class labels $1, 2, 3, 1$. If after cluster analysis has been applied clusters are as follows: $(a, d), (b), (c)$ where $()$ notes a cluster group, then results are good. One method to externally evaluate the quality is adjusted rand index. Adjusted rand index is also the method which is used for assessment in this thesis. Rand index computes a similarity measure between two clusters by considering all pairs of samples and counting pairs that are assigned in the same or different clusters in the predicted and true clusterings. Adjusted rand index is "adjusted for chance".
   The adjusted Rand index is thus ensured to have a value close to 0.0 for random labeling independently of the number of clusters and samples and exactly 1.0 when the clusterings are identical (up to a permutation) [Ped+11].

7.2.2 Categorical Data

Similar approach to dimensionality reduction has also been taken for cluster analysis. The difference is that formed clusters are directly compared with high-dimensional data set, no survey has been done.

7.3 Setting

In the following section evaluation criteria for visual analysis is defined and the selected algorithms presented.

7.3.1 Dimensionality Redution

Altogether four algorithms for evaluation has been selected, which use scikit[1] implementation.

1. *PCA* - represents linear dimensionality reduction techniques. The implementation uses singular value decomposition. Time complexity is $O(n^3)$

2. *LLE* - represents nonlinear dimensionality techniques that preserve local structure. The overall complexity of standard LLE is $\mathcal{O}(D * log(k) * N * log(N)) + \mathcal{O}(D * N * k^3) + \mathcal{O}(d * N^2)$ where N - number of training data points, D - input dimension, k - number of nearest neighbors, d - output dimension.
   Following tune parameters has been used *n_neighbors=5, n_components=2, reg=0.001,eigen_solver='auto', tol=1e-06, max_iter=100, method='standard', hessian_tol=0.0001, modified_tol=1e-12,neighbors_algorithm='auto', random_state=None*

3. *Isomap* - represents nonlinear dimensionality techniques that preserve global structure. The overall complexity of Isomap is $\mathcal{O}(D * log(k) * N * log(N)) + \mathcal{O}(N^2(k + log(N))) + \mathcal{O}(d * N^2)$ where N - number of training data points, D - input dimension, k - number of nearest neighbors, d - output dimension.
   Following tune parameters have been used *n_neighbors=5, n_components=2, eigen_solver='auto', tol=0, max_iter=None,path_method='auto', neighbors_algorithm='auto'*

4. *T-SNE* - a prize-winning[2] algorithm which represents state of the art dimensionality reduction techniques.
   Following tune parameters have been used *n_components=2, perplexity=30.0, early_exaggeration=4.0, learning_rate=1000.0,n_iter=1000, n_iter_without_progress=30, min_grad_norm=1e-07, metric='euclidean', init='random', verbose=0,random_state=None, method='barnes_hut', angle=0.5*

Please note that most of other dimensionality reduction algorithms are only variations of the selected ones.

[1] `www.scikit-learn.org` Machine Learning in Python
[2] http://blog.kaggle.com/2012/11/02/t-distributed-stochastic-neighbor-embedding-wins-merck-viz-challenge/

Metric Data

To evaluate the algorithms in a tangible way, scores as follows have been assigned

1. *Very Poor Visualization* – there aren't any recognizable patterns in the low-dimensional representation compliant with the high-dimensional manifold.

2. *Poor Visualization* – the majority of patterns in the low-dimensional representation isn't compliant with the high-dimensional manifold.

3. *Somehow Good, Somehow Poor* – local structure is well preserved, but global not or vice-versa.

4. *Good* – the majority of patterns in the low-dimensional representation is compliant with the high-dimensional manifold, but there are some visible inconsistencies.

5. *Very Good* – the majority of patterns in the low-dimensional representation are compliant with the high-dimensional manifold, without any visible inconsistencies.

Categorical Data

As already mentioned, a survey with alltogether 11 people has been done to evaluate the quality of dimensionality reduction algorithms for visualization enhancement of categorical data after user-defined map $f : \mathbb{S}tring \rightarrow \mathbb{N}umber$ has been applied. Since asking concrete questions about the low-dimensional data would have subjectified results much more, users have been asked to describe their interpretation of the data, patterns, structures, etc in a free text. Answers are then compared with high-dimensional representation for compliance.

Users had to answer three questions as follows

1. How would you classify the visual representation?

    a) *Very poor* - Most of the patterns that I recognize contradict reality (for instance, companies that are similar such as BMW and Audi are not in the same neighborhood or companies that are very different such as Turkish Airlines and Spotify are grouped together)

    b) *Poor* - I can't really recognize similarities, patterns or structures

c) *Somehow good, somehow poor* - I can recognize some patterns compliant with reality (for instance, companies that are similar such as BMW and Audi are in the same neighborhood or companies that are very different such as Turkish Airlines and Spotify are far away from each other), but I also recognize a fair amount of inconsistencies that contradict reality

d) *Good* - Most of the patterns that I recognize are compliant with reality, but there are some visible inconsistencies

e) *Very good* - Most of the patterns that I recognize are compliant with reality without visible inconsistencies

2. Did 3D representation enhance visual analysis?

3. Please explain your interpretation of the visualization (Patterns, Structures, Outliers, Relationships, Inconsistencies, Accuracies)

Furthermore, this is the mapping function $f : \mathbb{S}tring \rightarrow \mathbb{N}umber$ that has been used.

1. *Company Type*
   $f : e - business = 1$
   $f : non - e - business = 75$

2. *Segments*
   $f : B - to - B = 1$
   $f : B - to - B \quad and \quad B - to - C = 75$
   $f : B - to - C = 150$

3. *Value Proposition*
   $Low \quad cost = 1$
   $Differentiation \quad and \quad low \quad cost = 75$
   $Differentiation = 150$

4. *Relationships*
   $Low \quad level \quad service = 1$
   $Medium \quad level \quad service = 75$
   $High \quad level \quad service = 150$

5. *Channels*
   $Offline = 1$
   $Combination \quad of \quad online \quad and \quad offline = 75$
   $Online = 150$

6. *Partnerships*

$Not \quad many \quad important \quad partners = 1$

$Some \quad important \quad partners = 75$

$Many \quad important \quad partners = 150$

7. *Revenue model*

$Sales \quad of \quad access = 1$

$Sales \quad of \quad ownershipandaccess = 75$

$Sales \quad of \quad ownership = 150$

8. *Activities*

$Sales = 1$

$Services, Sales = 1$

$R\&D, Production = 30$

$R\&D, Production, Services = 30$

$R\&D, Production, Sales = 30$

$R\&D, Sales = 30$

$Programming, Sales = 80$

$Programming, Services, Sales = 80$

$Programming, Sales, Services = 80$

$Programming, Service = 80$

$Production, Programming = 80$

$Programming, Services = 80$

$Programming = 80$

$Logistics, Programming, Sales = 100$

$Programming, Logistics, Sales = 100$

$Production, Sales = 140$

$Production, Services = 140$

$Production = 140$

$Logistics, Sales = 140$

$Logistik, Production, Sales = 180$

9. *Resources*

$Physical = 1$

$Physical, Intellectual = 30$

$Physical, Human = 50$

$Physical, Human, Intellectual = 65$

$Human = 80$

$Human, IT = 130$

$Human, IT, Intellectual = 150$

10. *Industry field*

$Telecommunication = 1$
$Telecommunication = 1$
$Telecommunications = 1$
$Telecommunications \quad equipment = 1$
$Telecommunications \quad equipment, Semiconductors = 1$
$Geospatial, Transportation \quad and \quad Logistics, Telecommunications = 1$
$Internet, Computersoftware, Telecommunicationsequipment = 30$
$Financialservices = 30$
$Travel \quad Services = 40$
$Travel \quad Services, Internet = 40$
$ManagementConsulting = 50$
$ITservices, ITconsulting = 50$
$Videogames, interactiveentertainment = 60$
$Mobilegames, Computerandvideogames = 60$
$Videogames, interactiveentertainment, consumer \quad electronics = 60$
$Digitalimaging, Photography, Electronics = 60$
$Internet, electroniccommerce, online \quad auction \quad hosting = 65$
$Electronic \quad commerce = 65$
$Consumer \quad Electronics = 65$
$Electronics = 65$
$Onlineretailer = 70$
$Hardware, Software, Online \quad Retailing = 70$
$OnlineRetailing, Cloud \quad Computing = 75$
$Managed \quad cloud \quad computing = 80$
$Information \quad Technology = 80$
$Internet, Software = 80$
$Enterprise \quad software, Computerhardware = 80$
$Consumer \quad electronics = 80$
$Technology \quad Licensing = 80$
$Enterprise \quad software = 80$
$Software = 80$
$IT \quad infrastructure = 80$
$Internet = 80$
$Technology, Internet = 80$
$Semiconductors = 90$
$Technology \quad Consulting, EngineeringServices = 110$
$Consumer \quad electronics, Automotive, Licensing, Telematics = 120$
$Semiconductors, ComputerNetworks, LightingCircuitprotection = 150$
$Hardware = 150$
$Robots = 150$
$Hardware, Electronics = 130$

$Automotive, Renewable \quad Energy \quad Storage \quad Systems = 200$
$Automotive = 200$
$Automotive, Aviation, Telematics = 220$
$Motorcyclemanufacturer = 230$
$MachineConstruction = 240$
$Manufacturing, Distribution = 240$
$Airline = 280$
$AircraftManufacturer = 280$
$Aerospace, Defense = 280$
$Pharmaceuticals = 400$
$Pharmaceuticals, SyntheticMaterials, Plant \quad Protection = 400$
$Retail, Healthcare = 400$
$SanitaryFittings = 500$
$Clothing, Accessories = 500$
$Apparels, accessories = 500$
$Personal = 500$
$Retail = 500$
$Clothing, consumer \quad goods \quad manufacture = 500$
$Apparel, accessories = 500$
$Jewelers, silversmiths = 500$
$Restaurants = 500$
$Music = 600$
$Audio \quad encoding/compression, Audio \quad noise \quad reduction = 600$
$Coffee \quad Shop = 700$

Please note that one of the disadvantages of this strategy is that information loss occurs when two categorical properties are not in a transitive relation, because numbers are always transitive.

7.3.2 Cluster Analysis

Because of time constraints, only one clustering algorithm was selected for evaluation. The choice is $DBSCAN$ and Figure 7.1 shows why.

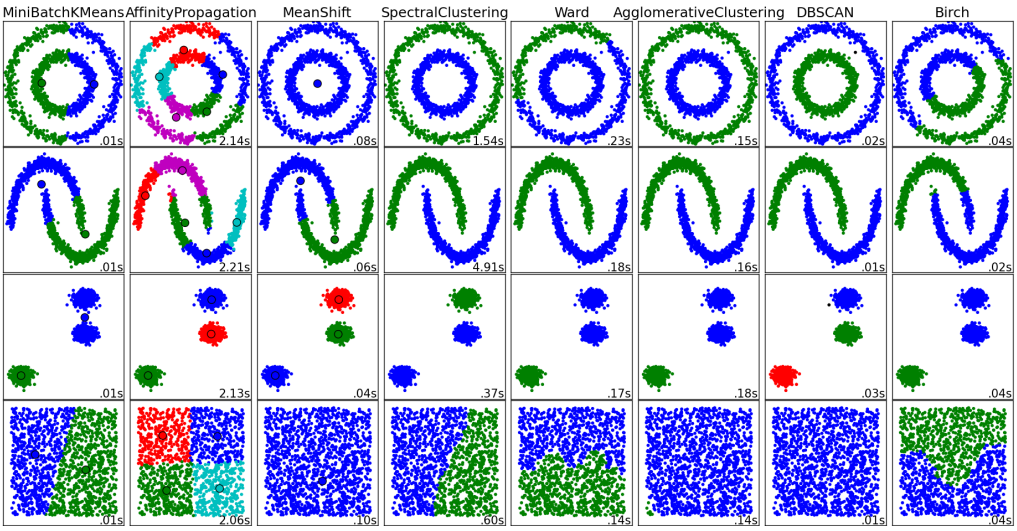


Fig. 7.1: A comparison of the clustering algorithms in scikit-learn. Adapted from [Ped+11]

This is the only algorithm that correctly, without any supervision, recognizes all clusters. Moreover, clusters in real world data most probably tend to have different density and $DBSCAN$ is density-based clustering algorithm. Admittedly, the choice of clustering algorithm is a pure speculation, because cluster can be defined in many different ways – it could be areas of different densities, it could be areas of different distribution or it could be areas that are far away from each other. As pointed in the paper [EC02] the definition of cluster is in the eye of the beholder. For the purposes of this paper clusters are defined as areas of different density and therefore $DBSCAN$ is chosen.
Following tune parameters have been used *eps=0.5, min_samples=5, metric='euclidean', algorithm='auto', leaf_size=30, p=None,random_state=None*

Metric Data

No specific settings. Cluster analysis applied, visualization yielded and adjusted rand index computed.

Categorical Data

Unfortunately, as previously mentioned, the used dataset has unknown high-dimensional structure (labels), therefore visual analysis is used instead of adjusted rand index. The same mapping function as for dimensionality reduction was used.

7.4 Evaluation Tool

One of the main requirements for this bachelor thesis was an evaluation tool where user without any technical knowledge is able to apply dimensionality reduction techniques on his own data sets.
The proposed solution is best of three worlds – state-of-the-art visualizations in a web browser, robust algorithms used directly from a machine-learning library and all the heavy lifting done on the server.

7.4.1 Functionality

1. User is able to choose between four state-of-the-art dimensionality reduction and one cluster analysis techniques.
2. User is able to fine-tune parameters of the implemented techniques.
3. User is able to apply the implemented techniques on default, prepared data sets.
4. User is able to apply the implemented techniques on his own data sets.
5. User is able to visualize 2D and 3D representation of the reduced data.
6. User is able to select only specific samples to be visualized.
7. User is able to see the initial features of a sample by hover with the mouse.
8. User is able to see general information about the data sets.
9. User is able to zoom in and rotate for better visualization.
10. User is able to define mapping $f : \mathbb{S}tring \rightarrow \mathbb{N}umber$ for categorical data.

7.4.2 Architecture

The evaluation tool is implemented as a web application, but could be deployed as standalone application for Mac, Linux and Windows by wrapping it with “Electron”.

These are the three main components

1. *Meteor* - meteor[3] is a full stack web application framework written in NodeJS. It integrates NoSQL database (MongoDB) and uses publish-subscribe pattern to automatically propagate data changes and synchronize server-client architecture. Both client and server code is written in JavaScript.

2. *Highcharts* - highcharts[4] is a data visualization library.

3. *Scikit* - scikit[5] project started as scikits.learn, a Google Summer of Code. This is a machine library written in python. All of the application's algorithms are based on scikit's implementation.

Communication flow

The high-level communication flow is defined follows

1. After the user has requested a visualization of his data set

2. Meteor spawns a child process which calls a python script to use scikit's functionality.

3. The child process returns with the result from scikit which contains the low-dimensional representation of the initial data

4. Meteor-server notifies meteor-client using the publish-subscribe design pattern.

5. Meteor-client gets the data and passes it to Highcharts.

6. Highcharts visualizes the data in the web view.

Fig. 7.2: Communication flow

Please note that the whole process is asynchronous, thus the main thread in the browser is not blocked at any time.

[3] `https://www.meteor.com/` Meteor is a full stack platform for web, mobile, and desktop.
[4] `http://www.highcharts.com/` MAKE YOUR DATA COME ALIVE
[5] `www.scikit-learn.org` Machine Learning in Python

Moreover, it is important to note that there are some papers [Scu] which try to optimize machine learning algorithms for web applications. In this case this is not needed, because the most computationally expensive part, namely 2, is run on the server.

7.5 Data Sets

Following datasets have been used for evaluation. Intentionally most of the examples represent real-world data sets. For completeness some toy data sets are also evaluated.

1. *Olivetti faces*

   Samples: 400
   Features: 10403
   Variables: Metric

   Real world dataset contains a set of face images taken between April 1992 and April 1994 at AT&T Laboratories Cambridge. As described on the original website: *There are ten different images of each of 40 distinct subjects. For some subjects, the images were taken at different times, varying the lighting, facial expressions (open / closed eyes, smiling / not smiling) and facial details (glasses / no glasses). All the images were taken against a dark homogeneous background with the subjects in an upright, frontal position (with tolerance for some side movement). [@Oli]*

2. *Images of a face*

   Samples: 698
   Features: 4096
   Variables: Metric

   Real world dataset representing images of a face with different left-right pose and brightness. The features correspond to the brightness of the pixels.

3. *COIL-20*

   Samples: 1440
   Features: 1024
   Variables: Metric

   As originally described in the technical documentation, *Columbia Object Image Library (COIL-20) is a database of gray-scale images of 20 objects. The objects were placed on a motorized turntable against a black background. The turntable*

was rotated through 360 degrees to vary object pose with respect to a fixed camera. Images of the objects were taken at pose interval of 5 degrees. This corresponds to 72 images per object. The database has two sets of images. The first set contains 720 unprocessed images of 10 objects. The second contains 1440 size normalized images of 20 objects [Nen+].

4. *Word-features*

   Samples: Unknown
   Features: Unknown
   Variables: Metric

   Real world datasets supplied by Andriy Mnih. Two words are similar if they have both a small pairwise distance and a similar color.

5. *Netflix Prize Dataset*

   Samples: 100480507
   Features: 17770
   Variables: Metric

   This is the official data set used in the Netflix Prize competition. The data consists of about 100 million movie ratings, and the goal is to predict missing entries in the movie-user rating matrix [Net09].

6. *MNIST dataset*

   Samples: 1000
   Features: 64
   Variables: Metric

   The MNIST database of handwritten digits, has a training set of 60,000 examples, and a test set of 10,000 examples. It is a subset of a larger set available from NIST. The digits have been size-normalized and centered in a fixed-size image. For the purposes of this bachelor thesis the first 1000 samples were fetched [@Mni].

7. *Iris Dataset*

   Samples: 150
   Features: 4
   Variables: Metric

   The dataset was introduced by Ronald Fisher in his paper [Fis36]. The data set consists of 50 samples from each of three species of Iris (setos, virginica

and versicolor). Four features were measured from each sample: the length and the width of the sepals and petals in cantimetres.

8. *Swiss Roll*

   Samples: 1500
   Features: 4
   Variables: Metric

   Data set which resembles swiss roll when plotted.

9. *Sphere*

   Samples: 1000
   Features: 4
   Variables: Metric

   Data set which resembles sphere when plotted.

10. *Business Data*

    Samples: 160
    Features: 10
    Variables: Categorical

    The data set represents about 160 companies with features and values as follows

    a) *Company type* - e-business; non-e-business

    b) *Segments* - B-to-B; B-to-C; B-to-B and B-to-C

    c) *Value Proposition* - Differentiation; low cost; differentiation and low cost

    d) *Relationships* - Low level service; medium level service; high level service

    e) *Channels* - Combination of online and offline; online; offline

    f) *Partnerships* - Not many important partners; some important partners; many important partners

    g) *Revenue model* - Sales of ownership and access; sales of access; sales of ownership

    h) *Activities* - Sales; Services, Sales; R&D, Production; R&D, Production, Services; R&D, Production, Sales; R&D, Sales; Programming, Sales; Programming, Services, Sales; Programming, Sales, Services; Programming,

Service; Production, Programming; Programming, Services; Programming; Logistics, Programming, Sales; Programming, Logistics, Sales; Production, Sales; Production, Services; Production; Logistics, Sales; Logistik, Production, Sales;

i) *Resources* - Physical; Physical, Intellectual; Physical, Human; Physical, Human, Intellectual; Human; Human, IT; Human, IT, Intellectual;

j) *Industry field* - Telecommunication; Telecommunication; Telecommunications; Telecommunications equipment; Telecommunications equipment, Semiconductors; Geospatial,Transportation and Logistics, Telecommunications; Internet, Computer software, Telecommunications equipment; Financial services; Travel Services; Travel Services, Internet; Management Consulting; IT services, IT consulting; Video games, interactive entertainment; Mobile games, Computer and video games; Video games, interactive entertainment, consumer electronics; Digital imaging, Photography, Electronics; Internet, electronic commerce, online auction hosting; Electronic commerce; Consumer Electronics; Electronics; Online retailer; Hardware, Software, Online Retailing; Online Retailing, Cloud Computing; Managed cloud computing; Information Technology; Internet, Software; Enterprise software, Computer hardware; Consumer electronics; Technology Licensing; Enterprise software; Software; IT infrastructure; Internet; Technology, Internet; Semiconductors; Technology Consulting, Engineering Services; Consumer electronics, Automotive, Licensing, Telematics; Semiconductors, Computer Networks, Lighting Circuit protection; Hardware; Robots; Hardware, Electronics; Automotive, Renewable Energy Storage Systems; Automotive; Automotive, Aviation, Telematics; Motorcycle manufacturer; Machine Construction; Manufacturing, Distribution; Airline; Aircraft Manufacturer; Aerospace, Defense; Pharmaceuticals; Pharmaceuticals, Synthetic Materials, Plant Protection; Retail, Health care; Sanitary Fittings; Clothing, Accessories; Apparels, accessories; Personal; Retail; Clothing, consumer goods manufacture; Apparel, accessories; Jewelers, silversmiths; Restaurants; Music; Audio encoding/compression, Audio noise reduction; Coffee Shop;

7.6 Results

Evaluation results of techniques for dimensionality reduction and data visualization are presented in this section.

7.6.1 Dimensionality Reduction

Firstly, let's explore dimensionality reduction and then cluster analysis is assesed.

Metric Data

In the following subsection every metric data set presented in section 7.5 is tested against each algorithm. Once again, scores are assigned as follows

1. *Very Poor Visualization* – there aren't any recognizable patterns in the low-dimensional representation compliant with the high-dimensional manifold.

2. *Poor Visualization* – the majority of patterns in the low-dimensional representation isn't compliant with the high-dimensional manifold.

3. *Somehow Good, Somehow Poor* – local structure is well preserved, but global not or vice-versa.

4. *Good* – the majority of patterns in the low-dimensional representation is compliant with the high-dimensional manifold, but there are some visible inconsistencies.

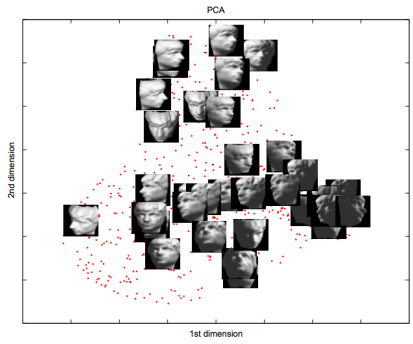
5. *Very Good* – the majority of patterns in the low-dimensional representation are compliant with the high-dimensional manifold, without any visible inconsistencies.

Now that scores assignment has been reminded, let's begin with the data sets.

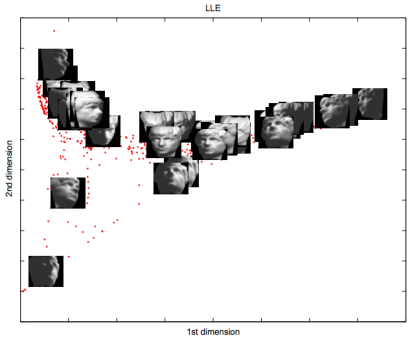
1. *Images of a face* - one could clearly see that PCA represents the data with visible incosistencies (bottom right) in both local and global structure. On the other hand, LLE has plotted the data pretty well locally and not so well globally. Isomap performs nearly perfectly. The authors of the cited paper didn't visualize the data with T-SNE.

   Final Scores

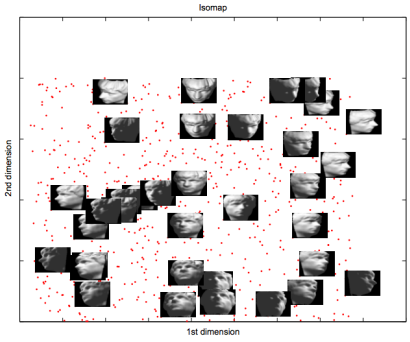
PCA:	3
LLE:	4
Isomap:	5
T-SNE:	X



(a) PCA



(b) LLE



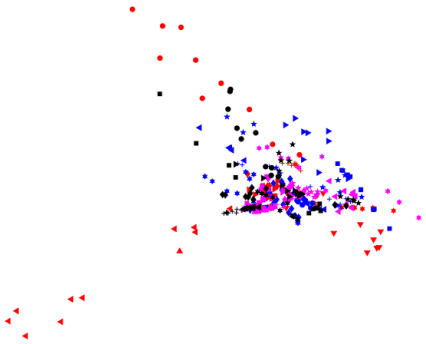
(c) Isomap

Fig. 7.3: Visualizations of the Images of a face data set. Adapted from [Li04]

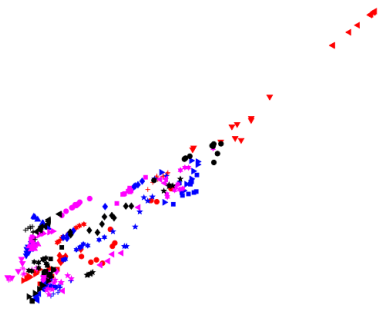
2. *Olivetti faces* - It is clear that LLE and Isomap don't really reveal the dependencies of the classes and thus the global structure of the data. The local structure is pretty good preserved, but there are some inconsistencies. T-SNE on the other hand deals pretty well with revealing both global and local structure of the data. The authors of the cited paper didn't visualize the data with PCA.

Final Scores

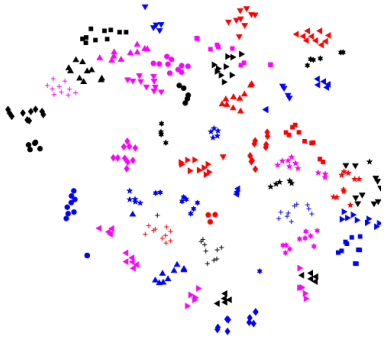
PCA:	X
LLE:	3
Isomap:	3
T-SNE:	5



(a) LLE



(b) Isomap



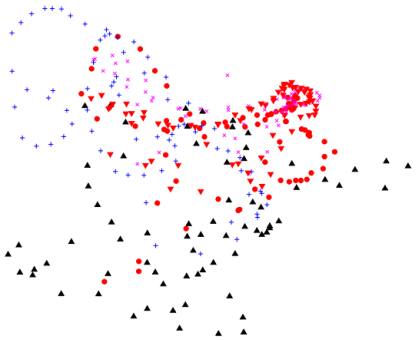
(c) T-SNE

Fig. 7.4: Visualizations of the Olivetti faces data set. Adapted from [Maa08a]

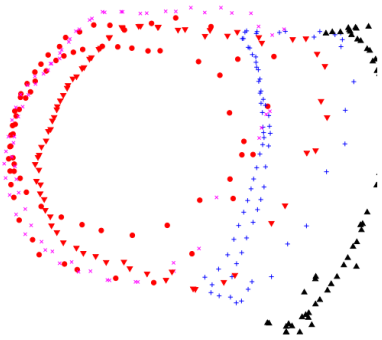
3. *COIL-20* - visualizations of Isomap and LLE are somehow poor. Local structure is once again consistent, but the relationships between the clusters are poorly presented. This is actually expected because LLE and Isomap are based on neighborhood and as stated in [Maa08a] those techniques are incapable of dealing with data sets consisting of widely separated submanifolds. On the other hand, t-sne performs once again pretty well revealing both local and global structure with some visible inconsistencies on bottom left.

Final Scores

PCA:	X
LLE:	3
Isomap:	3
T-SNE:	4



(a) LLE



(b) Isomap

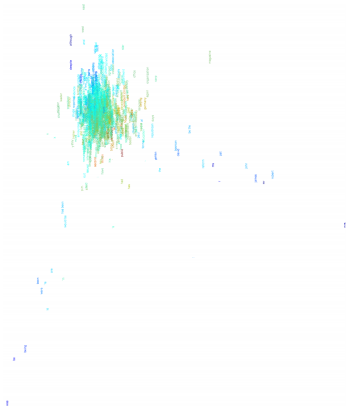
(c) T-SNE

Fig. 7.5: Visualizations of the COIL-20 data set. Adapted from [Maa08a]

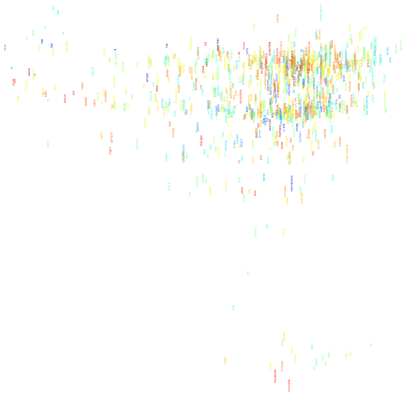
4. *Word-features* - note that words are similar if they have both similar color and are in the same neighborhood. LLE and Isomap obviously perform not so well retaining only local and accordingly global structure. However, T-SNE has some really good revelations – names of months are close together, also forms of verbs, professions, words related to time. But there are some inconsistencies presented also by T-SNE.

Final Scores

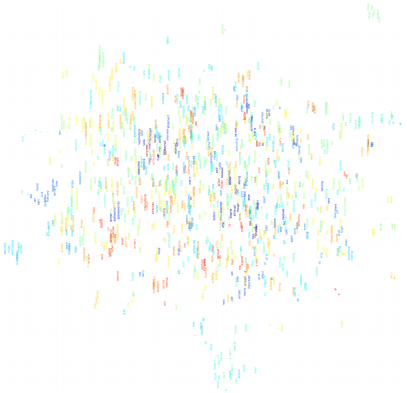
PCA:	X
LLE:	3
Isomap:	3
T-SNE:	4



(a) LLE



(b) Isomap



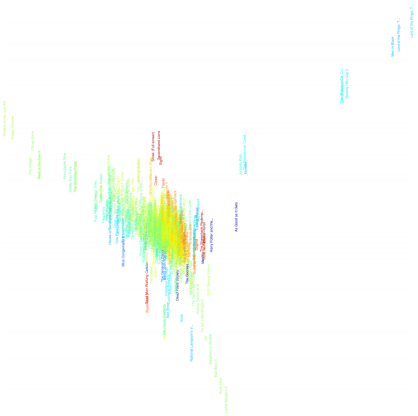
(c) T-SNE

Fig. 7.6: Visualizations of the Word features data set. Adapted from [Maa08a]

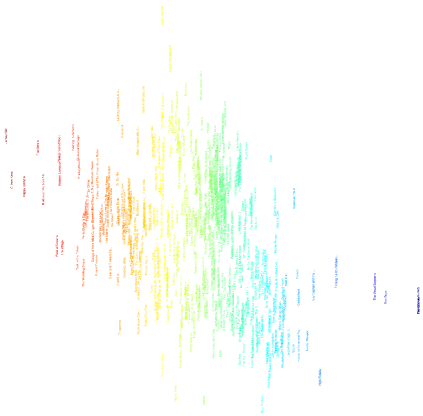
5. *Netflix Prize Dataset* - note that movies are similar if they have both similar color and are in the same neighborhood. As stated in the paper the visualization was adapted from, t-sne performs very well clustering together similar movies like Lord of the Rings & Harry Potter for instance, but global structure is poorly preserved. Results for LLE are also not good with poor global, but good local structure, however Isomap reveals the structure very well.

Final Scores

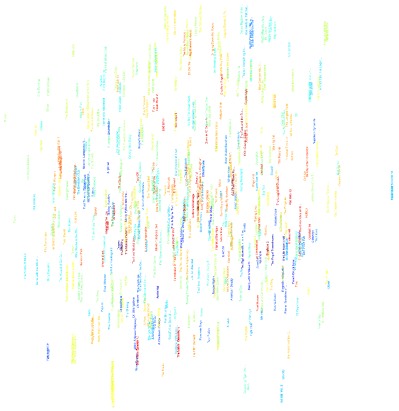
PCA:	X
LLE:	3
Isomap:	5
T-SNE:	3



(a) LLE



(b) Isomap

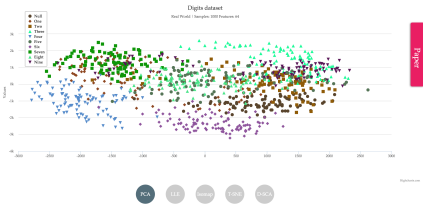


(c) T-SNE

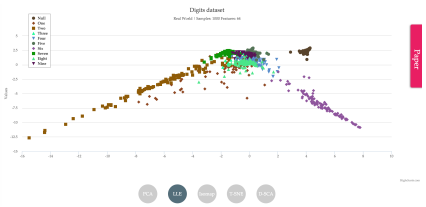
Fig. 7.7: Visualizations of the Netflix prize data set. Adapted from [Maa08a]

6. *MNIST Dataset* - obviously LLE preserves local structure extremely well, however global structure is somehow poorly revealed. Isomap also performs extremely well, although global structure is not perfect. On the other hand, T-SNE is almost perfect. PCA reveals both local and global structure somehow good, but also somehow poor.

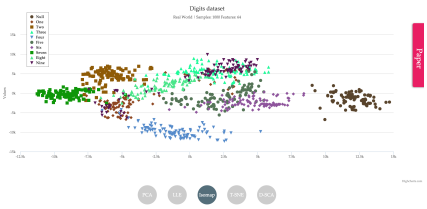
Final Scores	
PCA:	3
LLE:	3
Isomap:	4
T-SNE:	5



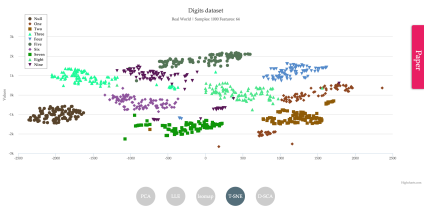
(a) PCA



(b) LLE



(c) Isomap



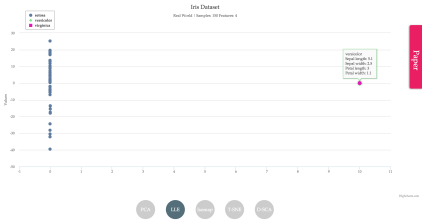
(d) T-SNE

Fig. 7.8: Visualizations of the MNIST data set.

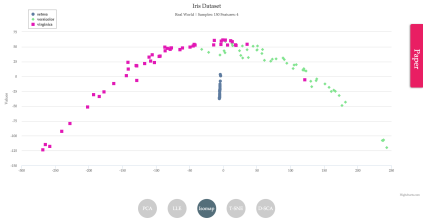
7. *Iris Dataset* - obviously Isomap, TSNE and PCA perform very well revealing both local and global data structure, while LLE fails to retain global structure at this level and groups setos and versicolor together.

Final Scores	
PCA:	5
LLE:	2
Isomap:	5
T-SNE:	4

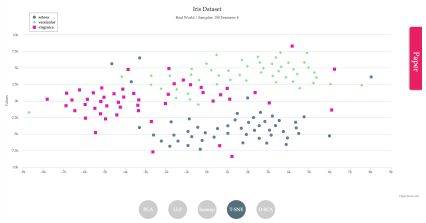
(a) PCA



(b) LLE



(c) Isomap



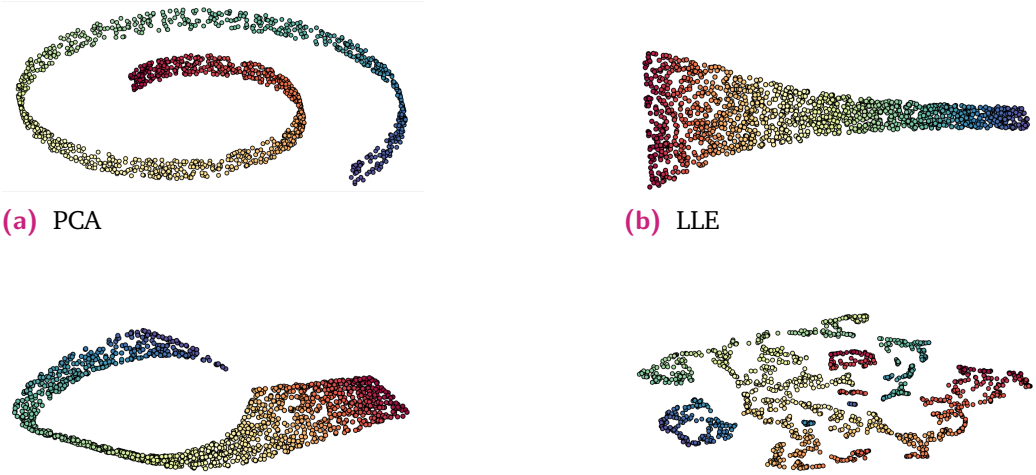
(d) T-SNE

Fig. 7.9: Visualizations of the Iris data set.

8. *Swiss Roll* - here PCA, LLE and Isomap preserve the swiss roll structure very well, whereby PCA performs better than both of them. T-SNE on the other hand fails to capture some patterns in both local and global structure.

Final Scores

PCA:	5
LLE:	4
Isomap:	4
T-SNE:	3



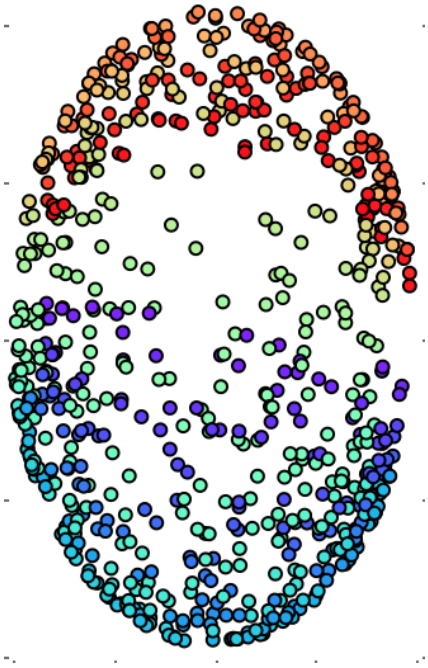
(a) PCA **(b)** LLE **(c)** Isomap **(d)** T-SNE

Fig. 7.10: Visualizations of the Swiss roll data set.

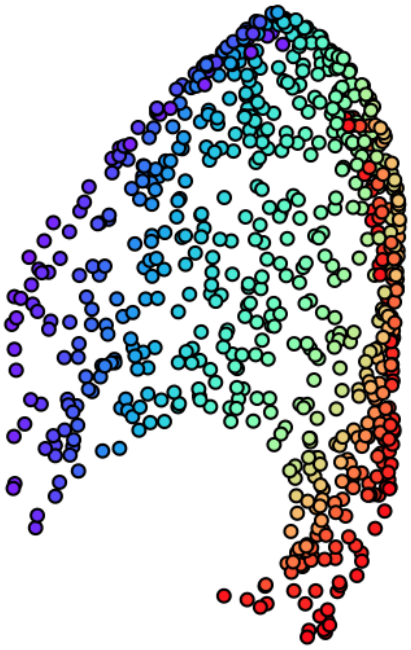
9. *Sphere* - PCA and LLE both reveal the sphere structure almost perfectly, while LLE and TSNE fail to retain global structure.

Final Scores

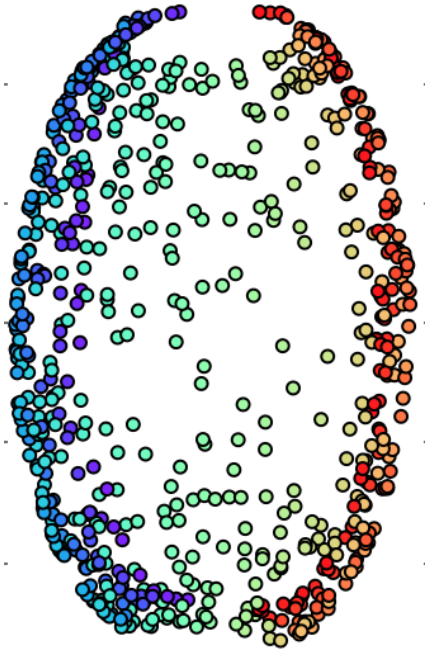
PCA:	5
LLE:	3
Isomap:	5
T-SNE:	3



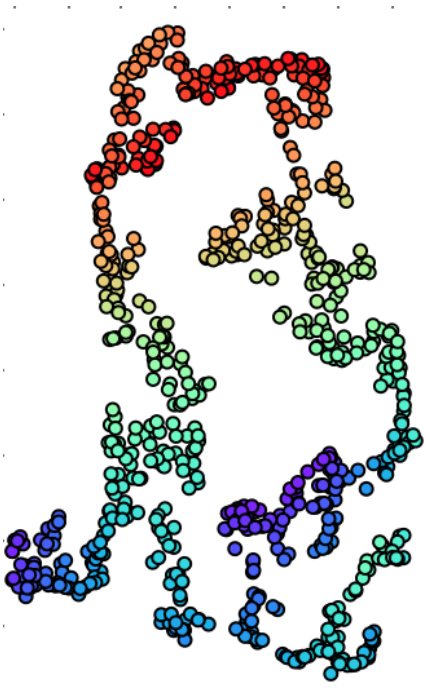
(a) PCA



(b) LLE



(c) Isomap



(d) T-SNE

Fig. 7.11: Visualizations of the Sphere data set.

There has also been made attempts to visualize some of the data sets in 3D, but results are far from satisfying, as shown on Figure 7.12

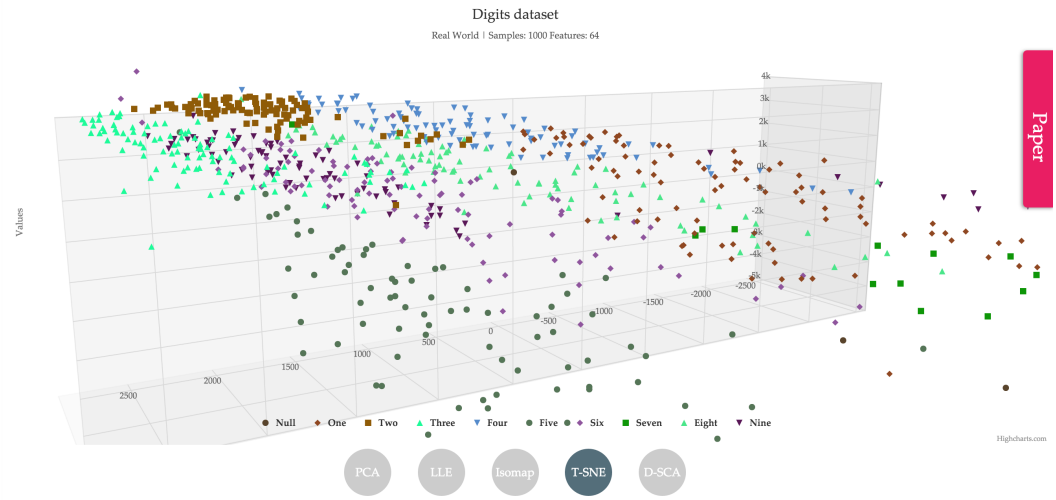


Fig. 7.12: 3D representation of MNIST reduced by T-SNE.

In Table 7.6.1 results of the evaluation for metric data are averaged

Final Average Scores	
PCA:	4.2
LLE:	3.11
Isomap:	4.11
T-SNE:	3.88

The evaluation clearly shows that dimensionality reduction algorithms reveal most of the manifold's structure, however in some cases there are inconsistencies and results are not always satisfying. Important to note is that PCA has turned to be the best unsupervised dimensionality algorithm according to this evaluation. Something that could also be confirmed in [Maa08a].

Categorical Data

There were alltogether 11 people available for the survey. Therefore, for more representitive results, after the first round of four algorithms being evaluated by four users, only two algorithms had to be selected based on the users' evaluation and the evaluation of the metric data set. In the first round visualization quality of LLE was marked somehow good, T-SNE was assesed as poor, however PCA and Isomap performed very well (more detailed results in Appendix 10). These results, except for T-SNE, were compliant with the performance of the algorithms tested against metric data, therefore PCA and Isomap were chosen for further evaluation. Moreover, T-SNE is non-deterministic and evaluation results wouldn't have been reproducible.

(a) LLE (b) T-SNE

Fig. 7.13: Visualizations of Business data for LLE and T-SNE

1. **PCA** - 5 people evaluated PCA. Without exception every and each one of them classified the visual representation as “Good - most of the patterns that I recognize are compliant with reality, but there are some visible inconsistencies”. The users were very consistent with their opinions and have identified all clusters - Healthcare, Retail, Automotive, Airlines and a smooth transition from Hardware to Semiconductors, Telecommunications, Software and Internet. The smooth transition from hardware to software and therefore the lack of clear separation in the big cluster that contains several sublusters on the right side was seen from people as something that is normal, because these industries overlap. Moreover, users identified Spotify, Starbucks and McDonalds as visible incosistencies. Nevertheless, this is because Spotify was categorized as a music company, but people think of it as software. This is not a weakness of PCA, but rather of the initial data set that has failed to describe companies in accordance to users’ perception.

Fig. 7.14: Visualizations of Business data for PCA

2. **Isomap** - 4 people evaluated Isomap, 3 of them classified the visual representation as “Good - most of the patterns that I recognize are compliant with reality, but there are some visible inconsistencies” and one as “Very Good -

most of the patterns that I recognize are compliant with reality without visible inconsistencies". The first impression of the people that has evaluated the algorithm was not very good, because clusters are not as clearly separated as in PCA. However, after careful visual analysis, all of them identified perfectly the underlying high-dimensional data structure. Again, all clusters have been recognized - Healthcare, Retail, Automotive, Airlines and a smooth transition from Hardware to Semiconductors, Telecommunications, Software and Internet. Certain users were more insightful to notice progression from low to high cost in regard to airline companies and from low to high service in the automotive industry. Visible inconsistencies here include also Rolls Royce Motor, Spotify, Starbucks and McDonalds, which as we said is not weakness of the algorithms itself, but of the initial data set.

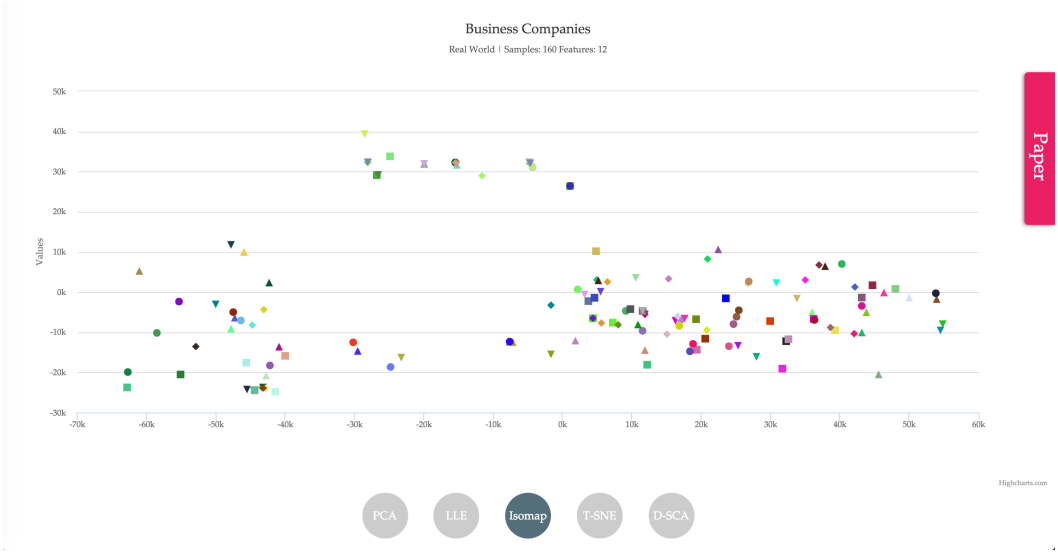


Fig. 7.15: Visualizations of Business data for Isomap

A lot of assumptions has to be made to apply dimensionality reduction on categorical data - for instance, that the mapping function $f : \mathbb{S}tring \rightarrow \mathbb{N}umber$ is "good" and, more importantly, that the attributes which describe the companies in the initial data set are consistent with how users perceive companies. Obviously the last assumption has made a negative impact and lead to information distortion in regard to Spotify, because most people consider Spotify as a software company and would categorize it similar to social media, but in the initial data set it was in the music industry. Nevertheless the results are very promising, since most people in the survey managed to describe the initial data set incredibly well and "would have grouped companies in similar way".

Note that 3D representation didn't enhance visual perception. Probably colors for the third dimension would have been more appropriate than adding another axis in

the space, as one of the evaluated users proposed. More detailed results of users' evaluation are attached in the Appendix 10.

7.6.2 Clustering

Metric and Categorical data

Cluster analysis was applied on handwritten digits, iris data set and business companies. Results were more than disappointing – the adjusted rand index was always nearly 0. Since further analysis wasn't encouraged, I decided to concentrate my efforts on dimensionality reduction algorithms. This probably implies that cluster analysis doesn't deal well with real-world data, real-world data is not really clustered or high-dimensional data is difficult to be clustered. Figure 7.16 show the results – items from different clusters have different y values. As it can be seen, cluster analysis shows that handwritten digits have just one cluster, which is very far from reality. An attempt to apply PCA prior cluster analysis was made in order to eliminate curse of dimensionality, but results were still extremely poor.

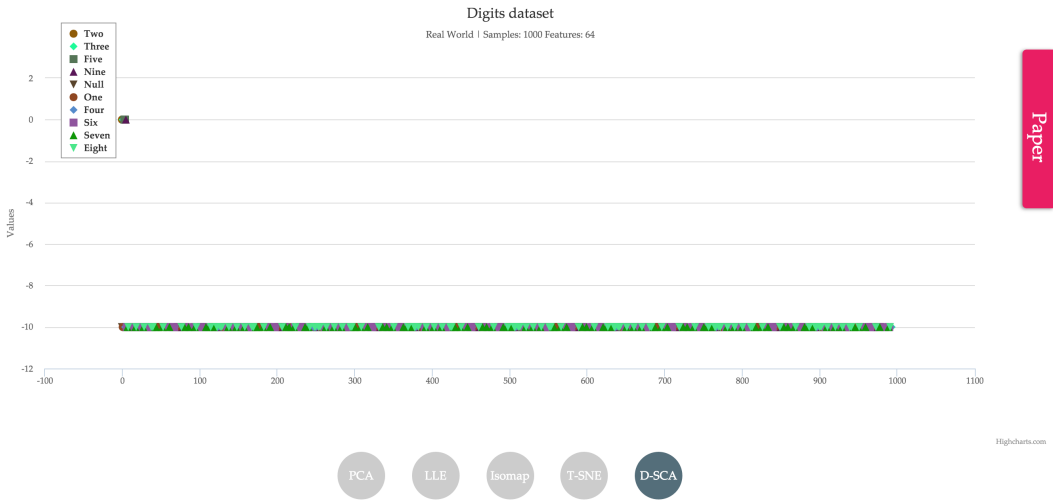


Fig. 7.16: DBSCAN of MNIST dataset. Different y values represent different clusters. Only one cluster detected. $AdjustedRandIndex = -1$

8 Discussion

In this paper different possibilities for recognizing complex patterns in high-dimensional data by dimensionality reduction and visual enhancement has been explored. They were mainly grouped in two categories – the first category concentrates on data adjustment, the second on visualization adjustment.

Techniques such as parallel coordinates and scatterplot matrices belong to the second group. The conclusion was drawn that current visualization techniques from this type couldn't be used to efficiently represent high-dimensional data due to scalability issues.

Then exploration of data adjustment techniques lead to the result that only feature extraction and cluster analysis will be taken into consideration, since only algorithms of such type were compliant with the requirements of this thesis. One of the main requirements was that the evaluated algorithms are unsupervised. As for dimensionality reduction, it became clear that a lot of techniques has been developed during the years, but most of them are only variations of one and the same technique, thus could be grouped in three categories

1. Linear
2. Locally nonlinear
3. Globally nonlinear

One algorithm of each group plus one novel algorithms were selected and evaluated. The evaluation of the selected algorithms for dimensionality reduction on metric data PCA, LLE, Isomap, TSNE has shown that in most cases fairly good results are achieved – at least local or global structure was recognizable by every algorithm. Moreover, despite of the fact that PCA was introduced about hundred years ago, it still performs at least as good as most of the novel techniques. Nevertheless, dimensionality reduction doesn't always lead to satisfying results. The problem is highlighted by the fact that there aren't any good mathematical means to evaluate the quality of such techniques. This is why heuristic approach by visual analysis has been taken to evaluate the techniques. An attempt has been made to visualize data in three instead of two dimensions, but no improvement was noticed.

These four algorithms were also tested against categorical data after user-defined mapping $f : \mathbb{S}tring \rightarrow \mathbb{N}umber$ has been applied. To evaluate results survey on the low-dimensional representation was done and people were able to identify crucial patterns in high-dimensional data from low-dimensional visualization. This implies that if the mapping function is good those techniques could be used to visualize categorical data. However, several shortcomings of this method were noticed: firstly, information loss occurs when two categorical properties are not in a transitive relation, because numbers are always transitive and secondly, sometimes there's inconsistency between how the initial data set describes data objects and how the user perceive them. Please consider the spotify example again.

Generally a lot of assumptions about the high-dimensional data have to be made when applying dimensionality reduction, because this is an ill-posed problem and the intrinsic dimensionality of the high-dimensional manifold is usually unknown, which can falsify results beyond what is desirable. In this sense, this is a very involved and complicated scientific field. However, if humans solve the challenges, that high-dimensional data brings, this would lead to a whole new world. Self-driving cars, smart cities, robots, predicting weather and human behavior wouldn't be only a taboo. This is why finding patterns in big data is such an important, exiting and at the same time difficult to master task.

Cluster analysis techniques has also been overviewed and one algorithm evaluated. Unfortunately the results from the evaluation of cluster analysis are not promising. PCA was applied also prior to cluster analysis to improve the results without any success. Either cluster analysis doesn't deal well with real-world data, clusters are difficult to identify in high-dimensional data or real world data is not really clustered.

9 Future Work

A basis for future work is of course appliance of dimensionality reduction for data visualization to categorical data after user-defined mapping is applied. The attempt, proposed in this thesis, to apply such algorithms on categorical data, has yielded very interesting and fairly good results. Namely, complex pattern in high-dimensional data were easily recognizable by visual analysis of reduced data. Unfortunately, there were also some inconsistencies. In future work such techniques could be applied on bigger number of data sets so that results are more representative. Moreover, this is an enormous field of science, therefore a lot of open questions arise, such as: Does combining dimensionality reduction techniques makes sense? Does dimensionality reduction helps to improve performance of other algorithms because of outliers removal and noise reduction? In which extend dimensionality reduction is affected by curse of dimensionality? How much is affected by noise in the data? How does supervised dimensionality reduction compare to unsupervised?

10 Conclusion

A subject of this paper is dimensionality reduction and data visualization algorithms. After comprehensive overview of those was proposed the most suitable of them for the purposes of this thesis were selected and evaluated. A novel attempt to apply suchlike techniques on categorical has also been successfully made. Although the algorithms yielded good results on most data sets, there were cases where information loss was far beyond what is acceptable. Moreover, cluster analysis completely failed to recognize clusters in real-world data. Generally the algorithms perform well, but not consistently, which implies that there is still a long way to go. However every effort is worth it, because this is a very mathematically involved and challenging subfield of computer science which brings a lot of opportunities. Its importance is highlighted by the fact that being able to analyze patterns in big data will push the development of human beings to another level. Behavior analytics, correct weather prediction based on past data, robots and extracting insights in neurology are only some of the use cases.

At last but not least, in the context of this thesis, a state-of-the-art web application for the enduser without any technical knowledge to apply dimensionality reduction and cluster analysis techniques on his own data sets in a simple and intuitive way was developed.

Bibliography

[BK] Guy Rosman Alexander M. Bronstein Michael M. Bronstein and Ron Kimmel. *Topologically Constrained Isometric Embedding*. Tech. rep. Technion Israel Institute of Technology (cit. on p. 21).

[BN] Mikhail Belkin and Partha Niyogi. *Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering*. Tech. rep. The University of Chicago (cit. on p. 20).

[BS98] Klaus-Robert Müller Bernhard Schölkopf Alexander Smola. „Nonlinear Component Analysis as a Kernel Eigenvalue“. In: *Neural Computation* 10 (1998), 1299–1319 (cit. on p. 17).

[CM98] Vladimir Cherkassky and Filip Mulier. *Learning from data: concepts, theory, and methods*. A Wiley-Interscience publication., 1998 (cit. on p. 6).

[DG] David L. Donoho and Carrie Grimes. *Hessian Eigenmaps:new locally linear embedding techniques for high-dimensional data*. Tech. rep. Stanford University (cit. on p. 19).

[Don] D. L. Donoho. „High-dimensional data analysis: The curses and bless-ings of dimensionality“. In: *Amer. Math. Soc. Lecture: Math challenges of the 21st century* () (cit. on p. 18).

[EC02] Vladimir Estivill-Castro. „Why so many clustering algorithms: a position paper“. In: *ACM SIGKDD Explorations Newsletter* 4 (2002), pp. 65–75 (cit. on p. 36).

[Fis36] R. A. Fisher. „The use of multiple measurements in taxonomic problems“. In: *Annals of Eugenics* 7 (1936), pp. 179–188 (cit. on p. 40).

[GR] Alexander M. Bronstein Ron Kimmel Guy Rosman Michael M. Bronstein. *Nonlinear Dimensionality Reduction by Topologically Constrained Isometric Embedding*. Tech. rep. Israel Institute of Technology (cit. on p. 11).

[HR] Geoffrey Hinton and Sam Roweis. *Stochastic Neighbor Embedding*. Tech. rep. University of Toronto (cit. on p. 21).

[HS89] T. Hastie and W. Stuetzle. „Principal curves“. In: *Journal of the American Statistical* (1989) (cit. on p. 16).

[I.T86] I.T.Jolliffe. *Principle Component Analysis*. Springer, 1986 (cit. on p. 15).

[Ize08] Alan Julian Izenman. *Modern Multivariate Statistical Techniques Regression, Classification, and Manifold Learning*. Springer, 2008 (cit. on pp. 6–8).

[Kog07] Jacob Kogan. *Introduction to Clustering LArge and High-Dimensional Data*. Cambridge University Press, 2007 (cit. on p. 29).

[Kra91] Mark A. Kramer. „Nonlinear principal component analysis using autoassociative neural networks“. In: *AlChe Journal* 37 (1991), 233–243 (cit. on p. 16).

[Li04] James Xinzhi Li. „Visualization of high-dimensional data with relational perspective map“. In: *Information Visualization* (2004), 49–59 (cit. on pp. 11, 20, 44).

[LKS] Sam T. Roweis Lawrence K. Saul. *An Introduction to Locally Linear Embedding*. Tech. rep. ATT Labs – Research, Gatsby Computational Neuroscience Unit, UCL (cit. on p. 19).

[Lon09] Tran Van Long. *Visualizing High-density Clusters in Multidimensional Data*. Tech. rep. Jacobs University, Dec. 2009 (cit. on p. 13).

[Maa] H.J. van den Herik L.J.P. van der Maaten E.O. Postma. *Dimensionality Reduction: A Comparative Review*. Tech. rep. Maastricht University (cit. on p. 11).

[Maa08a] Geoffrey Hinton Laurens van der Maaten. „Visualizing Data using t-SNE“. In: *Journal of Machine Learning Research* (2008), pp. 2579–2605 (cit. on pp. 11, 45–48, 52).

[Maa08b] Geoffrey Hinton Laurens van der Maaten. „Visualizing Data using t-SNE“. In: *Journal of Machine Learning Research* 9 (2008), pp. 2579–2605 (cit. on p. 21).

[Maa08c] H.J. van den Herik L.J.P. van der Maaten E.O. Postma. *Dimensionality Reduction: A Comparative Review*. Tech. rep. MICC, Maastricht University, P.O. Box 616, 6200 MD Maastricht, The Netherlands., Jan. 2008 (cit. on pp. 11, 15, 20, 28).

[MGMa] Dan Ventura Mike Gashler and Tony Martinez. *Iterative Non-linear Dimensionality Reduction by Manifold Sculpting*. Tech. rep. Brigham Young University (cit. on p. 11).

[MGMb] Dan Ventura Mike Gashler and Tony Martinez. *Iterative Non-linear Dimensionality Reduction by Manifold Sculpting*. Tech. rep. Brigham Young University (cit. on p. 21).

[Nen+] Sameer A. Nene, Shree K. Nayar, and Hiroshi Murase. *Columbia Object Image Library (COIL-20)*. Tech. rep. Columbia University (cit. on p. 40).

[Net09] Netflix. „Netflix Prize Data Set“. In: (2009) (cit. on p. 40).

[Non] Tech. rep. (cit. on pp. 16, 17).

[Ped+11] F. Pedregosa, G. Varoquaux, A. Gramfort, et al. „Scikit-learn: Machine Learning in Python“. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830 (cit. on pp. 6, 29, 36).

[RRC06] Stéphane Lafon Ronald R. Coifman. *Diffusion maps*. 2006 (cit. on p. 20).

[Sam69] J.W. Sammon. „A Nonlinear Mapping for Data Structure Analysis“. In: *Computers, IEEE Transactions on (Volume:C-18 , Issue: 5)* (1969), pp. 401–409 (cit. on pp. 18, 27).

[Scu] D. Sculley. *Web-Scale K-Means Clustering*. Tech. rep. Google (cit. on p. 39).

[SL] Michaël Aupetit Sylvain Lespinats. *False neighbourhoods and tears are the main mapping defaults. How to avoid it? How to exhibit remaining ones?* Tech. rep. Multisensor Intelligence and Machine Learning Laboratory (cit. on p. 27).

[SL09] P. Villemain J. Herault S. Lespinats B. Fertil. „RankVisu: Mapping from the neighborhood network“. In: *Neurocomputing* 72 (2009), 2964–2978 (cit. on p. 18).

[SLF07] Senior Member IEEE Alain Giron Sylvain Lespinats Michel Verleysen and Bernard Fertil. „DD-HDS: A Method for Visualization and Exploration of High-Dimensional Data“. In: *IEEE TRANSACTIONS ON NEURAL NETWORKS* 18 (2007), p. 1265 (cit. on p. 18).

[Suh] Diana D. Suhr. *Principal Component Analysis vs. Exploratory Factor Analysis*. Tech. rep. University of Northern Colorado (cit. on p. 16).

[Tor52] Warren S. Torgerson. „Multidimensional scaling: I. Theory and method“. In: *Psychometrika* 17 (1952), pp. 401–419 (cit. on p. 17).

[Ven07] Jarkko Venna. *DIMENSIONALITY REDUCTION FOR VISUAL EXPLORATION OF SIMILARITY STRUCTURES*. Tech. rep. Helsinki University of Technology, June 2007 (cit. on pp. 11, 25–28).

[Wan08] Jianzhong Wang. *Geometric Structure of High-Dimensional Data and Dimensionality Reduction*. Springer, Higher Education Press, 2008 (cit. on pp. 8, 19, 22).

[WS] Christian Walder and Bernhard Schölkopf. *Diffeomorphic Dimensionality Reduction*. Tech. rep. Max Planck Institute for Biological Cybernetics, Tubingen (cit. on p. 20).

[XX] Hans-Peter Kriegel Jörg Sander Xiaowei Xu Martin Ester. *A Distribution-Based Clustering Algorithm for Mining in Large Spatial Databases*. Tech. rep. University of Munich (cit. on p. 23).

[ZZa] ZHENYUE ZHANG and HONGYUAN ZHA. *PRINCIPAL MANIFOLDS AND NONLINEAR DIMENSION REDUCTION VIA LOCAL TANGENT SPACE ALIGNMENT*. Tech. rep. Zhejiang University, The Pennsylvania State University (cit. on p. 21).

[ZZb] Jing Wang Zhenyue Zhang. *MLLE: Modified Locally Linear Embedding Using Multiple Weights*. Tech. rep. Zhejiang University, Huaqiao University (cit. on pp. 11, 19).

Websites

[@JHO16] Michael Bostock Jeffrey Heer and Vadim Ogievetsky. *A Tour through the Visualization Zoo. A survey of powerful visualization techniques, from the obvious to the obscure*. Feb. 2016. URL: `http://delivery.acm.org/10.1145/1810000/1805128/p20-heer.html?ip=131.159.211.50&id=1805128&acc=OPEN&key=A8551CDF63AA2317%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35%2E6D218144511F3437&CFID=581459581&CFTOKEN=64737065&__acm__=1454945224_a0c1af144763ff6551b4a2f271256e12` (cit. on p. 25).

[@LW16] Johannes Lipp and Eric W. Weisstein. *Topological Space. @online*. Feb. 2016. URL: `http://mathworld.wolfram.com/TopologicalSpace.html` (cit. on p. 5).

[@Mni] URL: `http://yann.lecun.com/exdb/mnist/` (cit. on p. 40).

[@Oli] URL: http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html (cit. on p. 39).

[@SW16] Christopher Stover and Eric W. Weisstein. *Topological Space. @online*. Feb. 2016. URL: http://mathworld.wolfram.com/EuclideanSpace.html (cit. on p. 5).

[@Wei16] Eric W. Weisstein. *Map. @online*. Feb. 2016. URL: http://mathworld.wolfram.com/Map.html (cit. on p. 8).

List of Figures

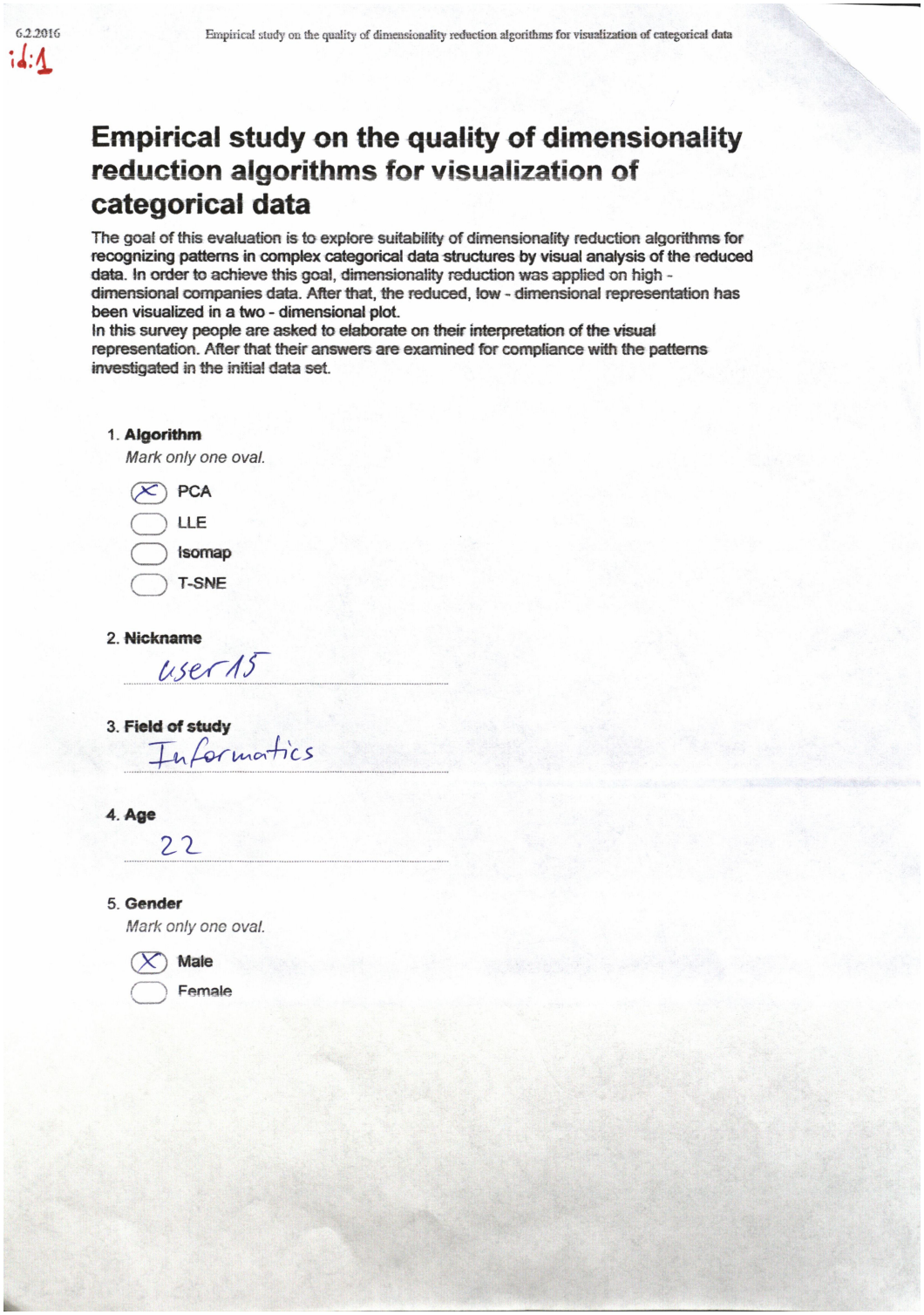
List of Tables

Colophon

This thesis was typeset with LaTeX 2_ε. It uses the *Clean Thesis* style developed by Ricardo Langner. The design of the *Clean Thesis* style is inspired by user guide documents from Apple Inc.

Download the *Clean Thesis* style at `http://cleanthesis.der-ric.de/`.

Appendix



6.2.2016 Empirical study on the quality of dimensionality reduction algorithms for visualization of categorical data

id:1

Empirical study on the quality of dimensionality reduction algorithms for visualization of categorical data

The goal of this evaluation is to explore suitability of dimensionality reduction algorithms for recognizing patterns in complex categorical data structures by visual analysis of the reduced data. In order to achieve this goal, dimensionality reduction was applied on high - dimensional companies data. After that, the reduced, low - dimensional representation has been visualized in a two - dimensional plot.
In this survey people are asked to elaborate on their interpretation of the visual representation. After that their answers are examined for compliance with the patterns investigated in the initial data set.

1. **Algorithm**
   Mark only one oval.
   - (X) PCA
   - ( ) LLE
   - ( ) Isomap
   - ( ) T-SNE

2. **Nickname**
   user15

3. **Field of study**
   Informatics

4. **Age**
   22

5. **Gender**
   Mark only one oval.
   - (X) Male
   - ( ) Female

6. **How would you classify the visual representation?**

Mark only one oval.

- ( ) Very poor - Most of the patterns that I recognize contradict reality (for instance, companies that are similar such as BMW and Audi are not in the same neighborhood or companies that are very different such as Turkish Airlines and Spotify are grouped together)
- ( ) Poor - I can't really recognize similarities, patterns or structures
- ( ) Somehow good, somehow poor - I can recognize some patterns compliant with reality (for instance, companies that are similar such as BMW and Audi are in the same neighborhood or companies that are very different such as Turkish Airlines and Spotify are far away from each other), but I also recognize a fair amount of inconsistencies that contradict reality
- (X) Good - Most of the patterns that I recognize are compliant with reality, but there are some visible inconsistencies
- ( ) Very good - Most of the patterns that I recognize are compliant with reality without visible inconsistencies

7. **Did 3D representation enhance visual analysis?**

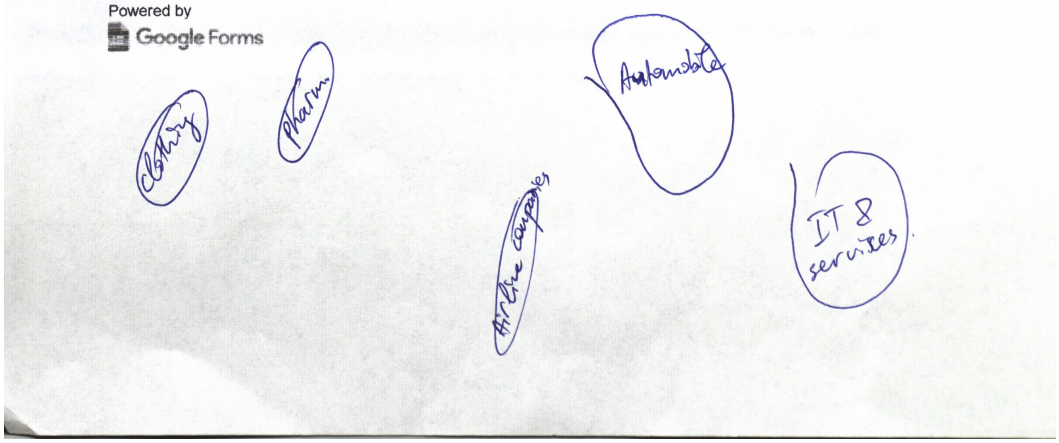
Mark only one oval.

- ( ) Yes
- (X) No
- ( ) Neutral

8. **Please explain your interpretation of the visualization (Patterns, Structures, Outliers, Relationships, Inconsistencies, Accuracies)**

In general the data clusters are well expressed and highly similar companies lie very close to each other. However, I find it strange that pairs of companies who share common relationships are sometimes displaced in one direction and sometimes in the other orthogonal.

Powered by Google Forms



id : 2

Empirical study on the quality of dimensionality reduction algorithms for visualization of categorical data

The goal of this evaluation is to explore suitability of dimensionality reduction algorithms for recognizing patterns in complex categorical data structures by visual analysis of the reduced data. In order to achieve this goal, dimensionality reduction was applied on high - dimensional companies data. After that, the reduced, low - dimensional representation has been visualized in a two - dimensional plot.
In this survey people are asked to elaborate on their interpretation of the visual representation. After that their answers are examined for compliance with the patterns investigated in the initial data set.

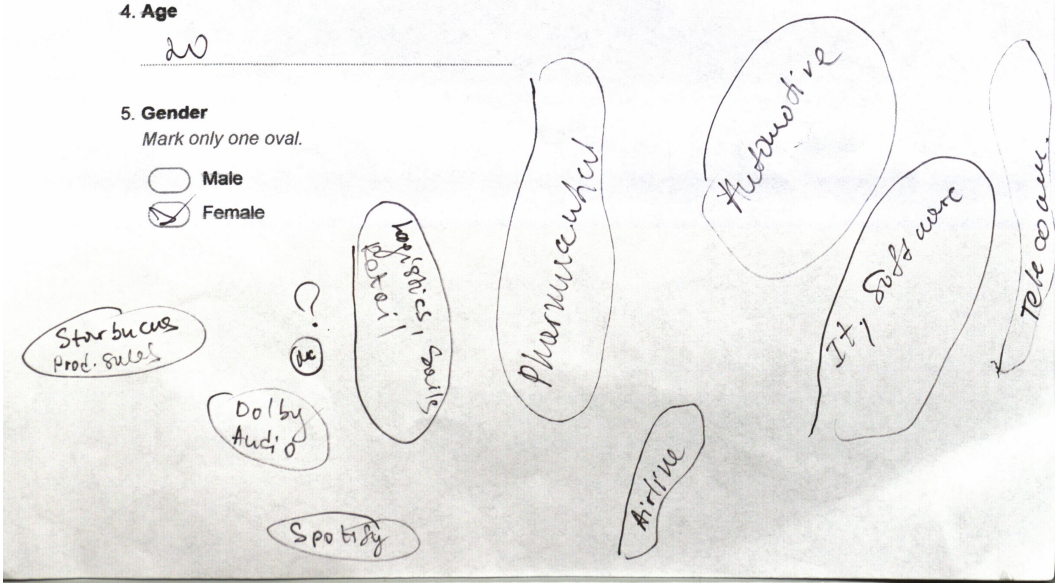
1. **Algorithm**
   Mark only one oval.
   - [x] PCA
   - [ ] LLE
   - [ ] Isomap
   - [ ] T-SNE

2. **Nickname**
   Dayana St

3. **Field of study**
   BWL

4. **Age**
   20

5. **Gender**
   Mark only one oval.
   - [ ] Male
   - [x] Female



6. How would you classify the visual representation?

Mark only one oval.

- ◯ Very poor - Most of the patterns that I recognize contradict reality (for instance, companies that are similar such as BMW and Audi are not in the same neighborhood or companies that are very different such as Turkish Airlines and Spotify are grouped together)
- ◯ Poor - I can't really recognize similarities, patterns or structures
- ◯ Somehow good, somehow poor - I can recognize some patterns compliant with reality (for instance, companies that are similar such as BMW and Audi are in the same neighborhood or companies that are very different such as Turkish Airlines and Spotify are far away from each other), but I also recognize a fair amount of inconsistencies that contradict reality
- ⊗ Good - Most of the patterns that I recognize are compliant with reality, but there are some visible inconsistencies
- ◯ Very good - Most of the patterns that I recognize are compliant with reality without visible inconsistencies

7. Did 3D representation enhance visual analysis?

Mark only one oval.

- ◯ Yes
- ⊗ No
- ◯ Neutral

8. Please explain your interpretation of the visualization (Patterns, Structures, Outliers, Relationships, Inconsistencies, Accuracies)

Most of the companies with similarities in Brandeon are concentrated in the same area (startup from left to right -> clothing companies (Production, Sales Retail), Pharmaceutical companies, Airline Comp., Automotive, It, Software, Telecommunications. What I find inconsistent is that MC Donalds is placed in the next to the group of companies in the clothing industry (~~cluster of outliers~~) and far away distanced from startups.

Powered by Google Forms

id:3

What I would categorize differently is that I would classify Spotify ~~and Netflix~~ to the group of online services such as Netflix, ebay. Over all I think that similar companies are grouped together without much information lost. It is also easy to identify clusters.

6. **How would you classify the visual representation?**

Mark only one oval.

- ( ) Very poor - Most of the patterns that I recognize contradict reality (for instance, companies that are similar such as BMW and Audi are not in the same neighborhood or companies that are very different such as Turkish Airlines and Spotify are grouped together)
- ( ) Poor - I can't really recognize similarities, patterns or structures
- ( ) Somehow good, somehow poor - I can recognize some patterns compliant with reality (for instance, companies that are similar such as BMW and Audi are in the same neighborhood or companies that are very different such as Turkish Airlines and Spotify are far away from each other), but I also recognize a fair amount of inconsistencies that contradict reality
- (✓) Good - Most of the patterns that I recognize are compliant with reality, but there are some visible inconsistencies
- ( ) Very good - Most of the patterns that I recognize are compliant with reality without visible inconsistencies

7. **Did 3D representation enhance visual analysis?**

Mark only one oval.

- ( ) Yes
- (✓) No
- ( ) Neutral

8. **Please explain your interpretation of the visualization (Patterns, Structures, Outliers, Relationships, Inconsistencies, Accuracies)**

The industry fealds are clearly separated as I would expect by the services they provied (food, cosmetics, healthcare, transport / separated in 3 groups - airline companies, automotive, motorcycle). What I found contradictory is that airlines are far away from automotive. Outher groups of companies that I've could see and belong to the same group are software, hardware, electronics groups.

Powered by Google Forms

id: 3

Empirical study on the quality of dimensionality reduction algorithms for visualization of categorical data

The goal of this evaluation is to explore suitability of dimensionality reduction algorithms for recognizing patterns in complex categorical data structures by visual analysis of the reduced data. In order to achieve this goal, dimensionality reduction was applied on high - dimensional companies data. After that, the reduced, low - dimensional representation has been visualized in a two - dimensional plot.
In this survey people are asked to elaborate on their interpretation of the visual representation. After that their answers are examined for compliance with the patterns investigated in the initial data set.

1. **Algorithm**
   Mark only one oval.
   - ( ) PCA
   - ( ) LLE
   - (√) Isomap
   - ( ) T-SNE

2. **Nickname**
   User 7

3. **Field of study**
   BWL

4. **Age**
   20

5. **Gender**
   Mark only one oval.
   - ( ) Male
   - (√) Female

id:4

Empirical study on the quality of dimensionality reduction algorithms for visualization of categorical data

The goal of this evaluation is to explore suitability of dimensionality reduction algorithms for recognizing patterns in complex categorical data structures by visual analysis of the reduced data. In order to achieve this goal, dimensionality reduction was applied on high - dimensional companies data. After that, the reduced, low - dimensional representation has been visualized in a two - dimensional plot.
In this survey people are asked to elaborate on their interpretation of the visual representation. After that their answers are examined for compliance with the patterns investigated in the initial data set.

1. **Algorithm**
   Mark only one oval.
   - ◯ PCA
   - ◯ LLE
   - ◯ Isomap
   - ⊗ T-SNE

2. **Nickname**
   User4

3. **Field of study**
   Informatik

4. **Age**
   21

5. **Gender**
   Mark only one oval.
   - ⊗ Male
   - ◯ Female

6. **How would you classify the visual representation?**

Mark only one oval.

- ( ) Very poor - Most of the patterns that I recognize contradict reality (for instance, companies that are similar such as BMW and Audi are not in the same neighborhood or companies that are very different such as Turkish Airlines and Spotify are grouped together)
- (X) Poor - I can't really recognize similarities, patterns or structures *in between*
- (X) Somehow good, somehow poor - I can recognize some patterns compliant with reality (for instance, companies that are similar such as BMW and Audi are in the same neighborhood or companies that are very different such as Turkish Airlines and Spotify are far away from each other), but I also recognize a fair amount of inconsistencies that contradict reality
- ( ) Good - Most of the patterns that I recognize are compliant with reality, but there are some visible inconsistencies
- ( ) Very good - Most of the patterns that I recognize are compliant with reality without visible inconsistencies

7. **Did 3D representation enhance visual analysis?**

Mark only one oval.

- ( ) Yes
- (X) No
- ( ) Neutral

8. **Please explain your interpretation of the visualization (Patterns, Structures, Outliers, Relationships, Inconsistencies, Accuracies)**

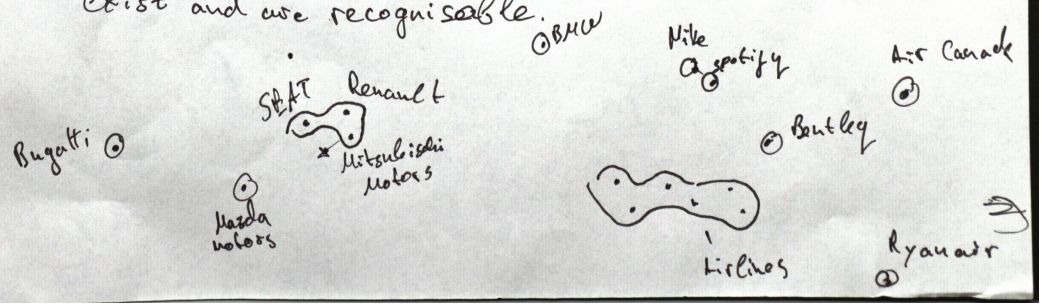
I would expect that companies like Bentley and Bugatti should be close together. Same goes for Ryanair and Air Canada, which are far away from the airline cluster and each other.

Conclusion: The algorithm doesn't accurately represent relations and similarities between companies that i would otherwise expect.

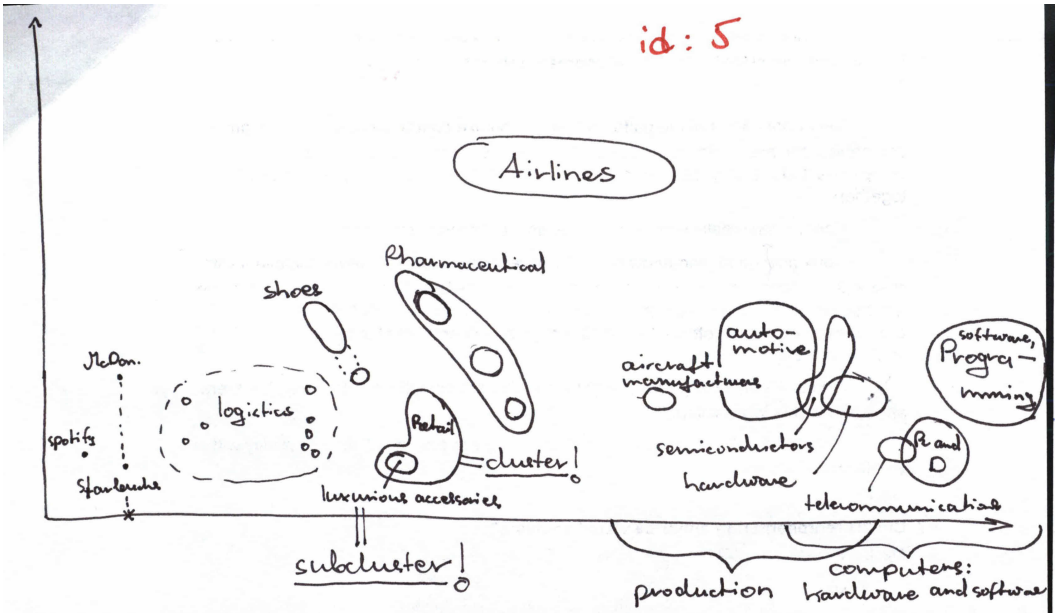
In my mind Spotify and Nike should be no where near each other.

The good part was that some partial patterns do exist and are recognisable.

Powered by Google Forms



id: 5


Airlines
Pharmaceutical
shoes
McDon.
logistics
spotify
Starbucks
Retail
cluster!
luxurious accessories
subcluster!
auto-
motive
aircraft
manufacturers
semiconductors
hardware
software,
Progra-
mming
R and
D
telecommunication
production
computers:
hardware and software

6. **How would you classify the visual representation?** id:5

Mark only one oval.

- ◯ Very poor - Most of the patterns that I recognize contradict reality (for instance, companies that are similar such as BMW and Audi are not in the same neighborhood or companies that are very different such as Turkish Airlines and Spotify are grouped together)
- ◯ Poor - I can't really recognize similarities, patterns or structures
- ◯ Somehow good, somehow poor - I can recognize some patterns compliant with reality (for instance, companies that are similar such as BMW and Audi are in the same neighborhood or companies that are very different such as Turkish Airlines and Spotify are far away from each other), but I also recognize fair amount of inconsistencies that contradict reality
- ⊗ Good - Most of the patterns that I recognize are compliant with reality, but there are some visible inconsistencies
- ◯ Very good - Most of the patterns that I recognize are compliant with reality without visible inconsistencies

7. **Did 3D representation enhance visual analysis?**

Mark only one oval.

- ◯ Yes
- ⊗ No
- ◯ Neutral

8. **Please explain your interpretation of the visualization (Patterns, Structures, Outliers, Relationships, Inconsistencies, Accuracies)**

- Companies are grouped in clusters according to their branche and/or activities (see on the back: sketch)
- Logically positioned clusters relatively to each other:
  semi conductors and hardware close to and at places covering with automotive → logical because semi cond. and hardw. are part of every up to date car
- One big cluster with production which includes subclusters automotive, aircrafts and hardware, semi-cond.

Powered by Google Forms

- Overall: Clusters that includes several sub clusters
  ↳ for example: Retail (luxurious & accessories)
- starbucks a bit oddly on the left, but at the same x-axis position as McDonalds
- spotify alone on the left, but only exception for software
- To sum up, Clusters grouping meets my expectations, (information) there are a few inconsistencies, (but) which don't lead to overall falsification and distortion

id: 5

Empirical study on the quality of dimensionality reduction algorithms for visualization of categorical data

The goal of this evaluation is to explore suitability of dimensionality reduction algorithms for recognizing patterns in complex categorical data structures by visual analysis of the reduced data. In order to achieve this goal, dimensionality reduction was applied on high - dimensional companies data. After that, the reduced, low - dimensional representation has been visualized in a two - dimensional plot.
In this survey people are asked to elaborate on their interpretation of the visual representation. After that their answers are examined for compliance with the patterns investigated in the initial data set.

1. **Algorithm**
   Mark only one oval.
   - ( ) PCA
   - ( ) LLE
   - (✓) Isomap
   - ( ) T-SNE

2. **Nickname**
   User 3

3. **Field of study**
   Mechanical Engineering (Biomedical Engineering)

4. **Age**
   21

5. **Gender**
   Mark only one oval.
   - (✓) Male
   - ( ) Female

6. **How would you classify the visual representation?** id:6

Mark only one oval.

- ( ) Very poor - Most of the patterns that I recognize contradict reality (for instance, companies that are similar such as BMW and Audi are not in the same neighborhood or companies that are very different such as Turkish Airlines and Spotify are grouped together)
- ( ) Poor - I can't really recognize similarities, patterns or structures
- ( ) Somehow good, somehow poor - I can recognize some patterns compliant with reality (for instance, companies that are similar such as BMW and Audi are in the same neighborhood or companies that are very different such as Turkish Airlines and Spotify are far away from each other), but I also see inconsistencies that contradict reality
- (✓) Good - Most of the patterns that I recognize are compliant with reality, but there are some visible inconsistencies
- ( ) Very good - Most of the patterns that I recognize are compliant with reality without visible inconsistencies

7. **Did 3D representation enhance visual analysis?**

Mark only one oval.

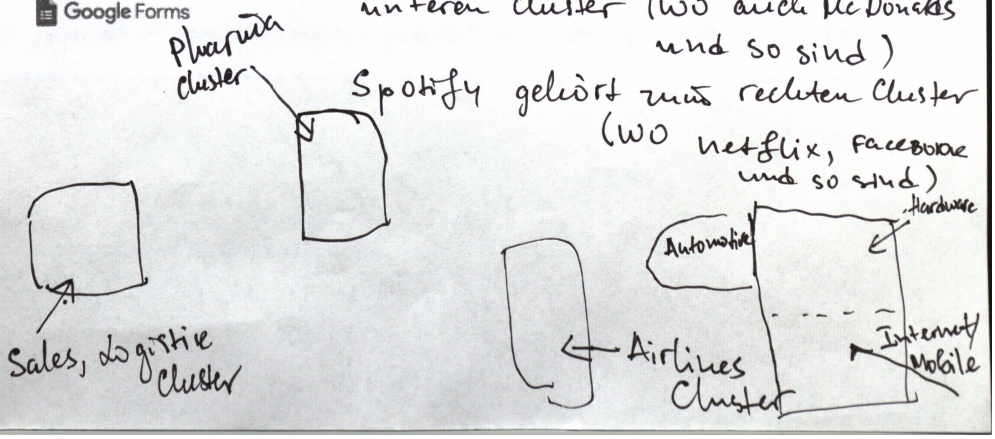
- ( ) Yes
- (✓) No
- ( ) Neutral

8. **Please explain your interpretation of the visualization (Patterns, Structures, Outliers, Relationships, Inconsistencies, Accuracies)**

- 5 Cluster
- 2 davon sehr eng zusammen
- 2 Cluster (~~Sales/Logistik~~ Internet/Mobile) zusammengewachsen
- Sales / Logistics cluster (links)
- ~~Software~~ cluster rechts
- Outliers: Starbucks gehört eigentlich zum unteren Cluster (wo auch McDonalds und so sind)

Spotify gehört zum rechten Cluster (wo Netflix, Facebook und so sind)

Powered by Google Forms



id: 6

Empirical study on the quality of dimensionality reduction algorithms for visualization of categorical data

The goal of this evaluation is to explore suitability of dimensionality reduction algorithms for recognizing patterns in complex categorical data structures by visual analysis of the reduced data. In order to achieve this goal, dimensionality reduction was applied on high - dimensional companies data. After that, the reduced, low - dimensional representation has been visualized in a two - dimensional plot.
In this survey people are asked to elaborate on their interpretation of the visual representation. After that their answers are examined for compliance with the patterns investigated in the initial data set.

1. **Algorithm**
   Mark only one oval.
   - [x] PCA
   - [ ] LLE
   - [ ] Isomap
   - [ ] T-SNE

2. **Nickname**
   User 1

3. **Field of study**
   Mathematics

4. **Age**
   26

5. **Gender**
   Mark only one oval.
   - [x] Male
   - [ ] Female

6. **How would you classify the visual representation?** id:7

Mark only one oval.

- ( ) Very poor - Most of the patterns that I recognize contradict reality (for instance, companies that are similar such as BMW and Audi are not in the same neighborhood or companies that are very different such as Turkish Airlines and Spotify are grouped together)
- ( ) Poor - I can't really recognize similarities, patterns or structures
- ( ) Somehow good, somehow poor - I can recognize some patterns compliant with reality (for instance, companies that are similar such as BMW and Audi are in the same neighborhood or companies that are very different such as Turkish Airlines and Spotify are far away from each other), but I also recognize a fair amount of inconsistencies that contradict reality
- ( ) Good - Most of the patterns that I recognize are compliant with reality, but there are some visible inconsistencies
- (✓) Very good - Most of the patterns that I recognize are compliant with reality without visible inconsistencies

7. **Did 3D representation enhance visual analysis?**

Mark only one oval.

- ( ) Yes
- (✓) No, but adding a moving camera can greatly improve data visual perception
- ( ) Neutral

8. **Please explain your interpretation of the visualization (Patterns, Structures, Outliers, Relationships, Inconsistencies, Accuracies)**

Companies are clustered by activity, which is logical and helps me to visualize patterns between the companies and to see which companies are "similar". The similarity criteria fits my expectation. However, it can be inaccurate, depending on how much it knows about the data. Possible improvements are to add more dimensions, but not spacial – rather colour and shape of data points with similar labels. Moreover, making the tool tip stationary will help browsing through data (make it faster) In the bottom right, the distance between companies from similar branches is also small, which is what I would expect. One more improvement is to show a particular label for all points (e.g. activity, Another issue is that I don't know perfectly how to cluster companies, because I don't Overall, I would group them in a similar way (or Value Proposition) have a goal in the like it is shown on the graphic below. clustering procedure.

Powered by Google Forms

id:7

Sales
logistics
Pharma
Airlines
Auto motive
IT
services
consulting
Internet

id: 7

Empirical study on the quality of dimensionality reduction algorithms for visualization of categorical data

The goal of this evaluation is to explore suitability of dimensionality reduction algorithms for recognizing patterns in complex categorical data structures by visual analysis of the reduced data. In order to achieve this goal, dimensionality reduction was applied on high - dimensional companies data. After that, the reduced, low - dimensional representation has been visualized in a two - dimensional plot.
In this survey people are asked to elaborate on their interpretation of the visual representation. After that their answers are examined for compliance with the patterns investigated in the initial data set.

1. **Algorithm**
   Mark only one oval.
   - ◯ PCA
   - ◯ LLE
   - ☑ Isomap
   - ◯ T-SNE

2. **Nickname**
   martosss

3. **Field of study**
   CSE - Computational Science and Engineering

4. **Age**
   24

5. **Gender**
   Mark only one oval.
   - ☑ Male
   - ◯ Female

id: B

Empirical study on the quality of dimensionality reduction algorithms for visualization of categorical data

The goal of this evaluation is to explore suitability of dimensionality reduction algorithms for recognizing patterns in complex categorical data structures by visual analysis of the reduced data. In order to achieve this goal, dimensionality reduction was applied on high - dimensional companies data. After that, the reduced, low - dimensional representation has been visualized in a two - dimensional plot.
In this survey people are asked to elaborate on their interpretation of the visual representation. After that their answers are examined for compliance with the patterns investigated in the initial data set.

1. **Algorithm**
   Mark only one oval.
   - ◯ PCA
   - ⊗ LLE
   - ◯ Isomap
   - ◯ T-SNE

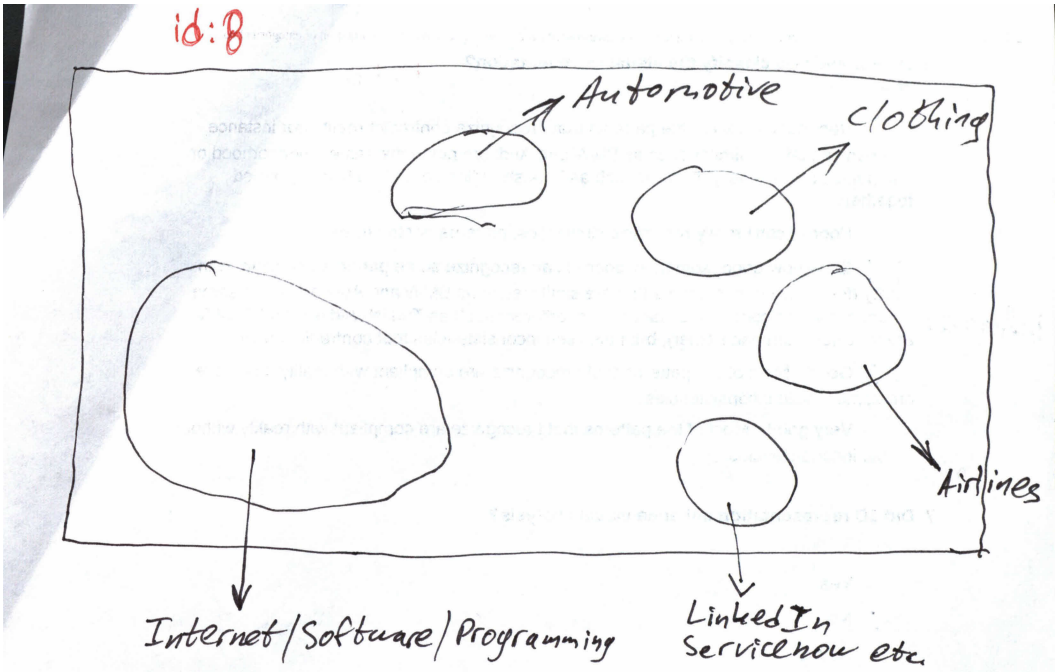
2. **Nickname**
   User 2

3. **Field of study**
   Law

4. **Age**
   23

5. **Gender**
   Mark only one oval.
   - ⊗ Male
   - ◯ Female

id:8


Automotive
clothing
Airlines
Internet/Software/Programming
LinkedIn
Servicenow etc.

6. **How would you classify the visual representation?** id: 8

Mark only one oval.

- ◯ Very poor - Most of the patterns that I recognize contradict reality (for instance, companies that are similar such as BMW and Audi are not in the same neighborhood or companies that are very different such as Turkish Airlines and Spotify are grouped together)
- ◯ Poor - I can't really recognize similarities, patterns or structures
- ⊗ Somehow good, somehow poor - I can recognize some patterns compliant with reality (for instance, companies that are similar such as BMW and Audi are in the same neighborhood or companies that are very different such as Turkish Airlines and Spotify are far away from each other), but I also see inconsistencies that contradict reality

inBetween

- ⊗ Good - Most of the patterns that I recognize are compliant with reality, but there are some visible inconsistencies
- ◯ Very good - Most of the patterns that I recognize are compliant with reality without visible inconsistencies

7. **Did 3D representation enhance visual analysis?**

Mark only one oval.

- ◯ Yes
- ⊗ No
- ◯ Neutral

8. **Please explain your interpretation of the visualization (Patterns, Structures, Outliers, Relationships, Inconsistencies, Accuracies)**

Most of the patterns that I identify on the plot are compliant with the reality such as the fact that all the Pharmaceuticals that I see are located in one and the same cluster on the right side of the plot. There is also a separate cluster for clothes but the inconsistency there is that the car producer BMW is situated in it as well. It makes a good impressi on me that there is a separate Airlines cluster but the inconsistency there is that the clothing consumer goods manufacture Puma ires in the same cluster too. The plot contains also a separa cluster with software companies but the problem there is that the distance is a little bit wider than expected. The cluster with the programming providers Servicenow, Liveperson, Groupon Inc. and LinkedIn match my expectations about reality. One more disadvantag of the plot is that both Telecommunications equipment producers Blackberry and Motorola are situated far away from each other. The same refers to both Starbucks and McDonalds that are active in almost the same field of commerce.

Powered by Google Forms

6. **How would you classify the visual representation?** 16:9

Mark only one oval.

- ◯ Very poor - Most of the patterns that I recognize contradict reality (for instance, companies that are similar such as BMW and Audi are not in the same neighborhood or companies that are very different such as Turkish Airlines and Spotify are grouped together)
- ◯ Poor - I can't really recognize similarities, patterns or structures
- ◯ Somehow good, somehow poor - I can recognize some patterns compliant with reality (for instance, companies that are similar such as BMW and Audi are in the same neighborhood or companies that are very different such as Turkish Airlines and Spotify are far away from each other), but I also recognize a fair amount of inconsistencies that contradict reality
- ☒ Good - Most of the patterns that I recognize are compliant with reality, but there are some visible inconsistencies
- ◯ Very good - Most of the patterns that I recognize are compliant with reality without visible inconsistencies

7. **Did 3D representation enhance visual analysis?**

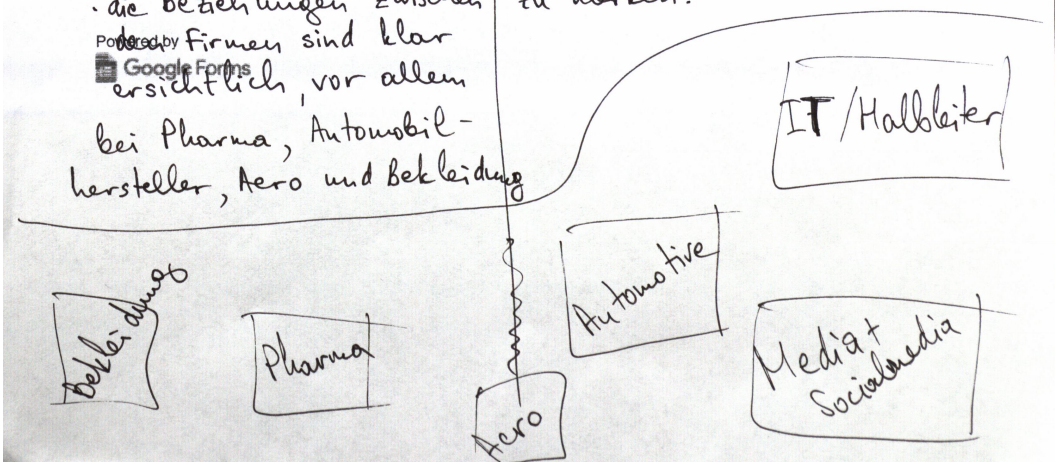
Mark only one oval.

- ◯ Yes
- ☒ No
- ◯ Neutral

8. **Please explain your interpretation of the visualization (Patterns, Structures, Outliers, Relationships, Inconsistencies, Accuracies)**

+ positiv	– negativ
• im Allgemeinen sehr gute Aufteilung der Bereiche	• einige Ausreißer vorhanden, wie z.B McDonalds liegt in der Bekleidung
• die meisten Firmen sind richtig zugeordnet	• Distanz zwischen Media und Social Media sollte vorhanden sein, momentan ist fast kein Unterschied zu merken.
• die Beziehungen zwischen Firmen sind klar ersichtlich, vor allem bei Pharma, Automobil-hersteller, Aero und Bekleidung	

Powered by Google Forms



id: 9

Empirical study on the quality of dimensionality reduction algorithms for visualization of categorical data

The goal of this evaluation is to explore suitability of dimensionality reduction algorithms for recognizing patterns in complex categorical data structures by visual analysis of the reduced data. In order to achieve this goal, dimensionality reduction was applied on high - dimensional companies data. After that, the reduced, low - dimensional representation has been visualized in a two - dimensional plot.
In this survey people are asked to elaborate on their interpretation of the visual representation. After that their answers are examined for compliance with the patterns investigated in the initial data set.

1. **Algorithm**
   Mark only one oval.
   - [x] PCA
   - [ ] LLE
   - [ ] Isomap
   - [ ] T-SNE

2. **Nickname**
   User 5

3. **Field of study**
   Automotive engineering

4. **Age**
   23

5. **Gender**
   Mark only one oval.
   - [x] Male
   - [ ] Female

16:10

Empirical study on the quality of dimensionality reduction algorithms for visualization of categorical data

The goal of this evaluation is to explore suitability of dimensionality reduction algorithms for recognizing patterns in complex categorical data structures by visual analysis of the reduced data. In order to achieve this goal, dimensionality reduction was applied on high - dimensional companies data. After that, the reduced, low - dimensional representation has been visualized in a two - dimensional plot.
In this survey people are asked to elaborate on their interpretation of the visual representation. After that their answers are examined for compliance with the patterns investigated in the initial data set.

1. **Algorithm**
   Mark only one oval.
   - [x] PCA
   - [ ] LLE
   - [ ] Isomap
   - [ ] T-SNE

2. **Nickname**
   user12

3. **Field of study**
   Computer Science, M.Sc.

4. **Age**
   22

5. **Gender**
   Mark only one oval.
   - [x] Male
   - [ ] Female

6. How would you classify the visual representation? id :10

Mark only one oval.

- ◯ Very poor - Most of the patterns that I recognize contradict reality (for instance, companies that are similar such as BMW and Audi are not in the same neighborhood or companies that are very different such as Turkish Airlines and Spotify are grouped together)
- ◯ Poor - I can't really recognize similarities, patterns or structures
- ◯ Somehow good, somehow poor - I can recognize some patterns compliant with reality (for instance, companies that are similar such as BMW and Audi are in the same neighborhood or companies that are very different such as Turkish Airlines and Spotify are far away from each other), but I also recognize a fair amount of inconsistencies that contradict reality
- ⊗ Good - Most of the patterns that I recognize are compliant with reality, but there are some visible inconsistencies
- ◯ Very good - Most of the patterns that I recognize are compliant with reality without visible inconsistencies

7. Did 3D representation enhance visual analysis?

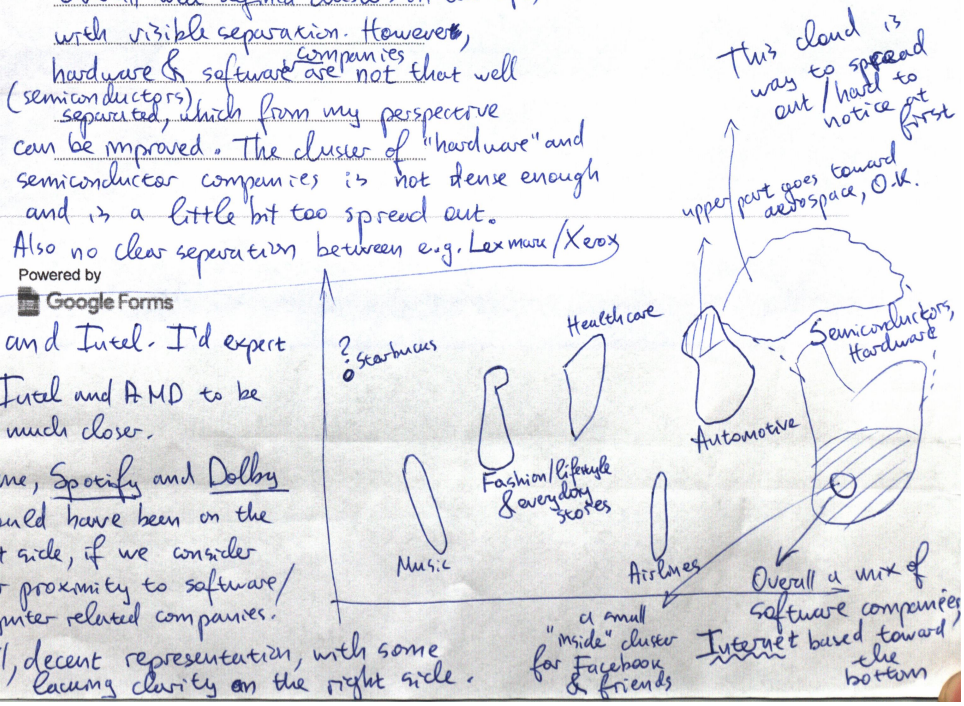
Mark only one oval.

- ◯ Yes
- ⊗ No
- ◯ Neutral

8. Please explain your interpretation of the visualization (Patterns, Structures, Outliers, Relationships, Inconsistencies, Accuracies)

Overall well-defined clusters on the left, with visible separation. However, hardware & software companies are not that well separated, which from my perspective (semiconductors) can be improved. The cluster of "hardware" and semiconductor companies is not dense enough and is a little bit too spread out.
Also no clear separation between e.g. Lexmark / Xerox and Intel. I'd expect Intel and AMD to be much closer.
To me, Spotify and Dolby should have been on the right side, if we consider their proximity to software / computer related companies.
Overall, decent representation, with some lacking clarity on the right side.

Powered by Google Forms



id: 11

Empirical study on the quality of dimensionality reduction algorithms for visualization of categorical data

The goal of this evaluation is to explore suitability of dimensionality reduction algorithms for recognizing patterns in complex categorical data structures by visual analysis of the reduced data. In order to achieve this goal, dimensionality reduction was applied on high - dimensional companies data. After that, the reduced, low - dimensional representation has been visualized in a two - dimensional plot.
In this survey people are asked to elaborate on their interpretation of the visual representation. After that their answers are examined for compliance with the patterns investigated in the initial data set.

1. **Algorithm**
   Mark only one oval.
   - ◯ PCA
   - ◯ LLE
   - ⊗ Isomap
   - ◯ T-SNE

2. **Nickname**
   user11

3. **Field of study**
   Computer Science

4. **Age**
   22

5. **Gender**
   Mark only one oval.
   - ⊗ Male
   - ◯ Female

6. **How would you classify the visual representation?** id: 11

Mark only one oval.

- ◯ Very poor - Most of the patterns that I recognize contradict reality (for instance, companies that are similar such as BMW and Audi are not in the same neighborhood or companies that are very different such as Turkish Airlines and Spotify are grouped together)
- ◯ Poor - I can't really recognize similarities, patterns or structures
- ◯ Somehow good, somehow poor - I can recognize some patterns compliant with reality (for instance, companies that are similar such as BMW and Audi are in the same neighborhood or companies that are very different such as Turkish Airlines and Spotify are far away from each other), but I also recognize a fair amount of inconsistencies that contradict reality
- ⊗ Good - Most of the patterns that I recognize are compliant with reality, but there are some visible inconsistencies
- ◯ Very good - Most of the patterns that I recognize are compliant with reality without visible inconsistencies

7. **Did 3D representation enhance visual analysis?**

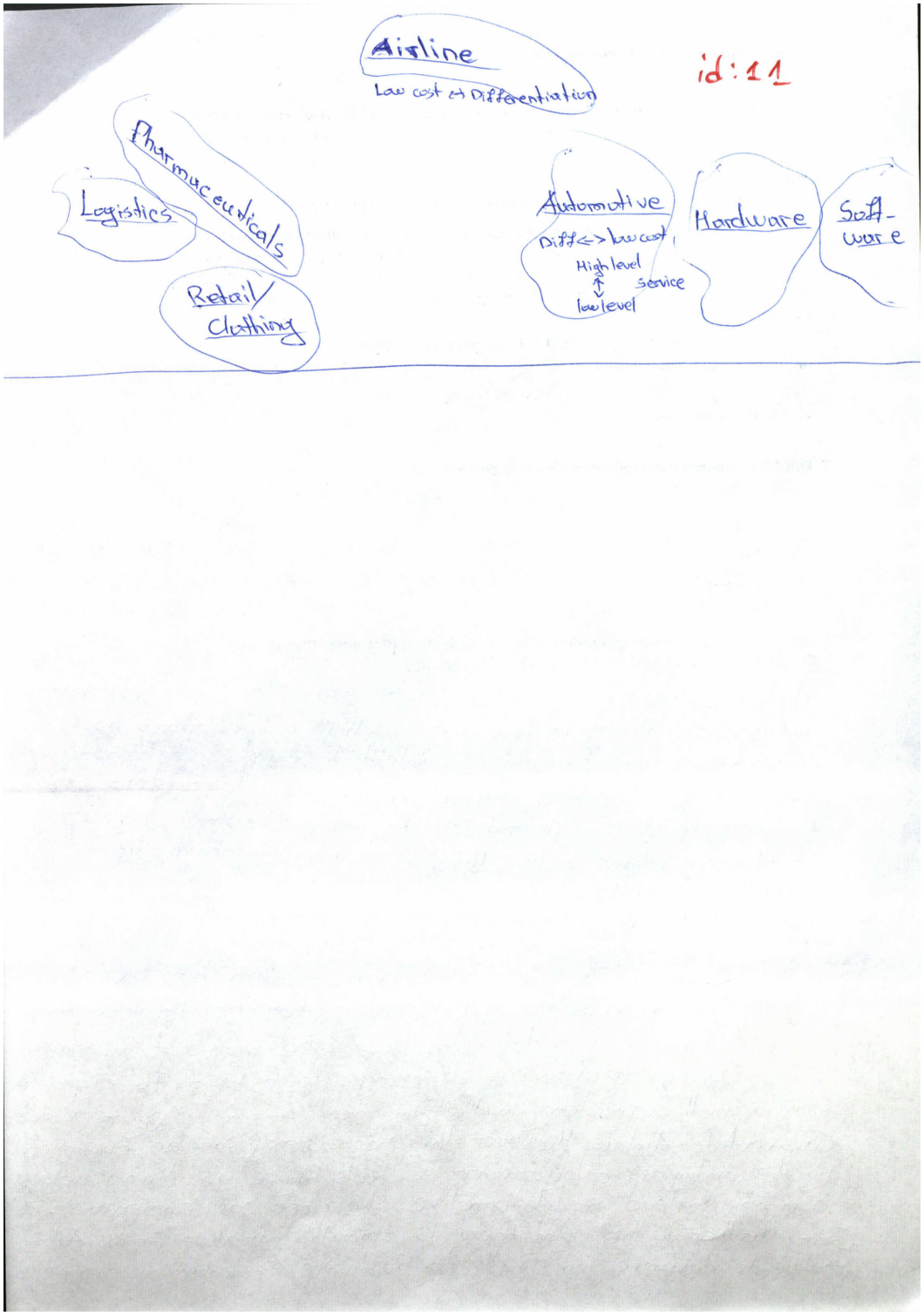
Mark only one oval.

- ◯ Yes
- ⊗ No
- ◯ Neutral

8. **Please explain your interpretation of the visualization (Patterns, Structures, Outliers, Relationships, Inconsistencies, Accuracies)**

Airline companies are clearly separated in the representation and are specializing in low cost to high cost offerings as clearly visible on the graphic from left to right.
On the bottom-left side there are recognizable clusters of Pharmaceutical, Retail and clothing companies with only small inconsistencies, (like IKEA in the near neighbourhood of PUMA and NIKE.
Going to the right there is a good separation of all the Automotive companies. The placement of these companies on the graphic representation has a clear trend of high level service to low level service companies from top to bottom and high cost/differentiation offerings to low cost offerings from left to right.
Only a small inconsistency in the placement of Rolls Royce Motor Cars next to Airbus. However this inconsistency can be explained by the ambigious categoration of Rolls Royce as Manufactoring company instead of Automotive, like other car companies.
Going further to the right there is smooth transition from Automotive to Hardware and finally to Software/Programming companies

Powered by Google Forms


Airline
Low cost vs Differentiation
id: 11
Pharmaceuticals
Logistics
Retail/ Clothing
Automotive
Diff <-> low cost
High level
service
low level
Hardware
Soft-ware