Technische Universität München

sebis

An Empirical Study on the Quality of Algorithms for Dimensionality Reduction and Visualization of High-Dimensional Data

Bachelor's Thesis Final Presentation

Lyubomir Stoykov 21.03.2015

Software Engineering for Business Information Systems (sebis)
Department of Informatics
Technische Universität München, Germany

wwwmatthes.in.tum.de

Structure

1. Motivation & Context
2. Thesis requirements
3. Approach
4. Evaluation – Results
5. Conclusion
6. Future Work
7. Web Application Architecture & Live Demo
8. Timeline

Motivation & Context

Enable visual analysis & pattern recognition.

Exponentially faster data processing with less resources.

Dimensionality reduction techniques mostly used “as is”, without evaluation.

No tool for easy appliance of dimensionality reduction techniques.

Thesis Requirements

sebis

First requirement – answer research questions

1. What kind of dimensionality reduction techniques do exist?

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

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

4. What results do they yield if used to visualize categorical data after a user-defined mapping f: String → Number has been applied?

Second requirement – develop a web application to apply dimensionality reduction

Approach

Overview of dimensionality reduction techniques → select best for evaluation

1) Linear (PCA)

2) Nonlinear, local properties (LLE)

3) Nonlinear, global properties (Isomap)

4) Recent algorithm (T-SNE)

Other techniques for high-dimensional data visualization → select best for evaluation

1) Cluster Analysis (DBSCAN)

Evaluation

Evaluation - Results

Cluster Analysis

3 Real-world datasets tested – Business data, MNIST, Iris.

1) Almost no clusters identified.

2) Rand Index nearly 0.

3) Applied also PCA prior to reduce curse of dimensionality – no improvement.

Implications

1) Real world data not really clustered.

2) Cluster analysis doesn't deal well with real world data.

3) Cluster analysis doesn't deal well with high-dimensional data.

Evaluation - Results

Dimensionality Reduction – Metric data

8 Data sets

1. Assign scores (Local vs Global structure)

    1 - Very Poor Visualization – ….
    ...
    5 - Very Good – ...

2. Average scores

Final averaged scores

PCA – 4.2

LLE – 3.11

Isomap – 4.11

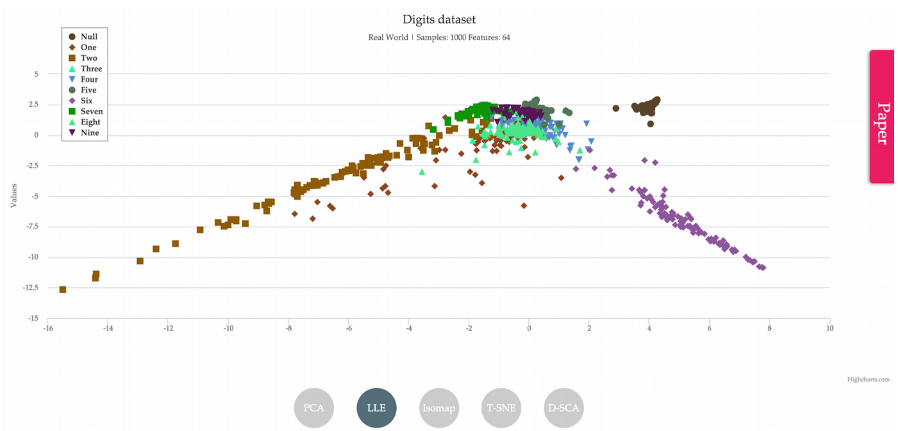
T-SNE – 3.88

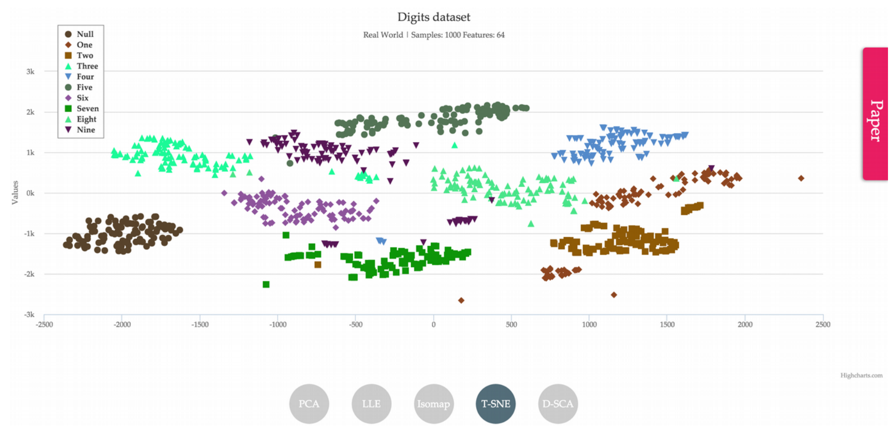
Evaluation - Results

sebis

Dimensionality Reduction – Metric data.

Example: MNIST dataset – Local vs Global Structure





Local structure: well preserved, digits grouped in clusters
Global structure: very poor representation, no distances between sub-manifolds
Score: 3 (or 2?)

Local structure: very well preserved, digits grouped in clusters
Global structure: very good representation, clear separation between sub-manifolds
Score: 5

Evaluation - Results

Dimensionality Reduction – Categorical data

No.	Company	Company type	Value proposition	Segments	Relationships	Channels	Activities	Resources	Partnerships	Revenue model	EBIT %	Branchen
1	[redacted]	non-e-business	Low cost	B-to-B and B-to-C	Medium level service	Combination of online and offline	Production, Sales	Physical, Human	Many important partners	Sales of ownership	3.19%	Consumer Electronics
2	[redacted]	non-e-business	Differentiation	B-to-C	Medium level service	Combination of online and offline	Production, Sales	Physical, Human, Intellectual	Many important partners	Sales of ownership	3.00%	Retail
3	[redacted]	non-e-business	Differentiation	B-to-B	Medium level service	Combination of online and offline	R&D, Production	Physical, Intellectual	Not many important partners	Sales of ownership	6.27%	Electronics
4	[redacted]	non-e-business	Differentiation	B-to-C	Medium level service	Combination of online and offline	R&D, Sales	Physical, Human, Intellectual	Many important partners	Sales of ownership	20.60%	Apparel, accessories
5	[redacted]	non-e-business	Differentiation	B-to-C	Medium level service	Combination of online and offline	Production, Sales	Physical, Human	Many important partners	Sales of ownership	4.90%	Retail
6	[redacted]	non-e-business	Differentiation	B-to-B and B-to-C	Medium level service	Combination of online and offline	Programming, Sales	Human, IT, Intellectual	Some important partners	Sales of ownership and access	9.95%	Software
7	[redacted]	non-e-business	Differentiation and low cost	B-to-C	High level service	Combination of online and offline	R&D, Production	Physical, Intellectual	Many important partners	Sales of ownership	-4.01%	Semiconductors
8	[redacted]	non-e-business	Low cost	B-to-C	Medium level service	Combination of online and offline	Services, Sales	Physical, Human	Some important partners	Sales of access	6.80%	Airline
9	[redacted]	non-e-business	Low cost	B-to-C	Medium level service	Combination of online and offline	Services, Sales	Physical, Human	Many important partners	Sales of access	15.77%	Airline
10	[redacted]	non-e-business	Low cost	B-to-C	Medium level service	Combination of online and offline	Services, Sales	Physical, Human	Some important partners	Sales of access	-7.00%	Airline
11	[redacted]	non-e-business	Differentiation	B-to-C	High level service	Combination of online and offline	Services, Sales	Physical, Human	Many important partners	Sales of access	3.41%	Airline
12	[redacted]	non-e-business	Differentiation	B-to-C	Medium level service	Combination of online and offline	Services, Sales	Physical, Human	Many important partners	Sales of access	6.90%	Airline
13	[redacted]	non-e-business	Differentiation	B-to-C	High level service	Combination of online and offline	Services, Sales	Physical, Human	Many important partners	Sales of access	3.80%	Airline
14	[redacted]	non-e-business	Differentiation	B-to-B	High level service	Offline	R&D, Production	Physical, Human, Intellectual	Many important partners	Sales of ownership	6.38%	Aircraft Manufacturer
15	[redacted]	non-e-business	Differentiation	B-to-B	High level service	Offline	R&D, Production	Physical, Intellectual	Some important partners	Sales of ownership	-30.00%	Machine Construction
16	[redacted]	non-e-business	Differentiation and low cost	B-to-C	Medium level service	Combination of online and offline	Services, Sales	Physical, Human	Many important partners	Sales of access	18.20%	Airline

Ask 11 students between 20 and 26 y.o.

1) PCA – 4. Patterns and clusters recognized.

2) Isomap – 4.2. Patterns and clusters recognized.

3) LLE – 3. Patterns and clusters recognized, but also a lot of inconsistencies.

4) T-SNE – 2. A lot of inconsistencies.

Evaluation - Results

sebis

Dimensionality Reduction – Categorical data

Example: Isomap

6.2.2016 id: 3 Empirical study on the quality of dimensionality reduction algorithms for visualization of categorical data

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

6.2.2016 id: 3 Empirical study on the quality of dimensionality reduction algorithms for visualization of categorical data

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

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 groued together without much information lost.
It is also easy to identify clusters.

Conclusion

High-Dimensional Real-World Data Visualization… (Research Questions 3 & 4)

1. Use Cluster Analysis?: No

2. Use Dim. Reduction for Metric Data?: Yes, with a grain of salt

3. Use Dim. Reduction for Categorical Data?: Yes, but see 2)

4. 3D visualization better than 2D?: No

5. Best algorithm?: PCA

Future Work

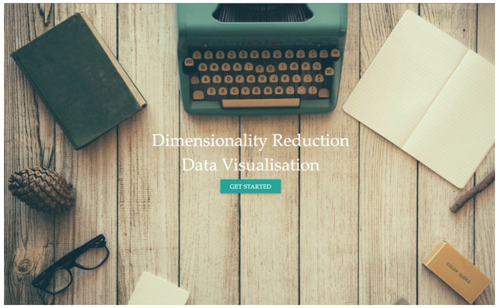
Dimensionality Reduction for Categorical Data

Combination of Techniques?

Supervised vs Unsupervised?

Curse of dimensionality?

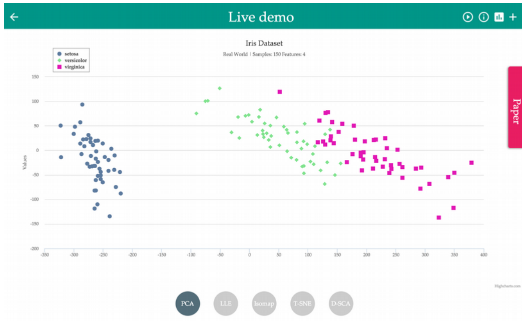
Web Application Architecture & Live Demo



1) Integrate Meteor and Scikit

2) Most computationally expensive calculations on server-side

→ Scalability and Fault Tolerance



Timeline

START: 10.15.2015

OCTOBER	NOVEMBER	DECEMBER	JANUARY	FEBRUARY

15.10.2015
Literature Review

15.11.2015
App Concept

20.12.2015
App Implementation

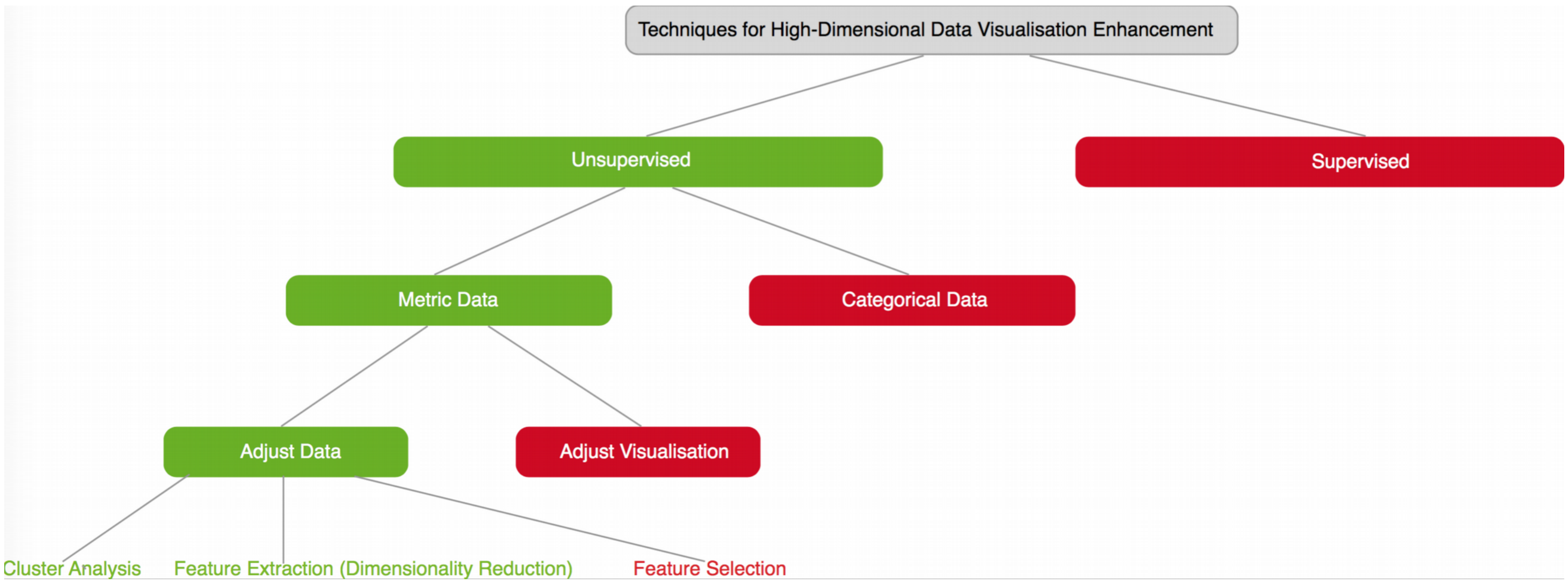
20.12.2015
Thesis Outline & Concept

15.01.2016
Evaluation and Thesis completion
15.02.2016

Thank you

Backup

Scope


Techniques for High-Dimensional Data Visualisation Enhancement
Unsupervised
Supervised
Metric Data
Categorical Data
Adjust Data
Adjust Visualisation
Cluster Analysis
Feature Extraction (Dimensionality Reduction)
Feature Selection

Map Function

```
# visual_mapping = {
#    "Company type":{
#       "e-business":"1",
#       "non-e-business":"75"
#    },
#    "Segments":{
#       "B-to-B":"1",
#       "B-to-B and B-to-C":"75",
#       "B-to-C":"150"
#    },
#    "Value Proposition":{
#       "Low cost":"1",
#       "Differentiation and low cost":"75",
#       "Differentiation":"150"
#    },
#    "Relationships":{
#       "Low level service":"1",
#       "Medium level service":"75",
#       "High level service":"150"
#    },
#    "Channels":{
#       "Offline":"1",
#       "Combination of online and offline":"75",
#       "Online":"150"
#    },
#
# …

# }
```