Multidimensional Projections and Tree-based Techniques for Visualization and Mining

Rosane Minghim
rminghim@icmc.usp.br

Closing Conference: Statistical and Computational Analytics for Big Data
Dalhousie University – Jun 12 2015
Where

Instituto de Ciências Matemáticas e de Computação

Universidade de São Paulo
Our research

• Visual Data Mining

• Visual Analytics

• Visualization

• Applications
Techniques

• Point placement: 2D or 3D similarity-based layouts

pairwise distances and/or dimensional embedding (feature space)
Techniques

• Projection-based
  • variations on MDS or other dimension reduction approaches
  • data mapped to low-dimensional visual space
  • preserving distances vs neighborhoods, global vs. local control, segregation

• fully interactive manipulation, dynamically adapting to user feedback
• massive data, sparse high-dimensional data, streaming data

• Tree-based
  • hierarchy of similarity relations
  • variations on tree layouts
Techniques

• Projection-based
  • PCA
  • MDS
  • Sammon Mapping

• Glimmer (distance)

• T-sne (segregation)
  - http://lvdmaaten.github.io/tsne/
LSP

- based on identifying samples (control points) and their neighborhoods
- distance matrices & spatially embedded data
- preserves data neighborhoods
- few thousand data items
PLMP


- spatially embedded data, more samples than dimensions
- millions of data items
- time varying and streaming data
- reduced amount of distance information
PLP


- local control

- flexibility in handling user interaction: users may change the layout based on previous knowledge/perception of similarity
NJ & PNJ Trees


...
Applications

Exploratory visualization of

- images
- text: news, scientific papers, web search results
- sensor measurements
- volumetric data: vector, scalar
- social networks
- neural fibers
- particle trajectories
- time series

Eler, Nakazaki, Paulovich, Santos, Andery, Oliveira, Batista Neto, Minghim, Visual analysis of image collections

*The Visual Computer*, 2009
Martins, Andery, Heberle, Lopes, Pedrini, Minghim; Multidimensional Projections for Visual Analysis of Social Networks (to appear), *JCST* 2012
Data from nanotech sensors & biosensors

finding good sensor configurations: segregation tasks on data

Moraes et. al, Detection of glucose and triglycerides using information visualization methods to process impedance spectroscopy data, *Sensors & Actuators B*, 2012
Data from nanotech sensors & biosensors

- Volpati et. al, Toward the optimization of an e-tongue system using information visualization: a case study with perylene tetracarboxylic derivative films in the sensing units, *Langmuir*, 2012

- Paulovich et al., Information visualization techniques for sensing and biosensing, *Analyst*, 2011

- Paulovich et al., Using multidimensional projection techniques for reaching a high distinguishing ability in biosensing. *Analytical and Bioanalytical Chemistry*, 2011


- Perinoto et al., Biosensors for efficient diagnosis of leishmaniasis: innovations in bioanalytics for a neglected disease, *Analytical Chemistry*, 2010
On studies on ecology and environment

• D.Sc. project: Visual exploration of feature spaces to support green algae taxonomic classification
• Classification based on features from imagens & other sources
• Collaboration with Dr. Armando Vieira, Department of Biology, UFSCar
• Time-varying images, feature extraction, representation and analysis
Open problems

• Metaphors: user interface, scalability, user control...
  • Handling text
  • Handling time-varying data
  • Going small: portable devices
  • Growing large: scalability issues

• Evaluation: user perception, quantitative & qualitative metrics

• Applications, reaching out to users: understanding their needs & tuning to specific profiles and application domains
Metaphors: clutter


Metaphors: clutter

Alencar, Paulovich, Börner, Oliveira, Time-aware visualization of document collections, *ACM SAC 2012, MMV track*
Open problems

• Metaphors: user interface, scalability, user control...
  • Handling text
  • Handling time-varying data
  • Going small: portable devices
  • Growing large: scalability issues

• Evaluation: user perception, quantitative & qualitative metrics

• Development software platform

• Recent developments and current work
Vispipeline
vicg.icmc.usp.br
Evaluation

• Numerical Evaluation
  • Distance Preservation
  • Neighborhood Preservation
  • Segregation

• User Understanding

• Visual Explanations

vicg.icmc.usp.br
Visual Coordinated exploration
Rule-based Topic Trees

(a) Topic Tree Control
(b) Document Map
More Applications – Word clouds

- Semantical ordering of keywords from projected points (new)
- Semantical filling of polygons over projections (new)

More Applications – Fiber Tracking

- Projection from fiber features
- Interaction through fast and reconfigurable projections (LAMP)
- Lines, Tubes and Surface Views

Plus

• Scalability
• Multiscale
• Understanding of feature spaces
• Time-varying volumes
• More evaluation
• Many more applications (molecular interactions, genome)
• Change of visual layouts
• Visual classification of images and other data

vicg.icmc.usp.br
MINING MEETS VISUALIZATION

Techniques and applications

Visual strategies to support data analysis/mining tasks

Problems regarding scale of data sets
Visual Data Mining

• Dimension Reduction

• Clustering Visualization

• Labeling

• Classification: sample selection, model creation and application, evolution of models

• Cooperation UNICAMP (Campinas), UFU (Uberlândia) and UFMG (Belo Horizonte)
Semi – Supervised Dimensional Reduction using PLS

- Sampling Procedure

Sampling Procedure

• Automatic Clustering Sampling
Layout Evaluation - NEWS

Original – 3731 dimensions

Reduced
Layout Evaluation - NEWS

NJ Tree

ISOMAP
Visual Improvement - NEWS

PLS – MultiClass

PLS – OAA
PLS Model Reuse

• Mapping of any size of data set bearing the same features
  • Fast model loading

• Growing data set mapping
  • Points groups kept in same regions of the layout: Mental Model Maintenance
ALL Data Set – 796 Instances
ALL Data Set – 1237 Instances
ALL Data Set – 1520 Instances
ALL Data Set – 1968 Instances
ALL Data Set – 2814 Instances
Visualization for Classification

• User: important role in building, applying and adjusting classifiers
  • Knowledge of the problem
  • Insertion of the classification process

• Insertion may be more effective: better data sets presentation
  • Data set structure and instances relationship understanding
  • Detection of specificities that justify classifiers behaviors
Task: Classification of Unlabeled Data set
Similarity Organization
Similarity Organization
Selection of Representative Instances

Instances selected to train classification model
Classification using Created Model
Classification Results
Evaluation: Classification Results
Evaluation: Classification Results
Evaluation: Classification Results
Evaluation: Classification Results
Classification Model Upgrading

• Several upgrade strategies: Layout also works as a guide
  • Example: relabeling of strategic instances: adjustment to specific scenarios

• Successive iterations: classifiers adaptation
  • Insertion of user knowledge on the classification model
  • Convergence to desired results
Misclassified Instances Relabeling
Misclassified Instances Relabeling
Classification Model Upgrading
Reclassification - Upgraded Model
Reclassification - Upgraded Model
Current Work: handling scalability?
The Visual Super Tree
Partners & collaborators

- Guilherme P. Telles, Hélio Pedrini IC-UNICAMP
- William Schwartz, UFMG
- Danilo Medeiros Eler, UNESP

- Lars Linsen, Jacobs University, Germany
- Alexandru Telea, University of Groningen, the Netherlands
- Charl Botha, T.U. Delft, the Netherlands
- Haim Levkowitz, University of Massachusetts Lowell, USA
- Evangelos Milios & team – Dalhousie U., Canada
- Stan Matwin – Big Data Institute – Dalhousie U., Canada

- Armando Vieira, Biology Department, UFSCar
The people

Maria Cristina F. Oliveira
ICMC

José Gustavo S. Paiva
UFU

Fernando V. Paulovich
ICMC

Guilherme P. Telles
UNICAMP

Luis Gustavo Nonato
ICMC

William Robson Schwartz
UFMG

Hélio Pedrini
UNICAMP
Some more people

Renato Oliveira

Henry Heberle

Sonia Castelo

Danilo Eler (UNESP)

Carlos Zampieri

Rafael Martins

Thanks!!!!

Fábio Rolli
Funding (current)

- CAPES / DAAD PROBRAL
- CAPES / NUFFIC
- FAPESP / CALDO
- CAPES 04/CII-2008, Network NANOBIO TEC-Brasil
- FAPESP – student grants
- CNPQ personal grants / student grants
- CNPq Universal 2012-2016

Thanks!!!!
VICG Group

vicg.icmc.usp.br (or http://vis.icmc.usp.br/vicg)

Papers: http://vis.icmc.usp.br/vicg/papers

Software mentioned in this talk soon to be made available at: http://vis.icmc.usp.br/vicg/software

Thanks!!!!