Real-Life Industrial Process Data Mining – Activities of the RISC-PROS project

Paavo Nieminen

Department of Mathematical Information Technology
University of Jyväskylä
Finland

Postgraduate Seminar in Information Technology
Jyväskylä 26th Feb 2009

UNIVERSITY OF JYVÄSKYLÄ
DEPARTMENT OF MATHEMATICAL
INFORMATION TECHNOLOGY

Outline

1. The RISC-PROS project
   - Data Mining
   - Clustering, Classification, Dimension Reduction
   - RISC-PROS
   - The Industrial Setting
2. Three Formerly Created Methods
   - Background 1/3: Robust Clustering
   - Background 2/3: Robust MLP Training
   - Background 3/3: Diffusion Geometries
   - Plans of Method Development within RISC-PROS
3. Current Issues
   - Project Status, Relation to Other Research
   - A Side Note About Openness of Algorithms

UNIVERSITY OF JYVÄSKYLÄ
DEPARTMENT OF MATHEMATICAL
INFORMATION TECHNOLOGY

Outline

1. The RISC-PROS project
   - Data Mining
   - Clustering, Classification, Dimension Reduction
   - RISC-PROS
   - The Industrial Setting
2. Three Formerly Created Methods
   - Background 1/3: Robust Clustering
   - Background 2/3: Robust MLP Training
   - Background 3/3: Diffusion Geometries
   - Plans of Method Development within RISC-PROS
3. Current Issues
   - Project Status, Relation to Other Research
   - A Side Note About Openness of Algorithms

UNIVERSITY OF JYVÄSKYLÄ
DEPARTMENT OF MATHEMATICAL
INFORMATION TECHNOLOGY

Data Mining Concepts Data Mining is a broad field of goals, methods, and techniques dealing with masses of numerical data.

- The “data” in data mining can be anything which can be numerically encoded, usually measurements from a real-world system (such as a human heart, lake ecosystem, or an industrial factory process).
- Objective of DM is to extract useful knowledge or patterns from the numerical bulk.

UNIVERSITY OF JYVÄSKYLÄ
DEPARTMENT OF MATHEMATICAL
INFORMATION TECHNOLOGY

Data – the Ore of Data Mining In the most confined case, DM deals with dataset matrices. That is, sets of row vectors of variables measured from a real-life system (or produced by simulation of such). A very traditional example is the Iris flower dataset:

```
Botanic measurements:
5.1, 3.5, 1.4, 0.2
4.9, 3.0, 1.4, 0.2
4.7, 3.2, 1.3, 0.2
4.6, 3.1, 1.5, 0.2
5.0, 3.6, 1.4, 0.2
5.4, 3.9, 1.7, 0.4
4.6, 3.4, 1.4, 0.3
5.0, 3.4, 1.5, 0.2
4.4, 2.9, 1.4, 0.2
...
```

UNIVERSITY OF JYVÄSKYLÄ
DEPARTMENT OF MATHEMATICAL
INFORMATION TECHNOLOGY

General Goals and Successes of DM So. We can collect data from plants, animals, humans, genes, chemicals, traffic, web shopping carts, Facebook, ... Successful DM can provide:

- A web search engine that quickly finds readings for you, based on search keywords.
- Protection against email spam or computer viruses.
- Automatic suggestions of which movies you'd like to watch.
- Targeted advertising, if you had something to sell.
- Automatic computation of populations based on digital photographs, if you were a biologist.
- Diagnoses, if you were a physician.

... and countless other things that improve life, humanity, or business.

UNIVERSITY OF JYVÄSKYLÄ
DEPARTMENT OF MATHEMATICAL
INFORMATION TECHNOLOGY

General Challenges in Data Mining

- When there is little data, there's nothing to mine.
- When the size and dimension of the data matrix increase, DM becomes useful but challenging
- Measurements are noisy
- Measurements are wrong
- Measurements are missing
- Measurements, by themselves, don't tell what they can or should be used for
- ...

UNIVERSITY OF JYVÄSKYLÄ
DEPARTMENT OF MATHEMATICAL
INFORMATION TECHNOLOGY

Some Specific Data Mining Activities Some examples of activities that could be considered DM are

- Clustering - finding groups in existing data
- Classification - finding a group “label” of new data, based on learning an existing one
- Dimension Reduction - mapping of data to a lower dimension “suitably”.

These three tasks are of special interest to our on-going research project.

UNIVERSITY OF JYVÄSKYLÄ
DEPARTMENT OF MATHEMATICAL
INFORMATION TECHNOLOGY

Clustering

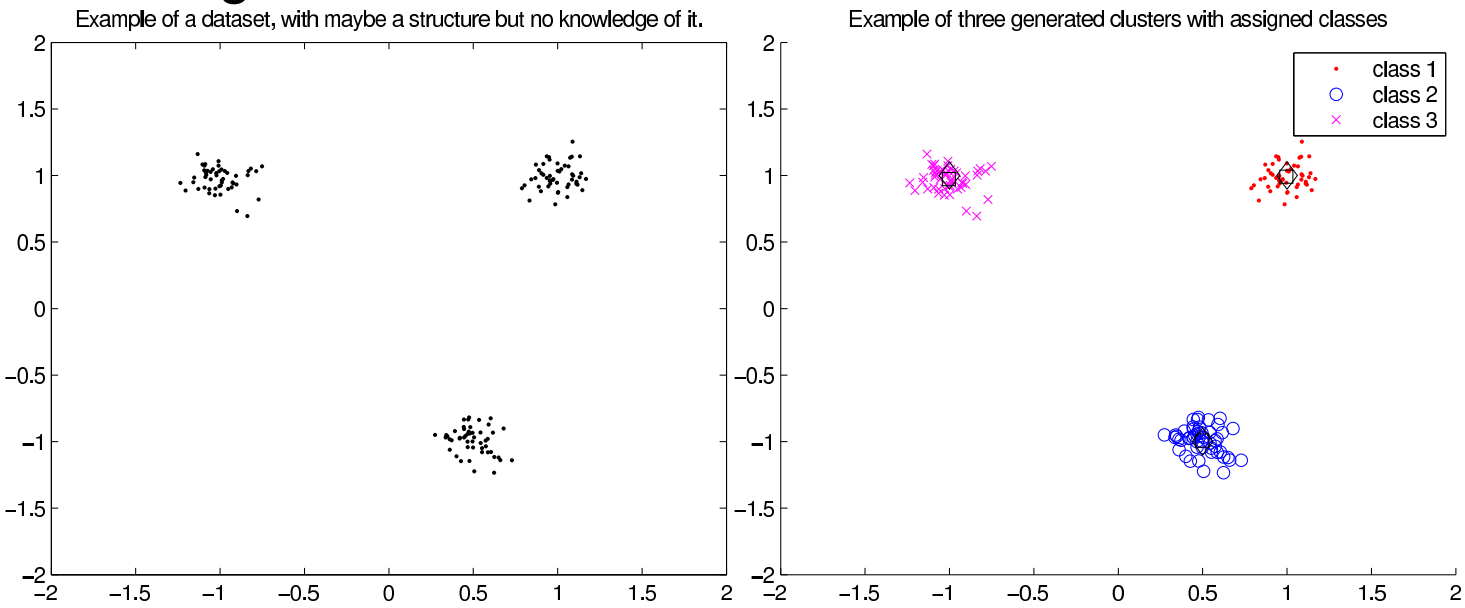


Figure: Clustering is the assignment of objects (observations) into categories (clusters) in a way that the objects in the same cluster are somehow similar to each other, and different from objects in other clusters.

UNIVERSITY OF JYVÄSKYLÄ
DEPARTMENT OF MATHEMATICAL
INFORMATION TECHNOLOGY

Classification

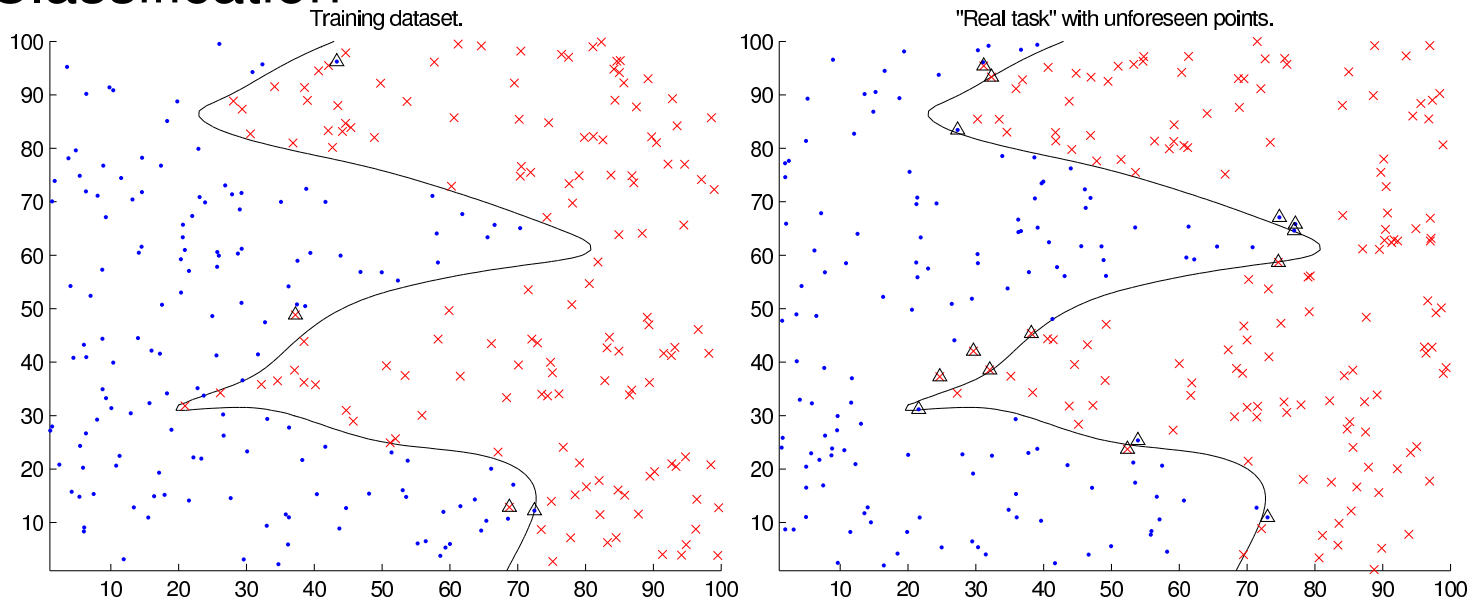


Figure: Classification is the assignment of a new, never before seen, object (observation) into a category (class); this can be called 'recognition'.

UNIVERSITY OF JYVÄSKYLÄ
DEPARTMENT OF MATHEMATICAL
INFORMATION TECHNOLOGY

Dimension Reduction

```
Botanic measurements:
5.1, 3.5, 1.4, 0.2
4.9, 3.0, 1.4, 0.2
4.7, 3.2, 1.3, 0.2
4.6, 3.1, 1.5, 0.2
5.0, 3.6, 1.4, 0.2
5.4, 3.9, 1.7, 0.4
4.6, 3.4, 1.4, 0.3
5.0, 3.4, 1.5, 0.2
4.4, 2.9, 1.4, 0.2
...
```

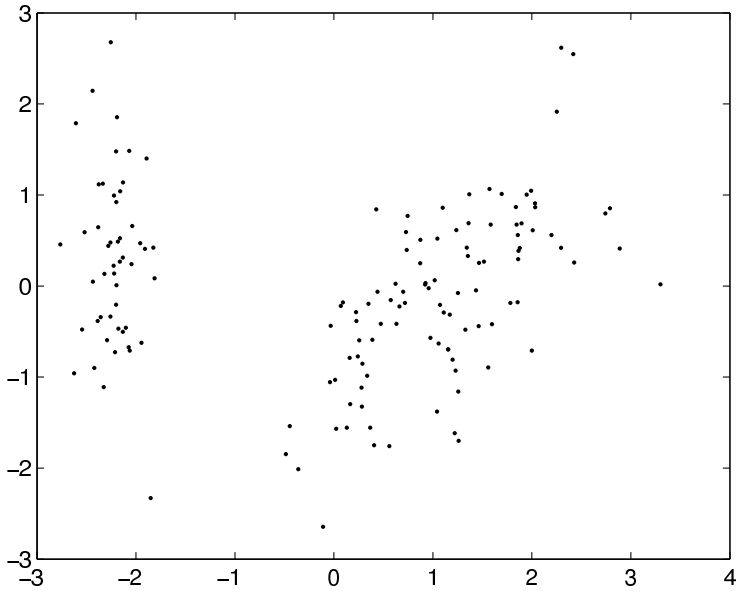


Figure: Dimension reduction is a mapping of high-dimensional data (observations with many variables) into a new dataset with less variables; a useful dimension reduction technique preserves indicative properties of the original observations.

UNIVERSITY OF JYVÄSKYLÄ
DEPARTMENT OF MATHEMATICAL
INFORMATION TECHNOLOGY

Main Goals and Collaboration We are now underway in a project called “RISC-PROS”.

- Longer title: “Real Time Industrial Process and Sensor Network Management Using Data Mining and Dimension Reduction”
- Researchers from Jyväskylä and Tel Aviv. (Me, Neta Rabin, Tommi Kärkkäinen, Amir Averbuch)
- Companies from Finland: Metso Paper, Fluidhouse, Patria Aviation
- Funding from the European Union, Regional Development Fund, via Tekes
- Total 2 years
- Goal: Business boost for the industry via applying and integrating novel data mining techniques in the products and development.

UNIVERSITY OF JYVÄSKYLÄ
DEPARTMENT OF MATHEMATICAL
INFORMATION TECHNOLOGY

The Industrial Setting: Lines of Business

- Metso Paper develops paper machinery; we look for new ways to help the engineer in interpreting measurements from a prototype machine.
- Fluidhouse produces fluid automation systems, for example hydraulic units in factories and dockyards; we look for automatic prediction of maintenance schedules.
- Patria Aviation develops airborne systems for civil and defence industries; our research deals with airborne radars.

UNIVERSITY OF JYVÄSKYLÄ
DEPARTMENT OF MATHEMATICAL
INFORMATION TECHNOLOGY

The Industrial Setting: Extents of Data In the industrial cases of the project, so far we seem to deal with:

- Measurements of ca. 900 variables from a continuously evolving physical process (sampling rate involved)
- Database storage that records point values only when there has been a change greater than a threshold (interpolation needed if we want a full data matrix)
- Processes spanning multiple scales of time: instantaneous to days, weeks or months.
- Gigabytes of numbers.

UNIVERSITY OF JYVÄSKYLÄ
DEPARTMENT OF MATHEMATICAL
INFORMATION TECHNOLOGY

Outline

1. The RISC-PROS project
   - Data Mining
   - Clustering, Classification, Dimension Reduction
   - RISC-PROS
   - The Industrial Setting
2. Three Formerly Created Methods
   - Background 1/3: Robust Clustering
   - Background 2/3: Robust MLP Training
   - Background 3/3: Diffusion Geometries
   - Plans of Method Development within RISC-PROS
3. Current Issues
   - Project Status, Relation to Other Research
   - A Side Note About Openness of Algorithms

UNIVERSITY OF JYVÄSKYLÄ
DEPARTMENT OF MATHEMATICAL
INFORMATION TECHNOLOGY

Research Background: Three Formerly Created Methods In addition to seeking a “business boost” RISC-PROS aims to continue the development of three computational methods that have been researched earlier at Jyväskylä and at Tel Aviv:

- Robust Clustering
- Robust MLP Training with Regularization
- Diffusion Geometry Framework

UNIVERSITY OF JYVÄSKYLÄ
DEPARTMENT OF MATHEMATICAL
INFORMATION TECHNOLOGY

Statistical robustness Define the l_q norm as follows:

$$\|\mathbf{u}\|_q = \left(\sum_{i=1}^{p} |(\mathbf{u})_i|^q \right)^{1/q}, \quad q < \infty. \tag{1}$$

The norm l_2 is the Euclidian distance of a point from the origin, l_1 is the city block distance. A "general K-means clustering" algorithm would minimize

$$\min_{\mathbf{c} \in \mathbb{N}^n, \mathbf{m}_k \in \mathbb{R}^p} \mathcal{J}(\mathbf{c}, \{\mathbf{m}_k\}_{k=1}^K) = \sum_{i=1}^{n} \|\mathbf{P}_i(\mathbf{x}_i - \mathbf{m}_{(\mathbf{c})_i})\|_q^\alpha \tag{2}$$

$$\text{subject to } (\mathbf{c})_i \in \{1, \ldots, K\} \text{ for all } i = 1, \ldots, n.$$

For notation key and details, please observe the public report of RISC-PROS, to be published on-line.

UNIVERSITY OF JYVÄSKYLÄ
DEPARTMENT OF MATHEMATICAL
INFORMATION TECHNOLOGY

A Traditional Multi-Layered Perceptron

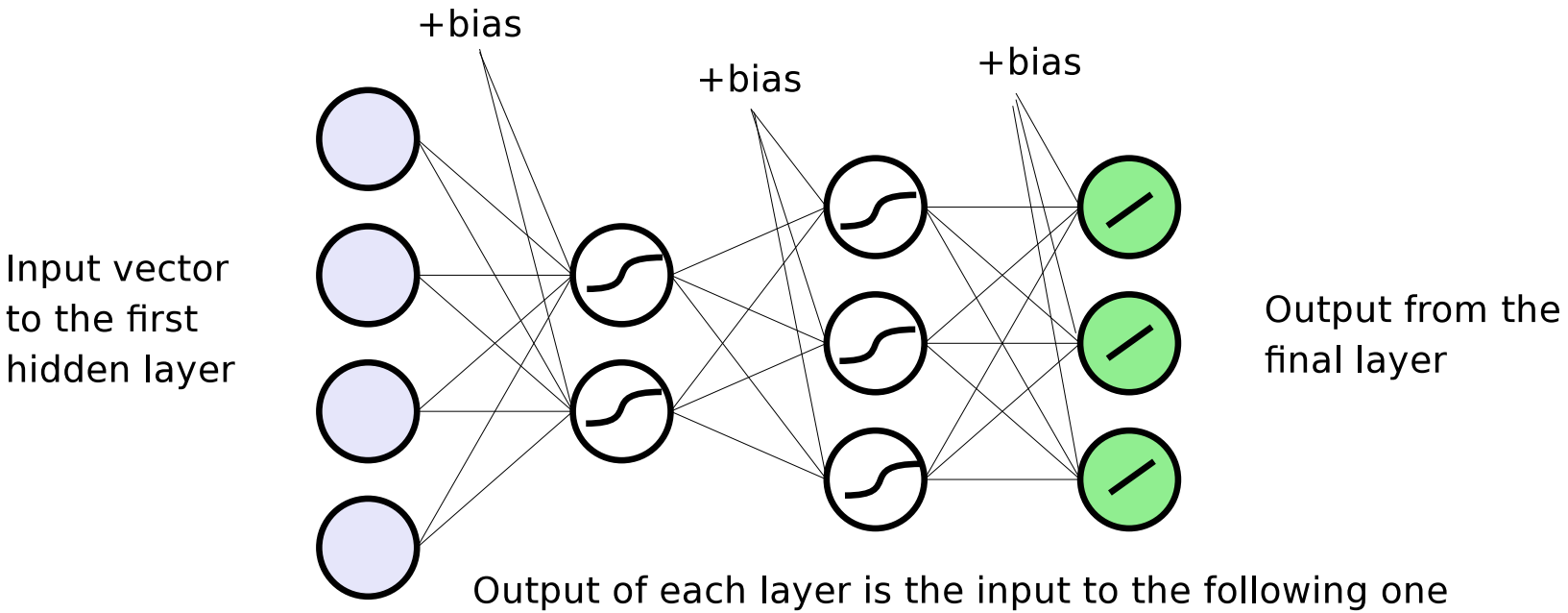


Figure: Neural architecture of a simple MLP neural network

UNIVERSITY OF JYVÄSKYLÄ
DEPARTMENT OF MATHEMATICAL
INFORMATION TECHNOLOGY

Novel Ideas for Training the MLP

Define the l_q norm as follows:

$$\|\mathbf{u}\|_q = \left(\sum_{i=1}^{p} |(\mathbf{u})_i|^q \right)^{1/q}, \quad q < \infty. \tag{3}$$

We choose to train our MLP by minimizing

$$\mathcal{L}^{\alpha}_{q,\beta}(\{\mathbf{W}^l\}) = \frac{1}{\alpha n} \sum_{i=1}^{n} \left\| \mathcal{N}(\{\mathbf{W}^l\})(\mathbf{x}_i) - \mathbf{y}_i \right\|_q^{\alpha} + \beta \sum_{l=1}^{L} \sum_{(i,j)\in I_l} \frac{1}{2S_l} |\mathbf{W}^l_{i,j}|^2 \tag{4}$$

for $\beta \geq 0$. For notation key and details, please observe the public report of RISC-PROS, to be published on-line.

UNIVERSITY OF JYVÄSKYLÄ
DEPARTMENT OF MATHEMATICAL
INFORMATION TECHNOLOGY

The Problem of Overfitting

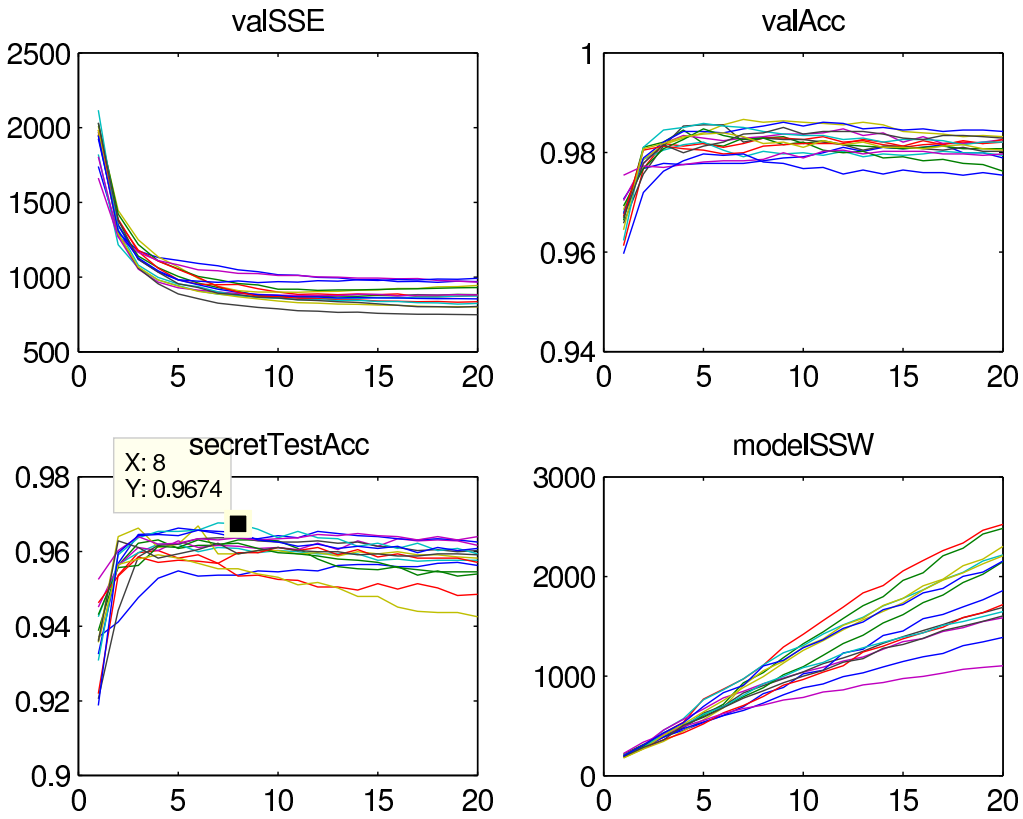


Figure: Progression of cost and accuracy with $\beta = 0$ (PenDigits 16-30-10).

UNIVERSITY OF JYVÄSKYLÄ
DEPARTMENT OF MATHEMATICAL
INFORMATION TECHNOLOGY

- When to stop training? As the cost function is minimized

Heuristic Search for a Good Regularization

- Basically subsampling and exhaustive search using a blind metric.
- First, use a subset of training data, and loose accuracy or quick iteration deadline for local search
- Later, use the whole training data, and tighter accuracy, but only once.
- More details to be published in the Proceedings of ICANNGA ’09 (LNCS series).

UNIVERSITY OF JYVÄSKYLÄ
DEPARTMENT OF MATHEMATICAL
INFORMATION TECHNOLOGY

Heuristic Search for a Good Regularization

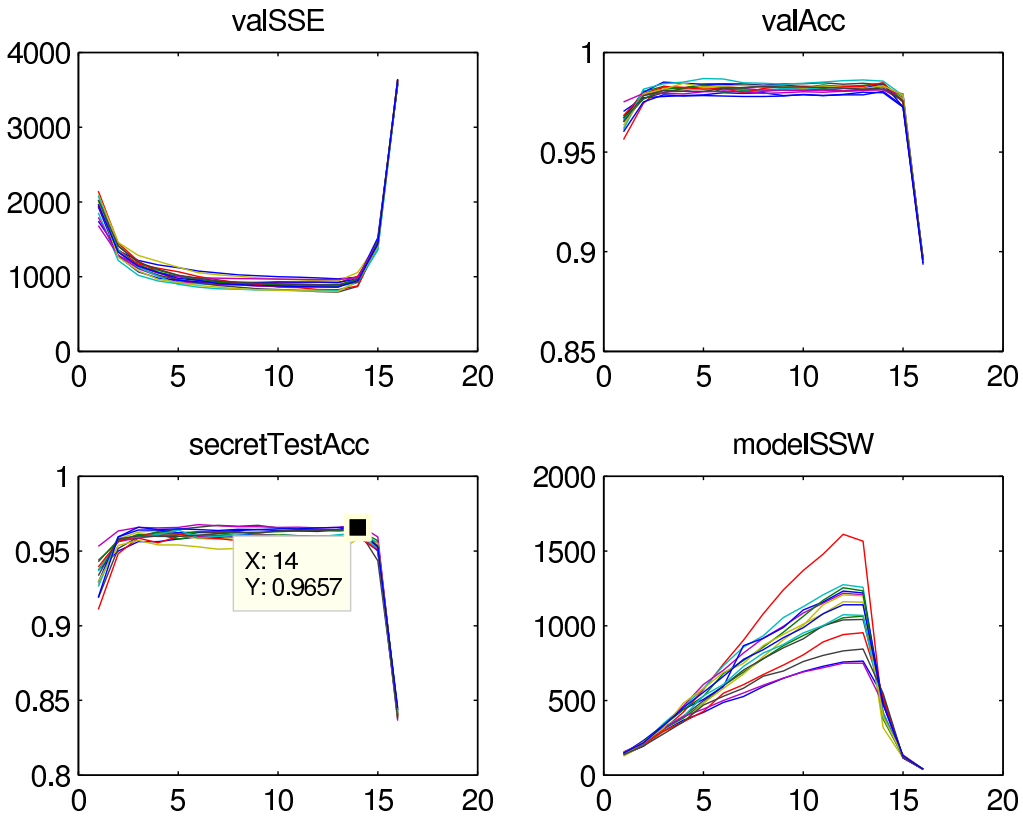


Figure: Progression of cost and acc. with varying β.

UNIVERSITY OF JYVÄSKYLÄ
DEPARTMENT OF MATHEMATICAL
INFORMATION TECHNOLOGY

Diffusion Geometry Framework Construct a graph $G = (\mathbf{X}, W)$ on a dataset with a kernel $W = w\left(\mathbf{x}_i, \mathbf{x}_j\right) = e^{-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\varepsilon}$ (which is only one of possible choices). Then form a Markov matrix $\mathbf{M} = \{m\left(\mathbf{x}_i, \mathbf{x}_j\right)\}$ as

$$m\left(\mathbf{x}_i, \mathbf{x}_j\right) = \frac{w\left(\mathbf{x}_i, \mathbf{x}_j\right)}{d\left(\mathbf{x}_i\right)} \text{ where } d\left(\mathbf{x}_i\right) = \sum_{\mathbf{x}_j \in \mathbf{X}} w\left(\mathbf{x}_i, \mathbf{x}_j\right).$$

By computing the eigenvalues of a conjugate matrix $\mathbf{A}$ given by $a(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{d(\mathbf{x}_i)} m(\mathbf{x}_i, \mathbf{x}_j) \frac{1}{\sqrt{d(\mathbf{x}_j)}}$. and using a couple of further steps, one obtains an embedding of the form

$$\Psi_t(\mathbf{x}_i) = \left(\lambda_1^t \psi_1(\mathbf{x}_i), \lambda_2^t \psi_2(\mathbf{x}_i), \lambda_3^t \psi_3(\mathbf{x}_i), \ldots\right), \quad (5)$$

that represents the dataset in a new Euclidean space.

UNIVERSITY OF JYVÄSKYLÄ
DEPARTMENT OF MATHEMATICAL
INFORMATION TECHNOLOGY

Diffusion Geometry Framework

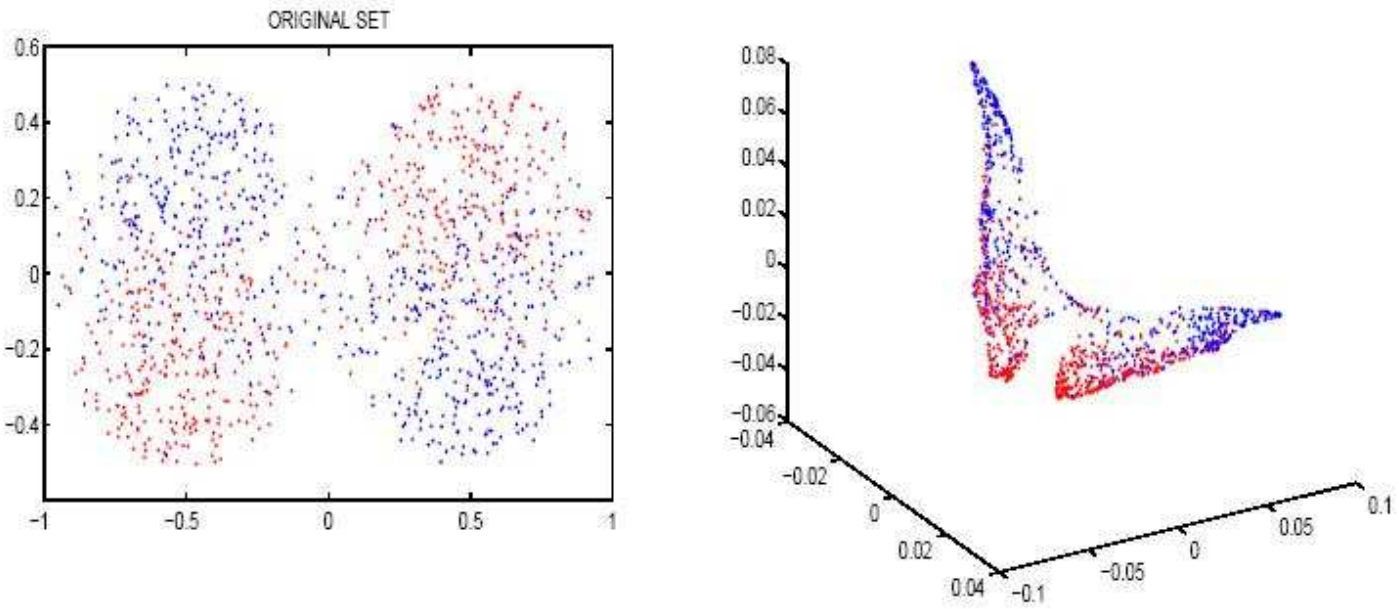
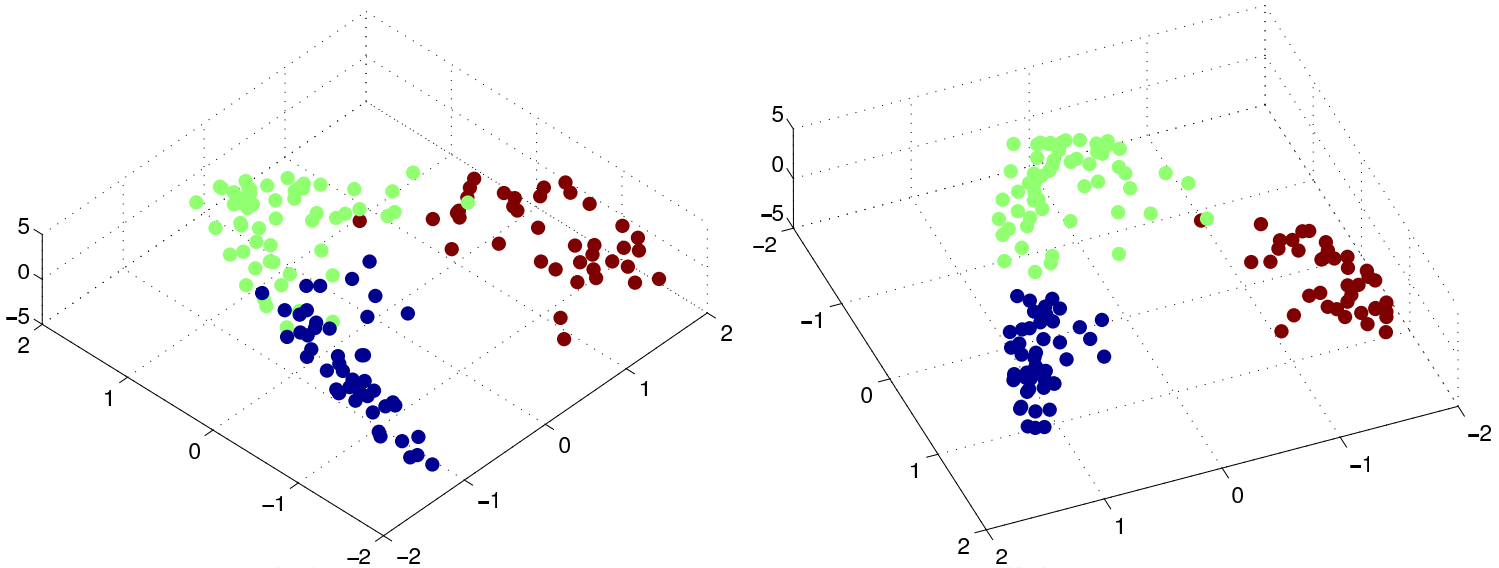


Figure: The original geometry is mapped as a butterfly shaped set, in which the red and blue phases are organized according to the diffusion they generate: the cord length between two points in the diffusion space measures the quantity of heat that can travel between these points.

UNIVERSITY OF JYVÄSKYLÄ
DEPARTMENT OF MATHEMATICAL
INFORMATION TECHNOLOGY

An Example Application

The case “Wine” from the UCI Machine Learning Repository: classification of wines into 3 categories based on 13 chemical measurements.



(a) $\omega = 1$ (b) $\omega = 0.9$

Figure: The 142 wine samples embedded using DM. The algorithm parameter ω controls the strength of labeled class separation in the embedding.

UNIVERSITY OF JYVÄSKYLÄ
DEPARTMENT OF MATHEMATICAL
INFORMATION TECHNOLOGY

What we plan to do with the methods Hybridization/piping of the methods:

- Visualize the datasets using DG projections
- Make the clustering and classification algorithms faster by pre-processing (to lower dimension) with DG?
- Incorporate missing data handling to all the methods. (Now implemented only for clustering.)

UNIVERSITY OF JYVÄSKYLÄ
DEPARTMENT OF MATHEMATICAL
INFORMATION TECHNOLOGY

Outline

1. The RISC-PROS project
   - Data Mining
   - Clustering, Classification, Dimension Reduction
   - RISC-PROS
   - The Industrial Setting
2. Three Formerly Created Methods
   - Background 1/3: Robust Clustering
   - Background 2/3: Robust MLP Training
   - Background 3/3: Diffusion Geometries
   - Plans of Method Development within RISC-PROS
3. Current Issues
   - Project Status, Relation to Other Research
   - A Side Note About Openness of Algorithms

UNIVERSITY OF JYVÄSKYLÄ
DEPARTMENT OF MATHEMATICAL
INFORMATION TECHNOLOGY

Current State of Affairs in RISC-PROS

- First report written; waiting for comments from the companies; then to publish in the dept. report series.
- Datasets have been aquired from the companies. This tends to take time!
- As of yet, we don't have enough information about the datasets and the actual goals that the companies have (and that we can help with, using our methodology)
- The plan is now to "just do some computations" and show some visualized results to the company people; we assume that seeing at least something will make it easier to get ideas about what the ultimate goals are.

Collaboration Opportunities Let me list some completely different topics that could be related to what is happening in RISC-PROS:

- We plan to try population-based optimization (genetic/memetic) with our cost function formulations in the future.
- Visualization methods developed originally for interactive multiobjective optimization could be useful for examining the reduced-dimension data. (And vice versa: suitable dimension reduction could make MOO visualizations more intuitive)?
- We already have ideas of incorporating, in a way, a multi-grid methodology in the MLP training.

UNIVERSITY OF JYVÄSKYLÄ
DEPARTMENT OF MATHEMATICAL
INFORMATION TECHNOLOGY

A Concern

- What happens to the method implementations I create while doing my work in an industrially inclined project?
- AFAIK, the university harvests all intellecual rights properties to whatever computer programs I write.
- It is simpler for those of us who are not controlled by outside funding; the researcher decides upon his/her code. But how many of us are? And does everybody who works here even know who owns their codes?
- My fear is that after some idea has been tried a couple of times in a project, the implementation is forgotten and needs to be re-made by the next people (or the whole idea gets forgotten if the project didn't produce visible-enough reports and journal articles)

UNIVERSITY OF JYVÄSKYLÄ
DEPARTMENT OF MATHEMATICAL
INFORMATION TECHNOLOGY

More Concerns

- Not all that is produced, especially in implementations, can be used in journals (we don't think source code is the research result to be published). But should a program be lost forever even if not published?
- Consider the research softwares WaveLab and WEKA (first ones that come to my mind): Would they be so popular and recognized (or even as good) as they currently are, if they were closed-source property of the researchers or universities?
- Also, "WaveLab and Reproducible Research" by Buckheit and Donoho was an inspiring reading for me; I recommend the paper warmly, if you haven't read it already.

UNIVERSITY OF JYVÄSKYLÄ
DEPARTMENT OF MATHEMATICAL
INFORMATION TECHNOLOGY

My Personal Salvation For the RISC-PROS project

- The agreement states that “all implementations of universally applicable mathematical algorithms” created as part of RISC-PROS shall be published publically as open source under the MIT License.
- For example, the implementations of ideas published in the forthcoming ICANNGA ’09 proceedings.
- Naturally the agreement states that “any code revealing industrial trade secrets shall not be made public”
- To make the University IPR people happy, we state “application code directly integrable in an industrial partner’s system shall not be open source”. (The last, unfortunate, pain that lessens my programming and overall motivation at work.)

UNIVERSITY OF JYVÄSKYLÄ
DEPARTMENT OF MATHEMATICAL
INFORMATION TECHNOLOGY

My Personal Salvation What I hope to be gained with open-sourcing the RISC-PROS method implementations?

- I assume it will add to the research group's publicity. Definitely making something more public will not have a negative effect on publicity!
- As for the concern about pre-mature publishing . . . If there are mistakes, maybe someone will spot them, and they get corrected before our next paper. I can take the blow of having computed wrongly; I'm sure everybody does that sometimes by accident.
- The anticipation of publicity has already made the codes a little bit better quality than what they had been if it was "just for the one project".

UNIVERSITY OF JYVÄSKYLÄ
DEPARTMENT OF MATHEMATICAL
INFORMATION TECHNOLOGY

One Project Is Not a General Solution

- Note that, as of yet, the RISC-PROS source codes are not published. . .
- We are still thinking about what is a good format, a publishing forum, and time for going public with the first version.
- It would be nice to have a general “recommended practice” about this – something that everybody knows about, and uses daily.
- Well, now we are attempting to build one!

UNIVERSITY OF JYVÄSKYLÄ
DEPARTMENT OF MATHEMATICAL
INFORMATION TECHNOLOGY

A More General Solution We have initiated a loose group (managed by Tero Tuovinen) to discuss the following agenda:

- Create a common source code database (with version control tools) where the mit.jyu.fi department staff would store their research source codes, and share them with others, maybe even the whole world.
- Possibility, maybe recommendation, not obligation!
- We are also aware of some difficulties, such as issues with IPRs & licenses and fears of pre-mature publicity.
- At the moment, technical implementation issues are being investigated.
- You're welcome to bring your input and ideas to our attention. More shall be announced.

UNIVERSITY OF JYVÄSKYLÄ
DEPARTMENT OF MATHEMATICAL
INFORMATION TECHNOLOGY

Acknowledgments Former research on clustering by Sami Äyrämö. Research is guided by Tommi Kärkkäinen and Amir Averbuch. The RISC-PROS project is funded by the European Union, European Regional Development Fund via Tekes.

European Union
European Regional Development Fund

UNIVERSITY OF JYVÄSKYLÄ
DEPARTMENT OF MATHEMATICAL
INFORMATION TECHNOLOGY

Summary In this talk, I presented:

- The EU-funded project RISC-PROS that I work with
- Some tasks in Data Mining, and specific methods being developed at our department
- Standing on barricades about open source software, and especially the benefits of producing such as part of scientific research.

UNIVERSITY OF JYVÄSKYLÄ
DEPARTMENT OF MATHEMATICAL
INFORMATION TECHNOLOGY