Conference ICDM 2013

Marketa Krmelova, Radim Belohlavek

International Center for Information and Uncertainty
Palacky University, Olomouc

Basic Information about Conference

- Institution:IEEE International Conference on Data Mining
- Location: Dallas, Texas, USA
- Date: 7.12. 10.12. 2013

Interesting facts

- Number of submitted papers: 809
- Number of accepted papers: 94 regular papers and 65 short papers
- Acceptance rate: 19.65%

Invited talks

- Alexander Tuzhilin (New York University): Opportunities and Challenges Facing Recommender Systems: Where Can We Go from Here?
- Joydeep Ghosh (University of Texas, Austin): Predictive Healthcare Analytics under Privacy Constraints
- Jianchang (JC) Mao (Microsoft): Large-scale Learning in Computational Advertising

R. Belohlavek, M. Krmelova: Beyond Boolean Matrix Decompositions: Toward Factor Analysis and Dimensionality Reduction of Ordinal Data

- Novel problem of matrix decomposition proposed.
  - novelty: ordinal data and matrix composition (generalizes Boolean case)
  - lattice-theoretical in nature
- New theoretical results.
  - optimal factors for decomposition
  - geometry of decompositions
  - "essential parts" of matrices (where to focus in computing decompositions?)
- New decomposition algorithms.
  - based on theoretical insight in geometry of decompositions
  - extending our previous algorithms for Boolean data, but new conceptual issues
- Examples and experimental evaluation.
  - factor analysis of sports data
  - several other data: dog breeds data, music data, questionnaire data . .

New matrix algebra

- matrix entries are elements of **complete residuated lattices**, i.e. algebras $\langle L, \wedge, \vee, \otimes, \to, 0, 1 \rangle$ where
  - $\langle L, \wedge, \vee, 0, 1 \rangle$ ... complete lattice,
  - $\langle L, \otimes, 1 \rangle$ ... commutative monoid,
  - $a \otimes b \leq c$ iff $a \leq b \to c$ (adjointness).

  used in (mathematical) fuzzy logic
  - $\otimes$ and $\to$ ... (truth functions of) conjunction and implication
  - Łukasiewicz as example: $a \otimes b = \max(0, a + b - 1)$, $a \to b = \min(1, 1 - a + b)$
  - $L = \{0, 1\}$: two-element Boolean algebra (Boolean matrices).
- **matrix composition**
  - $(A \circ B)_{ij} = \bigvee_{l=1}^{k} A_{il} \otimes B_{lj}$
  - Boolean matrix product is a particular example
  - transparent meaning of (de)composition retained: "object i has attribute j IFF there exists factor l s.t. i has l and j is one of particular manifestations of l"

Dog breeds example

- data from 151 dog breeds and their 11 attributes (petfinder.com)

	Energy	Playfulness	Friend. towards dogs	Friend. tow. strangers	Friend. tow. other pets	Protection ability	Exercise	Affection	Ease of training	Watchdog ability	Grooming
Labrador Retrievers	5	6	5	6	6	3	4	6	6	5	3
Golden Retrievers	4	6	6	6	6	3	4	6	6	4	4
Yorkshire terriers	5	5	3	4	3	2	2	4	3	6	5
German shepherds	4	3	2	3	4	6	5	4	6	6	3
Beagles	4	4	6	6	6	2	4	6	2	5	2
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

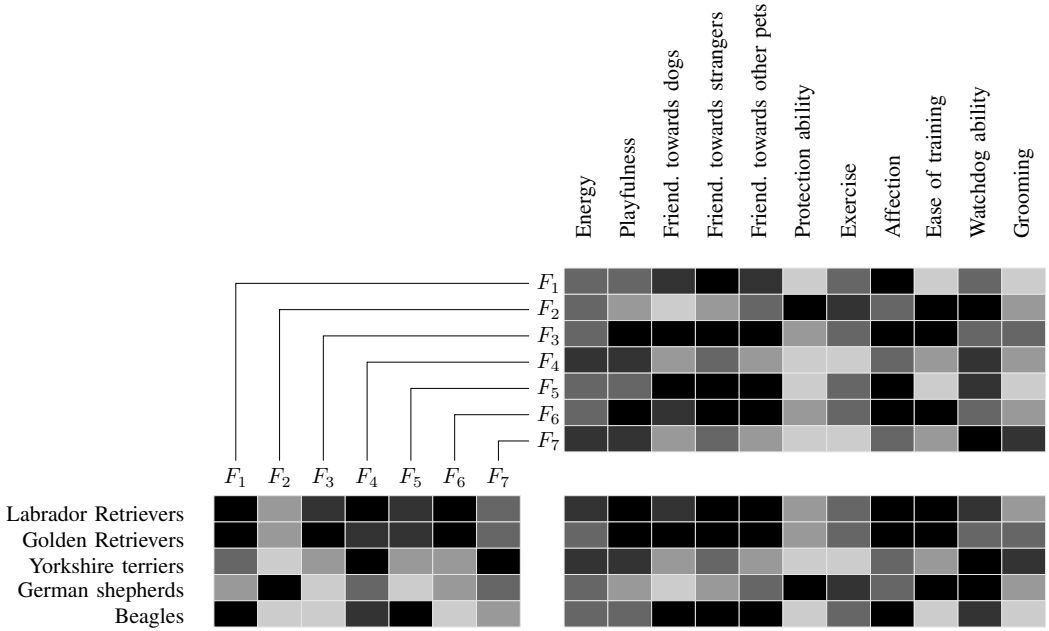

Figure : Decomposition $I = A_{\mathcal{F}} \circ B_{\mathcal{F}}$.

Three most important factors:

- friendliness
- guardian dog
- dogs suitable for kids

Regular papers

- Noise-Resistant Bicluster Recognition
  - *Huan Sun, Gengxin Miao and Xifeng Yan*
  - In this paper was presented new bicluster model - AutoDecoder. This model uses knowledges of machine learning and neural networks. They used this method on four real datasets and several synthetics datasets. This method works best on data with less then 15% of noise.
- Classifying Spam Emails using Text and Readability Features
  - *Rushdi Shams and Robert E. Mercer*
  - Recent methods for classifying spam emails use either header-based or content-based features. Spammers can bypass these methods easily. Authors in this talk present how to classify spam emails with using features based on email content-language and readability combined with previously used content-based task features. They present this method on real datasets.