evropský sociální fond v ČR
EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Název projektu: Mezinárodní centrum pro informaci a neurčitost

Registrační číslo: CZ.1.07/2.3.00/20.0060

Zpráva z účasti na konferenci

název konference:	IEEE ICDM 2013 (IEEE 13th International Conference on Data Mining)
datum konání:	7.12. - 10. 12. 2013
místo:	Dallas, Texas, USA
účastník konference:	Mgr. Jan Outrata, Ph.D.

Stručný popis konference:
IEEE ICDM je přední konferencí v mnoha oblastech data mining a machine leraning jako např. feature analysis, klasifikace, strojové učení, shlukování, sumarizace, doporučování big data, bioinformatika a medicínská informatika, bussiness intelligence, analýza časových řad, reprezentace dat a dolování dat ve videu, obrazu a textu. Pravidelně se zde setkávají hlavní a nejcitovanější osobnosti z této oblasti informatiky. Konference je vedle sdružení IEEE a IEEE Computer Society sponzorována grantovými agenturami a velkými formami (letos USA National Science Foundation a IBM a Toshiba). Mezi účastníky jsou jak vědci z akademické sféry, tak vědci z výzkumných oddělení vůdčích firem.

Zajímavá čísla

Acceptance rate:	19,65%
Počet účastníků:	přes 500
Počet prezentovaných příspěvků:	187 (94 regular papers, 65 short papers, 18 workshopů, 7 demonstrací, 3 tutoriály, 1 panel, 1 soutěž a 1 PhD fórum)
Počet zemí autorů příspěvků:	55
Počet plenárních přednášek:	3
Počet paralelních sekcí	4 (u workshopů 11)

Zajímavé přednášky

A. Plenární přednášky

Materiály k přednáškám jsou dostupné na http://icdm2013.rutgers.edu/keynotes.

Alexander Tuzhilin (NYU): Opportunities and Challenges Facing Recommender Systems: Where Can We Go From Here?.

Přednášející prezentoval některé málo probádané směry vývoje doporučovacích (recommender) systémů, které dle jeho názoru představují slibné příležitosti pro další výzkum této oblasti. Uvedl také výzvy, které je třeba zvládnout k překonání omezení stávajících systémů a vytvoření nové generace doporučovacích systémů.

Joydeep Ghosh (University of Texas, Austin): Predictive Healthcare Analysis under Privacy Constraints.
Přednášející nejdříve nahlédl otázku ochrany soukromí vs. využitelnost údajů v kontextu zdravotní péče a pak prezentoval dva jejich přístupy k prediktivnímu modelování zachovávajícímu ochranu soukromí, u nichž je malá degradace modelů navzdory omezením sdílení a analýzy údajů. První přístup predikuje z agregovaných údajů, zatímco druhý, nový, představuje generátor realistických, ale ne reálných, dat z původních chráněných údajů.

Jianchang (JC) Mao (Microsoft): Large-scale Learning in Computational Advertising.
Přednášející nastínil několik velkých problémů strojového učení z velkých dat (big data) v oblastech reklamy, porozumění dotazu, uživateli a dokumentu, modelování relevance výsledků dotazu, predikce uživatelovy reakce, doporučování klíčových slov apod. Potom prezentoval nedávná řešení několika problémů, zejména shlukování dotazu (EM metoda založená na KL-divergenci dotazů) a doporučování klíčových slov (multi-label random forest learning).

B. Řádné přednášky (výběr)

A. Dukkipati, G. Pandey, D. Ghoshdastidar, P. Koley, D.M.V. Satya Sriram: Generative Maximum Entropy Learning for Multiclass Classification
Na přednášce byla představena metoda klasifikace využívající maximální entropii s výběrem vlastností (feature selection). Pro redukci dimenzionality je využito „maximální diskriminace" (Jeffrey nebo Jensen-Shannon divergence) mezi hustotami podmíněné pravděpodobnosti tříd objektů.

L. Milli, A. Monreale, G. Rossetti, F. Giannotti, D. Pedreschi, F. Sebastiani: Quantification Trees
Na přednášce byly představny rozhodovací stromy optimalizované ne pro klasifikaci jednotlivých objektů, ale pro tzv. kvantifikaci – odhad distribuce objektů mezi třídy.

H. Sun, G. Miao, X. Yan: Noise-Resistant Bicluster Recognition
Na přednášce byla představena metoda biclusteringu založená na neuronové síti, která se naučí vlastnosti (features) dat. Pro potlačení šumu u genových dat je místo rekonstrukce celých dat pro zachycení bicluster vzorů použit přístup, při kterém se s nulovými a nenulovými částmi dat zachází různě.

E. Aksehirli, B. Goethals, E. Müller, J. Vreeken: Cartification: A Neighborhood Preserving Transformation for Mining High Dimensional Data
Na přednášce byla představena transformace objektů na vektory atributů, které reprezentují okolí objektu. Transformace zachovává podobnosti mezi objekty při více pohledech na data. Podle experimentů dosahují frequent itemset mining metody lepších výsledků nad takto transformovanými daty.

N. Tatti: Itemsets for Real-Valued Datasets

Na přednášce byla představena metoda dolování vzorů (pattern mining) pro reálně-hodnotová data pomocí transformace na binární data, kde prahy jsou chápány jako náhodné veličiny a support vzorů je průměrný.

Z. Jiang, S. Shekhar, X. Zhou, J. Knight, J. Corcoran: Focal-Test-Based Spatial Decision Tree Learning: A Summary of Results
Na přednášce byly představny prostorové (spatial) rozhodovací stromy, u kterých průchod stromem není založen jen na lokálních, ale i na fokálních, tj. okolních, vlastnostech objektu, a které redukují tzv. salt-and-pepper šum.

B. Micenková, R. T. Ng, X.-Hong Dang, I. Assent: Explaining Outliers by Subspace Separability
Na přednášce byla představena metoda určení možných vysvětlení nalezených tzv. outliers, tj. mimořádných objektů v datech, např. chyb. Vysvětlení, užitečná pro interpretaci či validaci outliers, jsou vyjádřená formou podprostorů, ve kterých jsou outliers separované od inliers.

R. Belohlavek, M. Krmelova: Beyond Boolean Matrix Decompositions: Toward Factor Analysis and Dimensionality Reduction of Ordinal Data
Na přednášce byla představena dekompozice matic s ordinálními daty založená na formální konceptuální analýze a algoritmus dekompozice. Technicky metoda spočívá v nahrazení dvouhodnotové Booleovy algebry pro binární data obecnějšími strukturami, které ale přináší nové problemy.

A. Krithara, G. Paliouras: TL-PLSA: Transfer Learning between Domains with Different Classes
Na přednášce byla představena metoda přenosového učení (transfer leaning), kdy cílová doména obsahuje pouze některé hodnoty tříd ze zdrojové domény, založená na Probabilistic Latent Semantic Analysis.

J. Tanha, M. Javad Saberian, M. Van Someren: Multiclass Semi-Supervised Boosting Using Similarity Learning
Na přednášce byl představen boosting algoritmus pro vícetřídní semi-supervised klasifikaci, využívající vícetřídní ztrátovou funkci pomocí souřadnicové gradientní metody.

N. Djuric, S. Vucetic: Efficient Visualization of Large-Scale Data Tables through Reordering and Entropy Minimization
Na přednášce byla představena metoda, EM-ordering, pro přeskládání řádků a sloupců tabulkových dat podle jejich podobností tak, aby tzv. heatmap (visuální inspekce dat) poskytovala hlubší pohled na data. Uspořádání minimalizuje entropii prediktivního kódování uspořádané tabulky dat. Pro řešení se využívá heuristické řešení problému obchodního cestujícího, kde řádky tabulky jsou města.

R. Shams, R. E. Mercer: Classifying Spam Emails with Text and Readability Features
Na přednášce byla představena klasifikace spamu používající atributy založené na přirozeném jazyku emailu (angličtině) a čitelnosti, v kombinaci s tradičními atributy založenými na obsahu emailu.

F. Chong Tat Chua, R. J. Oentaryo, Ee-Peng Lim: Modeling Temporal Adoptions Using Dynamic Matrix Factorization

Na přednášce byla představena metoda tvorby časových faktorizačních modelů predikce chybějících návrhů v uživatelově historii návrhů z doporučovacího systému. Metoda je rozšířením metody Non-negative Matrix Factorization.

Xu-Yao Zhang, K. Huang, Cheng-Lin Liu: Feature Transformation with Class Conditional Decorrelation
Na přednášce byl představen model transformace atributů, který diagonalizuje kovariantní matice různých tříd současně a nalezne společné hlavní komponenty mezi více třídami. Atributy jsou poté z hlediska tříd nekorelované, zlepšení klasifikace bylo ukázáno na nearest class mena metodě.

L. Du, Z. Shen, X. Li, P. Zhou, Yi-Dong Shen: Local and Global Discriminative Learning for Unsupervised Feature Selection
Na přednášce byla představena metoda učení integrující globální a množinu lokálně lineárních regresních modelů, která pro nalezení reprezentativních atributů potlačuje irelevantní a šumové atributy.

D. Erdős, P. Miettinen: Walk 'n' Merge: A Scalable Algorithm for Boolean Tensor Factorization
Na přednášce byl představen algoritmus pro Booleovskou tensorovou faktorizaci, tj. faktorizaci binárních dat na binární faktory při zachování binárních rekonstruovaného tensoru dat. Algoritmus provádí CP a Tucker dekompozici.

Další prezentace (panely, demonstrace, postery)
Zúčastnil jsem se následujících prezentací na některých workshopech a tutoriálu. Demonstrace, které jsem se chtěl zúčastnit, neproběhla. Podíval jsem se i na některé příspěvky na PhD fóru a zavítal jsem také na panel o dolování velkých dat (big data).

Z. Farzanyar: Accelerating Frequent Itemsets Mining on the Cloud: A MapReduce-Based Approach (workshop KDCloud: Knowledge Discovering Using Cloud and Distributed Computing Platforms)
- trendy v cloud computing, programových modelech a SW služeb s náhledem data mining a knowledge discovery přístupů je využívajících

R. Angryk: Solar Data Mining (workshop AstroInfo: Astroinformatics)
- aplikace data mining nástrojů na analýzu velkých astronomických repozitářů

P. Cintia, L. Pappalardo, D. Pedreschi: „Engine matters": a data driven study on cyclists' performance (workshop DMCS: Data Mining Case Studies and Practice Prize)
- implementace data mining mající za následek významné zlepšení business operací

E. M.L. Peters, G. Dedene, J. Poelmans: Empirical Discovery of Potential Value Leaks in Processes by means of Formal Concept Analysis (workshop EEML: Experimental Economics and Machine Learning)
- využití metod machine learning v teorii lidských strategických interakcí

P. Jiang, M. T Health: Pattern Discovery in High Dimensional Binary Data (workshop HDM: High Dimensional Data Mining)
- nové směry v řešení problémů zpracování dat s až miliony atributů, kde tradiční metody selhávají

F. Mendes, M. Y. Santos, J. Moura-Pires: Dynamic Data Analytics with an Incremental Clustering Approach (workshop IclaNov: Incremental Clustering, Concept Drift and Novelty)
- novinky v oblastech a aplikace analýzy informací proměnných v čase

J. Ye: Sparse Learning for Big Data, A. Saluja, M. Pakdaman, D. Piao, A. P. Parikh: Infinite Mixed Membership Matrix Factorization (workshop OEDM: Optimization Based Technique for Emerging Data Mining Problems)
- optimization a data mining a reálné aplikace

H. Tong, F. Wang, Ch. Ding: Applied Matrix Analytics: Recent Advance and Case Studies (tutoriál)
- přehled nových maticových data mining algoritmů s aplikacemi v sociální oblasti a zdravotní péči

Vlastní prezentace
Neměl jsem vlastní prezentaci.

Shrnutí konference (perspektivní témata apod.)

Na konferenci byly prezentovány průlomové příspěvky zejména z těchto oblastí: klasifikace a strojové učení (hlavně rozhodovací stromy), *feature selection a transformace atributů, doporučovací (recommender) systémy a prediktivní modelování. Významnými tématy byly také dolování vzorů (pattern mining) a dekompozice (faktorizace) matic.* Z příspěvků lze usuzovat, že perspektivními směry dalšího výzkumu jsou: klasifikace a strojové učení s učitelem (supervised) i bez něj (unsupervised, semi-supervised), feature selection a extraction a prediktivní modelování s aplikacemi v doporučovacích (recommener) systémech a částečně také v oblasti dolování vzorů (pattern mining) dekompozice matic. Tedy témata spíše z oblasti machine learning než data mining.

Navázání kontaktů
Barbora Micenková (Aarhus University, Denmark) – diskuze se slovenskou nadějnou vědkyní o jejím výzkumu, zkušenostech z jiných univerzit po světě a výzkumu u nás a na Slovensku

Fotografická dokumentace

Přílohy