



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: SDM 2013 (SIAM International Conference on Data Mining)
datum konání: 30.4. - 5.5. 2013
místo: Austin, Texas, USA
účastník konference: Mgr. Martin Trnečka

Stručný popis konference:

SDM je přední konferencí v oblasti data miningu. Konference je z většiny sponzorována společností Google a dalšími největšími firmami v oblasti informatiky jako jsou například NFS nebo IBM. Mezi účastníky jsou jak vědci z akademické sféry, tak vědci z výzkumných oddělení vůdčích firem v oblasti informatiky a prestižních vědeckých pracovišť (NASA).

Zajímavá čísla:

Acceptance rate:	25,57%
Submisí:	348
Počet účastníků:	704
Počet prezentovaných příspěvků:	89
Počet plenárních přednášek:	4
Počet paralelních sekcí:	4
Počet tutoriálů:	6
Počet aplikačních článků:	50%
Počet zemí:	26 (většina příspěvků z USA a Číny)

Zajímavé přednášky:

A. Plenární přednášky

Regional Climate Informatics: A Statistical Perspective

Doug Nychka

Přednáška se zabývala statistickým modelem pro analýzu klimatických změn a to především pro regionální předpovědi počasí. Tento matematický model, bere jako mezující podmínky

data s globálního klimatu. Díky tomu je předpověď efektivnější a rovněž je výpočet mnohem rychlejší protože, je možné ho soustředit na značně menší data, které jsou reprezentovány jako sparse matice. Dále bylo diskutováno rozšíření tohoto modelu o možné predikce regionálního klimatu.

Social Networks As Information Filters

Lada Adamic

Přednáška se zabývala tokem informací v sociálních sítích. Na šíření informace ovlivňuje celá řada sociálních aspektů, jako vzdělání, politická příslušnost, věk, zájmy a další. Díky tomu je možné v sociálních sítích pozorovat několik zajímavých fenoménů, které mají analogii v evoluční teorii druhů (mutace informace). Rovněž byl představen matematický model, který zachycuje danou skutečnost. Díky tomu je možné predikovat jak se bude informace šířit v sociálních sítích a jakou bude mít podobu určitém časovém intervalu, popřípadě, zda bude tato informace vůbec sítí putovat (filtrace informace).

Modeling Individual-Level Data in the 21st Century

Padhraic Smyth

Tato přednáška se věnoval individuálním datům různého charakteru (poslané e-maily, spánkové cykly, záznamy fyzické aktivity) a data miningem nad těmito data. Takováto data jsou velice specifická a odrážejí konkrétní znaky každého jedince. Analýzou takovýchto dat můžeme získat cenné informace, které je pak možné použít v mnoha dalších oblastech (predikce, studium chování, reklama, zdravotnictví, doporučování).

Mining Clinical Data to Build Predictive Models

Peter Szolovits

Tato zvaná přednáška diskutovala analýzu lékařských dat a to především pro tvorbu prediktivních modelů. Dále bylo prezentováno zajímavé srovnání ceny takovéto analýzy v průběhu posledních 30 let. Cena analýzy lidského genomu byla před 30 lety téměř nevyčíslitelná. V dnešní době se jedná o částku, kterou si může dovolit téměř každý člověk.

B. Řádné přednášky

Missing Or Inapplicable: Treatment of Incomplete Continuous Valued Features in Supervised Learning

Pang-Ning Tan, et. al.

Tento příspěvek se věnoval obecně známému tvrzení, že kvalita výstupu data miningových metod závisí především na kvalitě vstupu. Byla diskutována situace, kdy některé hodnoty jsou nedefinovány, respektive je nelze aplikovat (neaplikovatelná hodnota). Byla představena metoda pro práci s takovýmto typem dat, která byla prezentována na doporučovacím systému.

Patient Risk Prediction Model via Top-k Stability Selection

Jiayu Zhou, et. al.

Tento příspěvek se zabýval výběrem stabilních deskriptorů dat, které je možné použít pro predikci rizika u pacientů po operaci (například pooperační komplikace). Metoda Stability

selection, běžně používaná v oblasti machine learningu byla modifikována o výběr nejlepších top-k prvků. Výsledky získané touto metodou jsou pak mnohem kvalitnější.

Discriminative Feature Selection for Uncertain Graph Classification

Xiangnan Kong, et. al.

V tomto příspěvku byly prezentovány Uncertain graphs, které byly použity pro simulování vztahů mezi jednotlivými částmi mozku (data získaná z MRI). Díky tomu je možné studovat vztahy u "normálně" normálně fungujících mozků a mozků postižených degenerativním onemocněním, například roztroušenou sklerózou.

C. Tutoriály

Big Data Analytics for Healthcare

Jimeng Sun, Chandan Reddy

Tento tutoriál se věnoval zpracování rozsáhlých lékařských dat jako jsou například biomedická data, lékařské záznamy, kartotéky, finanční náklady na léčbu. Data mining takovýchto dat představuje obrovskou výzvu jelikož množství dat v lékařství narůstá téměř exponenciálně. Nabízí se tedy prostor pro využití data miningových metod jednak pro analýzu lékařských zpráv, analýzu procesů či predikce různého charakteru, ale i pro online monitoring, archivaci dat (elektronická karta) analýza genetických informací. Tutoriál představoval přehled stávajícího výzkumu právě v těchto oblastech.

Recent Advances in Applied Matrix Technologies

Fei Wang, Hanghang Tong

Tento tutoriál představil matice jako běžnou reprezentaci mnoha reálných dat a následně zde bylo prezentováno několik metod pro analýzu maticových dat. Zmíněny byly především metody Online learningu, Sparse learningu, dále pak metoda NMF a další low rank aproximační varianty NMF metody, metoda Graf Proximity a Eigenvalue optimalizace. Všechny metody byly prezentovány na aplikacích týkajících se analýzy lékařských dat různé povahy.

Další prezentace (panely, demonstrace, postery)

Paralelně s konferencí probíhal i několik workshopů z oblasti data miningu a zajímavých aplikací data miningu (zdravotnictví, sociální sítě, predikce počasí). Dále pak probíhala hlavní poster sekce, kde byly prezentovány články na které nebyly zařazeny do hlavních přednášek a poster sekce studentů místní univerzity.

Shrnutí konference (perspektivní témata apod.)

Na konferenci byly prezentovány příspěvky z těchto oblastí: clustering, big data mining, social networks mining, classification, learning. Dále byly diskutována analýza dat zejména ve zdravotnictví, meteorologii a analýza dat na sociálních sítích. Tyto tři témata pak představují aktuální nejperspektivnější směry vývoje v oblasti data miningu.

Fotografická dokumentace



Registrace



Konferenční místnost, přednáška

Dmitry