Multi-Task Data Mining toward Automating the KDD Process

Simplified version e.g., movie

Einoshin Suzuki
Kyushu University, Japan
suzuki@inf.kyushu-u.ac.jp

October 8, 2014 ICITEE 2014@Yogyakarta

Acknowledgement

Masamichi Shimura, Yves Kodratoff, Jan M. Zytkow, Shusaku Tsumoto, Yuu Yamada, Takeshi Watanabe, Hideto Yokoi, Katsuhiko Takabayashi, Masatoshi Jumi, Ning Zhong, Shin Ando, Daisuke Ikeda, Bin Tong, Thach Nguyen Huy, Hao Shao, Shigeru Takano, Asuki Kouno, Emi Matsumoto, Yutaka Deguchi, Daisuke Takayama, Vasile-Marian Scuturici, Jean-Marc Petit, Angdy Erna, Linli Yu, Kaikai Zhao, Wei Chen, Ryosuke Kondo
and many others

for the more-than-a-dozen papers I am going to present in this talk.

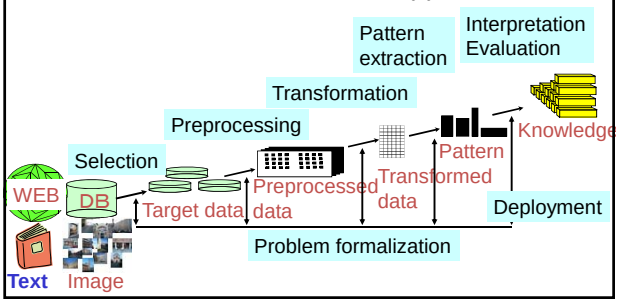
Massive Data

Birth of data mining in the early 90's

KDD Process Model

KDD: Knowledge Discovery in Databases

Model modified from [Fayyad 1996]



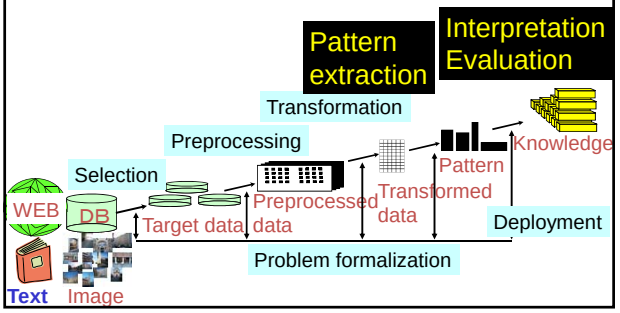
Toward Automating the KDD Process

Analogy to mining: impossible but a must

This talk: our 20-year efforts

- Part 1: Exception rule discovery
- Part 2: Medical data mining
- Part 3: Multi-task learning
- Part 4: Future direction

Part 1: Exception Rule Discovery



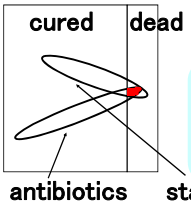
Rule Discovery

Fund type	5yrReturn	Diversity	Beta	Stocks	Yield	Xpnse Rat%	Turnover	Assets	Capgain
Growth	belowS&P	Low	Under1	Over75%	Under3%	Low	Low	Large	10to20%
Gth&Inc	belowS&P	High	Under1	Over75%	Under3%	Low	High	Small	10to20%
AggrGth	belowS&P	High	Over1	Over75%	Under3%	Low	High	Small	Under 10%

FundType=Gth&Inc, Beta=under1 → Yield=over3%
(25 out of 30 examples)
Stocks=under75%, XpnsRat=low
→ 5yrReturn=belowS&P (15 out of 19 examples)

Too many rules are discovered
Rules are not related (i.e., fragmented)

Exception Rule Discovery [Suzuki, Shimura KDD 96, Suzuki KDD 97]



Find all nonlinear interactions (few and interesting!)

antibiotics → cured
antibiotics, staphylococci → dead
(staphylococci ↛ dead)

$$\begin{cases} Y \rightarrow x \\ Y \wedge Z \rightarrow x' \\ Z \not\rightarrow x' \end{cases}$$

where

$$Y \equiv (a_1 = v_1) \wedge \cdots \wedge (a_\mu = v_\mu)$$
$$Z \equiv (a_1' = v_1') \wedge \cdots \wedge (a_\nu' = v_\nu')$$
$$x \equiv (a'' = v_1'') \quad x' \equiv (a'' = v_2'')$$

a : attribute, v : value

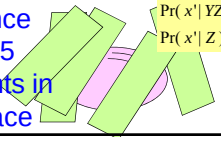
Difficulties and Invention

- High time complexity, $O(n^{2m+1})$ instead of $O(n^{m+1})$: branch-and-bound method, search pruning, bit-operation
- Many noisy patterns, weak exception or noise?: Analytical solution for the geometric problem

$\Pr(Y) \geq \theta_1^S$,
$\Pr(x \mid Y) \geq \theta_1^F$,
$\Pr(YZ) \geq \theta_2^S$,
$\Pr(x' \mid YZ) \geq \theta_2^F$,
$\Pr(x' \mid Z) \leq \theta_2^I$

Search tree

Confidence region + 5 constraints in a 7D-space



Analytical Solution

$$\left(1 - \beta \sqrt{\frac{1 - \hat{p}(Y)}{n\,\hat{p}(Y)}}\right) \hat{p}(Y) \geq \theta_1^S,$$

$$\left(1 - \beta \sqrt{\frac{1 - \hat{p}(Y,Z)}{n\,\hat{p}(Y,Z)}}\right) \hat{p}(Y,Z) \geq \theta_2^S,$$

$$\left(1 - \beta \sqrt{\frac{\hat{p}(\bar{x},Y)}{\hat{p}(x,Y)\left[(n+\beta^2)\hat{p}(Y) - \beta^2\right]}}\right) \hat{p}(x \mid Y) \geq \theta_1^F,$$

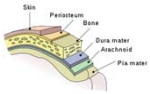
$$\left(1 - \beta \sqrt{\frac{\hat{p}(\bar{x}',Y,Z)}{\hat{p}(x',Y,Z)\left[(n+\beta^2)\hat{p}(Y,Z) - \beta^2\right]}}\right) \hat{p}(x' \mid Y,Z) \geq \theta_2^F,$$

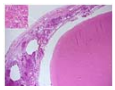
$$\left(1 + \beta \sqrt{\frac{\hat{p}(\bar{x}',Z)}{\hat{p}(x',Z)\left[(n+\beta^2)\hat{p}(Z) - \beta^2\right]}}\right) \hat{p}(x' \mid Z) \leq \theta$$

β: value related to the dimension of the ellipsoid and δ

$\hat{p}$: the ratio

Discovery from Meningitis Data [Suzuki, Tsumoto PAKDD 00]





High-quality data: aggregation at the patient level, collection by 2 physicians, one of whom is a data miner and the provider

83 ≤ CSF PRO ≤ 121 → CULT FIND = F
+ FOCUL = + → CULT FIND = T

4/5 validness, novelty, unexpectedness, usefulness

Attribute in the conclusion	#	Validness	Unexpectedness		
			Novelty		Usefulness
(all)	169	2.9	2.0	2.0	2.7
CULT FIND	4	3.3	4.0	4.0	3.5
CT FIND	36	3.3	3.0	3.0	3.2
EEG FOCUS	11	3.0	2.9	2.9	3.3

Discovery from Blood Test Data [Suzuki, Tsumoto ws 00]

20919 examples (1 hospital, 1 year), 135 attributes, many missing data

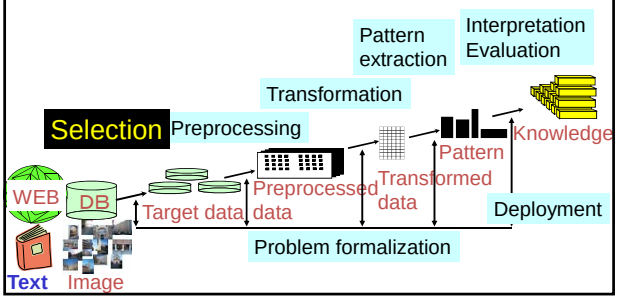
Relatively "hard" data: example = medical test (no aggregation)

Sex = M, LCMs = Sensitive → PCG = Resistant 316/598 examples

Sex = M, LCMs = Sensitive, Ward = Outpatient → PCG = Sensitive 34/46 examples

Ward = Outpatient ↛ PCG = Sensitive 1.6%

Part 2: Medical Data Mining



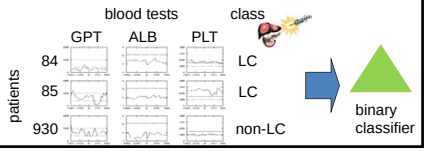
Chronic Hepatitis Data

Collected for 20+ years in a university hospital

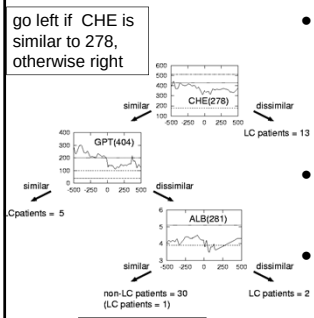
- Disease: virus inflame/harm liver cells
- Degree of fibrosis: index of progress F0(no fiber)→F1→F2→F3→F4 (liver cirrhosis: LC)
- Biopsy: to pick liver tissue for inspecting the degree of fibrosis (Invasive)
- Blood test: less invasive than biopsy
- Interferon: drug for killing virus. Has side effect

Selection/Formalization

- Laborious KDD process: preprocessing and meetings (endless?)
- 1st Problem: liver cirrhosis prediction (binary classification) with 14 attributes.
- Selection: remove acute hepatitis and other diseases. Use each 500 days before and after the first biopsy. At least 10 tests.



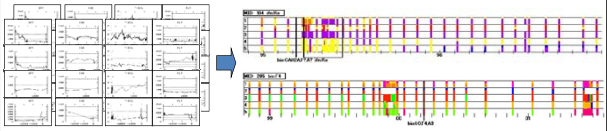
Time-Series Decision Tree [Yamada, Suzuki, Yokoi, Takabayashi, ICML 03]



- fits physician's interpretation, though we only specified attributes to use
- shows predictive accuracy 88.2 %
- is powerful in detecting exceptional cases

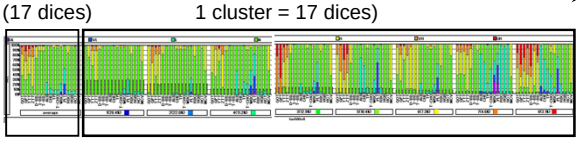
Less effective to patients with few data [Yamada, Suzuki, Yokoi, Takabayashi AM03]

PrototypeLines [Suzuki, Watanabe, Yokoi, Takabayashi ICDM 03]



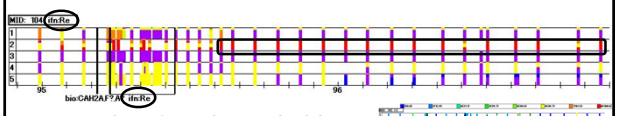
300 pages — **30 pages**

average (17 dices) ——— 8 clusters (best: purple to worst: red, 1 cluster = 17 dices) →



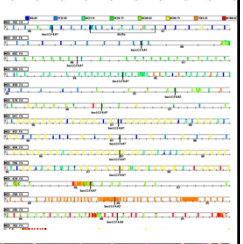
"Rareness" criteria for assigning colors to clusters

Data Cleaning, Clustering, Usability

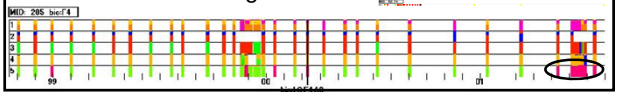


A CS student found a probable mistake in the medical record

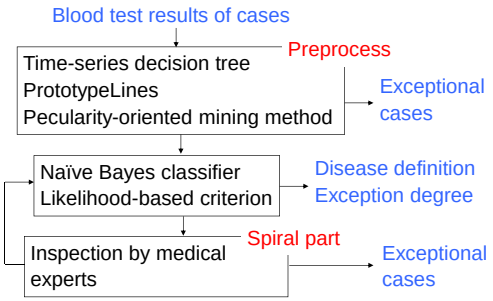
It allows intuitive clustering for diverse LC patients



Physician, in 5min, said "patient is anemia due to vomiting blood"



Spiral Exception Discovery [Jumi, Suzuki, Ohsima, Zhong, Yokoi, Takabayashi ISMIS 2005]



Automatic Discovery [Jumi, Ohsima, Zhong, Yokoi, Takabayashi , Suzuki NGC j. 07]

No expert involved except the final evaluation. 14 + 21 blood tests. 140 LC patients + 106 non-LC patients (at least one value for each patient)

Discovered a hypothesis $\frac{\Pr(\text{AMY} = \text{high} \mid \text{LC})}{\Pr(\text{AMY} = \text{high} \mid \text{non-LC})} > 3$

The odd was below 3 in the original data so we paid no attention until our system removed exceptional cases

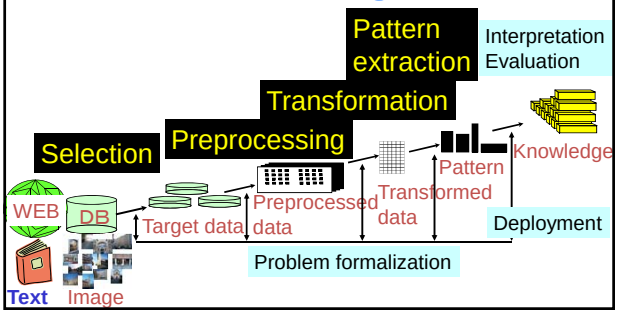
The hypothesis had been reported as the main result in Raffaele Pezzilli: ``Serum Pancreatic Enzyme Concentrations in Chronic Viral Liver Diseases'', Digestive Diseases and Sciences, Vol. 44, No. 2, pp. 350--355, 1999. (PMID: 10063922)

Anecdotes

- Medical criticism: Amy might be high due to operation and interferon → it is rare to do them before the first biopsy
- Astonishment: a pamphlet of MEXT says Amy is low → Dr. Takabayashi explained that the types of chronic hepatitis are different
- AI criticism: the discovery is just a good luck → Method shows an empirical evidence

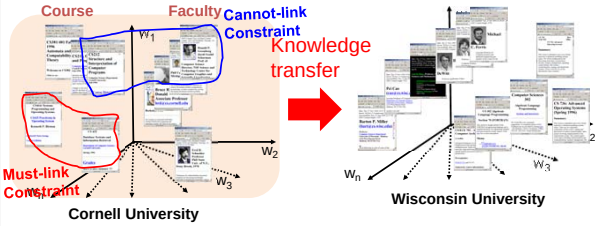
Difficulty of evaluating discovery methods [Suzuki LNCS State-of-the-Art Survey 2002]

Part 3: Multitask Learning



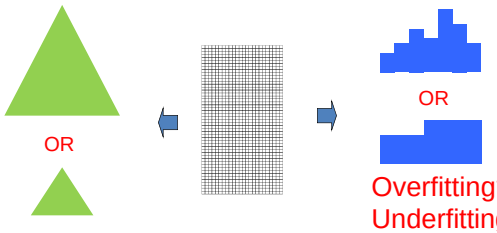
Multitask Learning

***aims to improve the performance of learning algorithms by learning patterns for multiple tasks jointly* (Weinberger's definition modified)**



[Tong, Shao, Chou, Suzuki, ECML/PKDD 2010]

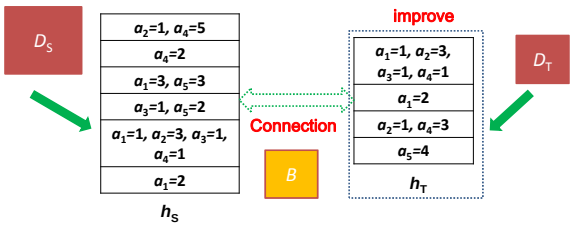
Overfitting and MDL [Rissanen 78-] for Single-Task Learning



$$h_{\text{MDL}} \equiv \arg\min_{h \in H}\left(-\log_2 P(D \mid h) - \log_2 P(h)\right)$$

$$= \arg\max_{h \in H} P(h \mid D)$$

Extended MDL for Multitask Learning [Shao, Suzuki SDM 11]



$$\underset{h_S,h_T,B}{\operatorname{argmin}}\left(-\log_2 P(D_S \mid h_S, B) - \log_2 P(D_T \mid h_T, B) - \log_2 P(h_S) - \log_2 P(h_T) - \log_2 P(B)\right)$$

$$= \underset{h_S,h_T,B}{\arg\max}\, P(h_S, h_T, B \mid D_S, D_T)$$

Experimental setting

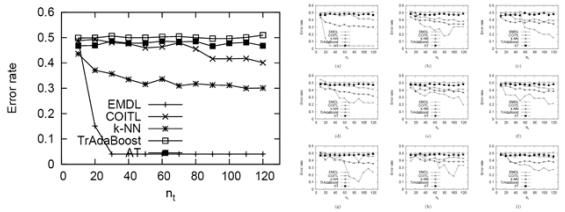
- Synthetic Datasets: Nine pairs (D_s, D_t) generated by the "target concepts" with 32 attributes. For example:

Index	μ	*common*	*shortest*	*longest*	*avg.*
(a)	3	1	1	3	1.33
(b)	3	1	2	3	2.33
(c)	3	2	2	3	2.33
(d)	4	1	2	4	3.24
(e)	5	1	3	4	3.60

- Real Datasets: Four data sets are used in the experiments in UCI repository. A pre-processing method [Y. Shi 09] is adopted on these data sets to split each data to the source and the target task.

Experimental Results (1)

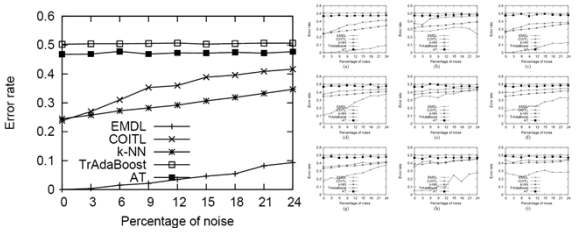
Results on synthetic datasets, for the different n_t of the target task, with 10% noise in both the source and the target tasks



Our method can find good features even n_t is small. By increasing n_t, EMDLP is able to achieve lower error rate.

Experimental Results (2)

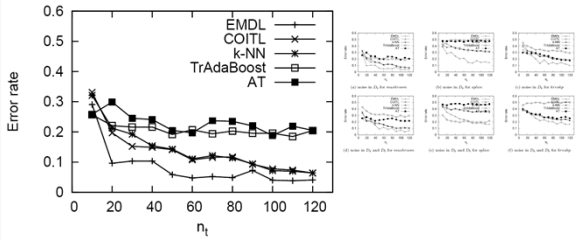
Results on synthetic datasets, for the different noise levels from 0% to 24% in both the source and the target tasks



Our EMDLP is robust to noise

Experimental Results (3)

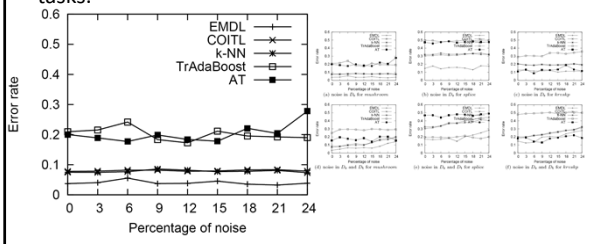
Results on real datasets, for the different n_t of the target task, with 10% noise in the source task or in both the tasks.



Our method is able to achieve lower error rate with few labeled information available.

Experimental Results (4)

Results on real datasets, for the different noise levels from 0% to 24% in the source task or in both tasks.



Even under severe noise levels, our EMDLP is still able to obtain the contemplated h_t

Information Distance [Li et al. 92, 97]

Universal distance between string *x* and *y*

$$d(x,y) \equiv 1 - \frac{K(x) - K(x \mid y)}{K(xy)} \approx \frac{K(x \mid y) + K(y \mid x)}{K(xy)}$$

Kolmogorov complexity $K(x|y)$: length of the shortest program that returns *y* given *x*

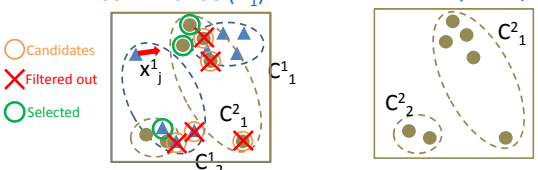
$K(xy)$: length of the shortest program that outputs *xy*

Not a computable function →Substitute K() by the compressed size

Can be used for single-task clustering

Extended Information Distance for Multi-Task Clustering [Nguyen Huy, Shao, Tong, Suzuki, ISMIS 11]

Task 1: Yahoo (D_1) Task 2: Wikipedia (D_2)



1. Clustering each domain separately
2. Find candidate related instance set $S_{x^1_j} = \{C^1_2, C^2_1\}$
3. Filter out unrelated instances $\Delta(x \mid y) = C(x) - C(x \mid y)$
4. Calculate distance matrix $LDCDM(x^i_j, x^i_l) = \frac{C(x^i_j \mid S_{x^i_j}) + C(x^i_l \mid S_{x^i_j})}{C(x^i_j x^i_l)}$

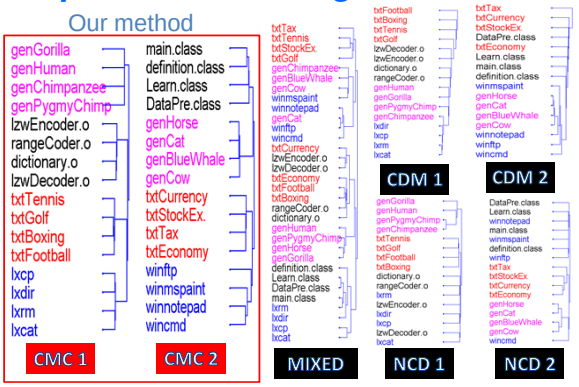
Go to Step 1

Experimental Setting

- 5 Data sets:
  - Heterogeneous: 8 DNAs, 8 compiled, 8 text, 8 executive files.
  - DNA, Music, Language: Declaration of Human Rights [a] in 30 languages: 10 Europeans, 10 Americans and 10 Africans.
  - Document: 20 Newsgroup data [b].
- Comparison to 10 methods: CDM [Keogh 04], NCD [Vitanyi 97], 4 Bernoulli methods [Steinbach 00], CLUTO [Karypis 98], Co-clustering [Dhillon 01], LSSMTC [Gu 09], MBC [Zhang 10].

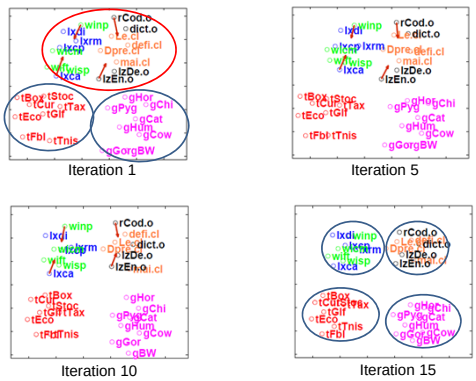
[a] http://www.un.org/en/documents/udhr/
[b] http://people.csail.mit.edu/jrennie/20Newsgroups/

Exp. Result: Heterogeneous Data

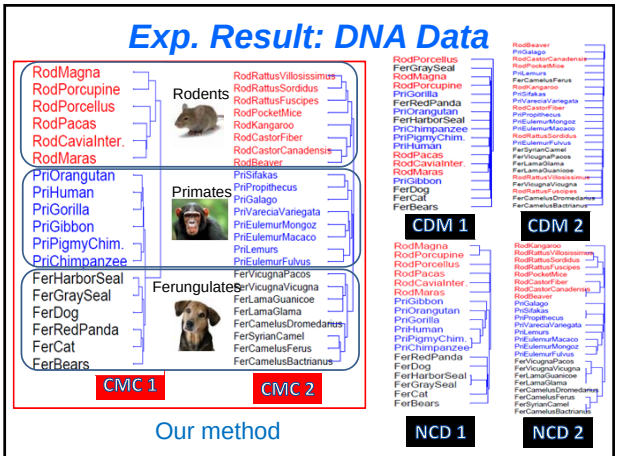


Supplement Website: http://tlas.i.kyushu-u.ac.jp/~suzuki/ISMIS2011/ISMIS2011.html

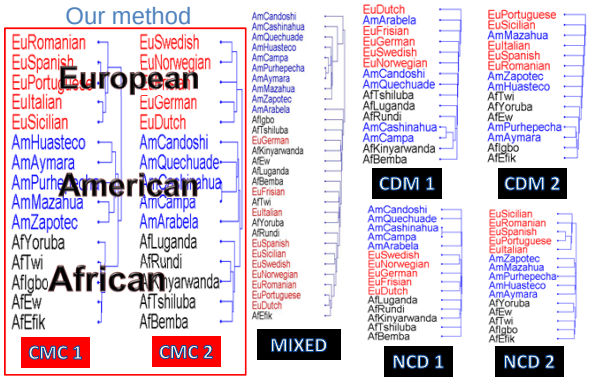
Exp. Result: Heterogeneous Data



Exp. Result: DNA Data

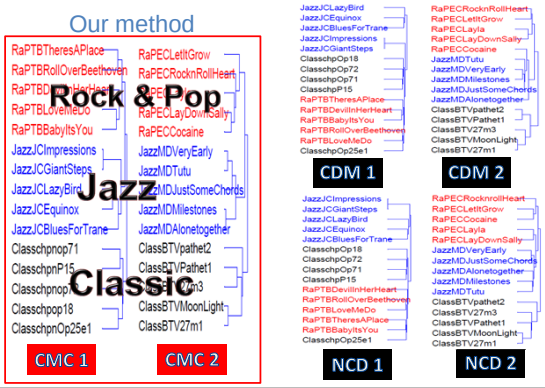


Exp. Result: Language Data



Supplement Website: http://tlas.i.kyushu-u.ac.jp/~suzuki/ISMIS2011/ISMIS2011.html

Exp. Result: Music Data



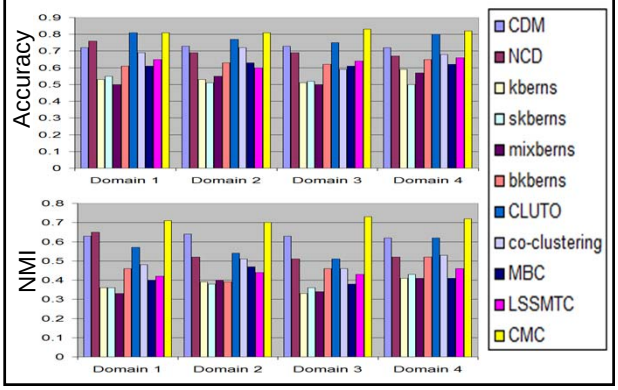
Exp. Result: 20 Newsgroup Data Define Domains [Steinbach 2000]

Comp.	Rec.	Sci.	Talk.	
Graphics	Auto	Crypt	Guns	Domain 1
Win.misc	Moto	Electronic	Mideast	Domain 2
Ibm.hw	Baseball	Med	Poli.misc	Domain 3
Mac.hw	Hockey	Space	Reli.misc	Domain 4

Each domain contains

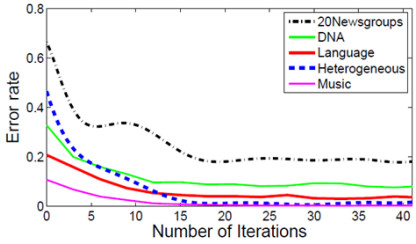
200 documents x 4 subcategories

Exp. Result: 20 Newsgroup data



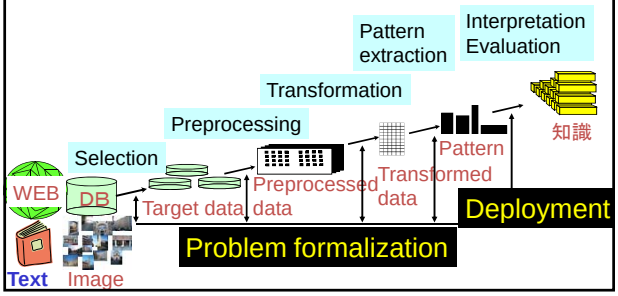
Convergence Property

$$Err = \frac{\sum Incorrectly\ clustered\ instances}{Total\ instances}$$



CMC converges after 17 iterations

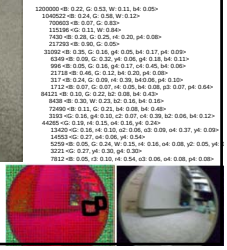
Part 4: Future Direction



Discovery Robot [Suzuki, Matsumoto, Kouno DS 12]

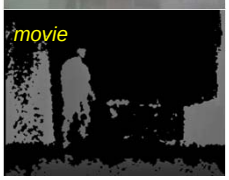
Clusters millions of subimages real-time + Detects anomalies

Condense subimages in a CF (cluster feature) tree [Zhang 97]



Report subimages that are dissimilar to the most similar leaf as anomalies

Kinect for Windows



Skeleton + color image, depth image, face image

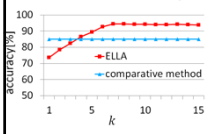
Now we have 5 new Kinects!

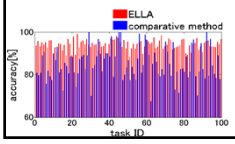
http://news.mynavi.jp/articles/2013/06/03/windows8report/001.html

Lifelong Learning [Ruvolo, Eaton ICML 13] and Application to 100-person Facial Expression Data [Tanoue, Suzuki dmc 14]

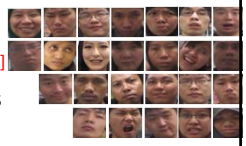
Sparse coding

$$e_T(\mathbf{L}) = \frac{1}{T}\sum_{t=1}^{T} \min_{\mathbf{s}^{(t)}} \left\{ \frac{1}{n_t}\sum_{i=1}^{n_t} \mathcal{L}\left(f\left(\mathbf{x}_i^{(t)}; \mathbf{L}\mathbf{s}^{(t)}\right), y_i^{(t)}\right) + \mu\|\mathbf{s}^{(t)}\|_1 \right\} + \lambda\|\mathbf{L}\|_F^2$$



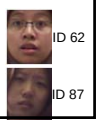


100-person data [Erna, Yu, Zhao, Chen, Suzuki AMT04]



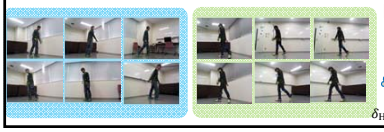
- Outperforms single-task learning in 83/100 tasks
- Over 20% gain in 4 tasks

Effective knowledge transfer

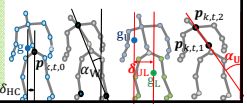


Fall Risk Dicovery [Deguchi, Suzuki ISMIS 14; Deguchi, Takayama, Takano, Scuturici, Petit, Suzuki PETRA 14 & AmI 14]

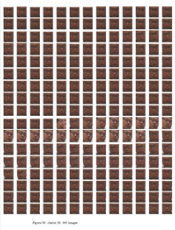
Examples of discovered fall postures (clusters)



Bilateral project with France (JSPS-CNRS)



Face Clustering [Kondo, Deguchi, Suzuki AmI 14]



Conclusions

- Overall: a brief tour over our 20-year efforts toward automating data mining
- Future: autonomous robots that conduct lifelong discovery for taking care of humans