Some problems and trends in data mining
データマイニングの課題とトレンド

Tu-Bao Ho
School of Knowledge Science
Japan Advanced Institute of Science and Technology

The talk aims to ...

→ Introduce to basic concepts and techniques of data mining (DM).
データマイニング（DM）の基本概念と技法を紹介する。

→ Present some challenging data mining problems, and kernel methods as an emerging trend in this field.
データマイニングのチャレンジ課題およびこの分野で興隆しつつあるカーネル手法について説明する。
What motivated data mining?
データマイニングの必要性？

We are living in the most **exciting** of times: Computer and computer networks
もっともエキサイティングな時代: コンピュータとコンピュータ・ネットワーク

- Much more data around us then before. They are collected and stored in huge databases (millions of records, thousands of fields).
前代未聞に膨大なデータに囲まれる生活。何百万件もの多岐に渡るデータデータは巨大データベースに格納されている。

- Many kinds of complexly structured data (non-vectorial).
多種多様な複雑に構造化されたデータ (非ベクトル型).

Astronomical data　天文学的データ

**Astronomy is facing a major data avalanche:**
天文学ではデータ崩壊の危機に瀕している

Multi-terabyte sky surveys and archives (soon: multi-petabyte), billions of detected sources, hundreds of measured attributes per source ...
何テラバイトもの天空観測データ、何十億もの観測源、観測源ごとに何百もある属性
Earthquake data

Japanese earthquakes
日本の地震1961-1994

1932-1996

04/25/92 Cape Mendocino, CA

Explosion of biological data
爆発的な生物学データ

DBGET Database Links

<table>
<thead>
<tr>
<th>KEGG</th>
<th>Pathway</th>
<th>Gene</th>
<th>Protein</th>
<th>Genome</th>
<th>LIGAND</th>
<th>Enzyme</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Protein interaction network
タンパク質相互作用網

EMBL Database Growth

10,267,507,282 bases in 9,092,760 records.
A portion of the DNA sequence, consisting of 1.6 million characters, is given as follows (about 350 characters, 4570 times smaller):

1600万文字からなるDNA配列の一部（4570分の一）

...TACATTGTTATTACATTGAGAAACCTTTATAATTAAAAAGATTCATGTAATTCTTATTGTTTTATAGGTTTTAATTTTCTAAGGGTTTGCTGGTTTCTATTGTTAATTTAAAACTTAAACATTTGTTAGAATTTTAACCTAAATTGGAATTTTTGAAAATTAGGATTAATTAGGTAAGTAATAAAAATTTCTCAGGATAAAAATACTACTCCTGGTTTTGTTTTATATATATGAAGTAGATTACCATTAAAGGTGATGAAGTGATGAAGTATATTATGT…

Many other kinds of biological data

Approximately 80% of the world’s data is held in unstructured formats (source: Oracle Corporation)

Example: MEDLINE is a source of life sciences and biomedical information, with nearly eleven million records

Example: MEDLINE is a source of life sciences and biomedical information, with nearly eleven million records

→ About 60,000 abstracts on hepatitis (そのうち6万件が肝炎)

Pathogenesis of autoimmune hepatitis.
Institute of Liver Studies, King's College Hospital, London, United Kingdom.

Autoimmune hepatitis (AIH) is an idiopathic disorder affecting the hepatic parenchyma. There are no morphological features that are pathognomonic of the condition but the characteristic histological picture is that of an interface hepatitis without other changes that are more typical of other liver diseases. It is associated with hypergammaglobulinaemia, high titres of a wide range of circulating auto-antibodies, often a family history of other disorders that are thought to have an autoimmune basis, and a striking response to immunosuppressive therapy. The pathogenetic mechanisms are not yet fully understood but there is now considerable circumstantial evidence suggesting that: (a) there is an underlying genetic predisposition to the disease; (b) this may relate to several defects in immunological control of autoreactivity, with consequent loss of self-tolerance to liver auto-antigens; (c) it is likely that an initiating factor, such as a hepatotropic viral infection or an idiosyncratic reaction to a drug or other hepatotoxin, is required to induce the disease in susceptible individuals, …
Web link data

Friendship Network
[Moody '01]

Over 3 billion documents

What is data mining?
データマイニングとは何か？

“Data-driven discovery of models and patterns from massive observational data sets”
大規模な観測データからのモデルおよびパターンのデータ駆動型発見

Languages, Representations
言語, 表現

Statistics, Inference
統計学, 推論

Engineering, Data Management
工学, データ管理

Applications
応用
**Example: mining associations in market data**

Super market data

“Young men buy diaper and beer together”

「紙おむつを買う男性は缶ビールを一緒に買うことが多い」

（解釈: 顧客像）紙おむつを買うように頼まれた男性がついでに自分用の缶ビールを購入していた → 今後の陳列に活かすことのできる知識。

---

**Convergence of thee technologies**

3つのテクノロジーの集合

- Increasing computing power
  - 計算力の増大

- Statistical and learning algorithms
  - 統計学的アルゴリズム/学習アルゴリズム

- Improved data collection and management
  - データ収集および管理の改良

---

KCGI, 13 July 2006
Improved data collection and management
データ収集および管理の改善

- Soon everything can be recorded and indexed
  すべてが記録され索引付けられる

- Most bytes will never be seen by humans
  ほとんどは人間が確認することはない

- What will be key technologies to deal with huge volumes of information sources?
  莫大な情報を取扱うための主要技法は何か？

["How much information is there?"
Adapted from the invited talk of Jim Gray (Microsoft) at KDD’2003]
Statistical and learning algorithms

**Principal Component Analysis (PCA)**
- Finds directions of maximal variance in Gaussian data (second-order statistics).
- 主成分分析 (PCA): ガウス分布データにおいて分散が最大となる方向の発見 (一次統計).

**Independent Component Analysis (ICA)**
- Finds directions of maximal independence in non-Gaussian data (higher-order statistics).
- 独立成分分析 (ICA): 非ガウス分布データにおいて独立性が最大となる方向の発見 (高次統計).

---

**Perform ICA**

- Speaker 1
- Speaker 2
- Speaker 3
- Speaker 4

- Mic 1
- Mic 2
- Mic 3
- Mic 4

- Terry
- Scott
- Te-Won
- Tzyy-Ping

**Play Mixtures**

**Play Components**
**Machine learning and data mining**

- **Machine learning** 機械学習
  - To build computer systems that learn as well as human does (science of learning from data).
  - 人間のように学習するコンピュータシステムを構築する（データからの学習の科学）
  - ICML since 1982 (23th ICML in 2006), ECML since 1989.
  - ECML/PKDD since 2001.

- **Data mining** データマイニング
  - To find new and useful knowledge from large datasets (data engineering).
  - 大きなデータベースから新しく有用な知識を見つける（データ工学）
  - ACM SIGKDD since 1995, PKDD and PAKDD since 1997
  - IEEE ICDM and SIAM DM since 2000, etc.

---

**Knowledge discovery in databases (KDD) process**

1. **Understand the domain and define problems** 領域を理解し問題を定義する
2. **Collect and preprocess data** データを収集し前処理する
3. **Data mining**
   - **Extract Patterns/Models** パターン・モデルを抽出する
4. **Interpret and evaluate discovered knowledge** 発見した知識を解釈し評価する
5. **Putting the results in practical use** 結果を実践に用いる

(KDD is inherently interactive and iterative)

(In many cases, viewed KDD as data mining)
**Types of data**
- Flat data tables 表形式データ
- Relational databases 関係DB
- Temporal & spatial data 時空間データ
- Transactional databases 取引データ
- Multimedia data マルチメディアデータ
- Genome databases ゲノムデータ
- Materials science data 材料データ
- Textual data テキストデータ
- Web data ウェブデータ
- etc.

**Mining tasks and methods**
- Classification/Prediction 分類/予測
  - Decision trees 決定木
  - Neural networks 神経回路網
  - Rule induction ルール帰納法
  - Support vector machines SVM
  - Hidden Markov Model 隠れマルコフ
  - etc.

- Description 記述
  - Association analysis 相関分析
  - Clustering クラスタリング
  - Summarization 要約
  - etc.

**Different data schemas**

**Dataset: cancerous and healthy cells**
データ例：がん細胞と健康な細胞

<table>
<thead>
<tr>
<th></th>
<th>color</th>
<th>#nuclei</th>
<th>#tails</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1</td>
<td>light</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>H2</td>
<td>dark</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>H3</td>
<td>light</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>H4</td>
<td>light</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>C1</td>
<td>dark</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>C2</td>
<td>dark</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>C3</td>
<td>light</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>C4</td>
<td>dark</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

**Unsupervised data**
**Supervised data**
Primary tasks of data mining
データマイニングの第一の課題

- **Predictive** mining tasks perform inference on the current data in order to make prediction or classification.
  
  （予測的マイニングの課題は，未知のデータの予測を目的として，現在のデータに関する推論を行うことである）
  
  **Ex.** IF “color = dark” and “#nuclei = 2”
  THEN cancerous

- **Descriptive** mining tasks characterize the properties of the data in the database.
  
  （記述的マイニングの課題は，データベース中のデータの全般的特性を特徴付ける記述を与えることである）
  
  **Ex.** “Healthy cells usually have one nuclei while cancerous ones have two”

Minning with decision trees
決定木によるマイニング

- Generalize classification models in form of trees.
  木構造の一般的な分類モデル

- Well known methods and systems:
  著名な手法とシステム:
  CART (Breiman et al.), C4.5, See5 (Quinlan)

Some problems:
問題点

- Learning decision trees from huge datasets (data access)
  大規模データからの学習

- Learning decision trees from complexly structured data
  複雑に構造化されたデータからの学習

- Decision tree ensembles: random forests (Breiman, 2001)
  組合わせによる決定木：ランダム・フォレストなど
**Decision trees with complex split tests**

**Study of liver cirrhosis (LC) in active data mining**

アクティブマイニングでの肝硬変の研究

the most dissimilar red curve to the blue curve
青い線ともっとも違う赤い線

(Suzuki group, Yokohama Nat. Univ. Accuracy 88.2%)

Motoda’s group, Osaka Univ., Prediction accuracy: 87.5% by 10-CV
Mining with neural networks
神経回路網によるマイニング

Some problems: 課題

- Difficult to understand the learned function (weights)
  学習結果の関数(重み)を理解することが困難
- Not easy to incorporate domain knowledge
  背景知識との組合わせは容易ではない

Multi layer neural networks
多層神経回路網

Well known methods: 著名手法
backpropagation, SOM, etc.

Mining associations
相関ルールによるマイニング

Some problems: 課題

- Database scan reduction: partitioning, hashing, sampling, find non-redundant rules
  データベース走査の低減：分割、ハッシュ、サンプリング、非冗長ルール
- New measures of association
  (Interestimgness and exceptional rules)
  相関に関する新指標

Associations among itemsets
アイテム集合間の相関

Apriori algorithm
(Agrawal, 1983). Variants: PF-growth, closed-frequent itemsets, multi level association rules, etc.
Mining clusters
クラスタリングによるマイニング

Some problems:
課題
- Find clusters with arbitrary shapes
  任意の形のクラスタを見つける
- Finding clusters from complex and huge datasets (e.g., Web communities)
  複雑で巨大なデータからクラスタを見つける

Challenges in data mining
データマイニングにおけるチャレンジ

Large data sets ($10^6$-$10^{12}$ bytes) and high dimensionality
(10^2-10^3 attributes)
規模と次元数

[Problems: efficiency, scalability?]
効率、スケーラビリティ

Different types of data in different forms
(mixed numeric, symbolic, text, image, voice, ...)
データの形式やタイプ

[Problems: quality, effectiveness?]
質、効果

Data and knowledge are changing
変化し続けるデータや知識

Human-computer interaction and visualization
人間－コンピュータのインタラクションと視覚化
**Numerical vs. symbolic data**

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Numerical</th>
<th>Symbolic</th>
</tr>
</thead>
<tbody>
<tr>
<td>No structure</td>
<td>≠</td>
<td>Places, Color Nominal (categorical)</td>
</tr>
<tr>
<td>Ordinal structure</td>
<td>= ≠ ≥</td>
<td>Age, Temperature, Taste, Rank, Resemblance Ordinal</td>
</tr>
<tr>
<td>Ring structure</td>
<td>= ≠ ≥ + ×</td>
<td>Income, Length Measurable</td>
</tr>
</tbody>
</table>

Combinatorial search in hypothesis spaces (machine learning)
仮説空間における組合わせ探索

Often matrix-based computation (multivariate data analysis)
通常は行列ベースの計算（多変量データ解析）

---

**Text mining**

Text Mining = Data Mining (applied to textual data) + Language Engineering

テキストマイニング = データマイニング (テキストへの応用) + 言語工学

Areas related to text mining: 関連分野
- Computational linguistics (NLP) 計算言語学
- Information extraction 情報抽出
- Information retrieval 情報検索
- Web mining ウェブマイニング
- Regular data mining 通常のデータマイニング

---

KCGI, 13 July 2006
A typical example of text mining

抜粋した文の断片を人間の医学専門知識を使って組合せ、文献にない新しい仮説を導き出す

Web mining

Web Mining = Data Mining (applied to Web documents and services) + Web technology

Areas related to Web mining:
- Information extraction
- Information retrieval
- Text mining
- Regular data mining

The contemporary robots travel on the Web and create maps like this.
(from http://www.caida.org/projects/internetatlas/galleries/index.xml)
Bioinformatics = Data Mining + Machine Learning + Biological Databases
バイオインフォマティクス = データマイニング + 機械学習 + 生物学データ

- **Sequence analysis**
  - Sequence alignment
  - DNA sequence analysis
  - Statistical sequence matching

- **Genomics**
  - Gene finding & prediction
  - Functional genomics
  - Structural genomics

- **Proteomics**
  - Functional proteomics
  - Structural proteomics
  - Structure, function relationship

- **Other problems**
  - Gene expression analysis
  - Pathway analysis
  - Protein-protein interaction

DNA \rightarrow RNA \rightarrow protein
Sequence \rightarrow Structure \rightarrow Function
Interaction \rightarrow Network \rightarrow Function

- **From Genes to Proteins**

The talk aims to...

- Introduce to basic concepts and techniques of data mining (DM).
  データマイニング（DM）の基本概念と技法を紹介する。

- Present some challenging data mining problems, and kernel methods as an emerging trend in this field.
  データマイニングのチャレンジ課題およびこの分野で興隆しつつあるカーネル手法について説明する。
A typical problem: Labeling sequence data

- $X$ is a random variable over data sequences
- $Y$ is a random variable over label sequences whose labels are assumed to range over a finite label alphabet $A$

**Problem:** Learn how to give labels from a closed set $Y$ to a data sequence $X$

<table>
<thead>
<tr>
<th>X:</th>
<th>Y:</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$: Thinking</td>
<td>$y_1$: noun</td>
</tr>
<tr>
<td>$x_2$: is</td>
<td>$y_2$: verb</td>
</tr>
<tr>
<td>$x_3$: being</td>
<td>$y_3$: noun</td>
</tr>
</tbody>
</table>

- POS tagging, phrase types, etc. (NLP),
- Named entity recognition (IE)
- Modeling protein sequences (CB)
- Image segmentation, object recognition (PR)
- etc.

1990s–2000s: Statistical learning

1980s: Standard resources and tasks

1970s: Kernel (vector) spaces

1960s: Representation transformation

Archeology of natural language processing (NLP)

Trainable FSMs

Trainable parsers
Trainable finite state machines

Hidden Markov Models (HMMs)
[Baum et al., 1970]
- Generative
- Need independence assumption
- Local optimum
- Local normalization

Maximum Entropy Markov Models (MEMMs)
[McCallum et al., 2000]
- Discriminative
- No independence assumption
- Global optimum
- Local normalization

Conditional Random Fields (CRFs)
[Lafferty et al., 2001]
- Discriminative
- No independence assumption
- Global optimum
- Global normalization

More accurate than HMMs

Machine learning and statistics in NLP

1992 ACL
24% (8/34)

1994 ACL
35% (14/40)

1996 ACL
39% (16/41)

1999 ACL
60% (41/69)

2001 NAACL
87% (27/31)

2005 ACL
96% (74/77)

(Marie Claire, ECML/PKDD 2005)
Finding “things” but not “pages”
情報抽出 (Information Extraction) vs. 情報検索 (Information Retrieval)

Information extraction: the process of extracting text segments of semi-structured or free text to fill data slots in a predefined template
情報抽出: テキストから事前定義したテンプレートを埋める部分的なテキストを抽出する

foodscience.com-Job2
JobTitle: Ice Cream Guru
Employer: foodscience.com
JobCategory: Travel/Hospitality
JobFunction: Food Services
JobLocation: Upper Midwest
Contact Phone: 800-488-2611
DateExtracted: January 8, 2004
Source: www.foodscience.com/jobs_midwest.html
OtherCompanyJobs: foodscience.com-Job1

Kernel methods and support vector machines
カーネル手法とサポートベクトルマシン

<table>
<thead>
<tr>
<th>Kernel methods &amp; SVM</th>
<th>166</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probabilistic, graphical models</td>
<td>146</td>
</tr>
<tr>
<td>Unsupervised learning, clustering</td>
<td>128</td>
</tr>
<tr>
<td>Statistical models</td>
<td>121</td>
</tr>
<tr>
<td>Language, Text &amp; web</td>
<td>68</td>
</tr>
<tr>
<td>Learning in bioinformatics</td>
<td>45</td>
</tr>
<tr>
<td>ANN</td>
<td>29</td>
</tr>
<tr>
<td>ILP</td>
<td>9</td>
</tr>
<tr>
<td>CRF</td>
<td>13</td>
</tr>
</tbody>
</table>

ICML 2006 (720 abstracts)
Kernel methods: the basic idea

Converting data into another high dimensional space can make data become linear separable

\[ \phi : \mathcal{X} = \mathbb{R}^2 \to \mathcal{Y} = \mathbb{R}^3 \]

\[ (x_1, x_2) \mapsto (x_1, x_2, x_1^2 + x_2^2) \]
Kernel methods: a bit of history
カーネル手法: 歴史を少し

- **Linear learning machines** (perceptrons, 1956) has one big problem of insufficient capacity. Minsky and Pappert (1969) highlighted the weakness of perceptrons.
- **Neural networks** (since 1980s) overcame the problem by glueing together many thresholded linear units (multi-layer neural networks: solved problem of capacity but ran into training problems of speed and multiple local minima).
- The **kernel methods** approach (since 2000s) is to stick with linear functions but work in a high dimensional feature space.

The data from $X$ into a (high-dimensional) vector space, the feature space $F$, by applying the feature map $\phi$ on the data points $x$.

Find a linear (or other easy) pattern in $F$ using a well-known algorithm (that works on the Gram matrix).

By applying the inverse map, the linear pattern in $F$ can be found to correspond to a complex pattern in $X$.

This implicitly by only making use of inner products in $F$ (kernel trick).
Kernel methods: math background

**Input space** $\mathcal{X}$ (入力空間 $\mathcal{X}$)

**Feature space** $\mathcal{F}$ (特徴空間 $\mathcal{F}$)

Kernel function $k: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$

Gram matrix $K_{nn} = \{k(x_i,x_j)\}$

Kernel-based algorithm on $K$

**Linear algebra, probability/statistics, functional analysis, optimization**

- **Mercer theorem**: Any positive definite function can be written as an inner product in some feature space.
- **Kernel trick**: Using kernel matrix instead of inner product in the feature space.
- **Representer theorem**: Every minimizer of $\min_{f \in \mathcal{F}} \{C(f,\{x_i,y_i\}) + \Omega \|f\|_p\}$ admits a representation of the form $f(\cdot) = \sum_{i=1}^{m} \alpha_i K(\cdot,x_i)$

Support vector machines: key ideas

- $\cdot$ denotes $+1$
- $\circ$ denotes $-1$

$f(x,w,b) = \text{sign}(w \cdot x + b)$

How would you classify this data?

(Pages 46-52 from Andrew Moore's SVM tutorial)
Support vector machines: key ideas

$\textbf{f}(\textbf{x}, \textbf{w}, b) = \text{sign}(\textbf{w} \cdot \textbf{x} + b)$

How would you classify this data?

- denotes +1
- denotes -1
Support vector machines: key ideas

- denotes +1
- denotes -1

\[ f(x, w, b) = \text{sign}(w \cdot x + b) \]

Any of these would be fine...

..but which is best?

Support vector machines: key ideas

How would you classify this data?

- denotes +1
- denotes -1

\[ f(x, w, b) = \text{sign}(w \cdot x + b) \]

How would you classify this data?

Misclassified to +1 class
Define the **margin** of a linear classifier as the width that the boundary could be increased by before hitting a datapoint.

The **maximum margin linear classifier** is the linear classifier with the maximum margin (maximum margin is equivalent to minimum \(1/\|w\|\))
Support vector machines

Soft margin problem

\[
\min_{w,b,\xi_1, \ldots, \xi_n} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi_i \\
\forall i, \begin{cases} 
\xi_i \geq 0 \\
\xi_i - 1 + y_i (w^T x_i + b) \geq 0 
\end{cases}
\]

Input space

Equivalent to dual problem

\[
W(\alpha) = -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j x_i^T x_j + \sum_{i=1}^{n} \alpha_i \\
\sum_{i=1}^{n} y_i \alpha_i = 0 \\
0 \leq \alpha_i \leq C \quad \text{for } i = 1, \ldots, n
\]

Feature space

Some of our recent results

- Improving prediction performance of CRFs (KDD’05, ACM Trans. ALIP’06)
- High-performance training of CRFs for large-scale applications (HPCS’06)
- Sentence reduction (in text summarization) by SVM (COLING’04)
- Model for emerging trend detection (PAKDD’06, KSSJ)
- Prediction and analysis of \(\beta\)-turns in protein structures (GIW’03, JBCB’05) and histone modifications by SVM (GIW’05) and CRFs (ICMLB’06)
- Simplifying support vector machines (ICML’05, IEEE Trans. Neural Network)
- Manifolds in imbalanced data learning (IEEE ICDM’06)
- Kernel matrix evaluation measure (IJCAI’07, submitted)
Prediction of $\beta$-turns and $\gamma$-turns by SVM

(GIW’03, JBCB’05)

Protein sequence

RPDFCLEPPYTGPKARI IRYFYNAKAGL
CQTFVYGGCRAKRNNFKSAEDCMRTCGGA

Predict by SVM

$\beta$-turns

RPDFCLEPPYTGPKARI IRYFYNAKAGL
nnnnnnnnnnnnnnnnnnnnnnnnnTTtttn
CQTFVYGGCRAKRNNFKSAEDCMRTCGGA
nnnnnnnnnnnnTTtttnnnnnnnnnnnnnnnnn

Histone modifications: Some amino acids of histone proteins (H3, H4, H2A, H2B) in nucleosomes are modified by added methyl group (methylation), acetyl group (acetylation), or other chemical groups.

146 pairs of DNA in nucleosomes are wrapped around a core of histone proteins.
Prediction of histone modifications in DNA
(GIW’05, BioMed Central 2006)

From DNA sequences

\[ \text{To computationally predict:} \]
- H3, H4 occupancy
- Acetylation state
- Methylation state

To find characteristics of areas at which H3, H4 occupancy, histone acetylation and methylation are at high and low levels.

The accuracy and correlation coefficient of qualitative prediction are consistent with experimental approach.

SVMs simplification
(ICML’05, IEEE Trans. Neural Networks 2006)

To replace original machine

\[ y = \text{sign} \left( \sum_{i=1}^{N_S} \alpha_i K(x_i, x) + b \right) \]

by a simplified machine

\[ y' = \text{sign} \left( \sum_{j=1}^{N_z} \beta_j K(z_j, x) + b \right) \quad (2) \]

with \( N_z < N_S \) \( (1) \) and \( (2) \) are similar

\( \{(z_j, \beta_j)\}_{j=1,...,N_z} \) – reduced vectors

Bottom-up approach that finds solution in a univariable function instead of multivariable ones in previous methods
Manifold for imbalanced data learning
(IEEE ICDM’06)

- Flexible assumption: Data having manifold structures.
- Up sampling data to make it exhibit manifold structures → give rise to patterns of interest.
- Our algorithms outperform SVMs and SMOTE (Chawla et al, JAIR’02).

Kernel matrix evaluation
(IJCAI 2007)

- Popular efficient measure of kernel matrix KTA (Kernel Target Alignment, Cristianini 2002) has fundamental limitations.
- A sufficient but not necessary condition.
- Proposed the new measure FSM (Feature Space-based Kernel Matrix Evaluation Measure) using the data distribution in the feature space that is efficient, having desirable properties.
- Implication of FSM is vast.

\[
KTA(K, y) = \frac{\langle K, y, y^T \rangle_F}{\sqrt{\langle K, K \rangle_F \langle y, y^T, y, y^T \rangle_F}}
\]

\[
FSM(K, y) = \frac{var_+ + var_-}{\|\phi_+ - \phi_-\|}
\]
Summary

- Data mining is a emerging interdisciplinary area with great interests from both research and industry.
- Many challenges in data mining, especially in mining complexly structured data.
- Kernel methods are a new emerging trend with mathematical foundations and high performance in solving hard problems of pattern analysis.

Thank you!