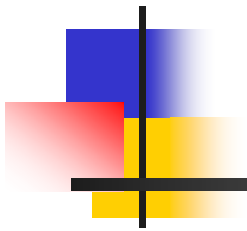


データマイニング環境整備の留意点とデファクトスタンダード

鷲尾隆
大阪大学産業科学研究所
2003/3/29
人工知能学会AIシンポジウム

問題：データ表現の現状

従来のデータマイニング：RDB対象が中心

表形式データ

天候	湿度	風速	Golf Play
晴れ	高い	強い	No
晴れ	普通	弱い	Yes
雨	高い	弱い	No
ー	ー	ー	ー

トランザクション

りんご	みかん	牛乳	パン	ー	ー	ー	ー	歯磨き粉
1	1	0	0	ー	ー	ー	ー	0
0	0	1	1	ー	ー	ー	ー	0
ー	ー	ー	ー	ー	ー	ー	ー	0

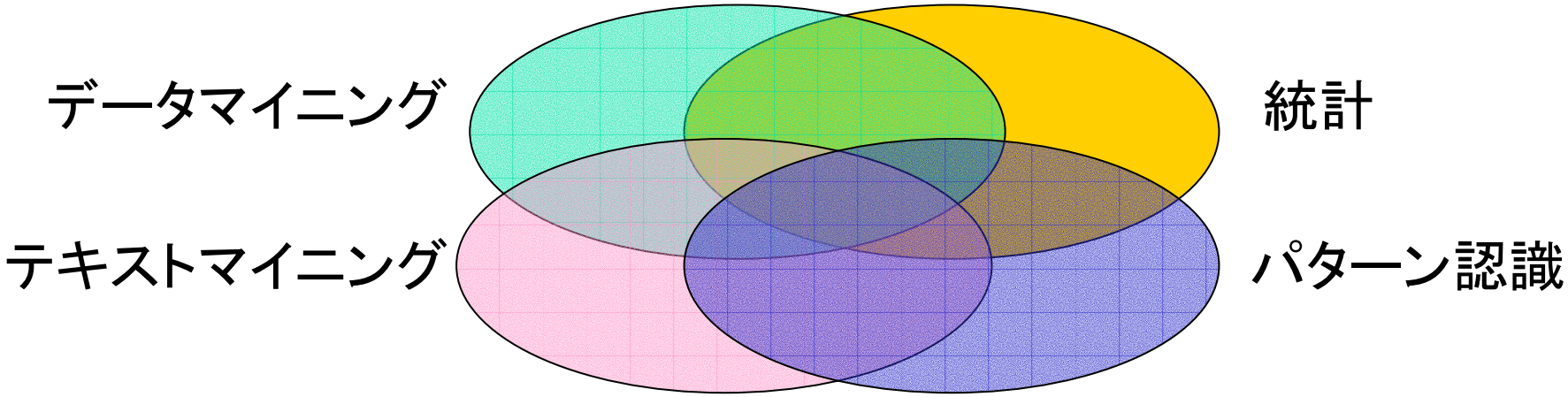
自然文テキスト，
画像や音声

固定された種類の特徴量に変換

平均照度	色成分比	ブロック1の明度変化率	など
54	0．7	1．2	ー
ー	ー	ー	ー

問題：関連分野の技術的重なり

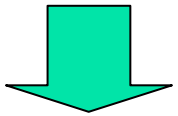
- データマイニング，テキストマイニング，パターン認識，統計解析など多分野で用いられる技術の重なり
  - パターン認識で用いられてきたサポートベクターマシン
    データマイニングでも用いられる．
  - 統計手法であるベイジアンネットがデータマイニングで用いられる．
  - デキストマイニングでもデータマイニング同様にナイーブベイズが用いられる．



いずれもデーター部分の特有な傾向把握，データ分類技術．
実適用の分野ではこれらを複合的に用いることが多い．

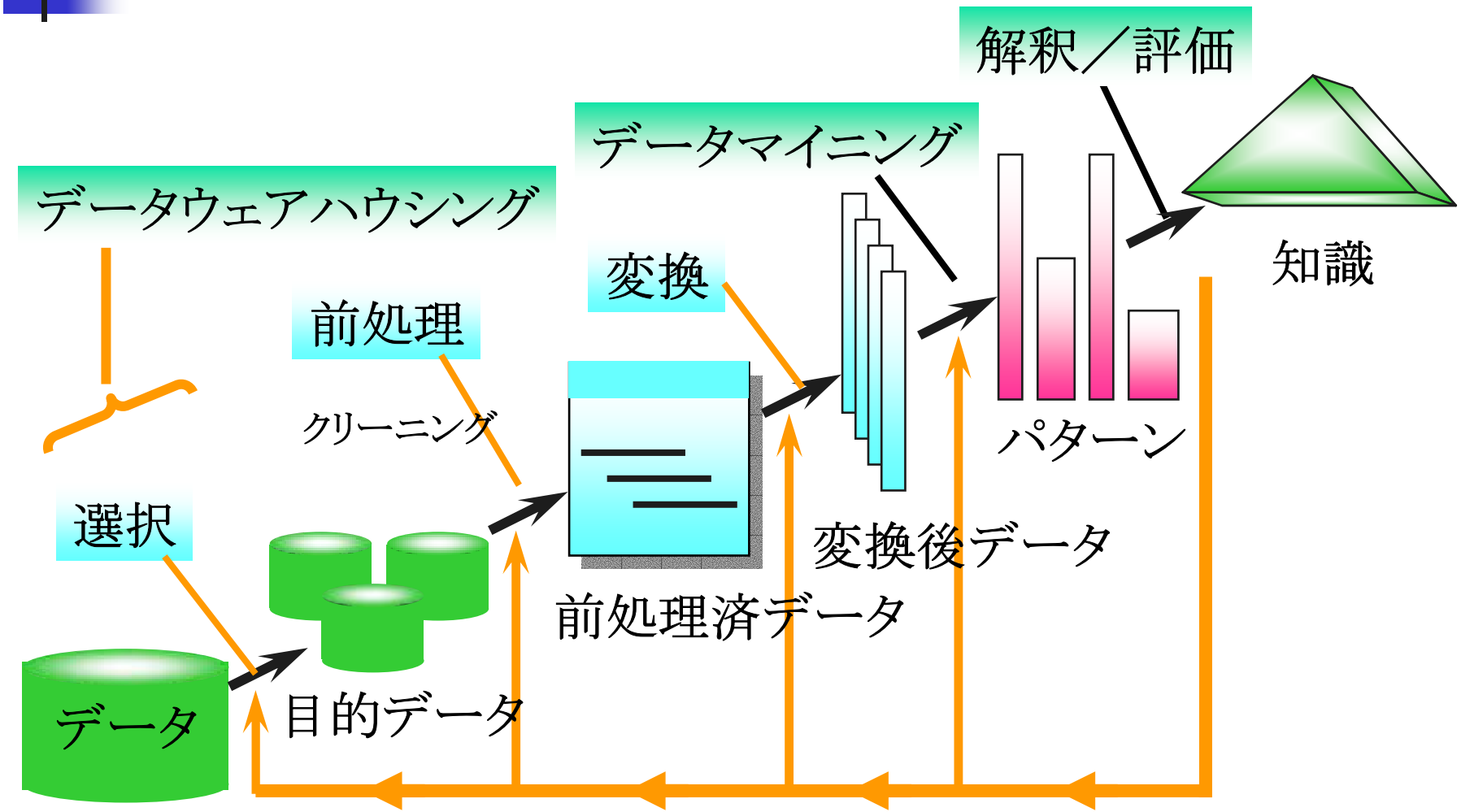
問題：データ表現と技術的重なり

- 問題の例
  - コンビニエンスストアのPOSレジの売り上げトランザクションデータは非常に疎な表になる．
  - 時系列やグラフなど複雑なデータ構造も，表形式では表しづらく，記憶容量効率が悪くなる．
  - どうしても画像とトランザクションデータを関連付けて解析せねばならない．



- 記憶容量効率の低下，データの読み込み・書き出し速度の低下など，膨大なデータを扱う際に深刻な問題
- 自然文や画像，音声を表形式の固定種類の特徴量に変換することによる情報の欠落が，マイニングの性能を劣化

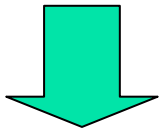
問題：複雑な処理



Fayyad(1996)が提案したデータマイニングの過程

問題：複雑な処理に伴う問題

- 中間段階毎にデータファイルが作成
- 探索的解析の結果として微妙に異なる似たような内容を持つデータファイルが各段階で多数蓄積
- 解析スキームに関するデータファイルや対応する解析結果ファイル，表示方式を表すファイルも多様



- 数ヶ月後の見直し，複数人による作業など現実的状況で，各ファイル内容やファイル相互関係の洗い出しに，多くの時間がかかる．
- 現状のデータマイニングツール環境では，解析内容が複雑化し解析回数が増えるほど，元データや中間ファイル，結果ファイルの一元的かつ体系的管理は難しくなる．

問題：手法の多様性

- 多様なマイニング手法（例）

決定木: ID3, C4.5, C5.0, CART, CHAID, QUEST, Pseudo Decision Tree, Option Tree, ファジイ決定木, 2次元領域の抽出決定木, 機能拡張決定木

分類器: サポートベクターマシン, K-nearest neighbor, 記憶ベース推論

相関ルール: Apriori, Generalized Rule Induction, 順序アソシエーション, 複数時系列, グラフマイニング

クラスタリング: コホーネン, K-means, Ward法, コンドルセ手法, 概念クラスタリング

関数ネット: 共分散構造分析, ニュラルネット(BP, MLP, RBF), ベイジアンネット

統計的手法: 重回帰分析, ロジスティック回帰分析, 判別分析, 主成分・因子分析

- 以上に加えて，欠測値・ノイズ処理など前処理，結果フィルタリング・結果表示などの後処理の多様性がある．

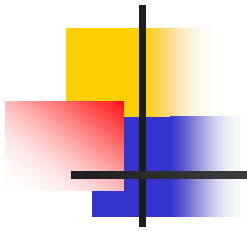
問題：ツールの多様性

- **現状の市販ツールベンダー（商用版）**
  SAS，SPSS，数理システム，スポットファイア，
  日本IBM，日本マイクロソフト，日本ユニシス，
  東芝，NEC，富士通，日立，日本SGI
- **オープンソースフリーウエア**
  - MUSASHI
    日本の大学関係者がコンソーシアムを組んで作成。情報系のみならず業務基幹系での使用も視野にいれ、膨大データの高速処理を行う。Linux OSのアークテクチャ上で構築。
  - Weka
    ニュージーランドワカイト大学のチームが開発。業務用ではなく、データマイニング研究者やデータマイニング試用時の中小規模検証向き。
  - R
    商用のS-Plusというデータマイニングソフトのフリーウエア版。中規模データまで扱える。データマイニングの教育向き。
- 学術的ツールを加えれば，100を軽く超える．

問題：手法とツールの多様性

- 各ツールには一長一短があり，すべての解析機能を網羅した万能なツールは現状存在しない．
- マイニングの目的によっては，複数のツールを適宜組み合わせて使用することが多い．

- ツール間のデータ引渡しや入出力データの形式互換性
  - 各ツールが独自のフォーマットを採用しているため，複数ツールを組み合わせるには，ツール間のデータ変換プログラムを逐一作成する必要
  - 作業効率を著しく低下
  - データマイニングの機能そのものの制約要因
    - あるツールが別のツールが必要とする情報を出力しないために可能な解析内容が限られるなど



背景

- 表形式データ中心主義による非効率性や情報欠落
- 複雑なデータの切り出しや選択，前処理，データ変換，解析結果の蓄積と表示
  - 中間段階や探索的解析途中での膨大なファイル生成
    ⇒ ファイル管理（内容や相互関係の把握）が困難
- 多様なデータマイニング手法により，中間・最終結果，データマイニングスキームデータも多様化
  - ファイル管理（内容や相互関係の把握）が困難
- データマイニングツールの種類の多さ
  （市販，オープン）
  - データ互換性がなく逐一変換プログラムが必要

問題解決

- 近年，データマイニングツール開発者の側から，データマイニング環境の共通化による問題解消ないしは軽減を図る努力が始められている．
- データファイルに関するソフトウエア工学上の留意点
  - 自己言及性：データファイルに内部に，そのファイルが含むフォーマット仕様やデータ仕様，作成履歴，説明などの情報が記述されており，ファイルの管理やプログラムによるファイル取り扱い方法の自動判定が容易に可能であること．
  - 柔軟性：データファイルの表現力が豊かであり，表形式，トランザクションはもちろん，時系列や木，グラフ，ルール，関数，クラスターなど，原データだけでなくマイニング結果データも柔軟に記述できること．
  - 拡張性：技術の発達に応じてより複雑なデータを表現可能なように容易に拡張可能であること．
  - 汎用性：オペレーティングシステムなどの計算機環境に依存せずに，幅広く利用可能なファイル形式に格納可能な表現であること．

データファイル形式：テキストファイル，XMLベース

ファイルサイズ，読み込み時間などは計算機性能向上でカバー可能

背景

- マイニングツールの製作者(ソフトウエア工学)の立場から,
  マイニングデータ形式の自己言及型標準化の動き
  - XMLデータベース研究の活発化
    ⇨ タグを柔軟に定義し種々のデータ表現が可能

⇩

- マイニング向けの表形式データ記述
  XML Table (MUSASHI)
- マイニング対象データや結果の
  XML表現世界標準規格PMML
  ⇨ データマイニングに関するあらゆるデータの表現を目途

PMML:Predictive Model Markup Language

PMMLの広がり

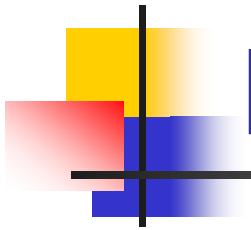
商用ツール

- DB2 Intelligent Miner (日本語対応):
  DB2の統合製品。PMML活用導入事例あり。
- SPSS Clementine 7.1 (日本語対応): Clementineの最新版。
- CART 5.0 Pro (日本語対応):
  決定木解析ツール。様々な拡張決定木解析が可能。
- OLE DB for Data Mining（英語のみ？）:
  MS SQL Server上のマイニングツール。
- SAS Enterprise Miner(英語のみ？): 英語版でPMML対応を確認。
- Oracle9i Data Mining (英語のみ): オラクルの9iのパーツの1つ。

オープンソースフリーウエア

- MUSASHI-1.0.3 (日本語対応)
  Linux･BSDベースのオープンソースのデータマイニングツール。
  決定木に関するPMMLに対応。

http://funahasi.com/pmml/ （舟橋氏のHPより）



PMMLとは

- XMLベースのデータマイニング用データ表現言語
  - 世界の主要なIT企業が企画作りに参加
    (SAS, SPSS, IBM, Oracle, MSなど)
  - XMLによる柔軟なデファクトスタンダード表現
    - データマイニング全体スキームやデータの流れの記述
    - 前処理変換の記述，統計解析過程・結果の表現
    - Taxonomyの指定記述
    - データやマイニング結果の表現
      Association Rules, Decision Trees,
      Center-Based and Distribution-Based Clustering,
      Regression, General Regression,
      Neural Networks, Naive Bayes, Sequences
- PMML開発は現在進行中
  - Verion 1.0から始まり現在Version 2.1がリースされ，
    更にVersion 3.0の作成が進められている.
  - SOURCEFORGE.NETというサイトのPublic Forumでオープン開発
    http://sourceforge.net/projects/pmml

PMMLはXMLのサブセット

- XML(PMML)は各タグを定義するスキーマ部，実際のデータ部から構成，記述される
- PMMLタグを定義するXMLスキーマ

```
<xs:element name='PMML'>
 <xs:complexType>
  <xs:sequence>
   <xs:element ref='Header'/>
   <xs:element ref='MiningBuildTask' minOccurs='0' maxOccurs='1'/>
   <xs:element ref='DataDictionary'/>
   <xs:element ref='TransformationDictionary' minOccurs='0' maxOccurs='1'/>
   <xs:sequence minOccurs='0' maxOccurs='unbounded'>
    <xs:choice>
     <xs:element ref='TreeModel'/> <xs:element ref='NeuralNetwork'/>
     <xs:element ref='ClusteringModel'/> <xs:element ref='RegressionModel'/>
     <xs:element ref='GeneralRegressionModel'/> <xs:element ref='NaiveBayesModel'/>
     <xs:element ref='AssociationModel'/> <xs:element ref='SequenceModel'/>
    </xs:choice>
   </xs:sequence>
   <xs:element ref='Extension' minOccurs='0' maxOccurs='unbounded'/>
  </xs:sequence>
  <xs:attribute name='version' type='xs:string' use='required'/>
 </xs:complexType>
</xs:element>
```

PMMLはXMLのサブセット

- マイニング結果モデル表現スキーマ

```
<xs:element name="ExampleModel">
 <xs:complexType>
  <xs:sequence>
   <xs:element ref="Extension" minOccurs="0" maxOccurs="unbounded"/>
   <xs:element ref="MiningSchema"/>
   <xs:element ref="ModelStats"minOccurs="0" maxOccurs="1"/>
   ...
   <xs:element ref="Extension" minOccurs="0" maxOccurs="unbounded"/>
  </xs:sequence>
  <xs:attribute name="modelName" type="xs:string" use="optional"/>
  <xs:attribute name="functionName" type="MINING-FUNCTION" use="required"/>
  <xs:attribute name="algorithmName" type="xs:string" use="optional"/>
 </xs:complexType>
</xs:element>
```

PMMLはXMLのサブセット

- データタイプスキーマ
  - 一般数値タイプ，整数値タイプ，高精度実数値タイプ，確率値タイプ，百分率タイプ，具体的な属性名を指定して値を定義するタイプなど

高精度実数値タイプ

```
<xs:simpleType name="REAL-NUMBER">
  <xs:restriction base="xs:double">
  </xs:restriction>
 </xs:simpleType>
```

配列タイプ

```
<Array n="3" type="int">
  1 22 3
</Array>
<Array n="3" type="string">
  ab  "a b"   "with ¥"quotes¥" "
</Array>
```

PMMLによる実データ記述

- アソシエーションルール記述の例(1)

```
<?xml version="1.0" ?>
<PMML version="2.1" >
<Header copyright=www.dmg.org
  description= "example model for association rules"/>
<DataDictionary numberOfFields="2" >
<DataField name="transaction" optype="categorical" />
<DataField name="item" optype="categorical" />
</DataDictionary>
<AssociationModel functionName="associationRules“
  numberOfTransactions="4" numberOfItems="3"
  minimumSupport="0.6" minimumConfidence="0.5“
  numberOfItemsets="3" numberOfRules="2">
  <MiningSchema>
    <MiningField name="transaction"/>
    <MiningField name="item"/>
  </MiningSchema>
```

つづく

PMMLによる実データ記述

- アソシエーションルール記述の例(2)

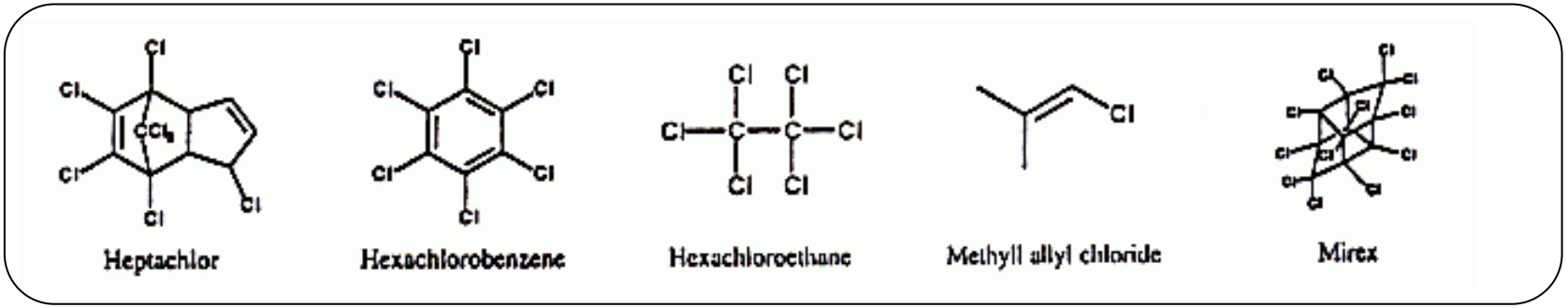
```
<!-- We have three items in our input data -->
<Item id="1" value="Cracker" />
<Item id="2" value="Coke" />
<Item id="3" value="Water" />
<!-- and two frequent itemsets  with a single item -->
<Itemset id="1" support="1.0" numberOfItems="1">
  <ItemRef itemRef="1" />
</Itemset>
<Itemset id="2" support="1.0" numberOfItems="1">
  <ItemRef itemRef="3" />
</Itemset>
<!-- and one frequent itemset with two items. -->
<Itemset id="3" support="1.0" numberOfItems="2">
  <ItemRef itemRef="1" />
  <ItemRef itemRef="3" />
</Itemset>
<!-- Two rules satisfy the requirements -->
<AssociationRule support="1.0" confidence="1.0"
  antecedent="1" consequent="2" />
<AssociationRule support="1.0" confidence="1.0"
  antecedent="2" consequent="1" />
</AssociationModel>
</PMML>
```

グラフ構造表現のPMML提案

光永悠紀,鷲尾隆,藤本敦,元田浩
大阪大学産業科学研究所

背景：我々のメイン研究テーマ

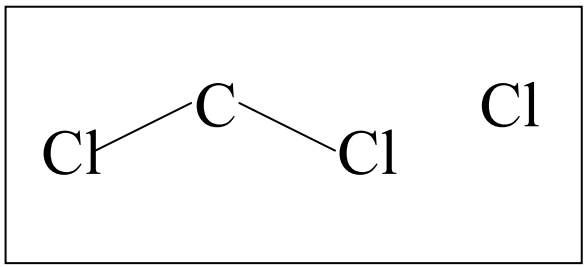
グラフ構造データマイニングでも同様の困難が存在



マイニングツール専用フォーマットのデータ

グラフマイニング実行

ルール抽出



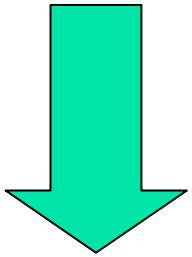
問題点

データ表現の標準規格がないため、各マイニングツール間でのデータの変換が必要

我々に研究紹介

現在マイニングデータのXML表現の世界標準規格としてPMMLの開発が進められている。

グラフ構造表現に関する標準規格は開発されていない。



グラフ構造表現可能なPMML規格を提案する。

グラフ構造表現のPMML提案

グラフモデルスキーマの定義

```
<?xml version="1.0" encoding="EUC-JP" ?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"
 targetNamespace="http://www.dmg.org/PMML-2_1"
 xmlns=http://www.dmg.org/PMML-2_1
 elementFormDefault="unqualified">
<xs:element name="GraphModel">
 <xs:complexType>
 <xs:sequence>
  <xs:element ref="Extension" minOccurs="0" maxOccurs="unbounded"/>
  <xs:element ref="MiningSchema"/>
  <xs:element ref="ModelStats" minOccurs="0" maxOccurs="1"/>
  <xs:element ref="Graph"/>
  <xs:element ref="Extension" minOccurs="0"  maxOccurs="unbounded"/>
 </xs:sequence>
 <xs:attribute name="modelName" type="xs:string" use="optional"/>
 <xs:attribute name="functionName" type="MINING-FUNCTION"  use="required"/>
 <xs:attribute name="algorithmName"  type="xs:string" use="optional"/>
 </xs:complexType>
</xs:element>
```

グラフ構造表現のPMML提案

グラフスキーマの定義

```
<xs:element name="Graph">
 <xs:complexType>
  <xs:sequence>
   <xs:element ref="Extension" minOccurs="0"  maxOccurs="unbounded"/>
   <xs:element ref="Vertex" minOccurs="0"  maxOccurs="unbounded"/>
   <xs:element ref="Edge" minOccurs="0" maxOccurs="unbounded"/>
  </xs:sequence>
  <xs:attribute name="graphId"  type="xs:string" use="optional"/>
  <xs:attribute name="graphType"  type="GRAPH-TYPE" use="optional"/>
  <xs:attribute name=“modelType"  type="MODEL-TYPE" use="optional"/>
  <xs:attribute name="recordCount"  type="NUMBER" use="optional"/>
 </xs:complexType>
</xs:element>
```

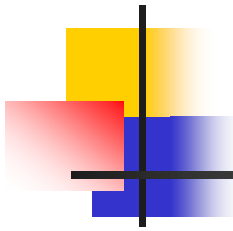
グラフ構造表現のPMML提案

（サブ）グラフのタイプを指定するタグの定義

```
<xs:simpleType name="GRAPH-TYPE">
 <xs:restriction base="string">
  <xs:enumeration value="original"/>
  <xs:enumeration value="induced"/>
  <xs:enumeration value="general"/>
 </xs:restriction>
</xs:simpleType>
```

マイニング結果の構造タイプを指定するタグの定義

```
<xs:simpleType name="MODEL-TYPE">
  <xs:restriction base="string">
   <xs:enumeration value="unRootedTree"/>
   <xs:enumeration value="rootedTree"/>
   <xs:enumeration value="orderedTree"/>
   <xs:enumeration value="path"/>
   <xs:enumeration value="graph"/>
  </xs:restriction>
</xs:simpleType>
```



グラフ構造表現のPMML提案

Vertexのタグ定義

```
<xs:element name="Vertex">
 <xs:complexType>
 <xs:sequence>
  <xs:element ref="VertexLabel" minOccurs="0" maxOccurs="unbounded"/>
 </xs:sequence>
  <xs:attribute name="vertexId" type="xs:string" use="required"/>
  <xs:attribute name="dimension" type="xs:int"/>
 </xs:complexType>
</xs:element>
```

グラフ構造表現のPMML提案

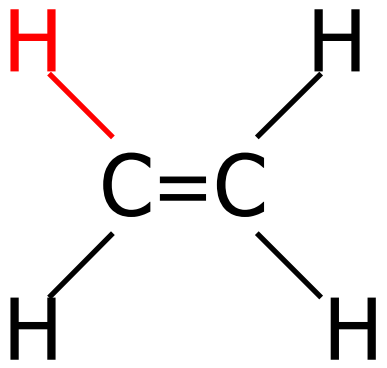
Edgeのタグ定義

```
<xs:element name="Edge">
 <xs:complexType>
  <xs:sequence>
   <xs:element ref="EdgeLabel" minOccurs="0" maxOccurs="unbounded"/>
  </xs:sequence>
  <xs:attribute name="edgeId" type="xs:string" use="required"/>
  <xs:attribute name="graphtype" type="EDGE-TYPE" default="undirected"/>
  <xs:attribute name="demension" type="xs:int"/>
  <xs:attribute name="bgnvertexid" type="xs:string"/>
  <xs:attribute name="endvertexid" type="xs:string"/>
 </xs:complexType>
</xs:element>
```

PMML記述例(エチレン)

開発したPMMLスキーマタグによるグラフ表現例

```
<?xml version="1.0" encoding="EUC-JP"?>
<PMML version="2.1" >
...
 <GraphModel modelName="sample">
  <Graph graphId="1">
   <Vertex vertexId="1" dimension="1">
    <VertexLabel field="atomy" value="H"/>
   </Vertex>
...
   <Edge edgeId="1" graphtype="undirected" dimension="1"
     bgnvirtexid="1" endvirtexid="3"/>
    <EdgeLabel field="bondtype" value="singlebond" />
   </Edge>
...
  </Graph>
...
 </GraphModel>
</PMML>
```



グラフマイニングツールへの実装

AGM形式データ

変換パーサー

PMMLデータ

変換パーサー

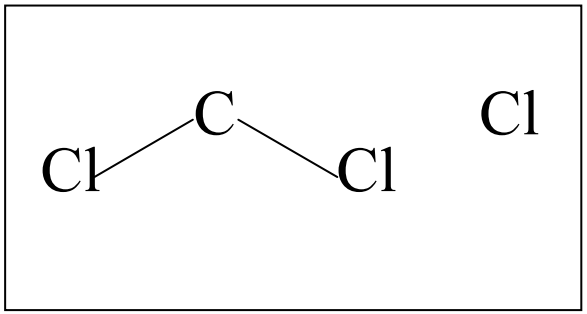
SDF形式
化学物質データ

グラフマイニングツール
AGM

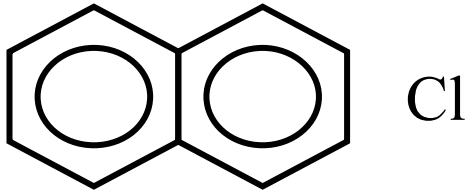
マイニング結果

PMMLデータ

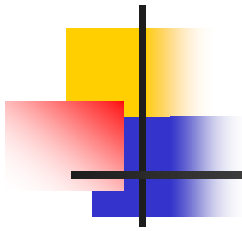
マイニング結果例



発ガン性あり



発ガン性あり



おわりに

- データマイニング環境整備上の課題と留意点について考察
- 自己言及性，高い表現能力と柔軟性，拡張性，汎用性の高いデータファイルのデファクトスタンダード規格の重要性
- 現在最有力候補であるPMLLを紹介
- PMML規格に則ってより高度なデータ構造であるグラフ構造の表現規格を構成した筆者等の研究を紹介

我々の課題

- グラフマイニングツールへのPMMLインターフェース実装
- ツールのMUSASHIへの実装
- 提案したグラフ表現規格の改良と拡張
- PMML Data Mining Group (DMG)への提案