- SAS共同企画セッション -

IMSTATに対する期待と現実，そしてアソシエーション分析，ときどきレコメンデーションシステム

塩野義製薬株式会社

藤原 正和，北西 由武，都地 昭夫，渡辺 秀章

- SAS joint planning session -

Consideration for IMSTAT by applying association analysis and recommendation system

Shionogi & Co., Ltd.

Masakazu Fujiwara, Yoshitake Kitanishi, Akio Tsuji, Hideaki Watanabe

要旨：

アソシエーション分析は，項目間の関連性を評価するために，主にマーケット分野を中心として古くから用いられているが，適用を工夫することで，様々な場面で有用であると考えられる．IMSTATに備わっているARM (Associative Rule Mining) ステートメントを用いることで，アソシエーション分析がSASで実装可能となった．本発表は，IMSTATの機能紹介，及びアソシエーション分析の理論的背景の説明に加えて，製薬業界における適用例として，医薬品副作用データベースであるJADER (Japanese Adverse Drug Event Report database) を適用して発現した副作用間の関連性評価を行った事例を紹介する．また，アソシエーション分析の考えをもとにした，レコメンデーションシステムについても同様に紹介する．

キーワード：IMSTAT，アソシエーション分析，ARMステートメント，Japanese Adverse Drug Event Report database

発表内容

1. ビッグデータと製薬企業，IMSTATの機能紹介
2. アソシエーション分析
3. JADERに対する適用結果
   - ARMステートメント，proc recommendの実装事例

1. ビッグデータと製薬企業，IMSTATの機能紹介

ビッグデータと製薬企業

Signal Detection
- 医薬品副作用データベース（JADER/FAERS）

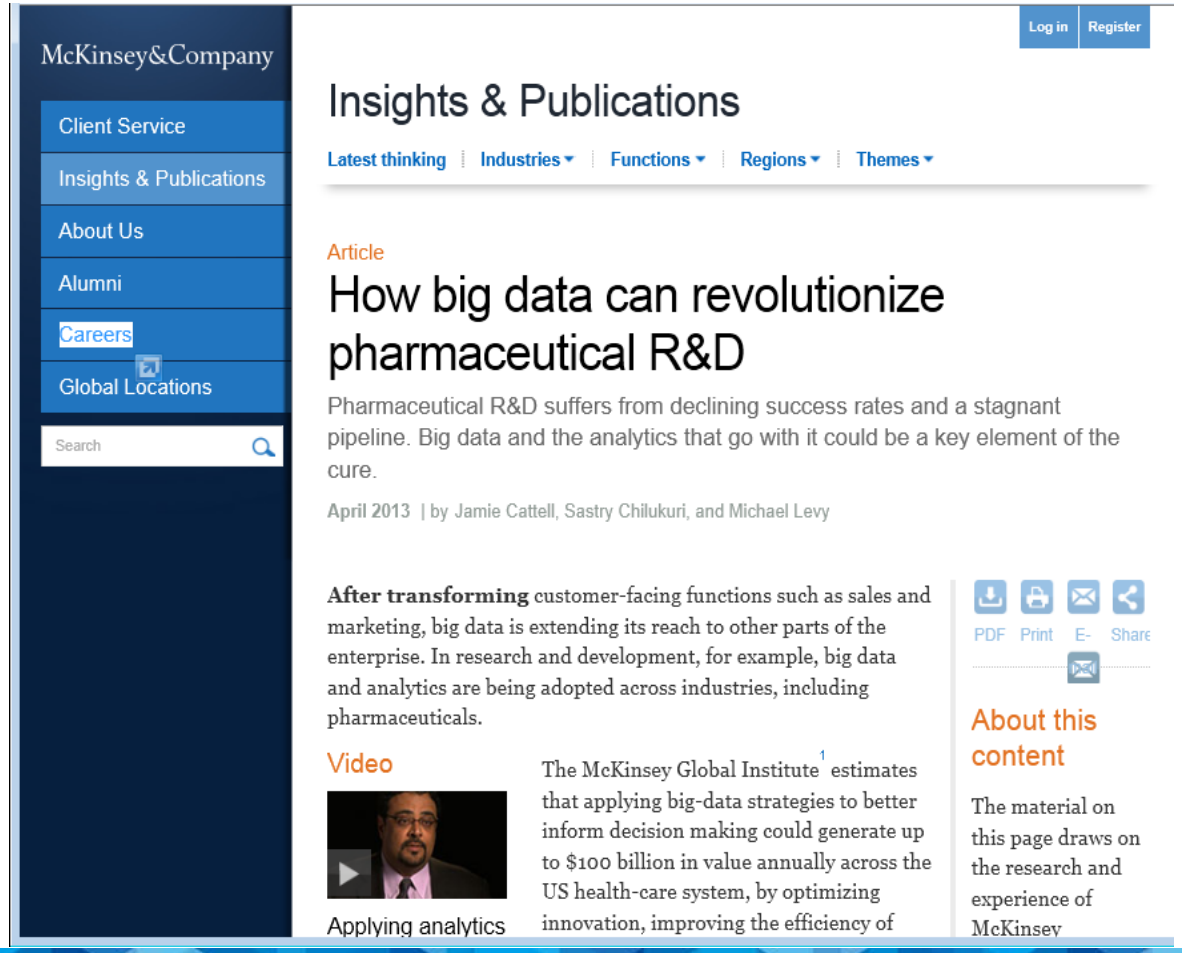
Chemoinformatics
- QSAR

Bioinformatics
- Pharmacogenomics
- System Biology

Epidemiology
- 医療DB

Text Mining
- 添付文書解析
- PubMed
- SNS解析
- 当局HP解析

Sales&Marketing



Big data解析のあるべき姿とSASプロダクト

- BigDataで機械学習を高速化.
- Data Scientistを量産.

Data Scientistの主要ツールを目指す.

IMSTAT (機械学習)

- 他DBとの接続を容易にし, データ収集を効率化.

Data Loader

社内外のデータを集めたData Hubの構築を目指す.

IMSTAT (base SAS)

演算性能を活かした活用を促す.

- 演算速度(性能)の増強.

Visual Analytics

Word CloudやPath解析等の新奇的な可視化を浸透させる.

- 機械学習の結果をシームレスに可視化.
- Spotfireとは違った視点での可視化を実現.

IMSTAT : In-Memory Statistics for Hadoop

探索＆記述統計

- 要約統計量 (summary)
- 集計表
- クロス集計表
- アソシエーション
- 相関係数
- 箱ひげ図(Box Plot)
- ヒストグラム
- パーセンタイル

など

モデリング ＆ 最適化

- 線形回帰
- 重回帰
- ロジスティック回帰
- 一般化線形モデル
- 一般線形モデル
- ディシジョンツリー
- ランダムフォレスト
- クラスタリング(k-means)
- クラスタリング(DBSCAN)
- 時系列予測
- 時系列予測(ゴール・シーク)
- 特異値分解（SVD）
- カーネル密度推定(KDE)
- ニューラルネットワーク

レコメンド計算

- 協調フィルタリング(KNN)
- アソシエーションルールタイプ
- 特異値分解(SVD)タイプ
- 回帰タイプ
- クラスタリングタイプ

テキスト解析

- 形態素解析
- クレンジング(ステミング,辞書)
- 語,文書の頻度
- 特異値分解（SVD)
- エンティティ抽出,トピック生成

SAS In-Memory Statistics for Hadoopライセンスが必要

本スライドはSAS社より提供いただいた

2. アソシエーション分析

アソシエーション分析適用のモチベーション

安全性に基づいた
医薬品の適切な使用

医薬品使用により
発現した副作用の評価

例えば，副作用を個別に評価するだけでなく，複数の副作用間の関連性について適切に考察することが必要

安全性に基づいた
医薬品の適切な使用

医薬品使用により
発現した副作用の評価

副作用情報の
関連性評価

背景情報,
薬剤情報を含む評価

目的：IMSTATを用いてアソシエーション分析をJADERへ適用して，これらの適用可能性を評価すること

・JADER (Japanese Adverse Drug Event Report) データベースの利用

・IMSTATを用いて，関連性を評価するためのアソシエーション分析の適用

アソシエーション分析

- 関連性を評価するためのデータマイニングの代表的な方法
  - ある関連ルールを満たすデータの組み合わせを抽出することが可能
    - Aprioriアルゴリズム

- 関連ルール
  - 副作用 X が発現した場合に，副作用 Y も発現すること
  - $X \Rightarrow Y$ で表す

Agrawal et al., 1993

関連ルールの指標 (Support, Confidence)

・データ

ID	副作用
1	幻覚，不安
2	転倒，発熱，悪寒
3	横紋筋融解症，発熱
4	倦怠感，悪寒，発熱
5	白血球数減少，悪寒

X:発熱

Y: 悪寒	有	無	合計
有	2	1	3
無	1	1	2
合計	3	2	5

$$\text{Support}(X \Rightarrow Y) = \frac{(X \Rightarrow Y\text{の例数})}{(\text{全例数})} = \frac{2}{5}$$

Support (支持度)
・全体に対する関連ルールの発現割合

$$\text{Confidence}(X \Rightarrow Y) = \frac{(X \Rightarrow Y\text{の例数})}{(X\text{有の例数})} = \frac{2}{3}$$

Confidence (確信度)
・副作用X が起きた場合における関連ルール X⇒Y の発現割合

Agrawal et al., 1993

Support(X⇒Y), Confidence(X⇒Y)の考察

- 値が高いほど関連性が強いと判断される
- 関連性の評価にはConfidenceだけでは不十分
  - X有が1例で，それがX有, Y有であったときにConfidenceは1になってしまうため，Supportと共に考えていく必要がある

- ただし，Support，Confidenceの値が高い場合でも，意味のないルールが抽出されてしまうことがある

例

X:発熱

Y: 悪寒

	有	無	合計
有	a	c	Y
無	b	d	N-Y
合計	X	N-X	N

Y/N の値が高い場合に，
a/X，a/Nの値も大きくなることがある

つまり，発熱に関係なく，悪寒を発現した例数が多い場合にも，Support(X⇒Y), Confidence(X⇒Y)は高い値を取ってしまう

関連ルールのその他指標

- 様々な指標が提案されている
  - Lift
  - Conviction
  - Bayes factor
  - information gain
  - implication index
  - least contradiction
  - truncated entropic intensity of implication
  - Loevinger など

本発表ではこの2指標を取り上げる

Lenca, P., Myer, P., Vaillant, B., Lallich, S. (2008)

関連ルールの指標(Lift, Conviction)

・データ

ID	副作用
1	幻覚，不安
2	転倒，発熱，悪寒
3	横紋筋融解症，発熱
4	倦怠感，悪寒，発熱
5	白血球数減少，悪寒

X:発熱

Y: 悪寒	有	無	合計
有	2	1	3
無	1	1	2
合計	3	2	5

$$\text{Lift}(X \Rightarrow Y) = \frac{\text{X⇒Yの確信度}}{\text{(Y有の例数)}/\text{(全例数)}} = \frac{(2/3)}{(3/5)}$$

$$\text{Conviction}(X \Rightarrow Y) = \frac{\text{(Y無の例数)}/\text{(全例数)}}{\text{X⇒Y無の確信度}} = \frac{(2/5)}{(1/3)}$$

Lift
・副作用Xの発現によって，副作用Yの発現がどの程度増加したかを示す指標

Conviction
・副作用Xの発現によって，副作用Yの発現しない割合が，どの程度減ったかを表す指標

Lift(X⇒Y), Conviction(X⇒Y)の考察

- 1以上であり，かつ値が高いほど関連性が強いと判断される

- LiftはLift(X⇒Y)＝ Lift(Y⇒X)という性質があるが，Convictionは成立しない
  - X⇒Y，Y⇒Xの違いについて考察したいときはLiftではなく，Convictionの使用が適切と考えられる

複数の副作用間の関連ルールについて

・データ

ID	副作用
1	幻覚，不安
2	転倒，発熱，悪寒
3	横紋筋融解症，発熱
4	倦怠感，悪寒，発熱
5	白血球数減少，悪寒，転倒

X:発熱，悪寒

Y: 転倒	有	無	合計
有	1	1	2
無	1	2	3
合計	2	3	5

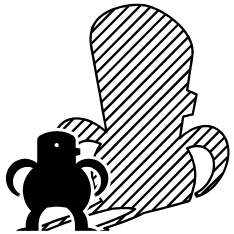
・Support(X⇒Y)，Confidence(X⇒Y), Lift(X⇒Y), Conviction(X⇒Y) は同様に計算可能

・X，Yを上記のように定義することで，複数の事象間の関連ルールについても考察が可能となる

3. JADERに対する適用結果
ARMステートメント,
PROC RECOMMENDの紹介

JADER

- JADER (Japanese Adverse Drug Event Report)
  - 医薬品医療機器総合機構 (PMDA) から公開されている医薬品副作用データベース
- データ内容 (csvファイル)
  - 症例一覧テーブル (性別，年齢，体重など)
  - 副作用テーブル (有害事象名，転帰，発現日など)
  - 医薬品情報テーブル (医薬品一般名，医薬品の関与(被疑薬，併用薬)，投与量など)
  - 原疾患テーブル (原疾患名など)
- 2015年5月時のJADERを使用
  - 約34万症例，副作用約54万件，医薬品情報テーブル約230万obs



IMSTATによるアソシエーション分析

- 抗うつ薬 に対して報告されている副作用を対象とする
  - 約5700症例，副作用 約10000件
- armステートメントを使用する

```
proc imstat data=LASRLIB.join2;
      arm item=PT_KANJI tran=ISR
         /  minItems=1 maxItems=5
itemsTbl
          support(LOWER=15)
          rules(confidence(LOWER=0.9)
                  numrhs(upper=5
lower=1) numlhs(upper=5 lower=1))
rulesTbl;
run;
```

X, Yのアイテム数の指定やSupport, Confidenceの値による出力の制御が可能

結果をSASデータセットで取ってくることが可能

JADERに対する適用結果

当日紹介します

応用：レコメンドシステム

被験者ID	AE_ID	AE_COUNT
000001	0043	1
000001	0044	2
000001	0002	1
000002	0043	1
:		:

rating

AE_ID	AE_NAME
0043	肝炎
0044	傾眠
0045	嘔吐
0046	腎障害
:	:

ae

レコメンドシステムの開発

demo

被験者ID	性別	年齢	薬剤Aの使用
000001	M	25	有
000002	F	34	有
:	:	:	:

アソシエーション分析の考え方を利用

応用：レコメンドシステム (proc recommend)

```
proc recommend  recom = LASRLIB.movielens;
   add LASRLIB.movielens/
   item = PT_CODE user = isr rating = count;
 addtable LASRLIB.rating  /type = rating
       vars=(isr PT_CODE PT_KANJI  count);
 addtable LASRLIB.ae       /type = item;
 addtable LASRLIB.demo   /type = user;
run;
 method knn / label = "knn" k = 20 positive
 similarity = pc seed = 1234;
run;
 predict / method = knn  label="knn" Num = 5
              users = ("000001“~"0XXXX");
run;
```

3つのデータセットを指定する

類似している被験者を探すアルゴリズム指定

ある被験者集団で起こる可能性が高い有害事象を5つピックアップする
(高齢者，女性など)

まとめ

- IMSTAT，アソシエーション分析を用いることで，大規模な副作用情報の中から，関連ルールを抽出することが可能となり，アソシエーション分析のJADERへの適用は有益であると考えられた
  - ただし，データベースの性質を把握しておく必要があり，解釈には注意が必要である
- また，アソシエーション分析の考え方は，レコメンドシステムへ応用することが可能であると考えられた

- 適用を工夫することで，アソシエーション分析は様々な場面で適用可能であると考えられた
- そのツールとして，IMSTATに備わっているARMステートメント，及びrecommendプロシジャは有用であると考えられた

参考文献

- Agrawal, R., Imielinski, T., Swami, A., 1993. Mining association rules between sets of items in large databases. In: Buneman, P., Jajodia, S. (Eds.), Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data. Washington, D.C., pp. 207–216.
- Philippe Lenca, Patrick Meyer, Benoît Vaillant, Stéphane Lallich. On selecting interestingness measures for association rules: User oriented description and multiple criteria decision aid. European Journal of Operational Research, 184(2):610-626, 2008.