3F2-03

類似度に基づくVisual Data Mining – 近未来チャレンジ2002の評価シートへの適用 –

Similarity-based Visual Data Mining and its Application to a Dataset on AI Challenge

津本周作[*1] 安田 晃[*1] 笠原 要[*2] 西浦敬信[*3] 額賀信尾[*4]
Shusaku Tsumoto Akira Yasuda Akira Yasuda Akira Yasuda Akira Yasuda

[*1]島根医科大学医療情報学講座
Department of Medical Informatics, Shimane Medical University

[*2]日本電信電話 (株) NTT コミュニケーション科学基礎研究所
NTT Communication Science Labs.

[*4]和歌山大学システム工学部
Faculty of System Engineering, Wakayama University

[*5]日立製作所中央研究所
Central Research Laboratory, Hitach Ltd.

Visual data mining is a growing area in data mining, which supports the data mining process of users via visual information. Especially, it has been reported that pattern generation methods by using similarities are one of the methods of (visual) data mining are useful for knowledge discovery, such as clustering, because patterns obtained show the similarity relations between samples. In this paper, we propose the combination of similarity-based methods to characterize a given dataset. For the combination, clustering (ward-method), correspondence analysis and multidimensional scaling (MDS) were adopted due to the following reasons. First, clustering forms groups with limited variabilities, which suggests that each group corresponds to a concept represented by each record. Correspondence analysis gives the similarity of records from the viewpoint of their response. Finally, MDS shows the arrangement of each sample on the two dimensional plane from the viewpoint of similarities. Thus, the combination is expected to discover knowledge about similarities from different viewpoints (multi strategy). We applied these three methods to the dataset on questionaire of AI challenge in JSAI-2002, which was originally used for the evaluation of competitions. The empirical results show that clustering and MDS gave interesting results, corresponding to the reaction of audiences in each challenge session.

1. はじめに

データに潜む背景構造を解析する手段として，データの視覚化は有効な手段であり，Visual Data Mining[VDM 02] としてデータマイニングにおいても大きく取り上げられるようになってきた．視覚化は，適用領域の知識をうまく取り入れることで，非常に有効な知識が発見できる可能性が高い．それをプロセスとして考察し，最も有効に技術化している分野としてチャンス発見をあげることができる [Osawa 2003]．また，アクティブユーザーリアクションという観点からも，視覚化の手法を人工知能の立場から探求していくことはきわめて重要である．

Visual Data Mining への期待は高いが，これまでの研究では，一般的でかつ有効な視覚化の獲得には，積極的な Domain Expert の介入が不可欠で，汎用の手段を開発を行うことは難しいとされており，単にナイーブな手法の開発では不十分で，今後，視覚化そのものとして，あるいはユーザーとのインターフェイスとしてより深く性質を調べていかなければならない．

一方，計量心理学の領域では，アンケートをはじめとした心理的な背景を探るためのデータの特性を視覚化する目的で，類似性を元にしたクラスタリング，多重尺度構成法 (Multidimensional Scaling: MDS) 等の視覚化の方法が開発されてきた．これらの方法は回答の値から類似度を計算し，それに関するパターンを視覚化する．例えば，クラスタリングでは類似度の近い値の回答についてグループを構成し，そのグループ間の類似性を視覚化する．一方，MDS では各回答の類似度から二次元平面上への各回答の付置を求め，各回答の類似性を二次元的に視覚化する．このように，異なる手法は，異なる観点から類似性に着目しており，視覚化された情報からは，異なる観点に関する情報を抽出することが期待できる．更に，これらの類似度による情報は，従来の条件付確率をベースとした分類知識の手法とはパターンの抽出方法が異なるため，分類知識と違う観点からのパターンを生成され，類似度の手法と分類知識の抽出法とを組み合わせることで，類似度情報の視覚化を経て，その背景に潜む分類知識から背景知識を効率的に発見できることが期待できる．

本論文では，類似度による視覚化と分類知識の抽出法とを組み合わせることで，効率的に知識を発見するプロセスについて考察し，実際のアンケートデータを利用し，このプロセスの有効性を検証した．

2. 方法

2.1 データ集合

データ集合としては，昨年度人工知能学会全国大会の近未来チャレンジの Evaluation Form に対する聴衆の回答を利用した．昨年度の AI チャレンジは表 1, 2 に示した 8 項目からなる．表 1 は昨年度からの継続課題である Survival of Challenge，表 2 は来年度からのチャレンジ参画を希望とするプレゼンテーションからなる New Challenge である． Evaluation Form は表 3 に示したごとく，3 つの質問からなり，チャレンジの実現性，貢献度についての質問から構成されて，得点は 5 段階評価とした．

連絡先: 津本周作，島根医科大学医療情報学講座，出雲市塩冶町 89-1，Phone: 0853-20-2172，FAX: 0853-20-2170，Email: tsumoto@computer.org

表 3: アンケート項目

Q1	この発表者のチャレンジは 5 年以内に達成できると思いますか。 (1) 達成は当分不可能である。 (2) 達成が当分困難である。 (3) 5 年以内の達成は困難だが、もう少し時間があれば達成できる。 (4) 5 年以内の達成の可能性は高い。 (5) 5 年以内に達成できる。
Q2	このチャレンジが成功すれば、社会に貢献すると思いますか。 (1) 全く貢献しない。 (2) 貢献の度合が小さい。 (3) 従来の人工知能研究の平均程度に貢献する。 (4) 大きく貢献する。 (5) 非常に大きく貢献する。
Q3	このチャレンジが成功すれば、人工知能に貢献すると思いますか。 (1) 全く貢献しない。 (2) 貢献の度合が小さい。 (3) 従来の人工知能研究の平均程度に貢献する。 (4) 大きく貢献する。 (5) 非常に大きく貢献する。

表 1: サバイバル・オブ・チャレンジ一覧

略称	チャレンジ名 , 提案者 (所属)
DSIU	ネット情報を使った意思決定支援 藤本 和則 (FRP)
RCT	高齢者・障害者の自立的移動を支援する Robotic Communication Terminals 矢入 郁子 (通信総合研究所) 他
危機	危機管理シミュレーションとその分析 石田 亨，福本 理人 (京都大学), 松原 仁 (公立はこだて未来大学) 他
LOS	日常言語コンピューティング 岩爪 道昭 (理化学研究所) 他

2.2 質問の独立性:属性選択

各チャレンジにおける Q1, Q2, Q3 の独立性はラフ集合による縮約生成 [Pawlak 91] と Fisher の直接確率法 [Fisher 34] を用いて解析した.

2.3 類似度による視覚化

類似度の視覚化については，Correspondence Analysis, クラスタリング, MDS を用いた．クラスタリングはセミパーシャル相関を用いる Ward 法 [Everitt 01]，MDS はアンケートの得点を数値として扱い，各点間の距離を算出，距離行列から内積行列を求め，特異値分解による空間への付置を与える Torgerson の方法 [Torgerson 52] を用いた．特に，MDS に関しては評点 1,2,3,4,5 について二値化するという前処理を行った．例えば，評点 3 についての場合，評点 3 は 1，それ以外は 0 と置き換えて，距離計算を行った.

3. 結果および考察

それぞれの回答数は，DSIU: 26, RCT: 34, 危機: 14, LOS: 27, 3B-01: 31, 3B-02: 32, 3B-03: 32, 3B-04: 32 であった.

3.1 独立性の解析

Fisher の直接検定の結果を表 4 に示した．直接検定では，p–値を直接計算するので，それぞれのセルの値が独立性の高さを示している．したがって，5%の危険率では，DSIU では Q2,Q3 に相関，RCT では，Q1,Q2,Q3 相互に相関，危機，LOS ではすべての質問が独立，3B-01,02 では Q2,Q3，3B-03 では，Q1,Q2 および Q2,Q3 に相関，3B-04 はすべての質問が独立であると示唆される．これらから，元のデータ集合に返って，さらに考察すると，危機，LOS，3B-04 では，Q1 から Q3 の項目が全く関連なく評価されたのに対し，他のチャレンジでは，質問の少なくとも二つが関連づけられて評価されていたと推定される．例えば，DSIU，3B-01，02 はも聴衆が貢献度に着目している．DSIU では，Q2,Q3 の評点が高く，3B-01,02 の評点が低かったことから，聴衆は 3B-01,02 について，貢献度が低いと判断していたことが推測される．一方，3B-03 では，Q1 の評価が高いものは Q2 の評価が低く，Q2 の評価が高いものは Q3 の評価が高いという傾向が見られ，貢献度が高いと評価したグループと実現性が高いと評価したグループに評価が分かれたことが推察される.

本データセットでのラフ集合による縮約生成の結果を表 5 に示した．この結果は，Fisher の直接検定とほぼ同じであるが，検定に比べて，結果が conservative でない.

3.2 クラスタリング

クラスタリングに関しては，ほとんどがクラスタ 2 つに分かれた (表 6) ．多くは，質問の一つが高く，他の質問の点数は低いグループとそうでないグループに分かれた．質問の連関の形式は同じであっても，実際のクラスタの性質が異なるチャレンジもあった．例えば，3B-01 と 3B-02 は Q2,Q3 に連関があったが，3B-01 は全体の総点が高いグループと低いものに分かれたが，3B-02 では，Q1 のみが高いグループと，Q2,Q3 が相対的に高いとに分かれた.

Dendrogram に関しては，3B-03 以外に Chain effect は見られなかった．Chain Effect は通常ばらついているデータを無理矢理グループ化する時に生まれるものであり [Everitt 01]，クラスタリングが有効でない場合が含まれている．したがっ

表 6: クラスタリングのまとめ

略称	クラスタ数	特徴	Chain-Effect
DSIU	2	Q1 のみに評価が高いものとそうでないもの	-
RCT	2	Q1-3 に 5 が含まれているものとそうでないもの	-
危機	3	総合点の大きさ	-
LOS	2	Q1 の評点が高いものとそうでないもの	-
3B-01	2	総合点の高いものと低いもの	-
3B-02	2	Q1 のみが高いもの，Q2, Q3 が高いもの	-
3B-03	2	Q1 のみが高いもの，Q2, Q3 が高いもの	+
3B-04	2	点数が低いものと Q2, Q3 の評価が高いもの	-

表 2: ニュー・チャレンジ一覧

略称	チャレンジ名 , 提案者 (所属)
3B4-01	インターネットからの情報獲得による 研究活動支援システム 砂山 渡, 谷内田 正彦 (大阪大)
3B4-02	Social Scheduler – P2P モデルを用いた 協調的リソースプランナの提案 大向 一輝 (1), 濱崎 雅弘 (2), 武田 英明 (3), 三木 光範 (1) (1) 同志社大 (2) 奈良先端大 (3) 国立情報学研究所
3B4-03	事例に基づくデザイン支援と評価基盤の構築 片寄 晴弘 (1), 平田 圭二 (2), 原田 利宣 (3), 平賀 瑠美 (4), 笠尾 敦司 (5) (1) 関西学院大, (1) 科技団さきがけ研究 21, (2)NTT, (3) 和歌山大 (4) 文教大, (5) 東京工芸大
3B4-04	記憶弱者の QOL (Quality of Life) を 補償する行動支援システム 山下 耕二 (1), 福原 知宏 (1), 松村 憲一 (1)(2), 寺田 和憲 (1), 久保田 秀和 (1)(4), 畦地 真太郎 (3), 西田 豊明 (1)(4) (1) 通信総合研究所, (2) 大阪大, (3) 北海道東海大, (4) 東大

表 4: Fisher の直接検定の結果

略称	(Q1,Q2)	(Q1,Q3)	(Q2,Q3)
DSIU	0.267	0.400	< 0.0001
RCT	0.00575	0.00196	0.00448
危機	1	0.767	0.413
LOS	0.341	0.270	0.657
3B-01	0.720	0.790	0.000621
3B-02	0.110	0.256	0.000709
3B-03	0.114	0.00588	0.00884
3B-04	0.0821	0.167	0.113

表 5: ラフ集合による縮約

略称	縮約
DSIU	{Q2,Q3}
RCT	{Q1,Q2},{Q1,Q3},{Q2,Q3}
危機	{Q1,Q2,Q3}
LOS	{Q1,Q3}
3B-01	{Q2,Q3}
3B-02	{Q1,Q2}, {Q2,Q3}
3B-03	{Q1,Q2}, {Q1,Q3}, {Q2,Q3}
3B-04	{Q1,Q2}, {Q1,Q3}, {Q2,Q3}

て，3B-03 は他に比べて，評価のばらつきが強かったことが示唆された．比較のため，図 1 と図 2 に 3B-02 と 3B-04 に関する dendrogram を示す．

3.3 MDS

MDS では，評点に関する回答者の評価を明確に図示することができた．ここでは，紙面の都合から，評点 2 についての MDS の結果を示す．図 3–6 は評点 2 から 5 までの MDS の結果を示す．これらの図からまず，評点 2 と 4 については，点が集中しているグループとある程度広がりをもったグループとに分かれていることがわかる．それぞれ，評点 2 をつけたものとそうでないもの，評点 4 をつけたものとそうでないものとに分かれている上，評点 2 をつけたものは評点 4 をつけていないという傾向があり，図 3 と図 5 とは相対の関係にある．したがって，このチャレンジに関しては，評点 2 と 4 とについて大きく評価が分かれたことを示唆している．評点 3 については，ばらついており，特定の傾向は見あたらないが，Q2 に対してのみ 3 をつけた回答者と，Q3 についてのみ 3 をつけた回答者のみが小グループを形成している．さらに，評点 5 についての MDS の結果から，ほとんどの回答者が 5 をつけず，第 1 象限に固まっており，ごく少数についてが 5 をつけている．第 2 象限のグループは Q1 に対して 5 をつけ，第 4 象限が Q3 について 5 をつけたグループである．以上の結果から，このチャレンジについては，評点が 2 のグループと 4 のグループに分かれ，比較的評価が大きく分かれた．さらに，Q1 について 5 をつけたグループが他の評点をつけたグループに比べて特異的な位置を占めていることが示唆される．

4. おわりに

本論文では，類似度による視覚化と分類知識の抽出法とを組み合わせることで，効率的に知識を発見するプロセスについて考察し，昨年度の人工知能学会全国大会の近未来チャレン

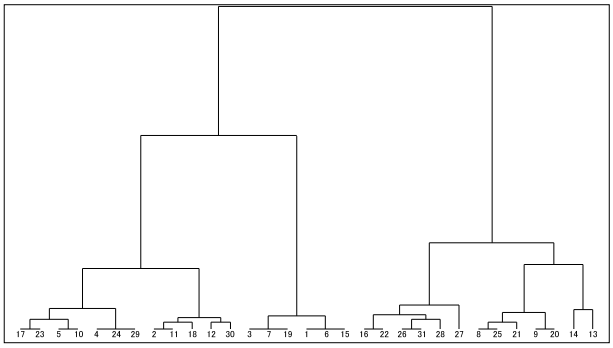

図 1: 3B-02 の Dendrogram

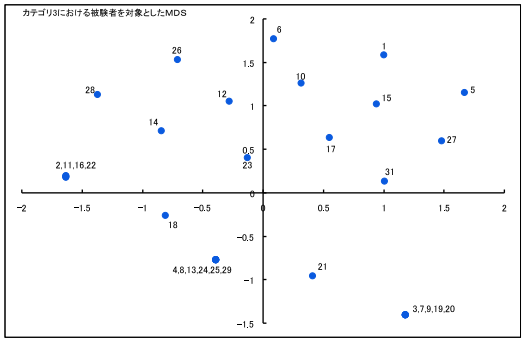

図 4: 3B-02: 評点 3 に関する MDS の結果

図 2: 3B-03 の Dendrogram

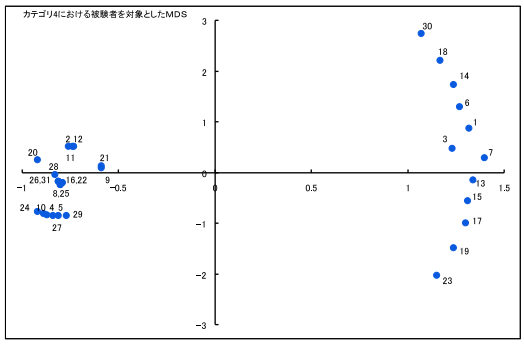

図 5: 3B-02: 評点 4 に関する MDS の結果

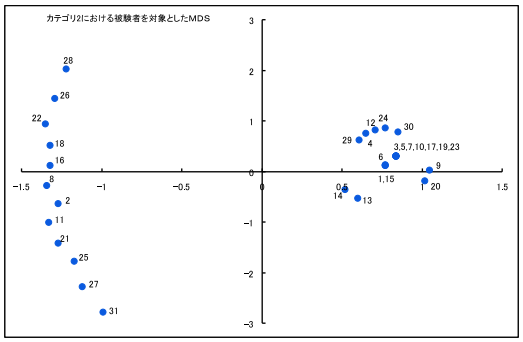

図 3: 3B-02: 評点 2 に関する MDS の結果

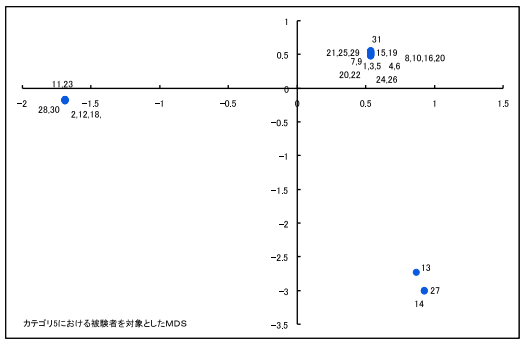

図 6: 3B-02: 評点 5 に関する MDS の結果

ジの評価データを用いて，このプロセスの有効性を検証した．それぞれの手法の特徴をいかすことで，聴衆の心理的構造を描出するとともに，聴衆の焦点を定量的に表現することが可能になった．今後，この手法を他のデータに適用し，その有効を検証していく予定である．

なお，演題発表においては，本抄録に掲載できなかった結果についても，供覧する予定である．

参考文献

[VDM 02] Simoff, S.J., Noirhomme-Fraiture, M. Boehlen, M.H. (eds.) (2002). *Proceedings of Second International Workshop on Visual Data Mining*, ECML/PKDD2002, Helsinki.

[Everitt 01] Everitt, B. S.(2001) *Cluster Analysis*, 4th Edition, John Wiley & Son, London.

[Fisher 34] Fisher, R. A. (1934) *Statistical Methods for Research Workers* (5th Ed.), Oliver&Boyd, Edinburgh.

[Osawa 2003] Osawa, Y. *Special Issue on Chance Discovery, New Generation Computing*, 2003.

[Pawlak 91] Pawlak, Z. (1991) *Rough Sets*. Kluwer Academic Publishers, Dordrecht.

[Torgerson 52] Torgerson, W. (1952) Mulitdimensional scaling: I Theory and Method. *Phychometrika*,**17**, 401–419.