分散表現を利用した特徴的相関ルールの抽出

Extraction of Characteristic Association Rules by Representation Learning

尾崎 知伸 *
Tomonobu Ozaki

日本大学 文理学部
College of Humanities and Sciences, Nihon University

Abstract: Frequent pattern mining is one of the most fundamental problems in data mining. While extensive research has been conducted over a long period, the derivation of huge number of low comprehensible patterns is widely recognized as an unsolved essential drawback in this area. To alleviate the drawback by ranking patterns from various aspects, in this paper, we propose several evaluation criteria for frequent patterns as well as association rules, using the embedding, or vector representations, of items, transactions and patterns. The proposed criteria were assessed by preliminary experiments using a real dataset in Japanese video-sharing site.

1 はじめに

データ集合中に現れる特徴的な組み合わせをパターンやルールとして抽出する頻出パターン・相関ルール発見 [1, 2] は，データマイニングの最も基本的な問題の一つであり，これまでに様々な拡張と幅広い分野での応用が展開されている．その一方で，頻出パターン・相関ルール発見が本質的に抱える問題点として，解釈が困難なパターン・ルールが大量に導出されるという点が指摘されている．この問題に対してこれまでに，頻度に着目した代表元に基づく圧縮表現 [3, 4] や，統計的テストに基づくフィルタリング [5, 6] などに加え，種々の観点からの評価関数 [7, 8] が提案されている．

本論文では，頻出パターン・相関ルール発見における大量かつ解釈困難な結果の導出という問題に対し，近年発展が著しい表現学習技術 [9] を援用することを提案する．具体的には，既存の表現学習技術を用い，アイテム，トランザクション，パターンそれぞれの低次元実数ベクトル表現（分散表現）を獲得するとともに，それらを基にした頻出パターン・相関ルールの新たな評価関数を開発する．

表現学習技術の頻出パターン発見への応用として，これまでに，パターンの分散表現を用いた頻出パターンの評価手法 [10] が提案されている．提案された手法は，既存の表現学習技術を用いて頻出パターンを実数空間に埋め込んだ上で，パターン間の距離を基に，各パターンの代表性や例外性，媒介性などの特徴を評価するというものである．これに対し本研究では，より多様な側面からの評価を目的に，パターンに加えて，アイテムとトランザクションの分散表現を利用することを提案する．また頻出パターンだけでなく，パターン組み合わせである相関ルールに対しても，新たな評価関数を導入する．

本論文の構成は，以下の通りである．2 章では，準備としていくつかの記号と用語を導入するとともに，頻出パターン・相関ルール発見問題を形式的に定義する．3 章では，頻出パターン，相関ルールそれぞれに対し，アイテム，トランザクション，パターンの分散表現を基にした評価関数を提案する．4 章で評価実験について述べ，最後に 5 章でまとめを行い，今後の課題を述べる．

*連絡先：日本大学文理学部情報科学科
〒 156-8550 東京都世田谷区桜上水 3-25-40
E-mail: tozaki@chs.nihon-u.ac.jp

2 準備

頻出パターン

$I = \{i_1, i_2, \cdots, i_{|I|}\}$ を全アイテムの集合とする．I の部分集合 $t \subseteq I$ をトランザクションと呼び，トランザクションの集合 $D = \{t_1, t_2, \cdots, t_{|D|}\}$ をデータベースと呼ぶ．

データベース D 中において，複数のトランザクションに共通して現れるアイテムの組み合わせ $P \subseteq I$ をパターンと呼ぶ．パターン P の要素数をサイズと呼び，$|P|$ と表記する．データベース D において，パターン P を支持する（包含する）トランザクションの集合を

支持集合と呼び

$$cov_D(P) = \{t \in D \mid t \subseteq P\}$$

と表記する．また，データベース D におけるパターン P の支持度 $sup_D(P)$ を，支持集合 $cov_D(P)$ を用いて

$$sup_D(P) = |cov_D(P)| \,/\, |D|$$

と定義する．最小支持度と呼ばれる閾値 $\sigma\,(0 < \sigma \leq 1)$ に対し，条件 $sup_D(P) \geq \sigma$ を満たすパターン P を頻出パターンと呼ぶ．また，頻出パターンの全体集合を

$$\mathcal{F}_D^\sigma = \{P \subseteq I \mid sup_D(P) \geq \sigma\}$$

と表記する．

頻出パターンの内，同一支持集合を持つ頻出パターンの集合を同値類としたときの（包含関係での）極大元，すなわち条件

$$sup_D(P) \geq \sigma \quad \wedge \quad \nexists Q \supset P \; s.t. \; cov_D(Q) = cov_D(P)$$

を満たすパターン P を頻出飽和パターン [4] と呼ぶ．一方，同一支持集合を持つ頻出パターンの集合を同値類としたときの極小元，すなわち条件

$$sup_D(P) \geq \sigma \quad \wedge \quad \nexists G \subset P \; s.t. \; cov_D(G) = cov_D(P)$$

を満たすパターン P を頻出極小パターン [3] と呼ぶ．また，頻出飽和パターンの集合を $\mathcal{C}_D^\sigma$，頻出極小パターンの集合を $\mathcal{G}_D^\sigma$ とそれぞれ表記する．

相関ルール

共通元を持たない二つのパターン A と $C\,(A, C \subseteq I, A \cap C = \emptyset)$ に対し，ルール $A \Rightarrow C$ を考える．ルール $A \Rightarrow C$ の支持度と確信度を以下の様に定義する．

$$\begin{aligned} sup_D(A \Rightarrow C) &= sup_D(A \cup C) \\ conf_D(A \Rightarrow C) &= sup_D(A \Rightarrow C) \,/\, sup_D(A) \end{aligned}$$

最小確信度と呼ばれる閾値 $\theta\,(0 < \theta \leq 1)$ と最小支持度 σ に対し，2 つの条件 $sup_D(A \Rightarrow C) \geq \sigma$ と $conf_D(A \Rightarrow C) \geq \theta$ を満たすルール $A \Rightarrow C$ を相関ルールと呼ぶ．また，相関ルールの全体集合を

$$\mathcal{R}_D^{\sigma,\theta} = \left\{ A \Rightarrow C \;\middle|\; \begin{array}{l} A, C \subseteq I, A \cap C = \emptyset, \\ sup_D(A \Rightarrow C) \geq \sigma, \\ conf_D(A \Rightarrow C) \geq \theta \end{array} \right\}$$

と表記する．

データベース D と閾値 σ, θ に対し，頻出パターンの全体集合 $\mathcal{F}_D^\sigma$ と相関ルールの全体集合 $\mathcal{R}_D^{\sigma,\theta}$ を求める問題をそれぞれ，頻出パターン発見問題，相関ルール発見問題と呼ぶ．

3 分散表現を用いた評価関数の提案

本章では，アイテム $i \in I$ の分散表現 $\vec{i}$ 及びトランザクション $t \in D$ の分散表現 $\vec{t}$，頻出パターン $P \in F_D^\sigma$ の分散表現 $\vec{P}$ を用いた頻出パターン，相関ルールに対する新たな評価関数を複数提案する．

なお各分散表現は，アイテムに関しては GloVe[11] や word2vec[12, 13]，fasttext[14]，トランザクションに関しては Trans2Vec[15] や doc2vec[16]，パターンに関しては文献 [17] や [10] で提案された手法などを含め，何らかの方法を用いて予め獲得済みであることを前提とする．また 2 つのオブジェクト O_i, O_j に対し，それらの分散表現 $\vec{O_i}$，$\vec{O_j}$ 間の距離（非類似度）を $dist\,(O_i, O_j)$ と表記する．

3.1 頻出パターンの評価

アイテムベクトルを用いた評価関数

本研究では，パターン P に対し，P が含むアイテム $i \in P$ の多様性を一つの評価関数とすることを提案する．具体的には，サイズ 2 以上の頻出パターン $P\,(|P| \geq 2)$ に対し，アイテムベクトルに基づく評価関数を

$$EP_I^{\min}(P) = \min_{i \neq j \in P} dist\left(\vec{i}, \vec{j}\right) \tag{1}$$

と定義する．評価関数 $EP_I^{\min}$ は，パターン中の 2 アイテム $i, j \in P$ 間の距離の最小値であり，この値が大きいほど，P は非類似なアイテムのみから構成されると判断できる．

一方で多様性を認めず，類似するアイテムのみから構成されるパターンを高く評価することも考えられる．この要求に対応するため，パターン中の 2 アイテム $i, j \in P$ 間の最大距離を基準とする評価関数

$$EP_I^{\max}(P) = \max_{i \neq j \in P} dist\left(\vec{i}, \vec{j}\right) \tag{2}$$

を提案する．$EP_I^{\min}$ とは逆に，$EP_I^{\max}$ の値が小さいほど，P が類似アイテムのみから構成されていると判断できる．また $EP_I^{\max}$ は，相互依存性の低いアイテムを排除するというハイパークリークパターン [18, 19] と同様の考えに基づいており，ハイパークリークパターンにおける相互依存性を，分散表現における実ベクトルの類似性で代用したものと見做すことも出来る．

トランザクションベクトルを用いた評価関数

トランザクションベクトルは，データベースにおける各トランザクションの特徴を表すと考えられる．従って，パターン P の支持集合 $cov_D(P)$ を考慮することで，P がどの様なトランザクションを含むのか（説明

するのか）といった新たな観点からパターンを評価することが期待できる．本論文では，トランザクションベクトルを用いたパターンの評価関数として，支持集合の差を利用することを提案する．具体的には，サイズ2以上の対象パターン $P\,(|P| \geq 2)$ の支持集合 $cov_D(P)$ と，P の部分集合パターン $G \subset P$ の支持集合 $cov_D(G)$ を利用する．以下に形式的な定義を示す．

$$EP^{bet}_{TR}(P) \;=\; \min_{G \in MSS(\mathcal{F}',P)} V^+(G,P) + V^-(G,P) \quad (3)$$

ここで，集合

$$MSS(\mathcal{F}',P) = \left\{ G \in \mathcal{F}' \;\middle|\; \begin{array}{l} G \subset P, \\ \nexists S \in \mathcal{F}'\ s.t.\ G \subset S \subset P \end{array} \right\}$$

は，頻出パターン集合 $\mathcal{F}' \subseteq \mathcal{F}^{\sigma}_D$ 中における P の部分集合パターンの極大元 G からなる集合である．また，

$$\begin{array}{l} V^+(G,P) \;=\; V(cov_D(G), cov_D(P)), \\ V^-(G,P) \;=\; V(cov_D(G), cov_D(G) \setminus cov_D(P)), \\ V(CovG, CovS) \;=\; \frac{|CovS|}{|CovG|} dist\left(\overline{CovS}, \overline{CovG}\right)^2 \end{array}$$

であり，

$$\overline{Cov} \;=\; \frac{1}{|Cov|} \sum_{t \in Cov} \vec{t}.$$

は，集合 Cov に含まれるトランザクションの分散表現 $\vec{t}$ の平均を表す．

評価関数 EP^{bet}_{TR} は，パターン G の支持集合 $cov_D(G)$ を二つのクラスタ $cov_D(P)$ と $cov_D(G) \setminus cov_D(P)$ に分けた場合のクラス間分散に相当する（$G \subset P$ より $cov_D(G) \supseteq covD(P)$ が成り立つ）．従ってこの評価値が大きいほど，パターン P は部分集合パターン G の支持集合 $cov_D(G)$ を選択的に分割していることとなり，部分集合パターンとの差の観点から，P が大きな情報を持つことが期待できる．

ところで評価関数 EP^{bet}_{TR} は，極小パターン [3] における考え方をベクトル表現へと拡張したものと見做すこともできる．（ノイズを考慮した）極小パターンでは，支持集合そのものの違いに着目しているが，提案手法では支持集合を構成する各要素のベクトルに着目しており，トランザクションが持つ情報をより積極的に利用した評価基準であると言える．

パターンベクトルを用いた評価関数

頻出パターン P の分散表現 $\vec{P}$ が，その構成要素の分散表現の合成結果とどの程度異なるのかを一つの基準とする．より具体的には，サイズ2以上のパターン $P\,(|P| \geq 2)$ を対象とし，P の分散表現 $\vec{P}$ と，P の最小構成要素であるサイズ1のパターン $\{x\} \subset P$ の分散表現 $\overrightarrow{\{x\}}$ の加重平均との差が大きい場合に，構成要素から P が持つ意味を推測することは難しい判断し，高い評価値を与える．この考えに従ったパターンベクトルに基づく評価関数 $EP^{\subset}_{\mathcal{F}}$ を以下の様に提案する．ここで w_x はパターン $\{x\}$ に対する重みを表す．

$$EP^{\subset}_{\mathcal{F}}(P) \;=\; dist\left(\vec{P}, \sum_{x \in P} w_x \overrightarrow{\{x\}}\right) \quad (4)$$

本論文では，パターンベクトルを用いた別の評価基準として，対象パターン P とその上位集合パターン $Q \supset P$ との関連性に着目した基準を提案する．

飽和パターン [4] やその発展であるノイズ許容飽和パターン [20] では，対象パターン P とその上位集合パターン $Q \supset P$ との支持度の差が大きい場合，P は Q とは異なる意味や役割を持つと考える．逆に言えば，P と Q の支持度の差が小さい場合，P の意味や役割は Q と大差ないとし，Q のみを考慮すれば十分であると考える．この考えに基づき，支持度の差を分散表現の差と置き換えた基準として，評価基準

$$EP^{\supset}_{\mathcal{F}}(P) \;=\; \min_{Q \in MGS(\mathcal{F}',P)} dist\left(\vec{P}, \vec{Q}\right) \quad (5)$$

を提案する．ここで，集合

$$MGS(\mathcal{F}',P) = \left\{ Q \in \mathcal{F}' \;\middle|\; \begin{array}{l} P \subset Q, \\ \nexists S \in \mathcal{F}'\ s.t.\ P \subset S \subset Q \end{array} \right\}$$

は，頻出パターン集合 $\mathcal{F}' \subseteq \mathcal{F}$ 中における P の上位集合パターンの極小元 Q からなる集合である．

評価関数 $EP^{\supset}_{\mathcal{F}}$ は，対象パターン P と最も類似する上位集合パターン Q との差であり，この値が大きいパターンのみを $\mathcal{F}$ から選択することで，相互に異なる役割や意味を持つパターンの効率的な収集が達成できると期待される．

3.2 相関ルールの評価

相関ルール $A \Rightarrow C$ は，2つの頻出パターン $A, C \in \mathcal{F}^{\sigma}_D$ 上に定義されるルールである．従ってその評価は，A と C の関連性をその基本とする．

アイテムベクトルを用いた評価関数

本研究では，相関ルール $A \Rightarrow C$ に対し，集合 A, C 間の距離を評価基準とすることを提案する．具体的には，アイテムベクトルに基づく評価関数を

$$ER^d_I(A \Rightarrow C) \;=\; Dist_d(A,C) \quad (6)$$

と定義する．ここで $Dist_d$ は，2つの集合 A, C 間の距離関数であり，本研究では，

$$\begin{array}{rcl} Dist_{min}(A,C) & = & \min_{a \in A, c \in C} dist\,(\vec{a}, \vec{c}) \\ Dist_{max}(A,C) & = & \max_{a \in A, c \in C} dist\,(\vec{a}, \vec{c}) \\ Dist_{avg}(A,C) & = & \frac{1}{|A| \times |C|} \sum_{a \in A, c \in C} dist\,(\vec{a}, \vec{c}) \end{array}$$

の 3 種（最短距離・最長距離・平均距離）を考える．これらの値，すなわち A, C 間の距離が大きいほど，A と C を持つ情報が異なり，A から C が推測しにくいと考えらえるので，より有益なルールであると判断する．

トランザクションベクトルを用いた評価関数

相関ルール $A \Rightarrow C$ のトランザクションに基づく評価として，A の支持集合 $cov_D(A)$ を，C を支持するトランザクションの集合 $cov_D(A) \cap cov_D(C)$ と，C を支持しないトランザクションの集合 $cov_D(A) \setminus cov_D(C)$ に分けた場合のクラス間分散を用いる．

$$ER^{bet}_{TR}(A \Rightarrow C) \quad = \quad V^{+}(A,C) + V^{-}(A,C) \quad (7)$$

ER^{bet}_{TR} の値が大きいほど，支持集合 $cov_D(A)$ を選択的に分割していると考えられ，ルールとして有益であると判断する．

パターンベクトルを用いた評価関数

パターンベクトルを用いた相関ルール $A \Rightarrow C$ の評価関数として，単純に前件 A と帰結 C のベクトル差を採用する．すなわち，

$$ER_{\mathcal{F}}(A \Rightarrow C) \quad = \quad dist\left(\vec{A}, \vec{C}\right) \qquad (8)$$

と定義する．

ER^{d}_{I} 同様，$ER_{\mathcal{F}}$ の値が大きなルールは，A から C が推測しにくく，また得られる情報が大きいと考えられるので，より有益なルールであると判断する．

4 評価実験

4.1 実験設定

実験には，国立情報学研究所が (株) ドワンゴおよび (株) 大百科ニュース社から提供を受けて研究者に提供している「ニコニコデータセット」[1] を用いた．

データセットから，(1)「スポーツ」タグが付与された動画集合と (2)「音楽」タグが付与された動画集合をそれぞれ抽出し，各動画に付与された（それぞれ「スポーツ」,「音楽」を除く）タグ集合をトランザクションとし，データセット D_{music} 及び D_{sports} を準備した．D_{music}, D_{sports} に対し，頻出パターン抽出器 Eclat[2] を用いて，最小支持度 $\sigma = 0.0001$ の飽和パターンの集合 $\mathcal{C}^{0.0001}_{D}$ を導出した．さらに最小確信度を $\theta = 0.75$ とし，条件 $A \cup C \in \mathcal{C}^{0.0001}_{D}$ を満たす相関ルール集合 $R \subset \mathcal{R}^{0.0001,0.75}_{D}$ を求めた．各データセットの統計量を，表 1 に示す．

[1] https://www.nii.ac.jp/dsc/idr/nico/nico.html
[2] http://www.borgelt.net/doc/eclat/eclat.html

表 1: 対象データの統計量

	$\|D\|$	$\|I\|$	$\|C\|$	$\|R\|$
D_{music}	456,659	545,219	24,748	20,244
D_{sports}	82,954	122,366	36,047	83,315

アイテムベクトル $\vec{i}$ の獲得には，GloVe[3] を用い，次元数 $d \in \{100, 200\}$ のベクトル集合（2 種類）を準備した．一方簡略化のため，トランザクション t の分散表現 $\vec{t}$ は，t を構成するアイテムの分散表現の平均値，すなわち $\vec{t} = \frac{1}{|t|}\sum_{i \in t} \vec{i}$ として算出した．またパターン P の分散表現 $\vec{P}$ の獲得には，文献 [10] に倣い，各パターンを単語，各トランザクションに支持される飽和パターン集合を文書と見做し，GloVe を用いて導出した．また，分散表現（実ベクトル）間の距離は,（値域を 0〜1 に正規化した）コサイン距離 $dist\,(\vec{p}, \vec{q}) = \frac{1}{2}\,(1 - cos(\vec{p}, \vec{q}))$ を利用する．

4.2 頻出パターン

提案した頻出パターンに関する評価関数 (式 (1) − 式 (5)) を評価するため，得られた評価値の統計値と，各評価値間の相関係数を算出した．結果を表 2 と表 3 に示す．なお $EP^{\subset}_{\mathcal{F}}(P)$ における重み w_x は，$(1) w_x = 1/|P|$ と $(2) w_x = sup_D(\{x\}) / \sum_{i \in P} sup_D(\{i\})$ の二種類を準備している．

表 2 より，次元数 ($d \in \{100, 200\}$) による大きな差は確認できない．また $EP^{\subset}_{\mathcal{F}}$ における重み設定 (avg, sup) に関しても大きな差は確認できないが，支持度を考慮した方 (sup) が平均，標準偏差に大きくなる傾向が読み取れる．ところで EP^{bet}_{TR} は，ほとんどのパターンで値が非常に小さく (0.000)，パターン間で差が付かないという点で識別能力に乏しいと考えられる．その一方で，$EP^{\supset}_{\mathcal{F}}$ の標準偏差は大きく，部分集合パターンとの比較するより，上位集合パターンと比較を行った方が差が出やすいことが想像できる．

表 3 より，特に D_{music} において，EP^{max}_{I} と $EP^{\subset}_{\mathcal{F}(sup)}$ の間に強い正の相関があることが確認できる．その一方で，EP^{bet}_{TR} は他の評価関数に対して負の相関を持つ傾向が見て取れる．また全体として，D_{music} より D_{sport} の方が若干相関が小さい結果となった．

4.3 相関ルール

提案した相関ルールに関する評価関数 (式 (6)−式 (8)) を評価するため，得られた評価値の統計値と，各評価値間の相関係数を算出した．結果を表 4 と表 5 に示す．

[3] https://nlp.stanford.edu/projects/glove/

表 2: パターンに対する各評価値に関する統計量

	min	1stQ.	median	3rdQ	max	mean	s.d.	min	1stQ.	median	3rdQ	max	mean	s.d.
D_{music}	$d = 100$							$d = 200$						
EP_I^{min}	0.012	0.056	0.070	0.083	0.618	0.325	0.105	0.014	0.056	0.071	0.086	0.618	0.342	0.099
EP_I^{max}	0.012	0.098	0.116	0.131	0.767	0.392	0.088	0.014	0.100	0.121	0.135	0.773	0.400	0.082
EP_{TR}^{bet}	0.000	0.000	0.000	0.000	0.006	0.000	0.000	0.000	0.000	0.000	0.000	0.006	0.000	0.000
$EP_{\mathcal{F}(avg)}^{\subset}$	0.071	0.129	0.139	0.149	0.748	0.349	0.088	0.072	0.130	0.144	0.154	0.747	0.358	0.085
$EP_{\mathcal{F}(sup)}^{\subset}$	0.070	0.132	0.151	0.160	0.713	0.401	0.093	0.072	0.137	0.154	0.164	0.702	0.410	0.089
$EP_{\mathcal{F}}^{\supset}$	0.027	0.067	0.079	0.085	1.000	0.647	0.396	0.025	0.072	0.082	0.087	1.000	0.649	0.393
D_{sports}	$d = 100$							$d = 200$						
EP_I^{min}	0.003	0.026	0.040	0.045	0.931	0.322	0.142	0.002	0.023	0.035	0.043	0.930	0.321	0.144
EP_I^{max}	0.030	0.107	0.138	0.154	0.972	0.454	0.104	0.028	0.107	0.135	0.152	0.977	0.455	0.105
EP_{TR}^{bet}	0.000	0.000	0.000	0.000	0.078	0.000	0.001	0.000	0.000	0.000	0.000	0.074	0.000	0.001
$EP_{\mathcal{F}(avg)}^{\subset}$	0.045	0.082	0.100	0.114	0.961	0.436	0.145	0.042	0.081	0.099	0.110	0.965	0.434	0.146
$EP_{\mathcal{F}(sup)}^{\subset}$	0.042	0.105	0.120	0.132	0.927	0.466	0.127	0.039	0.103	0.120	0.130	0.930	0.465	0.127
$EP_{\mathcal{F}}^{\supset}$	0.002	0.030	0.044	0.050	1.000	0.758	0.370	0.001	0.029	0.041	0.047	1.000	0.757	0.371

表 3: パターンに対する各評価値間の相関係数 (下：$d = 100$，上：$d = 200$)

	D_{music}						D_{sports}					
	EP_I^{min}	EP_I^{max}	EP_{TR}^{bet}	$EP_{\mathcal{F}(avg)}^{\subset}$	$EP_{\mathcal{F}(sup)}^{\subset}$	$EP_{\mathcal{F}}^{\supset}$	EP_I^{min}	EP_I^{max}	EP_{TR}^{bet}	$EP_{\mathcal{F}(avg)}^{\subset}$	$EP_{\mathcal{F}(sup)}^{\subset}$	$EP_{\mathcal{F}}^{\supset}$
EP_I^{min}		0.554	-0.049	0.264	0.328	0.010		0.373	-0.004	0.254	0.251	-0.079
EP_I^{max}	0.510		-0.232	0.554	0.649	0.075	0.371		-0.059	0.313	0.394	0.114
EP_{TR}^{bet}	-0.037	-0.222		-0.239	-0.309	-0.060	-0.005	-0.058		-0.069	-0.086	-0.086
$EP_{\mathcal{F}(avg)}^{\subset}$	0.193	0.530	-0.233		0.890	-0.183	0.260	0.319	-0.067		0.906	0.057
$EP_{\mathcal{F}(sup)}^{\subset}$	0.276	0.629	-0.294	0.894		-0.050	0.258	0.397	-0.084	0.907		0.056
$EP_{\mathcal{F}}^{\supset}$	0.016	0.084	-0.051	-0.178	-0.047		-0.078	0.116	0.006	0.056	0.055	

表 4: 相関ルールに対する各評価値に関する統計量

	min	1stQ.	median	3rdQ	max	mean	s.d.	min	1stQ.	median	3rdQ	max	mean	s.d.
D_{music}	$d = 100$							$d = 200$						
ER_I^{avg}	0.055	0.087	0.098	0.107	0.584	0.312	0.073	0.052	0.088	0.100	0.108	0.588	0.326	0.075
ER_I^{max}	0.058	0.103	0.124	0.143	0.767	0.393	0.083	0.055	0.105	0.126	0.146	0.773	0.400	0.078
ER_I^{min}	0.012	0.022	0.024	0.026	0.531	0.227	0.093	0.014	0.025	0.025	0.027	0.520	0.247	0.100
ER_{TR}^{bet}	0.000	0.000	0.000	0.000	0.014	0.000	0.000	0.000	0.000	0.000	0.000	0.015	0.000	0.000
$ER_{\mathcal{F}}$	0.000	0.000	0.000	0.000	0.738	0.363	0.110	0.000	0.000	0.000	0.000	0.724	0.367	0.109
D_{sports}	$d = 100$							$d = 200$						
ER_I^{avg}	0.041	0.117	0.138	0.152	0.869	0.349	0.079	0.039	0.114	0.136	0.150	0.872	0.349	0.080
ER_I^{max}	0.046	0.153	0.183	0.205	0.972	0.508	0.139	0.042	0.149	0.183	0.204	0.977	0.509	0.139
ER_I^{min}	0.003	0.006	0.006	0.010	0.818	0.195	0.112	0.002	0.004	0.005	0.007	0.823	0.195	0.113
ER_{TR}^{bet}	0.000	0.000	0.000	0.000	0.069	0.000	0.001	0.000	0.000	0.000	0.000	0.067	0.000	0.001
$ER_{\mathcal{F}}$	0.000	0.000	0.000	0.000	0.925	0.330	0.193	0.000	0.000	0.000	0.000	0.929	0.331	0.195

表 5: ルールに対する各評価値間の相関係数 (下：$d = 100$，上：$d = 200$)

	D_{music}					D_{sports}				
	ER_I^{avg}	ER_I^{max}	ER_I^{min}	ER_{TR}^{bet}	$ER_{\mathcal{F}}$	ER_I^{avg}	ER_I^{max}	ER_I^{min}	ER_{TR}^{bet}	$ER_{\mathcal{F}}$
ER_I^{avg}		0.807	0.835	0.072	0.418		0.630	0.531	0.177	0.179
ER_I^{max}	0.814		0.419	0.008	0.147	0.628		-0.210	0.046	-0.258
ER_I^{min}	0.806	0.387		0.114	0.489	0.528	-0.217		0.168	0.457
ER_{TR}^{bet}	0.090	0.012	0.140		0.071	0.175	0.041	0.169		-0.145
$ER_{\mathcal{F}}$	0.388	0.158	0.439	0.076		0.174	-0.263	0.457	-0.152	

表 4 より，提案した各評価関数の標準偏差が全体的に小さいことが分かる．またパターン評価における EP_{TR}^{bet} 同様，ER_{TR}^{bet} はほとんどのルールに対して値が小さく，識別能力に乏しいことが確認された．$ER_{\mathcal{F}}$ に関しては，第三四分位数までは ER_{TR}^{bet} と類似傾向が見られるが，平均，標準偏差ともに十分大きな値を示しており，少ないルールに対して大きな値が割り当てられていることが期待できる．

表 5 より，$ER_{\mathcal{F}}$ と ER_I^{min}，ER_I^{avg} との間にそれぞれ高い正の相関が確認できる．その一方で，D_{sports} においては $ER_{\mathcal{F}}$ と ER_I^{max} の間に負の相関が確認され，直観とは異なる結果となった．

5 まとめ

本研究では，分散表現技術を用いた頻出パターン・相関ルール発見の弱点の軽減と高性能化を目指し，アイテム・トランザクション・パターンの分散表現を用いた，頻出パターン・相関ルールに対する種々の評価関数を提案した．また，予備的な段階ではあるが，実データを用いて提案した評価関数が持つ傾向を調査した．

今後の課題としては，大規模かつ複数データを用いた提案手法の評価や，種々の分散表現獲得技術の援用，提案した評価関数の理論的な解析，構造データへの展開などが挙げられる．

謝辞： ニコニコデータセットを提供頂いた（株）ドワンゴと国立情報学研究所に感謝いたします．また，本研究の一部は JSPS 科研費 17K00315 の助成を受けたものです．

参考文献

[1] R. Agrawal, T. Imielinski and A.Swami : Mining association rules between sets of items in large databases, *Proc. of the 1993 ACM-SIGMOD International Conference on Management of Data*, pp.207–216, 1993.

[2] J. Han, H. Cheng, D. Xin and X. Yan : Frequent pattern mining: current status and future directions, *Data Mining and Knowledge Discovery*, Vol.15, No.1, pp.55–86, 2007.

[3] J.-F. Boulicaut, A. Bykowski and C. Rigotti : Free-Sets : A Condensed Representation of Boolean Data for the Approximation of Frequency Queries, *Data Mining and Knowledge Discovery*, Vol.7, No.1, pp.5–22, 2003.

[4] N. Pasquier, Y. Bastide, R. Taouil and L. Lakhal : Discovering frequent closed itemsets for association rules, *Proc. of the 7th International Conference on Database Theory*, pp.398–416, 1999.

[5] G. I. Webb : Discovering Significant Patterns, *Machine Learning*, Vol.68, No.1, pp.1–33, 2007.

[6] W. Hämäläinen and G. I. Webb : Statistically sound pattern discovery, *Proc. of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.1976, 2014.

[7] T. Wu, Y. Chen and J. Han : Re-Examination of Interestingness Measures in Pattern Mining: A Unified Framework, *Data Mining and Knowledge Discovery*, Vol.21, No.3, pp.371–397, 2010.

[8] P. Lenca, B. Vaillant, P. Meyer and S. Lallich : Association Rule Interestingness Measures: Experimental and Theoretical Studies, In F. J. Guillet and H. J. Hamilton (eds) *Quality Measures in Data Mining*, pp.51–76, 2007.

[9] Y. Bengio, A. Courville and P. Vincent : Representation Learning: A Review and New Perspectives, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.35, No.8, pp.1798–1828, 2013.

[10] S. Kawanobe and T. Ozaki : Experimental Study of Characterizing Frequent Itemsets using Representation Learning, *Proc. of the 2018 32nd International Conference on Advanced Information Networking and Applications Workshops*, pp.170–174, 2018.

[11] J. Pennington, R. Socher and C. D. Manning : GloVe: Global Vectors for Word Representation, *Proc. of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp.1532–1543, 2014.

[12] T. Mikolov, K. Chen, G. Corrado and J. Dean : Efficient Estimation of Word Representations in Vector Space, arXiv preprint, arXiv:1301.3781, 2013.

[13] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado and J. Dean : Distributed Representations of Words and Phrases and their Compositionality, *Advances in Neural Information Processing Systems 26*, pp.3111–3119, 2013.

[14] P. Bojanowski, E. Grave, A. Joulin and T. Mikolov : Enriching Word Vectors with Subword Information, arXiv preprint, arXiv:1607.04606, 2016.

[15] D. Nguyen, T. D. Nguyen, W. Luo and S. Venkatesh : Trans2Vec: Learning Transaction Embedding via Items and Frequent Itemsets, *Proc. of the 22nd Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp.361–372, 2018.

[16] Q. Le and T. Mikolov : Distributed representations of sentences and documents, arXiv preprint, arXiv:1405.4053v2, 2014.

[17] S. Kawanobe and T. Ozaki : Extraction of Characteristic Frequent Visual Patterns by Distributed Representation, *Proc. of the 2017 31st International Conference on Advanced Information Networking and Applications Workshops*, pp.525–530, 2017.

[18] H. Xiong, P.-N. Tan and V. Kumar : Mining Strong Affinity Association Patterns in Data Sets with Skewed Support Distribution, *Proc. of 3rd IEEE International Conference on Data Mining*, pp.387–394, 2003.

[19] H. Xiong, P.-N. Tan and V. Kumar : Hyperclique Pattern Discovery, *Data Mining and Knowledge Discovery*, Vol.13, No.2, pp.219–242, 2006.

[20] J. Cheng, Y. Ke and W. Ng: δ-Tolerance Closed Frequent Itemsets. *Proc. of the 6th International Conference on Data Mining*, pp.139-148, 2006.