2F2-05

観点に依存したクラスタリング手法に関する一考察

A Framework of Viewpoint-Based Clustering

大久保 好章　　原口 誠
Yoshiaki OKUBO　Makoto HARAGUCHI

北海道大学大学院情報科学研究科コンピュータサイエンス専攻
Graduate School of Information Science and Technology, Hokkaido University

In this paper, we propose a framework of Viewpoint-Based Clustering for the set of items in a transaction database. Our cluster, called a *star cluster*, is constructed based on *Self Mutual Information* (SMI) between items. Intuitively speaking, SMI evaluates a degree of correlation between two items e_i and e_j, where e_j is considered to be correlated with e_i in the sense that occurrence of e_j affects occurrence of e_i. Making a point of an item, called a *core*, our cluster consists of items each of which and the core gives their SMI higher than a threshold. That is, the items in the cluster are gathered under the core. Considering such a core as a viewpoint, we can extract clusters depending on a given viewpoint. Thus, we can find similarity among the items in the sense that each item is correlated with the core in a certain degree. This kind of similarity would be novel and increase the possibility of finding interesting clusters which cannot be obtained by traditional methods such as K-means and Nearest-Neighbor.

1. はじめに

本稿では，データマイニングの主要技術のひとつである**クラスタリング** [?] について考察する．

クラスタリングとは，所与のデータ集合から，類似したもの同士から成るグループ (**クラスタ**と呼ぶ) を抽出するタスクである．一般に，データの捉え方には多様性があり，ユーザの観点や目的に従って，データ間に多種多様な類似性を考えることができる．データを様々な角度から分析・解析・理解するためには，こうした多様な観点あるいは類似性を柔軟かつ陽に制御可能なクラスタリングの枠組が必要不可欠であるが，残念ながら，既存手法においてそれが実現されているとは言い難い．我々をとりまく膨大なデータを有効に活用するためには，この様な柔軟なクラスタリング手法の提案が強く望まれる．

以上の問題意識のもと，本稿では，所与のトランザクションデータベースに出現するアイテムの集合からクラスタを抽出する問題を考える．特に，**自己相互情報量**に基づくアイテム間の**相関度**に注目し，ある注目アイテム e^* とは δ 以上の相関を有し，それ以外のアイテムとは高々 γ の相関しか示さないアイテム群を，ひとつのクラスタとして抽出する枠組を提案する．すなわち，ここでのクラスタは，注目する e^* に依存して，構成アイテムが変化する動的なものとなり，e^* を観点と見立てることで，観点に依存したクラスタ抽出のひとつの実現となっている．

2. 準備

アイテムの集合を $\mathcal{I} = \{e_1, \ldots, e_n\}$ とした時，$I \subseteq \mathcal{I}$ なる I を**アイテム集合** (itemset) と呼ぶ．アイテム集合 I と識別子 id の組 $t = (id, I)$ を**トランザクション** と言い，トランザクションの集合 $\mathcal{D}$ を**トランザクションデータベース**と呼ぶ．

アイテム集合 I の**頻度**を $freq(I)$ で表し，

$$freq(I) = |\{t | t = (id, J) \in \mathcal{D} \wedge I \subset J\}|$$

と定める．以下では簡単のため，アイテム集合 $\{e_i\}$ や $\{e_i, e_j\}$ を e_i や e_ie_j と略記する．

連絡先: 大久保 好章
北海道大学大学院情報科学研究科コンピュータサイエンス専攻，〒 060-0814 札幌市北区北 14 条西 9 丁目, TEL : 011-706-7161，E-mail: yoshiaki@ist.hokudai.ac.jp

3. アイテム間の相関度

アイテムの集合 $\mathcal{I}$ とトランザクションデータベース $\mathcal{D}$ を考える．アイテム $e_i, e_j \in \mathcal{I}$ について，$\mathcal{D}$ における e_i と e_j の**相関度**を $correl(e_i, e_j)$ と表し，

$$\begin{aligned} correl(e_i, e_j) &= -\log_2 \frac{freq(e_i)}{|\mathcal{D}|} - (-\log_2 \frac{freq(e_ie_j)/|\mathcal{D}|}{freq(e_j)/|\mathcal{D}|}) \\ &= \log_2 \frac{freq(e_ie_j)|\mathcal{D}|}{freq(e_i)freq(e_j)}. \end{aligned} \quad (1)$$

と定める．いま，$\mathcal{D}$ から任意にトランザクションをひとつ取り出す試行を考える．取り出されたトランザクション中にアイテム e_i が出現する事象を E_i とすると，$\frac{freq(e_i)}{|\mathcal{D}|}$ は E_i の事前確率を，$\frac{freq(e_ie_j)/|\mathcal{D}|}{freq(e_j)/|\mathcal{D}|}$ は E_j が与えられたもとでの E_i の事後 (条件付き) 確率を表す．式 (**??**) は，E_i と E_j の**自己相互情報量** (*Self Mutual Information: SMI*) であり，この値をもって，e_i と e_j 間の相関の度合を測るものとする．E_j の生起が E_i の生起に何らかの影響を及ぼす場合は，E_i の事前・事後確率に差が観測されるであろう．特に，E_i がより生起しやすくなる場合は $correl(e_i, e_j)$ が正の値をとり，生起しにくくなる場合は負の値をとる．また，E_i と E_j が統計的に独立な場合は，$correl(e_i, e_j)$ は 0 となる．

次節では，こうした相関度に基づく新たなクラスタについて議論する．

4. スタークラスタリング

先に述べた通り，クラスタは，何らかの観点や目的のもとに集まったデータから成るはずであるが，従来のクラスタリング手法では，これら観点や目的は陽に扱われてこなかった．そのため，得られたクラスタを明確に解釈することは容易ではない．本節ではこうした点を改善するべく，新たなクラスタを提案する．特にここでは，所与のトランザクションデータベース $\mathcal{D}$ に出現するアイテムの集合 $\mathcal{I}$ からクラスタを抽出する問題を考える．

4.1 スタークラスタ

$\mathcal{I}$ 中のあるアイテム e^* に注目する．ここでは，e^* を観点に見立て，e^* に依存して集まるアイテムの集合をクラスタと考える．如何なる意味でアイテムが集まるかについては，様々な基準が考えられるが，本稿ではアイテム間の相関度に基づく基準を提案する．

簡単に述べると，クラスタを構成するアイテムは，その出現に関して互いにある度合以下の影響しか受けないが，e^* だけからはある度合以上の影響を受ける．観点となる e^* の変化に伴い，そのもとに集まることのできるアイテムも変化することから，観点に依存し，かつ，明確に解釈可能なクラスタの抽出が実現できる．クラスタ中のアイテムが e^* を中心として集まる様子をスター構造に見立て，こうしたクラスタを**スタークラスタ**と呼ぶ．アイテム間の相関度を用いてスタークラスタを以下の通り定義する．

観点として注目するアイテム $e^* \in \mathcal{I}$ を**核** (*core*) と呼ぶ．核との相関度下限値を δ, 核以外のアイテムとの相関度上限値を γ とする．$\mathcal{I}$ の部分集合 S ($\not\ni e^*$) が

$$\forall e \in S,\quad correl(e, e^*) \geq \delta \text{ かつ } \forall e_i, e_j \in S (i \neq j),\quad correl(e_i, e_j) \leq \gamma.$$

を満たす時，S を e^* **に関する** (δ,γ)-**スタークラスタ**と呼ぶ．

言うまでもなくこうしたクラスタは，従来の代表的な K-Means 法や Nearest-Neighbor 法では得ることのできないものである．また，データ間に明示的な距離を導入する必要がないという点においても，これらの手法とは大きく異なる．

4.2 スタークラスタリング問題

定義より，(δ,γ)-スタークラスタの部分集合は，同様に (δ,γ)-スタークラスタであることがわかる．よって，集合の包含関係の意味で極大なクラスタのみを抽出の対象とする．以上のことは，次のスタークラスタリング問題としてまとめられる．

Input： トランザクションデータベース $\mathcal{D}$ ($\mathcal{I}$), 核アイテム $e^* \in \mathcal{I}$, 許容相関度 δ および γ

Find： 極大 (δ,γ)-スタークラスタ.

実際には，すべての極大クラスタを求める場合や，サイズが最も大きな極大クラスタを求める場合等，様々な出力を考えることができる．

5. スタークラスタの抽出

本節では，スタークラスタの抽出手順の概略について述べる．

5.1 許容相関度に基づくグラフ構築

核との相関度下限値を δ，核以外のアイテムとの相関度上限値を γ とする．(δ,γ)-スタークラスタの抽出にあたっては，まず，アイテムの集合 $\mathcal{I}$ を頂点集合とするふたつのグラフ $G_\delta = (\mathcal{I}, E_\delta)$ と $G_\gamma = (\mathcal{I}, E_\gamma)$ を構築する．ここで，

$$E_\delta = \{(e_i, e_j) | e_i, e_j \in \mathcal{I} \wedge correl(e_i, e_j) \geq \delta\},$$
$$E_\gamma = \{(e_i, e_j) | e_i, e_j \in \mathcal{I} \wedge correl(e_i, e_j) \geq \gamma\}$$

とする．

5.2 クラスタの抽出

(無向) グラフ $G = (V, E)$ において，頂点集合 $V' (\subseteq V)$ に誘導される G の部分グラフを $G(V')$，頂点 v と隣接する頂点集合を $N_G(v)$ と表す．

核 e^* に関する極大 (δ,γ)-スタークラスタは，$N_{G_\delta}(e^*)$ に誘導される G_γ の部分グラフ，すなわち，$G_\gamma(N_{G_\delta}(e^*))$ における**極大安定集合** (*maximal stable set*) として与えられる．グラフの安定集合は，その補グラフの**クリーク**であることから，極大安定集合の抽出には，既存の極大クリーク抽出アルゴリズムが利用可能である．例えば，頂点数の意味で最大クリークを抽出する効率的な分枝限定アルゴリズム [**?**] や，すべての極大クリークを高速に列挙するアルゴリズム [**?**] が提案されている．

6. 課題

本稿で提案するスタークラスタは，文献 [**?**] で議論された同概念の特殊ケースに相当する．文献 [**?**] では,『各種ビール銘柄の味に対する印象』のアンケート結果をトランザクションデータベースと捉え，スタークラスタの抽出実験を行なっている．そこでは,『男性っぽい』と『高級な感じ』という一見相関がありそうに思えない印象が,『本格的』という印象をコアとして同じクラスタに属するという興味深い結果が得られている．こうしたクラスタは，製品開発における極めて重要な情報を与えると思われ，その有効性に期待が持てよう．本稿での枠組においても，計算機実験による有効性の確認を早急に行ないたい．また，因子分析や数量化理論といった統計的手法により得られるクラスタとの関連についても，さらに詳細な考察を行なう必要がある．

7. おわりに

本稿では，従来の手法では得られない新たなスタークラスタの概念を提案し，その抽出手順の概略を述べた．スタークラスタは，入力として与えるコアに依存して構成要素が変化する動的なクラスタであり，アイテムの集合を様々な角度から分析・解析するための柔軟な枠組みになるものと期待している．

現在，本手法の有効性を実験的に示すためのシステムを実装中である．種々の具体的データに対して得られるスタークラスタをより詳細に分析し，その有効性を確認したい．計算機実験の結果については稿を改めて述べたい．

参考文献

[Jain 88] A. K. Jain and R. C. Dubes: “Algorithms for Clustering Data”, Prentice Hall, 1988.

[Tomita 03] E. Tomita and T. Seki: “An Efficient Branch-and-Bound Algorithm for Finding a Maximum Clique”, Proc. of the 4th Int’l Conf. on Discrete Mathematics and Theoretical Computer Science, Springer-LNCS 2731, pp. 278-289, 2003.

[宇野 03] 宇野 毅明：“大規模グラフに対する高速クリーク列挙アルゴリズム”，電子情報通信学会技術研究報告, Vol. 103, No.31 (COMP2003 1-8)，pp. 55-62, 2003.

[松本 03] 松本 健太郎：“自己相互情報量を用いたスタークラスタリング法の提案”，北海道大学大学院工学研究科修士論文，2003.