



Title	タクソノミを取り入れた顕在パターンを利用したPOSデータの分析
Author(s)	森田, 裕之; 羽室, 行信
Citation	2010年度科学技術振興機構ERATO湊離散構造処理系プロジェクト講究録. p.437-439.
Issue Date	2011-06
Doc URL	http://hdl.handle.net/2115/48343
Type	conference presentation
Note	ERATO湊離散構造処理系プロジェクト: 2010年度初冬のワークショップ (ERATO合宿). 2010年11月29日 (月) ~ 12月1日 (水). 札幌北広島クラッセホテル.
File Information	13.morita.pdf



[Instructions for use](#)

タクソミを取り入れた顕在 パターンを利用したPOSデー タの分析

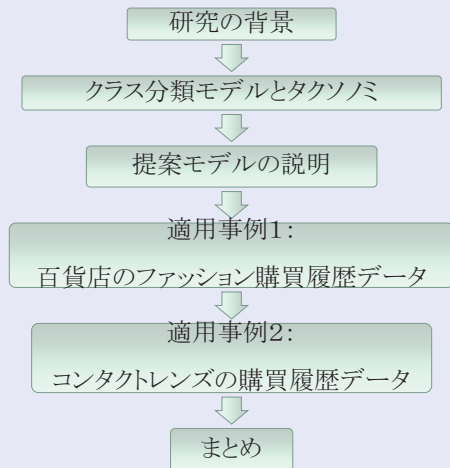
大阪府立大学経済学部 森田裕之
関西学院大学大学院経営戦略研究科 羽室行信

ERATO 湊離散構造処理系プロジェクト 2010年11月29日~12月1日合宿 (北広島)

はじめに

- ◆ LCMやZDDを利用することで、頻出パターンを活用した分類モデルを提案
- ◆ 出現する頻出パターンを上手に集約して、説明要因とすることは、実用的にも魅力的(タクソミの導入)
- ◆ 実際の百貨店のファッション購買、およびコンタクトレンズ購買履歴データに対して適用

発表の流れ



研究の背景

- ◆ LCMの高速な頻出パターン列挙は魅力的で分類モデルの説明要因となる候補パターンを増大
 - ◆ パターンの大量列挙は、より説明力をもつパターンを効率的に利用する方法を要求
 - ◆ その1つのアプローチとしてタクソミを取り入れたパターンの利用を検討
- ⇒解釈が容易でより強力な分類力を持つパターンを利用して分類モデルを構築

関連研究: 頻出パターンを応用した分類モデル

- 顕在パターン: サポート比の大きなパターン

$$GR_{C_1}(\pi) = \frac{SP(\pi, C_1)}{SP(\pi, C_2)}$$

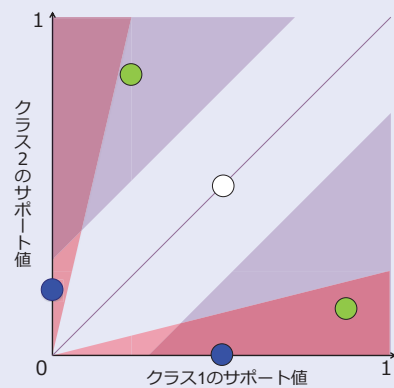
Guozhu Dong, Jinyan Li: Efficient Mining of Emerging Patterns: Discovering Trends and Differences. KDD 1999: pp.43-52.

- コントラストパターン: サポートの差の大きなパターン

$$DF(\pi) = |SP(\pi, C_1) - SP(\pi, C_2)|$$

Stephen D. Bay, Michael J. Pazzani: Detecting Change in Categorical Data: Mining Contrast Sets. KDD 1999: pp.302-306.

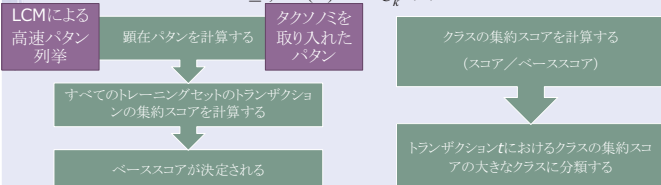
コントラストパターンと顕在パターン



顕在パターンを使った分類器: CAEP

- あるトランザクションt、クラスC_kの顕在パターン集合をE(C_k)とするとtのC_kにおけるスコアscore(t, C_k)を以下のように定める

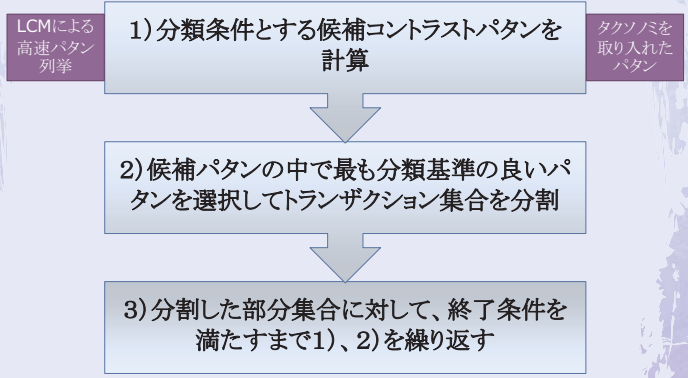
$$score(t, C_k) = \sum_{e \in t, e \in E(C_k)} \frac{GR_{C_k}(e)}{GR_{C_k}(e)+1} \times SP(e, C_k)$$



G. Dong, X. Zhang, L. Wong, J. Li. CAEP: Classification by Aggregating Emerging Patterns. Int'l Conference on Discovery Science, 1999.

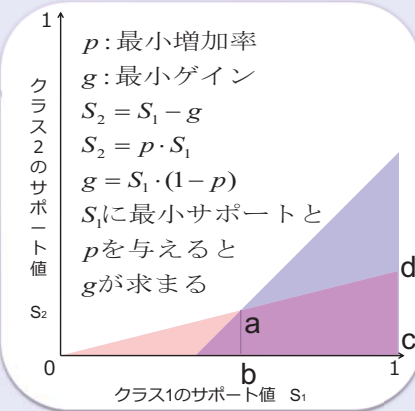
IGVcaep および IGVcaep改として開発

コントラストパターンを使った決定木モデル



IGVdtcp および IGVdtcp改として開発

CPとEPの関係

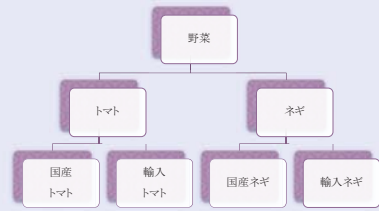


CPはLCMで直接計算できる

EPはLCMでCPを計算したのち不要な領域を削除する

タクソノミによるパターンの集約

- タクソノミとは: ツリー型で分類的な意味を持つ親子関係



- 方法 1) 既存タクソノミを専門知識から抽出してデータに与える(経験則と手法のハイブリッド)
- 方法 2) タクソノミを自動的に構築する

タクソノミの利用イメージ

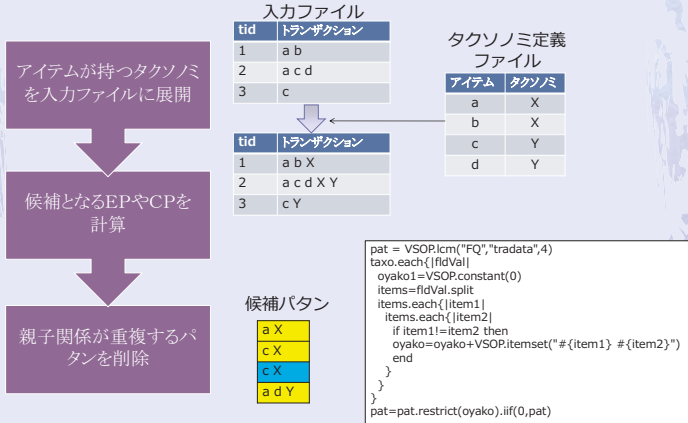
大分類	小分類	アイテム	お客さんクラス	レシート番号	同時購買アイテム	分類表現
野菜	国産野菜	A 国産トマト 1kg	普通のお客さん	1	G,H,M	β, δ
		B 国産トマト500g		2	B,N	α, δ
		C トマト (桃太郎) 300g		3	F,N	β, δ
		D トマト (桃太郎) 1箱	4	A,B,K	α, γ	
		E ハニーエイト (高糖度トマト) 350g	5	C,L	α, γ	
	輸入野菜	F 輸入トマトX 300g	よいお客さん	6	E,I	α, γ
		G 輸入トマトX 1kg				
		H 輸入トマトY 300g				
		I 岩津ねぎ 1袋 (300g)				
		J 青ねぎ (消費ねぎ) 大束 (200g)				
ねぎ	国産ねぎ	K 九条ねぎ 180g				
		L 白ねぎカット 1本分				
	輸入ねぎ	M 輸入ねぎZ 180g				
		N 輸入ねぎZ 360g				

各アイテムのパターンに共通性はないが、タクソノミのパターンでは共通性が見つかる場合がある

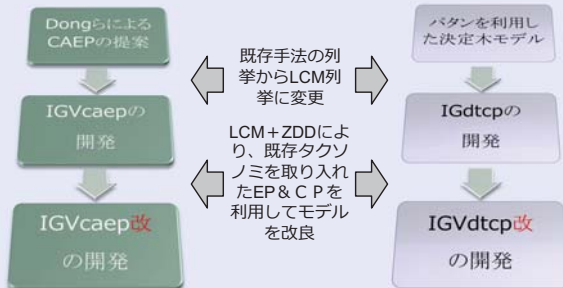
健在パターン
普通のお客さん: β, δ
よいお客さん: α, γ

個別のアイテムの購買が、良いお客さんの特徴ではなく、国産の野菜を購入することが特徴

タクソノミの処理方法



既存の手法と提案手法の違い



平成20年度、21年度情報大航海プロジェクトでビジネスマイニング研究センター
有限責任事業組合にて開発

アプリケーションの観点から

- ◆ 小売業では、初回または2回目の購買機会
で顧客レベルを判断することが重要(百貨店、
コンタクトレンズなど)
- ◆ 購買機会が少ない⇒購買行動をつかみにく
い
- ◆ 少ない購買行動から、より食欲に説明要因を
マイニングしたい
- ◆ 単純に増やすだけでは、説明力になりにくい
場合がある⇒タクソミの利用

百貨店の購買履歴データを 利用した適用事例

(分析内容のシートについては、データ提供元の関係から削除しています)

コンタクトレンズの購買履歴データを 利用した適用事例

(分析内容のシートについては、データ提供元の関係から削除しています)

まとめ

- ◆ 既存技術に、タクソミ取り入れたEPやCPを取
り入れた分類モデルを提案
- ◆ 2つの実データに適用して有効性を確認
- ◆ 問題によって、CPを利用した方がよいケースと
EPを利用した方がよいケースがある
- ◆ タクソミを自動的に生成して、取り入れるメカ
ニズムを導入したい
- ◆ 出現した顕在パターンやコントラストパターンを、より
効率的に利用したモデルを開発したい

参考文献リスト

- ◆ T. Uno, T. Asai, Y. Uchida, and H. Arimura, "LCM: An efficient algorithm for enumerating frequent closed item sets," in *In Proceedings of Workshop on Frequent Itemset Mining Implementations (FIMI'03)*, 2003.
- ◆ T. Nakahara and H. Morita, "Pattern mining in pos data using a historical tree," in *In Proceedings of Sixth IEEE International Conference on Data Mining - Workshops (ICDMW'06)*, 2006, pp. 570-574
- ◆ H. Alhammady and K. Ramamohanarao, "Using emerging patterns to construct weighted decision trees," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 7, pp. 865-876, 2006.
- ◆ Guozhu Dong, Jinyan Li: Efficient Mining of Emerging Patterns: Discovering Trends and Differences. KDD 1999: pp.43-52.
- ◆ Stephen D. Bay, Michael J. Pazzani: Detecting Change in Categorical Data: Mining Contrast Sets. KDD 1999: pp.302-306
- ◆ H. Morita, T. Nakahara, Y. Hamuro, S. Yamamoto, "Decision Tree-based Classifier Incorporating Contrast Pattern", *Proceedings of the 13th IEEE International Symposium on Consumer Electronics*, 2009.5.25-28, Kyoto, Japan
- ◆ S. Minato, T. Uno, and H. Arimura: "LCM over ZBDDs: Fast GeneSupn of Very Large-Scale Frequent Itemsets Using a Compact Graph-Based Representation," *In Proc. of 12th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2008)*, (LNAI 5012, Springer), pp. 234-246, May 2008