頻繁に更新されるデータのための形式概念解析を用いたマルチラベル分類手法

池田 真土里, 山本 章博
京都大学 情報学研究科

目標

- データが頻繁に更新されることなどを理由に，複数の異なる教師データに対してデータセットを繰り返し分類する際の全体の時間削減
  - Web上の文書，動画などのタグ付コンテンツ

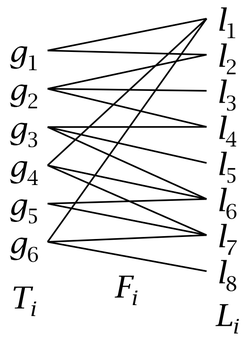
問題の定式化

- 形式文脈(2値ベクトルの集合)で表されるデータセット(G, M, I)と複数の教師データ(T_1, L_1, F_1), (T_2, L_2, F_2), …が与えられるとき，各教師データ(T_i, L_i, F_i)について未知データ$u \in G \setminus T_i$にラベル集合$\hat{F}_i(u) \subseteq L_i$を与える関数$\hat{F}_i : G \to 2^{L_i}$を求める
  - 形式文脈(G, M, I)
    - G: データ(オブジェクト)集合, M: 特徴(属性)集合, $I \subseteq G \times M$
  - 教師データ(T_i, L_i, F_i)
    - $T_i \subseteq G$, L_i: ラベルの集合, $F_i : T \to 2^{L_i}$

(G, M, I)

	m_1	m_2	m_3	m_4	m_5	m_6	m_7
g_1	×	×					
g_2	×	×		×			
g_3	×	×		×			
g_4		×			×	×	
g_5		×			×	×	
g_6			×		×		×
u		×		×		×	

(T_i, L_i, F_i)



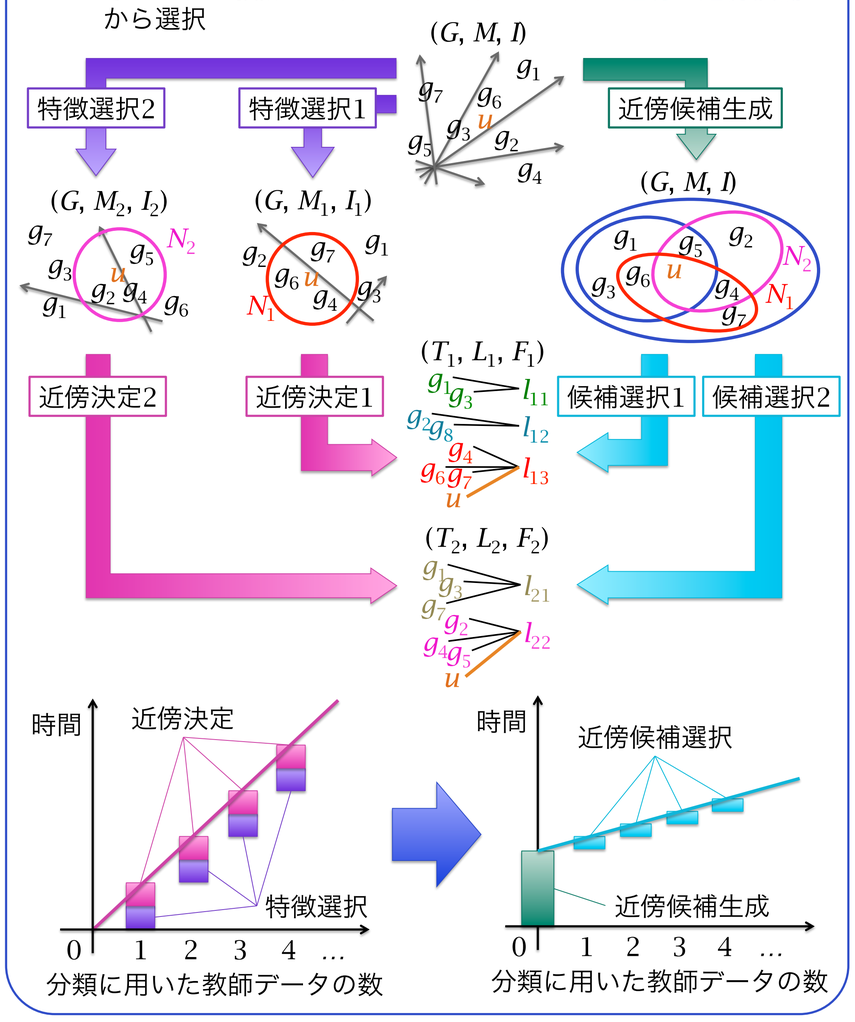
成果と今後の課題

- 成果
  - 形式概念解析を用いた近傍法によりマルチラベル分類問題を解く手法
    - 形式概念を近傍候補とみなし概念束により管理
    - 各教師データに対して行われる特徴選択を回避することで全体の時間を削減
  - 特徴選択と既存のマルチラベル分類手法の組み合わせと比べ，おおむね高精度であり，データによっては高速
- 今後の課題
  - 単調性のあるスコア関数の利用による高速化

提案手法

キーアイデア

- 各教師データ(T_i, L_i, F_i)について，データセット(G, M, I)に対して行われる特徴選択を回避することで全体の時間を削減
  1. 未知データ$u \in G \setminus T_i$の好ましい近傍$N_i \subseteq T_i$になりそうな近傍候補を教師データ(T_i, L_i, F_i)を用いずにあらかじめ生成
  2. 与えられた教師データについて未知データの近傍を近傍候補から選択



形式概念解析を用いたマルチラベル分類手法

1. データセット(G, M, I)から概念束$\underline{\mathcal{B}}(G, M, I)$を構築
2. 教師データ(T, L, F)を用いて，未知データ$u \in G \setminus T$の近傍候補を表す形式概念$c \in {\uparrow}\gamma u$のスコア$\sigma(c)$を計算

$$\sigma(c) = \begin{cases} 0 & \text{if } |\mathrm{Ex}(c) \cap T| = 0, \\ 1 & \text{if } |\mathrm{Ex}(c) \cap T| = 1, \\ \dfrac{\sum_{i=1}^{|\mathrm{Ex}(c) \cap T|} \sum_{j=i+1}^{|\mathrm{Ex}(c) \cap T|} \frac{|F(g_i) \cap F(g_j)|}{|F(g_i) \cup F(g_j)|}}{\binom{|\mathrm{Ex}(c) \cap T|}{2}} & \text{otherwise} \end{cases}$$

3. スコア$\sigma(c)$が最大の形式概念cの外延に含まれる既知データ$g \in \mathrm{Ex}(c) \cap T$からなる集合を未知データ$u$の近傍$N \subseteq T$とする
4. 近傍Nを用いて未知データuにラベル集合$\hat{F}(u)$を与える

$$\hat{F}(u) = \{l \in L \mid |\{g \in N \mid l \in F(g)\}| > |\{g \in N \mid l \notin F(g)\}|\}$$

形式概念解析 [3, 4]

- 形式概念$c = (X, Y)$
  - $X^I = Y$, $X = Y^I$ for $X \subseteq G$, $Y \subseteq M$
    - $X^I := \{m \in M \mid \forall g \in G.\ (g, m) \in I\}$ for $X \subseteq G$
    - $Y^I := \{g \in G \mid \forall m \in M.\ (g, m) \in I\}$ for $Y \subseteq M$
  - 外延$\mathrm{Ex}(c) := X$, 内包$\mathrm{In}(c) := Y$
  - オブジェクト概念$\gamma g := (\{g\}^{II}, \{g\}^I)$ for $g \in G$
- 概念束$\underline{\mathcal{B}}(G, M, I)$
  - $c \le c'$: 形式概念c, c'について$\mathrm{Ex}(c) \subseteq \mathrm{Ex}(c')$
  - ${\uparrow}c := \{c' \in \underline{\mathcal{B}}(G, M, I) \mid c \le c'\}$, $c \in \underline{\mathcal{B}}(G, M, I)$の上方集合

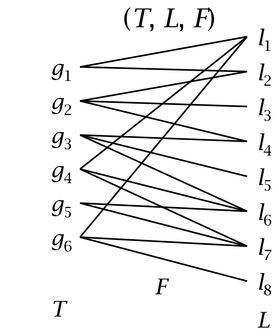
計算量

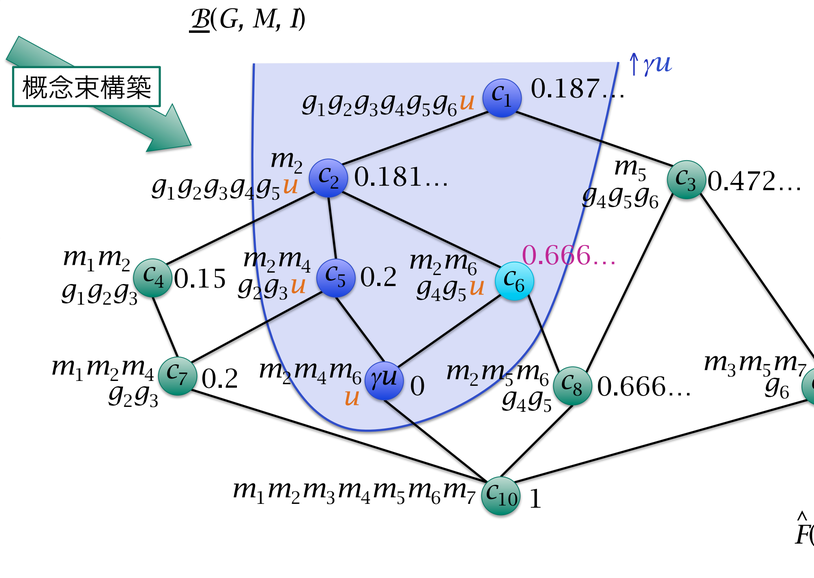
- 未知データuの$\hat{F}(u)$を求める計算量: $O(|{\uparrow}\gamma u| N_E^2 + N_E N_L)$
  - N_E: 外延の平均要素数, N_L: 既知データの平均ラベル数

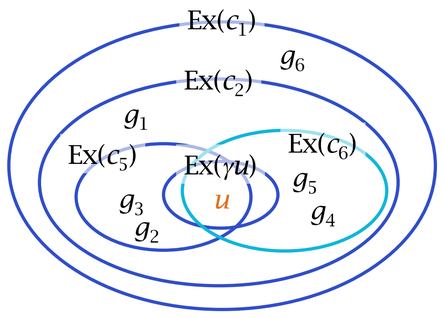
(G, M, I)

	m_1	m_2	m_3	m_4	m_5	m_6	m_7
g_1	×	×					
g_2	×	×			×		
g_3	×	×			×		
g_4		×				×	×
g_5		×				×	×
g_6			×		×		×
u		×		×		×	

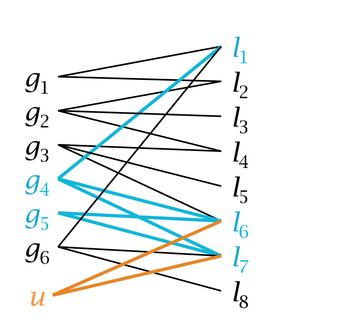
(T, L, F)







$\hat{F}(u) = \{l_6, l_7\}$



実験

比較手法

- kNN(k, f): 情報ゲインによる特徴選択とk近傍法に基づく分類 [6]
  - 上位f%の特徴, k個の近傍
- NB(f): 情報ゲインによる特徴選択とナイーブベイズ法に基づく分類 [8]
  - 上位f%の特徴

※パラメータは$k \in \{1, 2, \ldots, 30\}$, $f \in \{10, 20, \ldots, 100\}$から評価関数に対して最良のものを設定

- 情報ゲイン [2] を用いた特徴選択
  - 情報ゲインIGにより特徴をランキング，下位を削除

$$IG((G, M, I), m \in M) = H(G) - \left(\frac{|G_m|}{|G|} H(G_m) + \frac{|G_{\neg m}|}{|G|} H(G_{\neg m})\right)$$

※ $G_m = \{g \in G \mid (g, m) \in I\}$, $G_{\neg m} = \{g \in G \mid (g, m) \notin I\}$

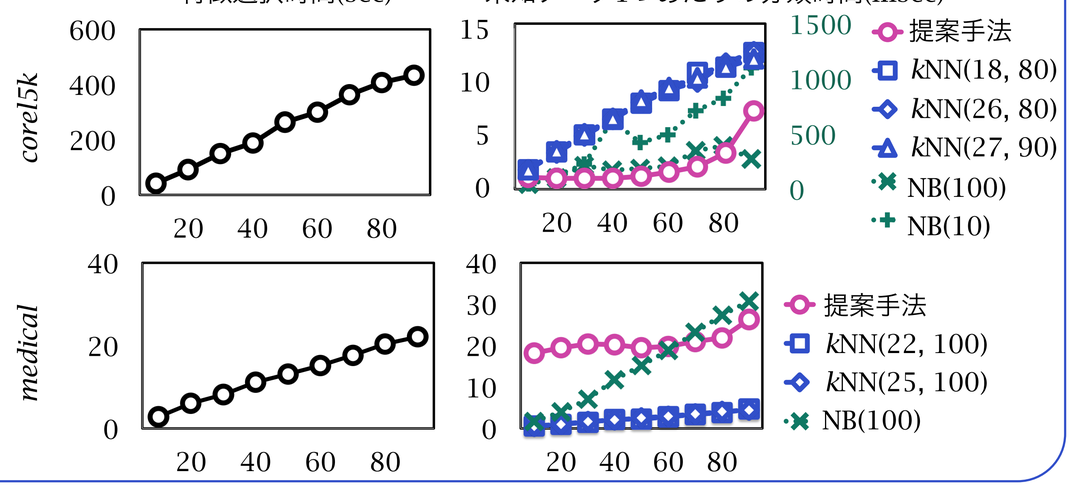
実行環境と実験データ

- 実行環境: Mac OS X 10.7.5, 2 × 2.66GHz 6-Core Intel Xeon, 64GB 800 MHz DDR3, Python 2.7.1, GCC 4.2.1
- データ: *corel5k*, *medical* [1]

(G, M, I), (G, L, F)	*corel5k*	*medical*
$\lvert G\rvert$	5,000	978
$\lvert M\rvert$	499	1449
$\lvert L\rvert$	374	45
$\Sigma_g \lvert g^I\rvert / \lvert G\rvert$	8.27	13.40
$\Sigma_g \lvert F(g)\rvert / \lvert G\rvert$	3.52	1.25
$\lvert\underline{\mathcal{B}}(G, M, I)\rvert$	56,199	45,649
$\Sigma_c \lvert\mathrm{Ex}(c)\rvert / \lvert\underline{\mathcal{B}}(G, M, I)\rvert$	3.98	6.27
$\Sigma_g \lvert{\uparrow}\gamma g\rvert / \lvert G\rvert$	44.72	292.51
概念束構築時間(sec)	1553.1	1333.8

※ Gから10, 20, …, 90%のデータを既知として教師データを作成し各教師データについて分類を行う

特徴選択時間(sec)　未知データ1つあたりの分類時間(msec)



精度

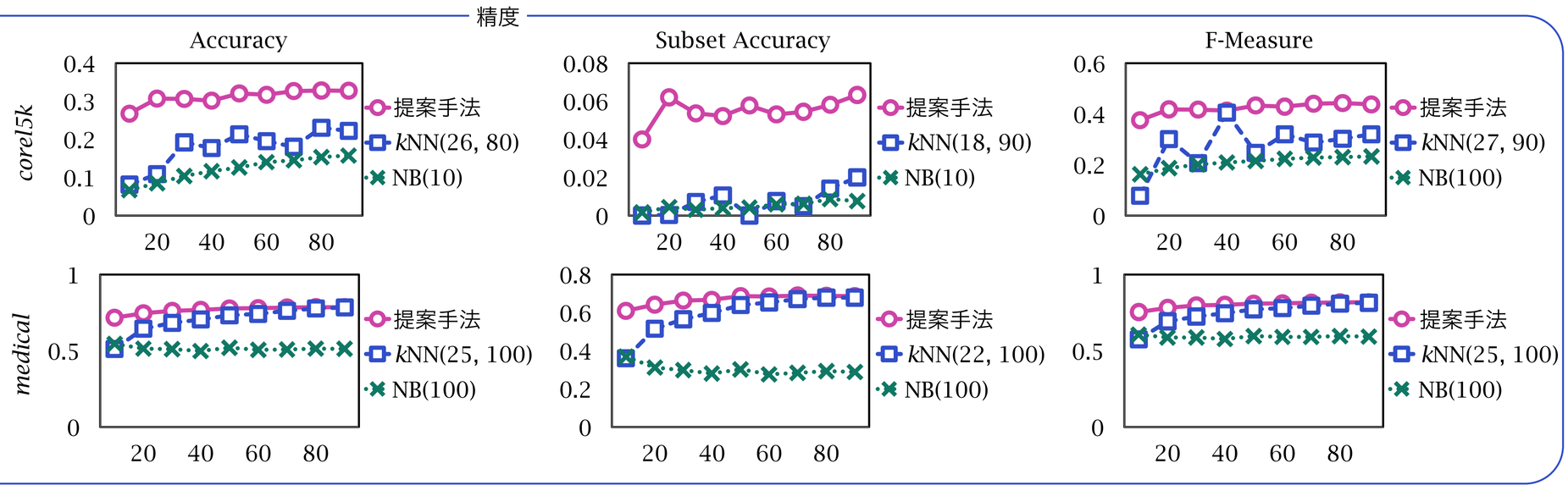
- 評価関数 [5, 7]

$$Accuracy(\hat{F}, G) = \frac{1}{|G|} \sum_{g \in G} \frac{|\hat{F}(g) \cap F_*(g)|}{|\hat{F}(g) \cup F_*(g)|}$$

$$Subset\_Accuracy(\hat{F}, G) = \frac{1}{|G|} |\{g \in G \mid \hat{F}(g) = F_*(g)\}|$$

$$F\text{-}Measure(\hat{F}, G) = \frac{1}{|G|} \sum_{g \in G} \frac{2|\hat{F}(g) \cap F_*(g)|}{|\hat{F}(g)| + |F_*(g)|}$$

※ $F_*(g)$: gの真のラベル集合



参考文献

1. Alcalá-Fdez, J., Fernandez, A., Luengo, J., Derrac, J., García, S., Sánchez, L., Herrera, H.: KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework. Journal of Multiple-Valued Logic and Soft Computing 17:2-3 255–287 (2011)
2. Clare, A., King, R., D.: Knowledge Discovery in Multi-label Phenotype Data. PKDD' 01 Proc. of the 5th European Conference on Principles of Data Mining and Knowledge Discovery, pp. 42–53 (2001)
3. Davey, B., A., Priestly, H., A.: Introduction to Lattice and Order. Cambridge University Press (2002)
4. Ganter, B., Wille, R.: Formal Concept Analysis: Mathematical Foundations. Springer-Verlag New York, Inc. (1999)
5. Spolaôr, N., Cherman, E., A., Monard, M., C., Lee, H., D.: A Comparison of Multi-label Feature Selection Methods using the Problem Transformation Approach. Electronic Notes in Theoretical Computer Science, vol. 292, no. 0, pp. 135–151 (2013)
6. Spyromitros, E., Tsoumakas, G., Vlahavas, I.: An Empirical Study of Lazy Multilabel Classification Algorithms. SETN '08 Proc. of the 5th Hellenic conference on Artificial Intelligence: Theories, Models and Applications, pp. 401– 406 (2008)
7. Tsoumakas, G., Katakis, I., Vlahavas, I.: Mining multi-label data. In Data Mining and Knowledge Discovery Handbook, pp.667–685 Springer (2010)
8. Wei, Z., Zhang, H., Zhang, Z., Li, W., Miao, D.: A Naïve Bayesian Multi-label Classification Algorithm With Application to Visualize Text Search results. International Journal of Advanced Intelligence, vol. 3, no. 2, pp. 173-188 (2011)