社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2007－MPS－63 (19)
2007／3／4

レート歪み学習に基づくマイノリティ集合の分布推定

安藤 晋†

† 横浜国立大学工学研究院

本稿は混合推定を基にしたデータベース中の例外的な事例集合を検出する手法を提案する．提案手法は全体からのダイバージェンスを最大となるような部分集合の分布を推定する．提案手法は代表事例や平均から遠い事例を例外とみなす従来手法と異なり特徴空間内での明示的な分離が困難な場合にも有効で，テキスト分類と画像分類の実験において有望な結果を示した．

Estimation of Minority with Rate-Distortion Learning

Shin ANDO†

† Faculty of Engineering, Yokohama National University

This paper addresses the problem of detecting minority instances using mixture estimation. The proposed method maximizes the divergence between the distributions of the minority and the majory in the dataset. The proposed method is effective for detecting an atypical subset which overlaps with the majority. Good performances are observed in the text classification and image classification benchmarks.

1 はじめに

近年，様々な分野で大規模なデータベースが構築され，大量のデータから知識を抽出する手法の重要性が高まっている．クラスタリングを始めとする教師無し学習はデータセットの特徴を把握する上で有力な手段であるが，大規模なデータセットの全ての事例を明確に特徴付けることは必ずしも必要ではなく，計算量の面から困難な場合も多い．

サーチエンジンでは数ページ程度の関連文書を抽出することが求められるが，扱うドキュメントの総数が膨大であるため，細かい分析をデータベース全体に適用することは難しい．このようなデータセットを扱う場合，その部分集合を取り出すことで明確な特徴・傾向を把握することが容易になる．

本研究はこのようなマイノリティとみなせる部分集合の分布を推定する混合分布推定問題を考える．マイノリティは大部分の事例と大きく異なる特徴を持つ部分集合を指す．便宜的にマイノリティ以外の事例はマジョリティと呼ぶ．ただし，観測事例数が大きく異なる混合では最尤法によるパラメータ推定が難しい．本質的に，小さな部分集合のモデルを学習する場合汎化が強いと検出対象が無視されてしまい，汎化が弱いと過適合した場合との見分けが難しくなるという問題がある．そこで本研究では，尤度に替えてマジョリティからのダイバージェンスの最大化に基づくマイノリティ分布の推定を提案する．分布の差を最大化することでマジョリティ事例と大きく重複するようなマイノリティ集合も検出することが可能になる．提案手法の理論的な基礎となるのは，不可逆データ圧縮を扱うレート歪み理論の枠組みである．レート歪み問題を基にダイバージェンスと学習結果の冗長性に関するペナルティを併せた最小化する分布を求める．上記の問題は一ステップの時間・空間計算量がマイノリティの濃度に線形な繰り返しアルゴリズムで解けるため，大規模データベースへの可用性が高い．

2 レート歪み理論に基づく混合推定

確率分布に基づく混合推定では観測値 $X = \{\mathbf{x}_i\}_{i=1}^N$ の要素分布への割り当てを隠れ変数 $Y = \{y_i\}_{i=1}^N$ とし，尤度を最小化するような条件付確率 $P(Y|X)$ を求める．マイノリティ分布推定問題は二要素の混合推定とみなせる．

レート歪み理論は不可逆データ圧縮を扱う理論であるが，クラスタリングのようなデータの要約を生成する教師無し学習問題と多くのコンセプトを共有する．レート歪み理論の中心的な問題であるレート歪み問題は教師無し学習の文脈では

$$\min_{P(Y|X)} I(Y;X) + \beta \langle d(\mathbf{x},y) \rangle_{X,Y} \qquad (1)$$

と定式化される．$I(Y;X)$ は X と Y の相互情報量であり，$d(\mathbf{x},y)$ は情報理論では歪み関数と呼ばれ，

$\mathbf{x}$ の y による記述が正確なほど減少する関数が適宜選択される．(1) の最小化はデータの簡潔かつ正確な確率的記述を与える分布を求めることに当たる．第二項は記述の冗長性に対するペナルティを表し，β でその重みを決定する．

この枠組みは適切な類似度を歪み関数として選択することで様々な学習問題を扱うことができ，歪み関数を各事例の条件付対数尤度とし，$\beta=1$ とした場合は最尤推定と等価である [1]．

レート歪み理論の拡張に基づく学習の方法論は情報理論的枠組み (Information Theoretic Framework) と呼ばれ，トップダウンに与えられるドメイン情報を教師無し学習に取り入れる情報ボトルネック (Information Bottleneck) [4] 等がある．

3 マイノリティ分布推定の定式化

マイノリティ検出問題では観測値 $\mathbf{x}_i$ に対し $y_i \in \{l_n, l_s\}$ のいずれかのラベルが割り当てられる．$y_i = l_n$ ならばマジョリティ，$y_i = l_s$ ならばマイノリティに属す．それぞれのラベルの観測値に対する条件付確率はそれぞれ $\hat{\pi}, \theta$ によってパラメータ化される．

$$P(\mathbf{x}|l_n) = P(\mathbf{x}|\hat{\pi}), P(\mathbf{x}|l_s) = P(\mathbf{x}|\theta).$$

また，問題設定からデータセットの大部分はマジョリティ事例であるため，$P(\mathbf{x}|l_n) \approx P(\mathbf{x})$ という近似が有効である．すると，$P(\mathbf{x}|l_n)$ の分布パラメータ $\hat{\pi}$ は X に基づく最尤推定値で与えられる．このとき，相互情報量 $I(Y;X)$ は

$$\sum_{\mathbf{x}\in X} P(\mathbf{x}, l_n) \log \frac{P(\mathbf{x}|l_n)}{P(\mathbf{x})} + \sum_{\mathbf{x}\in X} P(\mathbf{x}, l_s) \log \frac{P(\mathbf{x}|l_s)}{P(\mathbf{x})}$$

と展開されるが，$P(\mathbf{x}|l_n) = P(\mathbf{x}|\hat{\pi}) \approx P(\mathbf{x})$ から

$$\sum_{\mathbf{x}\in X} P(\mathbf{x}, l_s) \log \frac{P(\mathbf{x}|l_s)}{P(\mathbf{x})} \equiv I(l_s; X) \qquad (2)$$

と第一項は無視して整理できる．

(1) を基にした定式化には歪み関数の定義が必要である．ここでの目的は特異な分布を持つ事例の発見なので下式のような二つの分布間の類似度が自然な歪み関数となる．

$$d(\mathbf{x}, y) = -\log P(\mathbf{x}|y) + \log P(\mathbf{x}|\hat{\pi}) \qquad (3)$$

$d(\mathbf{x}, l_n)$ は 0 となるため，$d(\mathbf{x}, y)$ の期待値は

$$\langle d(\mathbf{x}, l_s) \rangle = -\sum_{\mathbf{x}\in X} P(\mathbf{x}, l_s) \log \frac{P(\mathbf{x}|\theta)}{P(\mathbf{x}|\hat{\pi})} \qquad (4)$$

となる．(4) は二つの分布間の情報幾何的な距離を表し，$\mathbf{x}$ がマイノリティ分布によって対数尤度の意味で正確に記述されるほど，マジョリティ分布の記述から乖離するほど減少する．

(4) と (2) を (1) に代入し，目的関数 F とする．

$$F = I(l_s; X) + \langle d(\mathbf{x}, l_s) \rangle \qquad (5)$$

ここで，マイノリティ事例の条件付確率が $P(y_i|\mathbf{x}_i) = \{0, 1\}$ となる解のみを考えるものとする．このとき $\langle d(\mathbf{x}, l_s) \rangle$ は

$$-P(\mathbf{x}) \sum_{\mathbf{x}\in X} P(l_s|\mathbf{x}) \log \frac{P(\mathbf{x}|\theta)}{P(\mathbf{x}|\hat{\pi})} \propto -\sum_{\mathbf{x}\in Z} \log \frac{P(\mathbf{x}|\theta)}{P(\mathbf{x}|\hat{\pi})}$$

と整理される．ただし，$Z = \{z_i\}_{i=1}^{M}$ は $y_i = l_s$ となる事例の集合を表す．

さらに，Y と X の条件付エントロピーが 0 となり，相互情報量 $I(l_s; X)$ は下のように整理される．

$$P(\mathbf{x}) \sum_{\mathbf{x}\in X} P(l_s|\mathbf{x}) \log \frac{P(l_s|\mathbf{x})}{P(l_s)} \propto \sum_{\mathbf{x}\in Z} \log \frac{1}{P(l_s)}$$

上の二式を (5) に代入し，さらに $\beta=1$ とすると目的関数は下式で表される．

$$F'(Z, \theta, \hat{\pi}) = \sum_{\mathbf{x}\in Z} \log \frac{P(\mathbf{x}|\hat{\pi})}{P(l_s) P(\mathbf{x}|\theta)} \qquad (6)$$

F' は第一項が学習結果の冗長性，第二項が二つの分布による記述の乖離を表し，最小化によって分布のダイバージェンスが大きく，同時にデータを簡潔に記述する $P(\mathbf{x}|\theta)$ が得られる．また，F' はマイノリティ事例に関する総和で表されることからアルゴリズムの計算量を抑えることにも貢献する．

β の値は二つの分布の境界において，データ密度が等しくなる面とするという条件を与えることで決定できる．詳細は本稿では割愛する．

4 アルゴリズム

目的関数 (6) は以下のようなアルゴリズムで極小解に収束することが保障される．

初期分布 $P(\theta')$ から Z を初期化した後，F' が収束するまで以下の手順を繰り返す．

- 確率密度 $P(\mathbf{x}|\theta)$ が最小となる $\mathbf{x}_i \in Z$ について F' が増大するようにラベル y_i を更新する．
- 更新されたラベル Y に基づき θ を更新する．

1 ステップの最大の計算量は Z の要素の条件付確率を求める操作であり，統計量の逐次更新が可能なモデルを用いた場合は $O(M)$ となる．

初期化された Z がマイノリティ事例を含まない場合，Z は $\emptyset$ に縮退するが，実用的には空集合まで収束させる必要はないため次節の実験では，マイノリティの最小濃度 $c_{\min}$ を与え，M が $c_{\min}$ 未満となった時点で終了とした．

表 1: 人工データの分類結果

次元数	相対平均密度	相対精度 (基準精度)	感度	特異度
2	0.054	1.9 ± 0.38 (0.16)	0.69 ± 0.098	0.96 ± 0.013
3	0.052	1.9 ± 0.26 (0.18)	0.60 ± 0.16	0.98 ± 0.0019
5	0.054	2.1 ± 0.67 (0.19)	0.65 ± 0.15	0.99 ± 0.0016
8	0.061	1.7 ± 0.34 (0.25)	0.70 ± 0.20	0.99 ± 0.0013

5 実験結果

5.1 シミュレーション実験

テストデータを一辺の長さ 2 の d 次元の超立方内に一様分布する 1000×2^d 個の点とガウス分布から生成した 100 点より構成する．ガウス分布の平均 (μ_i), 標準偏差 (σ_i), および相関係数 $[\rho_{ij}]$ はそれぞれ $[-0.5 : 0.5]$, $[0.1 : 0.2]$, $[-0.5 : 0.5]$ の範囲からランダムに決定する．次元数は $2, 3, 5, 8$ とする．

ガウス分布から生成された事例の検出数を真陽性とし, 精度 p, 感度 Sn, 特異度 Sp を評価する．ただし, 分類精度の最大値はデータセットの密度によりそれぞれ異なるため, 精度の指標として相対精度 p_r を以下のように定める．ガウス分布から生成した事例の共分散行列を基にそれらを全て包含するような超楕円を考える．その曲面によってテストデータを分類した場合の精度を p_z とし, $p_r = p/p_z$ とする．また, マイノリティの相対的な密度が小さい程検出が難しいため, 指標として R_Δ を定義する．ガウス分布の 2 標準偏差範囲内の平均データ密度を Δ_G, 一様分布から生成したデータの密度を Δ_U とし, $R_\Delta \equiv \Delta_G/\Delta_U$ とする．

提案手法は各データに異なる初期条件で 20 回適用し, 最良スコアの結果を評価する．Z の初期濃度は 100×2^d とし, 最小濃度は 50 とする．

実験結果を表 1 に示す．提案手法は高い相対精度と感度を示し, 8 次元までほぼ同等の安定した分類性能を示す．高次元のガウス分布から生成された事例は密度の高い集合であるため検出が容易なことが多いが, ここでは相対密度 R_Δ を基準に, 一様分布の密度を大きくしている．このため, マイノリティ事例 100 に対して総事例数は最大数十万となっており, 提案手法の高い検出精度は大規模なデータセットにおける可用性の高さを示すといえる．

5.2 実データ実験

5.2.1 ドキュメント抽出問題

多くのテキスト分類研究で利用されているロイター 21578 コーパスを元に実験データを構成する．このコーパスは 10,789 の新聞記事からなり, 各文書は内容に応じた一つ以上のトピックを持つ．

文書の生成モデルは多項分布とする [3]．辞書を $W = \{w_i\}_{i=1}^k$ とし, 文書 $\mathbf{x}$ を語 w_i の出現回数 v_i により $\mathbf{x} = (v_1, v_2, \ldots, v_n)$ と表すとき, 文書 $\mathbf{x}$ の条件付確率は下式で与えられる．

$$P(\mathbf{x}|y) = V! \prod_{j=1}^{k} \frac{P(w_j|y)^{v_j}}{v_j!}$$

ただし, $V = \sum_{i=1}^k v_i$ は語の総出現頻度である．語の条件付確率 $P(w_j|l_n)$, $P(w_j|l_s)$ は以下のとおり．

$$\frac{1 + \sum_{\mathbf{x} \in X} v_j}{k + \sum_{i=1}^{k} \sum_{\mathbf{x} \in X} v_j}, \quad \frac{1 + \sum_{\mathbf{x} \in Z} v_j}{k + \sum_{i=1}^{k} \sum_{\mathbf{x} \in Z} v_j}$$

ここで, Z は $y = l_s$ となる X の部分集合である．

また, 辞書 W は $I(W; X)$ への寄与,

$$I(w_i) = p(w_i) \sum_{\mathbf{x} \in X} p(\mathbf{x}|w_i) \log \frac{p(\mathbf{x}|w_i)}{p(\mathbf{x})}$$

が大きい k 語とする．辞書サイズ k は 200 とした．

実験に用いるデータセットを以下の手順で構成する． まず, 要素数の非常に大きいトピック *acq* と, その $\frac{1}{10}$ 以下の要素数のトピック *oilseed*, *money-supply*, *sugar*, *gnp* を持つ文書の部分集合を抽出する．それぞれの文書数は 2448, 192, 190, 184, 163 である． *acq* とそれ以外のトピックの文書を個別にマージした文書集合をテストデータ $C1 \sim C4$ とする．それぞれからマイノリティにあたる文書を検出し, その内 *acq* でないトピックを持つ文書を真陽性として精度, 感度, 特異度を求める．

比較のため, 一クラス SVM と多項式モデルに基づく EM 法 [3] を同じデータセットに適用する．一クラス SVM の設定は [2] に基づき, 正規化した語の頻度によって文書を表し, 提案手法と同じ辞書を用いる．ν パラメータは 0.01 から 0.1 まで 0.005 刻みで変化させ, その中で精度が最大となる結果を評価した．カーネルは線形, 多項式, RBF を用いた．提案手法と EM 法は各テストデータに異なる初期条件で 20 回適用し, 最もスコアの良い結果を評価した．

表 2: Reuters21578 文書の分類結果

	精度	感度	特異度
提案手法	0.95±0.025	0.94±0.047	1.0±0.0020
OCSVM	0.34±0.39	0.31±0.38	0.99±0.019
EM	0.25±0.23	0.97±0.022	0.64±0.21

表 2 にテストデータ $C1 \sim 4$ における分類性能の平均と標準偏差を示す. ただし, 一クラス SVM については最も精度の高かった多項式カーネルを用いた結果を示す. 提案手法は精度, 感度, 特異度ともに安定して最も高い値を示した. 一クラス SVM の性能はデータによって異なり, *oilseed*, *gnp* では比較的高い感度と精度を示したが, 他のデータでは 0 となり, 平均では感度・精度の値は提案手法を大きく下回る. EM 法による分類性能も不安定で, *gnp* では比較的高い精度・特異度を示したが, それ以外のデータでは非常に低い値となった.

5.2.2 衛星画像分類問題

続いて UCI Machine Learning Repository で公開されている衛星画像データを用いた実験を示す. 82×100 ピクセルの領域があり, ピクセル毎に 4 つのスペクトル強度が測定され, 各事例は隣接 8 近傍のスペクトルを含めた 36 次元の整数ベクトルで表される. 各データは土壌の種類によって赤土, 灰土, 湿灰土等に対応するクラス 1, 2, 3, 4, 5, 7 のいずれかに割り振られている. 実験に使用したクラス 1, 3, 7 のピクセル数は 1533, 1358, 1508 でクラス 2, 4, 5 のピクセル数は 224, 211, 237 である. 次の組合せでデータセット $D1 \sim 5$ を用意した. $D1 = \{1,2\}$, $D2 = \{3,2\}$, $D3 = \{3,4\}$, $D4 = \{3,5\}$, $D5 = \{7,2\}$. 各組合せの後のクラスがマイノリティに当たる. 検出された事例の内, マイノリティのクラスに属すピクセルを真陽性として精度, 感度, 特異度を求める.

提案手法は各データセットに異なる初期条件で 10 回適用し, 最もスコアの良い結果を評価した. マイノリティ, マジョリティの生成モデルとしてはガウス分布を用いた. 一クラス SVM では ν パラメータは 0.1 から 0.2 まで 0.01 ずつ変化させ, 精度が最大となった結果を評価した. カーネルは線形, 多項式, RBF を用いた.

3 に $D1 \sim 5$ からの検出結果の平均と標準偏差を示す. ただし, 一クラス SVM については最も精度の高かった多項式カーネルを用いた結果を示す. 提案手法の性能は全てのデータセットで安定して高い精度, 感度, 特異度の値を示した. 一クラス SVM による分類はデータセット毎に大きく異なり, $D2$, $D3$ では比較的高い精度と感度を示したが, $D1$, $D5$ では 0 かそれに近い値となった.

表 3: 衛星画像データの分類結果

	精度	感度	特異度
提案手法	0.79±0.051	0.99±0.028	0.93±0.14
OCSVM	0.43±0.37	0.93±0.060	0.49±0.42

6 おわりに

本稿で扱ったマイノリティ検出問題は, 大規模なデータベースから知識を獲得する上で重要な問題である. マイノリティ事例を検出する上では複数の困難が存在するが, われわれはレート歪み理論の枠組みを基にいくつかの近似を導入することで上記の問題に対処する分布推定手法を提案した. 目的関数がマイノリティ分布のダイバージェンスと学習結果の冗長性に対するペナルティを含む点が大きな特徴である. この目的関数はマイノリティ事例に関する総和となるため, 大規模データベースにおける可用性も高いアルゴリズムが実現できる. 問題に依存するパラメータを持たない点も手法の重要な特徴である. 小さな事例集合を検出するタスクでは一クラス分類や外れ値検出が主に用いられるが, 提案手法では発生源の重複を前提とした問題での新しい応用も期待される.

参考文献

1) Banerjee, A., Dhillon, I. S., Ghosh, J. and Merugu, S.: An information theoretic analysis of maximum likelihood mixture estimation for exponential families., *Machine Learning, Proceedings of the 21st International Conference*, ACM (2004).

2) Manevitz, L. M. and Yousef, M.: One-Class SVMs for Document Classification., *Journal of Machine Learning Research*, Vol. 2, pp. 139–154 (2001).

3) Nigam, K., McCallum, A., Thrun, S. and Mitchell, T. M.: Text Classification from Labeled and Unlabeled Documents using EM., *Machine Learning*, Vol. 39, No. 2/3, pp. 103–134 (2000).

4) Tishby, N., Pereira, F. C. and Bialek, W.: The information bottleneck method, *Computing Research Repository(CoRR)*, Vol. physics/0004057 (2000).