局所的構造に着目したアンサンブル学習の分類精度向上に関する研究

1X13C039-7　業天 大貴
指導教員　後藤 正幸

1　研究背景と目的

インターネット上や企業システムで膨大な量のデータが扱われるようになったことを背景に，自動分類技術の重要性が高まっている．機械学習における自動分類では，カテゴリが既知の学習データから分類規則の学習を行い，カテゴリが未知のデータの所属カテゴリを推定することが多い．分類問題においてカテゴリ数が 2 の場合は二値分類問題と呼ばれ，解決のための様々な学習法が提案されている．本研究ではこのうち，アンサンブル学習法の 1 つである Adaboost [1] に着目する．Adaboost は，一般に高い精度を示す手法として知られているが，特徴空間上のデータ分布に局所的な構造の差異がある場合，その局所構造を反映できないという問題がある．そこで，データの局所的構造を考慮して Adaboost を用いた学習を行う手法として Cluster-Based Boosting [2](以下，CBB) が提案されている．

CBB では，k-means 法により予め類似したデータのクラスタリングを行い，クラスタごとの分類の難易度に応じて，Boosting を用いた学習をするべき領域か否かの判別を行う．分類の難易度の高いクラスタに対しては Adaboost を用いて重点的な学習を行い，難易度の低いクラスタに対しては簡易的な手法で学習法を行う．このように，クラスタごとに異なる学習法を適用することで，複雑な構造を持つデータに対しても局所的な特性を考慮した分類器を構築することができる．すなわち，データの分類の難易度を適切に表現することのできるクラスタリングを行うことで分類精度の向上が示唆される．

しかし，CBB では k-means 法によるクラスタリングの過程で類似したデータ構造をもつ領域のデータが 1 つのクラスタとして形成されなかった場合，その後の学習プロセスがうまく機能しない可能性がある．k-means 法では単純にユークリッド距離を用いてクラスタリングを行うため,「特徴量間に相関がない」「クラスタの分散共分散行列が等しい」という構造が暗に仮定されている．従って，この仮定に合致しないデータ群に対しては過度にサイズの小さいクラスタを多数生成してしまう可能性がある．

そこで本研究では，上記の k-menas 法の条件を緩和した混合ガウス分布を用いたクラスタリング手法を CBB に適用することで，データの構造の難易度を適切に表現可能で，従来の CBB よりも多様なデータの分布に適応できる手法を提案する．さらに，UCI 機械学習レポジトリのデータを用いて提案手法の有用性を示す．

2　Adaboost を用いた二値分類

学習データ集合を $\mathcal{D} = \{(\boldsymbol{x}_n, y_n)\}_{n=1}^{N}$ とする．ここで，$\boldsymbol{x}_n$ は I 次元特徴ベクトル，$y_n \in \{1, -1\}$ は $\boldsymbol{x}_n$ の所属するカテゴリとする．このとき，カテゴリが未知の新規入力データ $\boldsymbol{x}_{new}$ の属するカテゴリを推定することを考える．

Adaboost は，T 個の分類器 $g_1, g_2, ..., g_T$ と重み $\alpha_1, \alpha_2, ..., \alpha_T$ により $\boldsymbol{x}_{new}$ の予測カテゴリ $\hat{y}_{new}$ を

$$\hat{y}_{new} = \mathrm{sign}\left(\sum_{t=1}^{T} \alpha_t g_t(\boldsymbol{x}_{new})\right) \tag{1}$$

として求める手法である．ここで，$\mathrm{sign}(x)$ は，$x \geq 0$ のとき 1，$x < 0$ のとき -1 をとる関数である．分類器 $g_t(\boldsymbol{x})$ は，$g_{t-1}(\boldsymbol{x})$ で誤分類したデータ $(\boldsymbol{x}_n, y_n)$ を正しく分類するように学習が行われ，これを繰り返すことで T 個の分類器が得られる．また，分類器 g_t に対する重み α_t は，分類器の誤り率が低いほど大きくなるように決定される．分類器 g_t での重み α_t は分類器 g_t における誤り率 ε_t とパラメータ η を用いて以下の式のように表される．

$$\alpha_t = \eta \ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right) \tag{2}$$

ここで，η は学習率と呼ばれ，この値を大きくすることで，よりでデータにフィットする分類器が構築される．

Adaboost の問題点として，特徴空間上のデータ分布に局所的な構造の差異がある場合その局所構造を反映できないことや，カテゴリにノイズが含まれる場合には過度に影響を受けてしまう可能性が挙げられる．

3　Cluster-Based Boosting

3.1　概要

データの局所的な特性を考慮した分類器を構築する手法として CBB が提案されている．CBB では，あらかじめ k-means 法を用いて学習データのクラスタリングを行い，得られたクラスタに所属するデータの傾向と初期分類器 f_1 の分類精度を基に，クラスタのタイプ分類を行う．このタイプごとに Adaboost を含む異なる学習法を適用することで，複数の分類器を構築する．

具体的には，はじめに学習データ全体を二値分類する初期分類器 f_1 を構成する．次に，学習データ全体に k-means 法を適用し複数のクラスタを生成する．得られたそれぞれのクラスタに所属する学習データに対し，初期分類器 f_1 の精度，クラスタ内のカテゴリ比率を基にクラスタを 4 つのタイプに分け，それぞれのクラスタに所属しているデータ構造に適した学習を行う．新規入力データ $\boldsymbol{x}_{new}$ の分類では，$\boldsymbol{x}_{new}$ と各クラスタの中心までの距離が最小となるクラスタを $\boldsymbol{x}_{new}$ の所属クラスタとし，そのクラスタに付与されている分類器を用いて予測を行う．このようなクラスタごとに異なる学習法を適用することで，データの局所的構造を考慮した分類器を構築することが可能である．

3.2　クラスタのタイプとその学習

k-means により得られたクラスタ集合を $\mathcal{C} = \{c_1, \cdots, c_K\}$ とする．CBB では，得られたクラスタをその特性に基づいて表 1 に示す 4 種類のタイプのいずれかに分類し，クラスタのタイプ別にその特性を考慮した学習を行う．表 1 に示したように各クラスタの特性を決める基準には，クラスタ c_k に所属するデータに対する (1) 初期分類器 f_1 の分類精度, (2) クラスタ内の少数派カテゴリの比率という 2 つの属性を用いる．

表 1．クラスタタイプ分け

(2) \ (1)		f_1 の分類精度 高	f_1 の分類精度 低
少数派の比率	少	HOP	HOS
	多	HEP	HES

表 1 は，各クラスタに所属しているデータの分類の難易度を示している．前述の通り，CBB ではこれらのタイプごとに以下の通り異なる学習を行う．HEP, HES は，カテゴリが混在しているクラスタで，分類難易度の高い分類境界が想定されるため，Adaboost を用いた重点的な学習を行う．

- HOP ; そのクラスタに対して特別な学習は行わず，分類を行う際は f_1 のみを用いる．
- HOS ; クラスタ内のデータを分類することのできる分類器を 1 つ学習し，分類を行う際は，その分類器を用いる．
- HEP ; 学習率 η の低い Adaboost により学習を行う．分類を行う際には，初期分類器 f_1 と Adaboost により構築された T 個の分類器による多数決を行う．
- HES ; 学習率 η の高い Adaboost により学習を行う．分類は HEP と同様に行う．

4　提案手法

4.1　概要

CBB では，データの分類の難易度を適切に表現するクラスタリングを行えるか否かが分類精度に影響する．従来の

CBB で用いられている k-means 法は，単純にユークリッド距離に基づいてクラスタを構築している．そのため，特徴間に相関がなく，かつ分散が等しい正規分布を暗に仮定したクラスタリングを行っている．よって，特徴間に相関がある場合にはクラスタリングの過程で類似した統計的構造を持つ領域のデータが 1 つのクラスタを形成せず，適切なサイズのクラスタと十分な精度が得られない可能性がある．一方，各クラスタに固有の平均ベクトルと分散共分散行列で定められる正規分布を仮定し，クラスタリングを行う手法として混合ガウス分布が知られている．混合ガウス分布を用いることにより，各クラスタに所属するデータの特徴間に相関があることを許容し，かつクラスタ間で分散共分散行列が異なる正規分布を仮定することができる．これにより，k-means 法を用いる場合より適切なクラスタリングが可能となるため，分類精度の向上につながると考えられる．以上の議論により，CBB におけるクラスタリングに混合ガウス分布を用いることで，k-means 法を用いる場合よりも多様なデータ分布に対応可能な手法を提案する．

4.2 混合ガウスモデルと CBB への導入

混合ガウス分布では，潜在変数 $z_k(1 \leq k \leq K)$ のもとでのデータの出現確率 $p(\boldsymbol{x}_n|z_k)$ に正規分布を仮定したモデルである．潜在変数 z_k における平均ベクトルを $\boldsymbol{\mu}_k$ と分散共分散行列を $\boldsymbol{\Sigma}_k$ とすると，データ $\boldsymbol{x}_n$ の生起確率は以下の式のように表せる．

$$p(\boldsymbol{x}_n) = \sum_{k=1}^{K} p(\boldsymbol{x}_n|z_k)p(z_k) \tag{3}$$

$$p(\boldsymbol{x}_n|z_k) = \frac{1}{(2\pi)^{\frac{I}{2}}|\boldsymbol{\Sigma}_k|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)\right) \tag{4}$$

上記の式におけるパラメータ $p(z_k)$, $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$ は EM アルゴリズムを用いて求められる．ここで，式（4）の $\boldsymbol{\Sigma}_k$ に着目すると，潜在変数ごとに異なる値をとることを許容しているため，複雑な構造を持つデータクラスタリング手法として有効であることが知られている．

提案手法では，CBB におけるデータのクラスタリング手法として，k-means 法に代えて混合ガウス分布を適用することにより，各クラスタに所属するデータの特徴間に相関があることを許容し，かつクラスタ間で分散共分散行列を持つことを考慮したクラスタリングを行う．混合ガウス分布によるクラスタリングは，データ $\boldsymbol{x}_n$ が複数のクラスタに属することを許容するソフトクラスタリングの手法であるが，提案手法では CBB との整合性を保つため，各データ $\boldsymbol{x}_n$ に対し，$p(z_k|\boldsymbol{x}_n)$ が最大となる潜在クラス z_k のみに所属させることで K 個のクラスタを構築する．ここで得られた K 個のクラスタについて 3.2 節と同様に表 1 の 4 種のクラスタタイプを割り当て，それぞれのクラスタに対してタイプ別の学習を行う．また，新規入力データ $\boldsymbol{x}_{new}$ の分類を行う際には $p(z_k|\boldsymbol{x}_{new})$ が最大となるクラスタ k の分類器と初期分類器 f_1 による多数決を行う．

以上を踏まえ，提案手法のアルゴリズムを以下に示す．

【学習フェーズ】

Step1) 学習データ $\{\boldsymbol{x}_1, \boldsymbol{x}_2 ..., \boldsymbol{x}_N\}$ から混合ガウス分布のパラメータ推定を行う．

Step2) 所属確率 $p(z_k|\boldsymbol{x}_n)$ をもとに所属クラスタを 1 つ選定しハードクラスタリングを行う．

Step3) 全ての学習データ $\{(\boldsymbol{x}_n, y_n)\}$ に対して初期分類器 f_1 を学習する．

Step4) クラスタごとに初期分類器 f_1 の分類精度とカテゴリの少数派の割合を算出し，表 1 の 4 種のクラスタタイプを付与する．

Step5) 各クラスタに 3.2 節と同様にタイプ別の学習を行い分類器を構築する．

【分類フェーズ】

Step1) 新規入力データ $\boldsymbol{x}_{new}$ の各クラスタへの所属確率 $p(z_k|\boldsymbol{x}_{new})$ を算出する．

Step2) 所属確率が最大となるクラスタ k の分類器と初期分類器 f_1 による多数決を行い，得られたカテゴリを $\boldsymbol{x}_{new}$ の所属するカテゴリの予測値 $\hat{y}_{new}$ する．

5 実験

提案手法の有用性を検証するため，ベンチマークデータセットを用いた分類実験を行った．

5.1 実験条件

データセットには，UCI 機械学習レポジトリから 3 種類のデータセット (ionosphere, bupa, sonar) を用いた．全データから 8 割をランダムに選択して学習を行い，残りの 2 割でテストを行う操作を 10 回行い，その平均を精度とした．また，比較手法として従来の k-means 法によるクラスタリングを行う CBB を用いた．データセットの概要を以下に示す．

表 2．データセット概要

データセット	次元数	学習データ数	テストデータ数
ionosphere	34	243	65
bupa	5	304	76
sonar	60	183	46

事前実験により，Adaboost の分類器数 $T = 5$, HEP の学習率 $\eta = 0.5$，HES の学習率 $\eta = 1$ とした．混合ガウス分布の混合数 K については ionosphere と sonar は 2，bupa は 5, とした．また，分類器には決定木を用いた．評価指標は式 (5) で表される分類精度を用いるものとした．

$$\text{分類精度} = \frac{\text{正しく分類したテストデータ数}}{\text{テストデータ数}} \tag{5}$$

5.2 実験結果と考察

実験結果を図 1 に示す．ここで，図 1 における*は平均の差が 5%有意, **は 1%有意であることを示している．

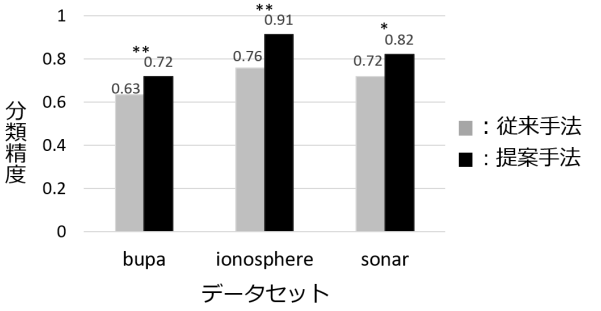

図 1．各データセットの実験結果

図 1 より，すべてのデータセットにおいて，従来手法よりも提案手法の分類精度が優れていることがわかる．また，クラスタのタイプに着目すると，提案手法では同一カテゴリのデータが所属するクラスタである HOP, HOS の割合が多くなり，カテゴリが混在しているクラスタである HEP, HES が少なくなった．これは，クラスタの特性を考慮した適切なクラスタリングを行えているためと考えられる．このことにより，データ構造を適切に表現し，分類の難易度が高い局所構造を重点的に学習することができたため，分類精度の向上につながったと考えられる．

6 まとめと今後の課題

本研究では，CBB におけるクラスタリング手法に混合ガウス分布を用いることにより，データの分布形を考慮しデータの局所的構造をより適切にとらえる方法を提案し，実験によりその有用性を示した．今後の課題としては，多値分類問題への拡張，パラメータの自動推定などがあげられる．

参考文献

[1] Y. Freund and R. Schapire, “A desicion-theoretic generalization of on-line learning and an application to boosting,” *Computer and System Sciences* , Vol. 55, pp. 119–139, 1997.

[2] L. Miller and L. Soh, “Cluster-Based Boosting,” *IEEE Trans. Knowledge and Data Engineering*, Vol. 27, No. 6 , pp. 1491–1504, 2015.