論 文 の 内 容 の 要 旨

論文題目　Identifying Discriminative Features from High-Dimensional Data using Support Vector Machines
（サポートベクターマシンを用いた高次元データからの有効分類属性の特定）

氏名　　嶋　幸太郎

序

インターネット上で公開される情報量の増加に伴い、テキスト等のデータを効率的に分類する技術が重要となってきている。しかし、これらのデータは一般的に極めて高次元であるため、元データをそのまま扱うと、精度の高い分類器を構築することが困難、計算コストが膨大、モデルの解釈が困難などといった問題が生じる。したがって、膨大な属性の中から、分類に寄与する属性（本研究では有効分類属性と呼ぶ）を属性選択・抽出手法によって正確に特定し、分類に適したデータ表現に次元を圧縮させることが肝要となる。これまで提案されてきた次元圧縮手法は比較的少数の属性が想定されてきたが、テキスト分類等への応用の際には属性数が数万以上にも上るため、従来手法では十分に対処し切れない。したがって、高次元データから有効分類属性を効率的に特定する手法の開発が望まれている。

近年、サポートベクターマシン（SVM）と呼ばれる分類アルゴリズムが大変注目を集めている。この手法は構造リスク最小化という原理に基づいているため、高次元データに対しても過学習を起こすことなく高い汎化能力を示すこと

が報告されている。最近になって同手法を属性選択に用いることが提案されたが、同手法の高次元データへの有効性は属性選択においても発揮されると期待される。そこで、本研究では、高次元データからの有効分類属性の特定に関して、SVM を用いた属性選択法の有効性を示すことを目的とする。

SVM 属性選択法

SVM では「マージン」という量を定義し、マージンを最大化する線形分類器を求める。この問題は次の凸二次計画問題に帰着される。

目的関数：$\sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j \rightarrow$ 最大化 (1)

制約条件：$0 \le \alpha_i \le C\,(i=1,\mathrm{K}\,,m), \sum_{i=1}^{m}\alpha_i y_i = 0$

ただし、ここで$\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_m\}\in\boldsymbol{R}^n$は訓練データ、$\{y_1,\cdots,y_m\}\in\{-1,1\}$はクラスラベル、$\alpha_i(\geqq 0)$はラグランジュ乗数、$C$ は誤分類とマージンのトレードオフを決定するソフトマージンパラメータである。上の最適化問題を解くことで、判別関数は下記のように求まる。

$$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b,\ \ \mathbf{w} = \sum_{i=1}^{m}\alpha_i y_i \mathbf{x}_i \qquad (2)$$

ここで、多くのα_iは 0 となり、訓練データの一部のみが判別境界に寄与する。

最近になって、判別関数の重み $\mathbf{w}$ を用いた属性選択法が提案された。この手法は、属性の重みw_k^2がほぼ 0 である属性は判別境界へほとんど影響を及ぼさないと考えられることから、w_k^2の大きさを属性選択の指標として用いるものである。既往の研究で同手法の高次元データへの有効性が報告されているが、次に挙げる点に関しては未だ十分な知見が得られていない。まず第１に、同手法はテキスト分類における従来手法と異なり多変量手法であるので属性間の関連も考慮されるが、その効果については十分な解析がなされていない。第２に、分類精度を低下させることなく属性数をどの程度まで削減できるかについての決定手法が確立されていない。第３点として、ソフトマージンパラメータ Cは属性選択の効果に大きな影響を及ぼすと考えられるが、既往の研究では十分に大きな Cが用いられ、その影響については解析がなされていない。第４点として、同手法はこれまで元の属性のみに適用されてきており、抽出属性への適用例は

報告されていない。本研究では以上の点に着目し分析を行った。

テキスト分類問題における次元圧縮

SVM 属性選択法をテキスト分類において標準的に用いられている Reuters-21578、20 Newsgroups データセットに適用した。分類精度の評価には、適合率と再現率の調和平均で定義される F1 値を用いた。

まず、SVM 属性選択法の多変量性の効果について分析を行った。比較対象としたのは、テキスト分類において広く用いられ、その性能の高さが実証されている情報利得及び χ^2 指標である。属性数の減少に伴い、他の二手法は若干 F1 値が変動したのに対し、本手法は F1 値が極めて安定であった。さらに、分類精度と w_k^2 の累積寄与率との間に強い相関が見られた。そこで、本研究では w_k^2 の累積寄与率を不要属性数の推定に用いることを提案する。累積寄与率の閾値を設定し、各クラス毎に閾値に達する属性数を求め、F1 値の閾値依存性を調べた。全クラスに対して計算を行った結果、本手法による F1 値は他の二手法と比べ、平均は高く、標準偏差は小さいという結果が得られ、本手法はクラスに依らず頑健に不要属性を除去できることが示された。これは、他の二手法が属性を個々に評価しているために組み合わさって初めて分類に寄与する属性を除去してしまう危険性があるのに対し、本手法は全属性の関連を考慮した上で属性を評価していることに起因すると考えられる。また、w_k^2 の累積寄与率が不要属性数の推定に有効な指標であることが示された。

次に、C が属性選択に与える影響について分析を行った。訓練誤差が小さくなるように C を大きくした場合、汎化誤差が小さくなるように 10-fold cross-validation で決定した C を用いた場合、C を十分に小さくした場合の 3 ケースに対して比較を行った。その結果、属性数が多い領域では大きな C が、属性数が少ない領域には小さな C が適していることが明らかになった。

テキスト分類においては、特異値分解に基づいて属性抽出を行い次元圧縮すると、単語間の関連が考慮され分類精度が向上することが報告されている（Latent Semantic Indexing, LSI）。しかし、この手法によって抽出された属性の重要度は分類性ではなく分散で評価されるため、必ずしも分類に適したデータ表現になっている保証が無い。そこで本研究では、より分類に適したコンパクトな LSI 部分空間を求めるために SVM 属性選択法を適用した。抽出属性 $\mathbf{t}_k$

の分類性は次式で評価される。

$$\left(\mathbf{t}_k^T\mathbf{w}\right)^2 = \sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j \left(\mathbf{t}_k^T\mathbf{x}_i\right)\left(\mathbf{t}_k^T\mathbf{x}_j\right) \qquad (3)$$

属性抽出において、クラス内に複数のクラスタが存在する場合、分類性とデータ表現性の両方を考慮する必要性が指摘されている。そこで、本研究では(3)式の値と分散の調和平均を属性選択指標として用いた。その結果、本手法で特定された LSI 部分空間は通常の LSI と同様に分類精度の向上が見られたが、最大の分類精度を実現するのに必要な属性数は通常の LSI よりも少ないことが分かった。大規模データに LSI を適用する際には、訓練・テストにかかる計算コストが大きな問題となるが、本手法を用いることで計算量の軽減が可能となる。

情報フィルタリングにおけるユーザの興味の学習

ウェブページの情報フィルタリングに SVM 属性選択法を適用した。ユーザが検索エンジンを用いて研究テーマに関する検索を行い、各自の興味の有る/無しを評価したウェブページを分析に用いた。まず、SVM 属性選択法で不要属性を除去した上で、LSI によって属性抽出を行い、その中から SVM 属性選択法を用いて有効分類属性を特定した。その結果、比較的少数の属性数で全属性とほぼ同等の分類精度が実現できるという結果が得られた。そして、抽出された属性の中でどの属性が寄与しているかをユーザ自身に評価してもらった結果、SVM 属性選択法で上位に評価された属性と大きな相違は無かった。したがって、本手法はユーザの興味を表す因子を特定するのに有効である可能性が示唆された。

非線形属性への拡張

データが線形には分類されない場合には非線形な有効分類属性を特定することが重要となる。そこで、上記の SVM 属性選択法を非線形に抽出された属性に拡張した。

SVM は基本的には線形分類器を学習するが、目的関数(1)式、判別関数(2)式共に訓練データには内積形でしか依存しないという性質を利用することで、非線形の場合にも容易に拡張することができる。適切な非線形関数 φ を用いて元の空間よりも高次元な空間に写像すれば線形分離性を高めることができるので、写像した先の空間で線形 SVM を学習できる。そこで、$\mathbf{K}(\mathbf{x},\mathbf{y}) = \mathbf{\Phi}(\mathbf{x})^T\mathbf{\Phi}(\mathbf{y})$ を

満たすカーネル関数を用意し、式(1)、(2)の内積形をこの関数で置換して解けば元の空間で非線形分類器を求めることができる。

Latent Semantic Kernel(LSK)は、カーネルを用いて非線形な属性抽出を行う手法である。具体的には、カーネル行列に対して固有値分解を行う。

$$\mathbf{K} = \mathbf{D} \cdot \mathbf{\Lambda} \cdot \mathbf{D}^T$$

ここで、D は固有ベクトルからなる行列、$\mathbf{\Lambda}$は対角成分が固有値である対角行列である。LSK も LSI と同様に属性の重要度は分類性ではなく分散で評価されるので、本研究では SVM 属性選択法を LSK に適用することを提案する。各属性に対応する固有値を λ_k、固有ベクトルを $\boldsymbol{d}_k$ とすると、非線形属性 $\mathbf{t}_k$ の分類性の評価指標は次式で与えられる。

$$\left(\mathbf{t}_k^T \mathbf{w}\right)^2 = \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j \lambda_k d_{ik} d_{jk}$$

属性選択指標としては、LSI と同様に分類性と分散の調和平均を用いた。非線形な分類器が必要であることが知られている USPS データセットに本手法を適用した結果、本手法は通常の LSK よりも少数の属性で高い分類精度が実現できることが示され、本手法は線形属性のみならず非線形属性に対しても有効であることが分かった。

結論

高次元データからの有効分類属性の特定に関し、SVM 属性選択法の有効性を明らかにした。特に、同手法の多変量性に由来するクラスに依存しない頑健性を実証し、不要属性数を決定する指標を提案した。また、属性選択の効果はソフトマージンパラメータに大きく依存することを示した。さらに、同手法が元属性のみならず、線形および非線形に抽出された属性にも有効であることを示した。そして、同手法を情報フィルタリングに適用し、ユーザの興味を表す因子を特定できる可能性が示唆された。