複雑な構造を持つ高次元データに対する機械学習

Machine learning for structured high-dimensional data

1071012

研究者代表　東京工業大学大学院情報理工学研究科　准教授　杉山　将

[研究の目的]

高度情報化社会の現代，コンピュータの性能は飛躍的に向上している。一方でコンピュータはどんどん複雑になり，コンピュータを使える者と使えない者の差，いわゆるディジタルディバイドが大きな社会問題となっている。この状況を改善するためには，多くの人がコンピュータを使えるよう教育を施すよりも，逆に個々のユーザのレベル・趣向に合わせるようコンピュータを学習させるのが有効なアプローチである。そのためには，様々なユーザから抽出された複雑なデータを用いてコンピュータを学習させる必要がある。本研究課題では，そのような複雑な構造を持つ高次元データに対する機械学習法を開発する。

近年，様々な機械学習の手法が提案されているが，データの次元が非常に高い場合，次元の呪いと呼ばれる現象のため，学習の効率が致命的に低下してしまうことが知られている。この問題を回避するためには，データに含まれている本質的な構造・情報をできるだけ維持したまま，データの次元数を削減することが有効である。本研究課題の目的は，有効な次元削減の手法開発することである。

[研究の内容，成果]

従来の次元削減の手法では，正規性などの単純な構造はうまく保持することができる。しかし近年の現実的なデータは非常に複雑な構造を持っており，それを失うことなく如何にして次元削減を行うかが本研究課題の技術的な課題である。そこで我々は，データのマルチモーダリティと呼ばれる構造に着目する。マルチモーダリティとは，ある一種類の情報が更に異なるタイプに階層的に分類されるような構造のことであり，自然言語，脳波信号，遺伝子・たんぱく質情報など多くの現実データで見受けられる構造である。本研究課題では，マルチモーダル構造を有効に保持できる次元削減法の開発を行う。

次元削減は，1936年に提案されたフィッシャーの判別分析など統計学の分野で非常に古くから研究され，数学的にも実用的にも非常に重要な研究課題である。また，近年の高度情報化によるデータ量の爆発的増加に相まって，機械学習の分野でも次元削減の研究が盛んに行われている。現在よく用いられている代表的な次元削減法は，複雑な最適化処理を行うため計算に非常に時間がかかる。また，必ずしも最適解が求まるとは限らないため，出力結果の信頼性が低い。更に，マルチモーダル構造に着目した研究は皆無である。従って，マルチモーダル構造を保持できる高速かつ信頼性の高い次元削減法を開発すれば，機械学習，統計学及び様々な応用分野に多大な貢献ができると期待される。

本研究課題においてマルチモーダリティを保持するための鍵となるアイデアは，データマイニング分野で盛んに研究されているデータクラスタリング技術である。これは，類似したデー

タを自動的にグルーピングする技術であり，データの局所構造を積極的に利用している。我々はクラスタリング法におけるデータの局所構造保存技術に着目し，これを次元削減に応用することによりマルチモーダリティを維持できる次元削減法を開発した。

提案法は局所フィッシャー判別分析と名づけられ，その有効性を様々なデータを通して実証した。図1，図2に甲状腺データの分析結果を示す。局所フィッシャー判別分析法は，レイリーの原理と呼ばれる最適化原理に基づいており，解を高速かつ安定に計算することができる。従って，提案法は大規模なデータに対しても適用することができるという特徴を持っている。

我々は，人工知能学会「データマイニングと統計数理研究会」にて局所フィッシャー判別分析法に関する講演を行ない，研究会優秀賞を受賞した。また，複数の民間企業に招待され，関連する講演を行なうなど，手法の有効性が客観的に認められた。

[今後の研究の方向，課題]

局所フィッシャー判別分析法は，ラベルの付いたデータの分析に有効な次元削減法である。一方，近年の応用例では，ラベル付きデータに加えて，無数のラベル無しデータが与えられることがある。このような状況は，準教師付き学習と呼ばれ，機械学習分野で近年盛んに研究されている。

このように部分的にラベル付けされているデータを解析できるように，局所フィッシャー判別分析の概念を拡張することが本研究の自然な流れである。現在，我々は準教師付き局所フィッシャー判別分析法を開発しており，その第一報が国際会議に採録されたところである。今後ますます発展させていく予定である。

[成果の発表，論文等]

1） Sugiyama, M.: Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis. Journal of Machine Learning Research, vol. 8 (May), pp. 1027 – 1061 (2007)

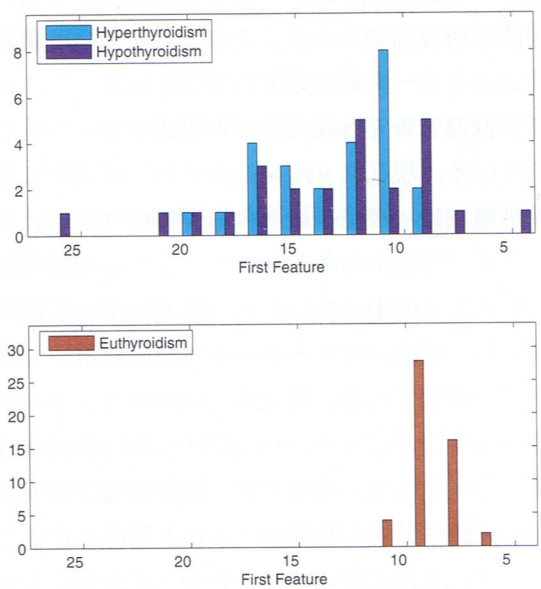

図1　通常のフィッシャー判別分析による甲状腺疾患データの解析結果。正常な患者と疾患を持つ患者のデータはうまく分離されるが，異なる疾患（機能亢進と機能低下）の患者は混ざってしまう。

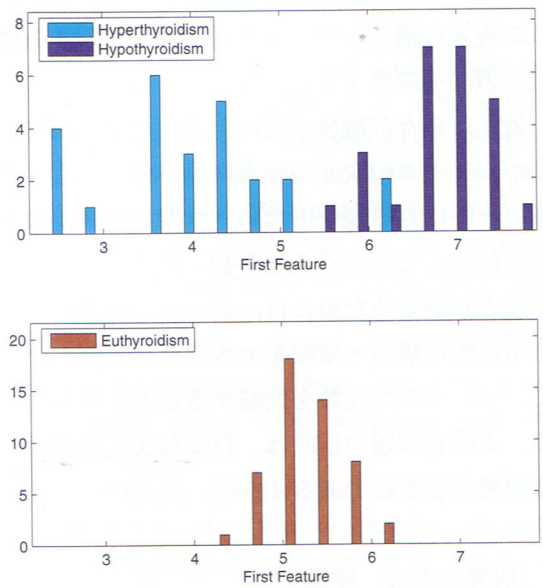

図2　局所フィッシャー判別分析による甲状腺疾患データの解析結果。正常な患者と疾患を持つ患者のデータはうまく分離され，異なる疾患（機能亢進と機能低下）の患者も分離される。

2) Sugiyama, M., Idé, T., Nakajima, S. & Sese, J.: Semi-supervised local Fisher discriminant analysis for dimensionality reduction. The 12th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2008), Osaka, Japan, May 20 - 24 (2008)

3) Kitamura, Y. & Sugiyama, M.:Dimensionality reduction of partially labeled multimodal data. In Proceedings of The 21st Annual Conference of The Japanese Society for Artificial Intelligence (JSAI 2007), no. 3 D 6 -1, Miyazaki, Japan, Jun. 18 - 22 (2007)

4) Sugiyama, M.: Local Fisher discriminant analysis for dimensionality reduction. In Proceedings of the Japanese Society for Artificial Intelligence, 3rd Meeting of Special Interest Group on Data Mining and Statistical Mathematics, SIG-DMSM-A 603 - 04, pp. 19 - 26, Kobe, Japan, Feb. 27 - 28 (2007)