1J1-OS9-2

データ間類似性判定のためのアクティブユーザフィードバックデザイン

Active User Feedback Design for Interactive Constraints Selection in Distace Metric Learning

岡部正幸*1　山田誠二*2
Masayuki OKABE　Seiji YAMADA

*1豊橋技術科学大学　*2国立情報学研究所／総合研究大学院大学
Toyohashi University of Technology　National Institute of Informatics

This paper describes an interactive system for distance metric learning that helps users to select effective constraints efficiently during the learning process. This system has some functions such as 2-D visual arrangement of a data set and constraint assignment by mouse manipulation. Moreover, it can execute distance metric learning and k-means clustering. In this paper, we introduce the overview of the system and how it works, especially in the functions of display arrangement by multidimensional scaling and incremental distance metric learning.

1. はじめに

近年，機械学習の分野ではデータ対に関する制約を利用してデータ集合内の距離尺度（または距離行列やカーネル行列）を学習する方法に関する研究が広く行われている．ここでいう制約とは，あるデータ対が同じグループに属するか否かに関する単純な知識で，クラスの存在が事前に明らかでないクラスタリングにおいても有効である．制約には 2 種類あり，同じグループに属すべきデータ対は must リンク，同じグループに属すべきでないデータ対は cannot リンクと呼ばれ，それぞれのデータ対間の距離が目標値になるようデータ集合全体の距離尺度を学習することが目的となる．

このような枠組みは距離学習と呼ばれ分類学習に有効であることが多くの研究から明らかになっているが，同時に制約が分類精度にもたらす効果にはばらつきがある（制約によっては性能低下を招く場合もある）ことも分かってきた．各制約において，データ対が must/cannot リンクであるかどうかを実際に判定するのは人間なので制約として大量のデータ対をランダムに選択することは，人手による判定作業コストの増大を招き，また制約の “品質” についてもバラつきを生むため得策とは言えない．このため，高い効果の期待できるデータ対のみを選択できるような戦略的な選択方法の獲得が望まれる．

本研究では，このような制約選択を行うためのアプローチの一つとして，人が制約を逐次的に選択・追加し，その効果を対話的に確認しながら分類学習を行えるシステムを試作した．このシステムでは主に以下のことを目的としている．

1. データ間の近接性を GUI を通して視覚的に確認することができ，マウス操作によってデータ対に容易に制約を付与することができるようにする．
2. 制約追加による分析結果を逐次的に確認することができる対話的環境の提供により，選択に関する戦略のヒントを得やすくさせる．

この試作システムは，多次元尺度構成法によってデータ集合を 2 次元の GUI 領域に描画し，GUI 上のデータをクリックすることで簡単に制約を追加することができる．また，逐次的な距離学習とクラスタリング（K-Means を実装）をバックグラウンドで走らせることができ，制約の効果を確認しながら戦略的な制約選択と分類学習を同時に進めることが出来る．

以降の章では，システムの概要と主要な構成モジュールである多次元尺度構成法によるデータ分布の描画および逐次的距離学習方法について説明する．

連絡先: 岡部正幸，豊橋技術科学大学情報メディア基盤センター
〒 441-8580 豊橋市天伯町雲雀ヶ丘 1-1
okabe@imc.tut.ac.jp

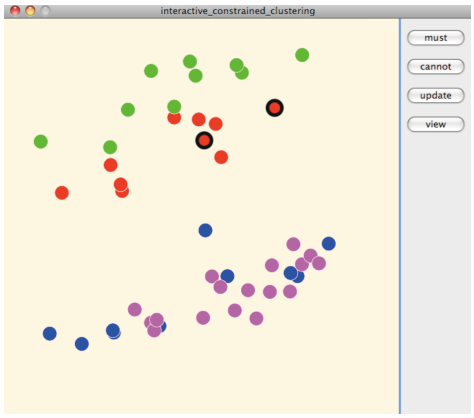


図 1: システムインタフェース

2. システムの概要

図 1 は我々が作成した試作システムの GUI 画面である．データを描画する領域と制約の追加，座標軸の変更および距離学習とクラスタリングを行うためのいくつかのボタンで構成されている．データは初期ロード時に制約の無い状態で一度クラスタリングが行われる．この結果に基づいて各データは色分けされ，多次元尺度構成法によって計算された座標に基づいて描画される．図中の色付けされた丸が各データを表している．

多次元尺度構成法を用いることで，各データ間の近接関係が反映されており，これを基にユーザは次に追加する制約を選択する．データの選択はマウスクリックで行うことができ，選択されたデータは黒の太丸で囲まれて強調表示される（図中では 2 つの赤丸のデータが選択されている）．データ対を選択した後，must/cannot ボタンを押すことで，そのデータ対が must/cannot リンクのどちらであるかをシステムに教えるこ

とができる．ちなみに選択データを再度クリックすることで選択解除ができる．追加する制約が決まったら，update ボタンを押すことで新しく追加した制約と既存の制約を利用した距離学習とクラスタリングが行われ，その結果に基づきデータが描画される．

ユーザは以上の作業を繰り返しながら，対話的に分類学習を進めていくことができる．

3. 多次元尺度構成法によるデータ配置

データ分布はユーザが制約を選択する際の重要な手がかりとなりうる．例えば，あるデータ対が cannot リンクであるにもかかわらずそのデータ間距離が小さければ，距離学習により修正することで全体の分類性能が向上，結果として制約として高い効果が得られる場合がある．このため，本システムでは 3 次元以上の属性を持つ データ集合においても集合内の近接関係をできるだけ保存した形でユーザに提示できるよう，多次元尺構成法 (Multidimensional Scaling, MDS) [1] を用いて描画を行っている．多次元尺度構成法は，データ集合の距離行列（非類似度行列）からデータの布置を決定する方法としてよく用いられており，距離学習を伴う本システムには都合がよい．

MDS は技術的には行列の固有分解に基づいている．S を正方行列，V を S の固有ベクトルを各列に並べた行列とすると，一般に S は V とその逆行列および対応する固有値を対角成分に並べた行列 Λ によって分解できるが，本システムでは S に距離行列（対称行列）を当てはめるので，逆行列を計算することなく以下のように表すことができる．

$$\begin{aligned} S &= V\Lambda V^T \\ &= V\Lambda^{1/2}(V\Lambda^{1/2})^T \end{aligned}$$

$V\Lambda^{1/2}$ の各行が MDS によって得られる座標となる．ただし，座標軸はクラスタ分割数に応じて対応する固有値の大きい順に選択しておき，view ボタンを押すことで軸の組み合わせを変更できるようにしておく．図 2(a) と (b) は軸の組み合わせの例を示している．データ集合は同じものであるが，MDS によって得られた座標軸の異なるものを組み合わせている．図 2(a) では紫と青のクラスタが重なっているが，図 2(b) では，軸の組み合わせが (a) の場合と異なるため分離されている．

4. 距離学習アルゴリズム

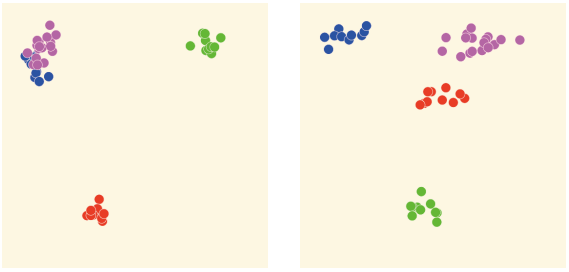
距離学習アルゴリズムはこれまでにいくつか提案されているが，本システムでは Jain らによって提案されたアルゴリズム [2] を用いている．このアルゴリズムはオンライン学習の枠組みに基づいており，制約を追加するごとに逐次的に距離行列を更新することができる．バッチ処理的に最適化計算を行う他の方法と比較して性能は多少低下するものの，ユーザによる制約付与から学習・再描画までの即応性を重要視する本システムには適した手法と言える．

学習の目的は，以下のようなマハラノビス距離におけるパラメータ行列 A を与えられた制約を満たすように学習することである．

$$d_A(\mathbf{u},\mathbf{v}) = (\mathbf{u}-\mathbf{v})^T A(\mathbf{u}-\mathbf{v})$$

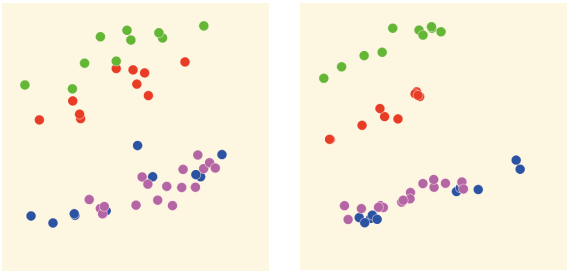
A の初期値は単位行列（すなわち最初の距離行列はユークリッド距離）とし，制約が与えられる度に更新される．Jain らは，パラメータ行列 A の更新式を以下のように定式化した．

$$A_{t+1} = \arg\min_{A\succ 0} D(A, A_t) + \eta l(d_A(\mathbf{u}_t,\mathbf{v}_t), y_t)$$



(a) 組み合わせ例 1　(b) 組み合わせ例 2

図 2: 座標軸の組み合わせによるデータ配置の違い



(a) 1 ステップ後　(b) 5 ステップ後

図 3: 距離学習の効果

ここで，A_t は t ステップ後（t 個の制約追加後）に更新されたパラメータ行列，$(\mathbf{u}_t,\mathbf{v}_t,y_t)$ を t ステップで与えられた制約，また，$D(A,A_t)$ と $l(d_A(\mathbf{u}_t,\mathbf{v}_t),y_t)$ はそれぞれ正規化関数と損失関数である．この式の意味するところは，現在と大きな差異のない範囲でできるだけ制約を満たすようなパラメータ行列 A を求めることであると言える．上式は更に近似処理などが加えられ，最終的に解析的に求めることが可能である．

この距離学習アルゴリズムをもとに制約付きクラスタリングを逐次的に行った結果を図 3 に示す．(a) では緑と赤のクラスタがわずかに重なっているが (b) では完全に分離されている様子が分かる．

5. まとめ

本論文では，ユーザが対話的に制約を追加しながら分類学習を行うためのシステムを提案した．このシステムは多次元尺度構成法によるデータ分布の描画，マウス操作による制約付与，また逐次的な距離学習とクラスタリング機能を持っており，ランダムに選択された制約を利用する場合よりも少なくかつ効果の高い制約選択を行うことを目的としている．また，対話的な操作を通してユーザが制約選択のためのより良い戦略を発案することを促す効果も期待している．

本システムの有効性については今後様々なデータを用いて検証していく予定である．

参考文献

[1] I. Borg and P. Groenen, “Modern multidimensional scaling,” *Theory and applications*, 1997.

[2] P. Jain and et al., “Online Metric Learning and Fast Similarity Search,” *NIPS’08*, pp.761-768, 2008.