視覚的でインタラクティブな分類器の構築手法

尉林 暉 * 五十嵐 健夫 †

概要. 近年機械学習の分野で研究されている様々な分類器には，分類器の内部での挙動を数値的に把握することはできても，直感的に見て取ることは難しいという問題がある．また，分類器の構築に必要なラベル付きデータを作成するのに手間がかかるという問題がある．我々はデータを 2 次元平面上に表示する可視化インタフェースを活用することで，一般ユーザでも容易にラベル無しのデータのみからユーザの好みに応じたラベルを持つ分類器を構築できる手法を提案する．システムは，まずラベルなしデータ間の類似性に基づいた 2 次元マップを生成する．その 2 次元マップ上にアノテーションすることで，ユーザがラベル付けを行う．システムは，新規データが入力されると，それを同じ 2 次元マップに投影することで分類を行う．本論文ではこの手法の紹介とともに，我々が行った実装例についても紹介する．

1 はじめに

近年様々な領域で機械学習が広がりを見せており，画像，文書，音声など様々なデータを学習し，分類，識別できるようになっている．その一方でこういった分類器を構築する際，分類器がどのようなデータを学習しどのような学習パフォーマンス出すのか，直感的に把握しながら構築するのは難しい．

本論文では画像や動画，文書のような目で見て扱えるデータの分類器を，視覚的で方法で構築，使用する手法を提案する．本手法はラベル付きの教師データを一切必要とせず，データのラベル付けから分類までの全てを一貫して 2 次元マップ上で視覚的に行える新しい概念である．一般ユーザにもラベル付けが可能であるが故に，ユーザの好みに応じた自由なラベル付けができる点にも新規性がある．

2 関連研究

可視化技術を用いたインタラクティブな機械学習システムにはいくつかの先行研究が知られている．画像検索に一般ユーザの嗜好を反映させる小規模なシステムとして Fogarty らの提案があり，画像をタイル状に並べる可視化が行われている [1, 3]．また，ユーザーの補助のもと大規模な学習器を作成する手法としてアクティブラーニングという概念があり [8]，これに可視化インタフェースを導入する試みもなされている [7].

これらと比べたときに，過去の提案が学習手法と可視化手法をそれぞれ別個で扱っていたのに対し，我々の提案する手法は可視化手法そのものを分類器として利用する点に新規性がある．それ故ラベル付けから分類までを同様の 2 次元マップ上で一貫して行うことができるという利点がある．また全てのラベル付けがユーザによって行われるので，ラベル付きのデータが一切必要ないという点も明確な違いである．

3 提案手法

以下に本提案手法を，ラベル無しのデータ群（以下サンプルデータ）から分類器を構築し，テストデータの分類を行うまでの，具体的なワークフローを記すことによって説明する．

サンプルデータの可視化 まず第一段階として，サンプルデータを 2 次元マップ上に可視化する．ここで用いる可視化手法は，サンプルデータを類似度に基づいて 2 次元で表示できるものであればどのようなものでも使用可能である．これらは次元圧縮技術として様々に研究がなされてる可視化技術であり，代表的なものに PCA[4] や t-SNE[5] が知られている．

このように可視化されたサンプルデータを見ることで，分類器が学習するデータ群の全体像を直感的に把握できるという点が本提案の利点である．

ユーザによるラベル付き領域構成 次にこうして表示されたサンプルデータの配置を元に，ユーザがラベル付き領域を構成する．図 1 中央のように，ユーザは似たデータが密集している部分を丸で囲い，そこにラベル名を付けていく．

この操作は一般的な教師データを作成する際のラベル付け操作に対応するものである．しかし，教師データに一つずつでラベルを与えていく一般的ラベル付けと比べ，近くにあるものはまとめてラベル付けできるため，データが多い場合でも効率よくラベル付けを進められる．また，ユーザがサンプルデータの全体像を把握しているが故，ラベルの種類や粒度をユーザの好みで変えられるのも利点である．

* University of Southern California
† 東京大学

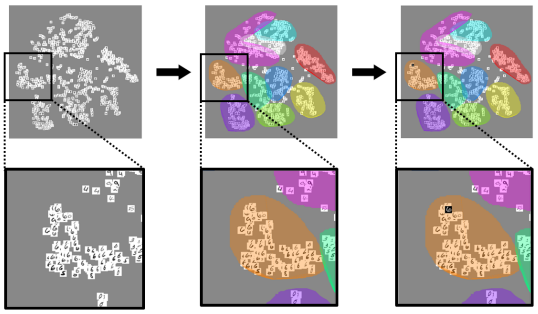

図 1. 本手法のワークフロー．可視化されたサンプルデータの分布に基づき，ユーザがラベル付き領域を作成する ．その上にテストデータが投影されることで，視覚的に分類が行われる．(テストデータは反転色で表示)

テストデータの分類 分類は，同様の可視化技術を用いて，作成した 2 次元領域上にテストデータを投影することによって行う．

こうして行われる分類はラベル名の情報以外に，テストデータの投影位置という視覚的な情報も含んでいるため，ユーザにとってより示唆に富むものである．具体的には，テストデータの位置とサンプルデータの全体像を比較することで，他のどのサンプルデータと似た（あるいは異なる）データなのかをデータ間の距離から見て取ることができる．これは分類器の挙動を直感的に把握できる本提案手法の利点である．

4 結果

上で述べた提案手法の実装例およびその使用結果を示す．ここでは対象のデータとして MNIST の手書き文字データを使用し，可視化手法としては t-SNE を使用する（図 1）．ここでは，Maaten らが提案している高速化された t-SNE[6] を用いている．

以下に筆者の手によってこのシステムを使用した結果を記すが，分類器としての性能に絞って実験を行うため与えるラベルは 0 から 9 の 10 種類に固定してラベル付けインタラクションを行った．実験環境として，Amazon Web Service で提供されている，8 コア CPU メモリ 32GiB の VPS(EC2 インスタンス,t2.2xlarge) を用いた．この実験では MNIST からランダムに選んだ 1000 個のデータを用いて実験を行い，10 分割交差検証によって F 値を推定した．t-SNE の主なパラメタとして Perplexity を 15 最大 Iteration 数を 1000 回の設定で実験を行った．

実験の結果 MNIST データの識別における本分類器の F 値は 0.6 と推定された．また，t-SNE による可視化及びラベル付け操作まで含めて分類器作成にかかった時間は平均して 2 分程度であった．

5 議論

提案システムの現状での課題として，既存の手法と比べて識別精度が低いという点がある．これは実用面では問題であるが，本論文ではラベル付けから分類までの全てを視覚的に行う新しい分類器の概念を提案するものであり，実験は予備的なものである．提案概念は CNN など既存の学習アルゴリズムと組み合わせによって精度を向上させることが期待でき，また Brown らが提案している手法を用いれば [2]，ユーザインタラクションによって距離関数を更新することで精度をさらに高めることができると予想している．本手法の，ユーザの嗜好を元に自由にラベル付けでき，一貫して視覚的に構築を進められる，という利点を損なうことなく，高い識別精度を実現する方法を探ることが今後の課題である．

その他に本提案概念には本質的な制約も存在する．まず，通常の機械学習のような大量のデータを使用することはできないという点である．本手法の肝はデータを分布させた 2 次元マップであるが，当然ながらユーザが視認できない程大量のデータは使用することができない．それゆえ，近年のニューラルネットワークで用いられているような１００万個規模の教師データを使用することはできない．使用可能なサンプルデータは最大数万個程度であると考えている．また，音声データなどの視覚で捉えられないデータには適用できないという点も制約の一つである．

参考文献

[1] S. Amershi, J. Fogarty, A. Kapoor and D. Tan. Overview based example selection in end user interactive concept learning. *UIST* 2009.

[2] E.T. Brown, J. Liu, C.E. Brodley and R. Chang. Dis-function: Learning distance functions interactively. *VAST* 2012.

[3] J. Fogarty, D. Tan, A. Kapoor and S. Winder. CueFlik: interactive concept learning in image search. *CHI* 2008.

[4] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology* 1933.

[5] L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research* 2008.

[6] L. van der Maaten. Accelerating t-SNE using tree-based algorithms. *Journal of Machine Learning Research* 2014.

[7] C. Seifert and M. Granitzer. User-based active learning. *International Conference on Networked Digital Technologies* 2010.

[8] B. Settles. Active learning literature survey. *University of Wisconsin, Madison* 2010.