6C-02

特徴訓練に基づいた分類器 FTApproach の提案

鄭　弯弯†　　金　明哲‡

同志社大学文化情報学研究科†　　同志社大学文化情報学研究科‡

1. はじめに

分類問題は外的基準ありとなしに分かれる。本研究では外的基準ありの分類問題を対象とする。分類に用いるデータは少なくとも個体と変数により構成され、次のように四種類に分けることができる。①個体と変数が少ないデータ；②個体が多く変数が少ないデータ；③個体が少なく変数が多いデータ；④個体と変数が多いデータ。一般的に、高精度の分類結果を得るためには十分なデータが必要であるといわれている (Zhu et al., 2015; Halevy et al., 2009; Mathur and Foody, 2008)。しかし、大量なデータがあっても、分類の精度が高くなるとは限らない。SVM (Support Vector Machine)と RF (Random Forest) は優れたアルゴリズムであり、現時点の機械学習領域において最も推奨されている分類器である。SVM はすべての変数を分類に用いるため、ノイズに影響されやすく、データの次元数が高い時、精度があまりよくない。一方、RF は個体と変数をランダムサンプリングするという点から、個体と変数が少ない時、精度があまりよくない。

本研究は、SVM と RF の欠点を改善した特徴を訓練するモデル FTApproach (Feature Traning Approach)を提案する。

2. FTApproach

FTApproach は主に三つの部分（特徴選択の部分；特徴訓練部分；SVM 多数決の部分）に構成される。その全体構造を図 1 に示す。

本研究はベースの特徴選択方法として IG (Information Gain)を用いる。IG は一つの代表的な特徴選択方法として、その有効性が検証されている (Geurts et al., 2018; Chinnaswamy et al., 2017; Shen et al., 2015; Wosaiak and Dziomdziora, 2015)。また、IG は学習ありの特徴選択方法であり、主成分分析のような学習なしの方法とカイ二乗のような距離ベースの特徴選択方法より、学習サンプル数に影響されにくい。特徴訓練はこの点を利用し、学習サンプルを徐々に増やして選ばれた特徴リストを更新していく。FTApproach は個体数の 2/3 をランダムサンプリングすることで訓練を始め、SVM を用いて予測することで終わる。同じプロセスを k 回行い、最後に多数決で各個体にラベルを決定する。

Feature training-based classifier FTApproach

† Wanwan Zheng, Graduate School of Culture and Information Science, Doshisha University

‡ Mingzhe Jin, Graduate School of Culture and Information Science, Doshisha University

3. 分析

3.1 分析データ

今回用いたすべてのデータはネット上で公開されているベンチマークデータであり、生物データ、画像データ、音声認識データ、物理データと人工データが含まれている。変数が少ないと多いデータは 10 組ずつを用いた。また、個体が少ないデータと個体が多いデータを作成するため、10 回ずつランダムサンプリングした。これで、個体と変数が少ないデータ（個体の数 5；変数の数 13－40）、個体が多く変数が少ないデータ（個体の数 40－100；変数の数 13－40）、個体が少なく変数が多いデータ（個体の数 5；変数の数 294－3,645）、個体と変数が多いデータ（個体の数 100；変数の数 294－3,645）をそれぞれ 10×10 組を作成した。分類は主に二群分類と三群分類を行った。

3.2 分析結果

分類器において、最もチャレンジになるデータは個体が少なく変数が多いデータである。変数の増加に伴い、過剰適合の可能性も高くなる。少数の学習サンプルであるにも関わらず、データの特徴を表せる学習モデルの作成が求められている。表 1 に個体が少なく変数が多いデータの結果を例として示す。FTApproach でデータの次元数を最小約 83.60%、最大 93%減少した。SVM の平均マクロ F 値は 0.64、RF は 0.62、FTApproach は 0.91 である。また、精度が上回った平均回数は 10 回の中に、SVM は 1 回、RF は 0.7 回、FTApproach は 9.7 回である。更に、多重比較検定を行ったところ、FTApproach のマクロ平均 F 値と SVM、RF に有意の差が見られた。

また、個体と変数が少ないデータに対して、どの分類器にも分類しにくい点があるが、FTApproach は最も高い精度を得た。個体が多く

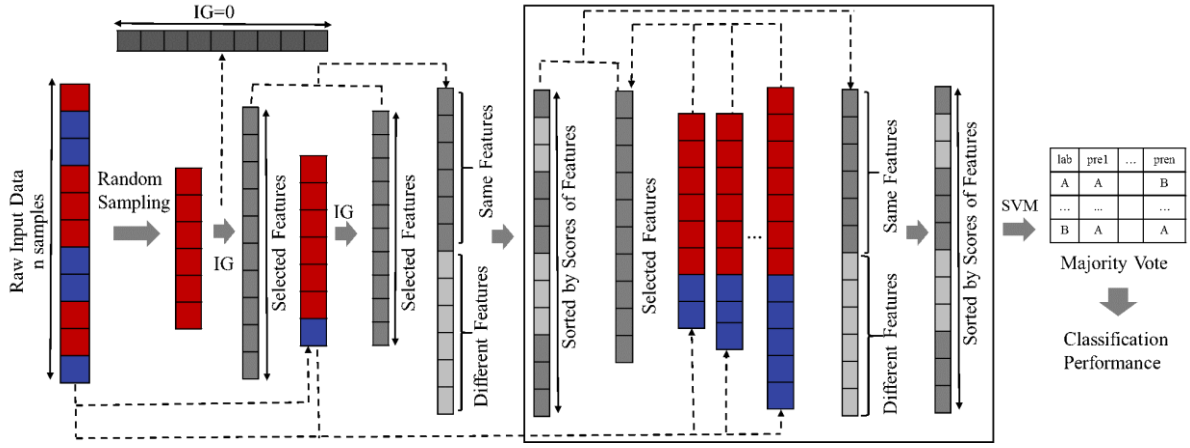

図 1　FTApproach の全体構造

表 1 個体が少なく変数が多いデータの分類結果

データ	Leukemia	Bioresponse	Gina	Scene	Eating	Isolet	Speech	Robert	Christine	Madelon	mean
Min (次元数減少)%	75	93	92	63	91	61	97	89	86	89	83.60
Max (次元数減少)%	88	99	96	91	95	77	99	94	93	98	93.00
Mean(SVM)	0.81	0.54	0.61	0.74	0.39	0.96	0.77	0.59	0.53	0.41	0.64
Mean(RF)	0.80	0.57	0.60	0.67	0.42	0.95	0.73	0.52	0.50	0.42	0.62
Mean(FTA)	0.98	0.93	0.92	0.94	0.58	0.98	0.82	0.95	0.95	0.99	***0.91***
Win(SVM)	0	0	0	2	1	7	0	0	0	0	1.00
Win(RF)	0	0	0	1	1	5	0	0	0	0	0.70
Win(FTA)	10	10	10	10	8	9	10	10	10	10	***9.70***
p(SVM-RF)											
p(SVM-FTA)	***	***	***	**	**		*	***	***	***	
p(RF-FTA)	***	***	***	***	*		***	***	***	***	

*** p < 0.001, ** p < 0.01, * p < 0.05,† p < 0.1

変数が少ないデータには、ノイズが少なく、また学習サンプル数の増加は SVM に有利である。一方、RF は個体をランダムサンプリングするため、学習サンプルの増加という点では RF に有利である。このような SVM と RF に有利なデータに対して、RF は最も高い精度を示し、続いては FTApproach、SVM になる。個体と変数が多いデータは、学習サンプルの増加は SVM に有利であるが、変数の増加は不利点になる。一方、このようなデータは RF が得意であるが、多くの場合には FTApproach は精度が最も高く、続いては RF と SVM である。

4. まとめ

本研究は、特徴学習に基づいた分類器 FTApproach を提案した。ベンチマークデータ用いた比較分析の結果、分類器に対して最も分類しがたい 2 種類データ、個体と変数が少ないデータと個体が少なく変数が多いデータにおいては SVM、RF より高い精度を得た。

その理由としては以下の点が考えられる。

- FTApproach は特徴選択があるため、変数が多い場合にノイズに影響されやすい SVM の欠点を改善することが期待できる。
- 学習サンプルを徐々に増やして選ばれた特徴リストを更新することは、異なる学習データで繰り返し学習させることと同様の効果が得られ、学習サンプルが小さい時に効果がないと言われている RF の欠点を克服することが期待できる。
- 多数決によるラベル付けの導入は、高精度を得ることを“保証”することができる。

参考文献

[1] Mathur, A. and Foody, G. M., Crop classification by a support vector machine with intelligently selected training data for operational application, International Journal of Remote Sensing, 29, 2227-2240, 2008.
[2] Zhu, X., Vondrick, C., Fowlkes, C. C., and Ramanan, D., Do we need more training data?, International Journal of Computer vision, 119(1), 76-92, 2016.