操作性･汎用性の高い データマイニングツール FlexibleDataMiner の開発

越前　陽祐(15602018) 、　櫻沢　研一 (15802036)
原田研究室

１．はじめに

近年、膨大なデータから意味ある情報を見つけ出すデータマイニングの技術が注目を集めている。その中でよい結果を導くことが可能な、ニューラルネット、決定木、SVM は、様々な分野に幅広く用いられるようになってきている。

このような需要と共に、データマイニングを用いるためのソフトウェア（データマイニングツールと呼ぶ）が数多く出回るようになった。しかし、現状のデータマイニングツールには様々な問題が存在する。例えば、専門外の人がコマンドラインから用いるソフトを使いこなせるようになるには時間がかかる、パラメータなどの設定が難しいなどの利便性の問題が存在する、などである。また、データマイニングツールに内蔵されている決定木や SVM 等の機能を自由に改変できないため、応用プログラムに組み込むことが難しいという問題も挙げられる。そこで、本研究ではこれらの問題を解決するため、独自のデータマイニングツール FDM の開発をすることとした。

図-1 FDM の実行画面

２．プログラムの構成要素

本研究では、FDM にフィルタリング機能、および決定木と SVM による学習・予測機能を実装した。

2.1 フィルタリング

フィルタリングとして、属性の削除、数値の離散化、欠損値補充、標準化、名義属性から数値属性への変換などを実装した。これを行う際には、データを図表示したり、フィルタリングの説明を載せたりして、初心者にも分りやすいような工夫をこらした。

2.2. 決定木

FDM においては、既存の C4.5 というプログラムにおける学習アルゴリズムを C 言語から C#言語に書き直し、GUI 環境で扱えるようにした。これにより、コマンドラインからコマンドを打ち込む手間が省け、かつどのような機能のオプションがあるのかが、一目瞭然となった。

2.3. SVM(Support Vector Machine)

SVM とは、2 値判別の手法の一つである。FDM においては既存の SMO アルゴリズムを実装した。また、これを拡張し多値判別も行うことができるようにした。さらに、非線形 SVM の実験を何回も行った経験より一番適切そうなパラメータを求め、これをデフォルト設定とし、初心者にも使いやすいようにした。

3. 結果

具体的なデータで評価実験を行ったところ以下の結果を得た。

3.1 自然言語処理のプログラムへの応用

本研究の SVM を自然言語処理のキーワード抽出に応用したところ去年使用した tinySVM の結果を上回った。

表-1 キーワード抽出の評価実験の結果

	クローズドテスト		オープンテスト	
	去年	今年	去年	今年
正解率(%)	75.9	94.0	77.6	93.3
適合率(%)	73.1	93.7	76.7	94.9
再現率(%)	91.2	97.1	92.1	95.8
F値	81.2	95.4	83.7	95.4

3.2 決定木と SVM の比較

決定木と SVM の比較をするために以下のようなデータを学習し、評価テストを行ったところ下表のような精度(誤り率)と速度となった。

表-2　SVM と決定木の比較

データ情報					誤り率(%)			
データ名	クラス	属性数	学習データ数	テストデータ数	クローズテスト		オープンテスト	
					決定木	SVM	決定木	SVM
breast-cancer	2	9	200	77	19.0	9.0	28.6	29.8
german	2	20	700	300	21.3	1.1	29.3	23.0
image	2	18	1300	1010	3.5	3.1	3.6	3.4
ringnorm	2	20	400	7000	1.5	0.0	14.8	1.8
splice	2	60	1000	2175	1.8	0.0	6.5	10.9
titanic	2	3	150	2051	20.0	19.3	22.6	22.9
twonorm	2	20	400	7000	2.0	0.0	20.9	4.3
waveform	2	21	400	4600	0.5	0.3	17.9	14.0

表-3　分類時間の比較

手法	モデルの生成時間	分類時間
決定木	2分13秒	27秒718
SVM	17時間8分	37分22秒

決定木は速く属性が少ない場合に有効であり、SVM は時間がかかるが属性が多い場合に有効である。