科学研究費助成事業　　研究成果報告書

平成　２６　年　　６　月　　９　日現在

機関番号：　３４４２７
研究種目：　基盤研究(C)
研究期間：　2011～2013
課題番号：　２３５００２８４

研究課題名（和文）ファジィｃ平均識別器の高精度・高機能化

研究課題名（英文）Improvement of fuzzy c-means classifier with respect to precision and functions

研究代表者
市橋　秀友（Ichihashi, Hidetomo）

大阪経済法科大学・経済学部・教授

研究者番号：３０１５１４７６

交付決定額（研究期間全体）：（直接経費）　　4,000,000円、（間接経費）　　1,200,000円

研究成果の概要（和文）：クラスタリングとパラメータ最適化による簡便な識別器であるファジィc-平均識別器(FCMC)の識別時間（テスト時間）と訓練時間（事前計算時間）の大幅な改善手法を開発した．そして，FCMCの訓練時間を高性能なサポートベクターマシンとして知られているLibSVMと比較した．四つのパラメータのうちの二つを自動最適化する．LibSVMのパラメータ数は２である．改良されたFCMCではLibSVMと同等の識別精度が得られ，100万件以上の大量データでの訓練時間は，LibSVMに比べて100倍から1000倍の高速化を実現した．

研究成果の概要（英文）：Fuzzy c-Means based Classifier (FCMC) is a simple approach to classification based on the clustering and parameter optimization methods. The training time and testing time of FCMC are significantly improved. FCMC and the state of the art classifier: LibSVM are compared. The two parameters are automatically optimized by the revised random search approach. When the number of training samples is more than a million, the total training time for FCMC is estimated to be two to three orders of magnitude smaller than LibSVM, while FCMC achieves the same level of classification accuracy with LibSVM.

研究分野：　総合領域

科研費の分科・細目：　情報学、感性情報学・ソフトコンピューティング

キーワード：　識別器　クラスター分析　画像処理

１．研究開始当初の背景

(1) 当研究グループでは，クラスター分析法を前処理として用いるのではなく，直接的に用いることでパターン識別器の性能向上を目指した研究を進めていた．またその実システムへの応用として駐車場管理システムに用いたものが実用化された．

(2) 既存の優れた識別器として，サポートベクトルマシン(SVM)が世界的にもよく知られている．開発してきた識別器と SVM との相違点を明らかにし，研究の意義を明確化する必要があった．

２．研究の目的

(1) 種々の大量データの識別問題に適用し，識別精度が SVM より優れていることを明らかにする．

(2) 種々の大量データの識別問題に適用し，計算機での訓練時間と識別時間が優れていることを明らかにする．

(3) 上記の性能をさらに改善する．

３．研究の方法

(1) ファジィ c-平均識別器（Fuzzy c－Means Based Classifier，FCMC）は，訓練データが大量であればクラスター数を多くすることで訓練データに対する精度を向上できると考えられる．大量訓練データを用いて訓練データ数が変化した場合の性能を比較する．クラスター数をあまり多くせずに，逆に訓練データに対して最適化するパラメータ数を多くした場合の比較とする．世界的に実用的なツールとして認められている LibSVM を用いて，テストデータの識別精度を比較する．

(2) ファジィ c-平均識別器(FCMC)の訓練時間の改善方法を検討する．まず，メモリー不足が発生しないように訓練データを分割して読み込み，訓練データのクラスターへのメンバシップもデータとしてハードディスクに分割して書き出す．ランダムサーチによるパラメータ最適化についても高速化を図る．そして，FCMC の訓練時間を LibSVM と比較する．

(3) MATLAB での計算速度を高速化するためにプログラムの部分的な C 言語への組換えを行う．

４．研究成果

(1) いくつかの大量ベンチマークデータを用いて訓練データ数の識別精度への影響を比較した．パラメータ最適化は訓練データに対して行い，c と r を自由パラメータとしてテストデータの精度が良くなるように選択した．最適な自由パラメータは既知であるとして LibSVM と比較した．

表１：性能比較に用いたデータ

データ	特徴量	訓練用	テスト用
Parking	50 (1,024)	135,000	135,000
USPS	50 (676)	266,079	75,383
Checker	2 (2)	1,000,000	2000
KDD-CUP	5 (127)	2,449,216	311,029
MNIST	50 (576)	60,000	10,000

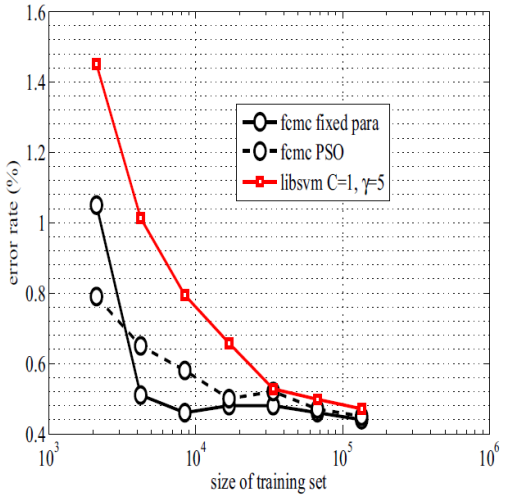

図１：駐車場の画像データでのテスト（識別）精度の比較

図１はパラメータを固定して識別を行った結果を実線と○で示している．訓練データ数が少なくなるほど LibSVM のテスト精度が悪くなった． LibSVM は訓練データ数が小さくなった場合に二つのパラメータを選びなおしてもほとんど改善されなかった．FCMC は汎化能力に優れ，少ない訓練データで高い精度が得られることを示している．

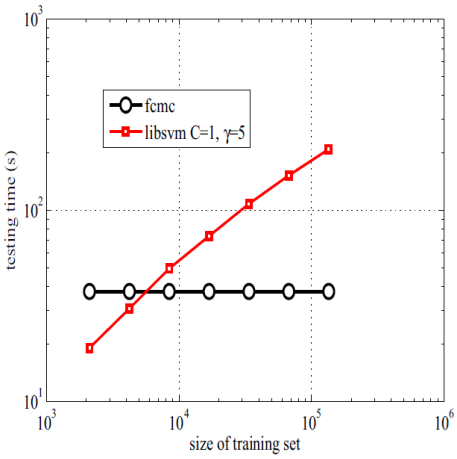

図２：駐車場の画像データでのテスト（識別）時間の比較

図２は，訓練データが増加してもテストデータに対する FCMC の識別時間は変化しないが

LibSVM では，急増することを示している．この傾向は他のデータを用いた場合も同様であった．

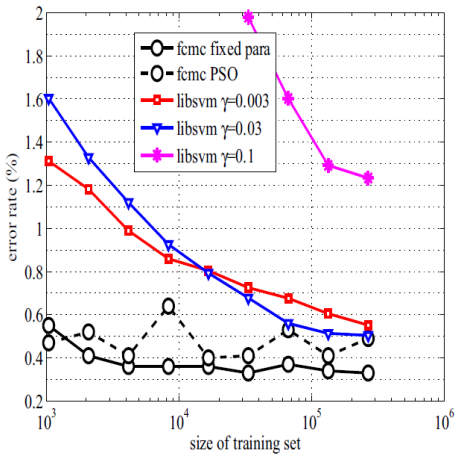

図３：郵便番号データでのテスト（識別）精度の比較

図３は，FCMC が郵便番号データでも訓練データ数が小さい場合にテスト（識別）精度に優れていることを示している．

(2) FCMC の計算機による訓練時間の改善を図った場合の訓練時間の比較を表２に示す．

表２：駐車場データでの訓練時間の比較

training sample 200,000, test sample 70,000, feature dimension : 50		
	FCMC (c=16)	LibSVM
best hyper-parameter	m=0.1929, γ=14.4254	C= 3.58, g= 5.40
test error	0.51%	0.39%
total training time	85.8+32.2+37.5=**155.5**s	estimated training time (50 times) **78531.0**s [21.8h]

	FCMC (c=32)	FCMC (c=64)
best hyper-parameter	m=0.1219, γ=8.5049	m=0.1913, γ=15.4249
test error	0.31%	0.36%
total training time	154.6+63.4+62.3=**280.3**s	293.1+127.7+113.6=**534.4**s [0.15h]

データは 20 万件程度であるが，LibSVM に比べて非常に高速であり，件数が増えるほどその差が大きくなった．

表３：改善された FCMC の訓練時間の比較

	LibSVM	kd(c=32)	PCA(c=32)
hyper-parameter	C=1.34 g=4.96	m=0.1131 γ=8.4473	m=0.1569 γ=11.6914
test error	0.39%	0.45%	0.35%
total time	52006s [14.4h]	76.0s	87.0s

(3) 表３は，MEX のよる部分的な C 言語の採用などで改善された FCMC での駐車場データでの訓練時間を LibSVM と比較したもので，1.5 倍高速な計算機を用いた場合，FCMC は３倍高速化されている．総訓練時間は 357 秒で，LibSVM より 146 倍高速である．

表４：KDD データでの訓練時間の比較

Training Time(KDD)

C ∈ [0, 10]
g ∈ [0, 10]

	SVM	kd(c=8)	PCA(c=8)
hyper-parameter	C=100 g=11	m=0.6916 γ=3.2813	m=0.6114 γ=2.8762
test error	5.24%	4.98%	4.93%
total time	618000s [17.1h]	22.9s	23.2s

表４は KDD データの約 240 万件を用いた場合の比較で，FCMC の総訓練時間の推定値は約 47 秒であるのに対して，データ数が大きいために LibSVM では約 1 万倍の時間がかかっている．正確な訓練時間の見積もりは困難であるが，LibSVM のハイパーパラメータの最適化（探索）は区間[0,1]内では正解を見つけられず，c=100 である．すなわち，LibSVM で前提としている 50 回の探索では到底不可能であると考えられる．以上は訓練時間の改善結果であるが，表５は駐車場データでのテスト

表５：駐車場データでのテスト（識別）１件当たりの時間の比較

Testing Time(Parking)

	FCMC(c=32)	LibSVM
best hyper-parameter	m=0.1569 γ=11.6914	C=3.58 g=5.40
test error	0.35%	0.39%
test time(MATLAB)	0.883ms	
test time(revised code)	0.221ms	1.510ms

Compare to LibSVM: **7** times faster

Compare to MATLAB: **4** times faster

時間の改善結果を示している．FCMC のコードに MEX を用いて部分的に C 言語化したことで MATLAB のみに比べて４倍高速化でき，LibSVM に比べて７倍高速になった．ただし，LibSVM は C 言語のコードが用いられている．

表６：KDD データでのテスト（識別時間）

Testing Time (KDD)

	FCMC(c=8)	LibSVM
best hyper-parameter	m=0.6144 γ=2.8762	C=100 g=11
test error	4.93%	5.40%
test time(MATLAB)	0.109ms	
test time(revised code)	0.003ms	0.573ms

LiBSVM: **190** times faster

MATLAB: **36** times faster

表６の結果はデータベースからの知識発見に用いられた KDD データを用いた場合である．MEX による C 言語化で 36 倍高速化できた．その結果，LibSVM に比べて識別段階での計算時間（テスト時間）は 190 倍高速である．

FCM 識別器の訓練（計算）時間と識別（テスト）時間の改善を行い，世界的にもその有効性が認識されている識別器であるサポートベクトルマシンに比べて，識別精度は同等であるが，計算時間は格段に高速であることを明らかにした．

(4) 現在駐車場の管理システムや，ごみ焼却場のごみ量を図るレベルセンサーや，燃焼状態の識別器など，種々の実用化が進行中である．また，研究成果を報告した論文「ファジィｃ－平均識別器の訓練時間の改善」は，2012 年度日本知能情報ファジィ学会論文賞を受賞した．

５．主な発表論文等

〔雑誌論文〕（計　６件）

① 市橋秀友，本多克宏，野津 亮，多数のパラメータを用いるファジィｃ-平均識別器の訓練データ数による性能比較，知能と情報（日本知能情報ファジィ学会誌），査読有，Vol.23，pp. 254-263 (2011)

② 市橋秀友，本多克宏，野津 亮，ファジィｃ－平均識別器の訓練時間の改善，知能と情報（日本知能情報ファジィ学会誌），査読有，Vol. 23，pp. 783-793 (2011)

③ T. Ogita, H. Ichihashi, A. Notsu, K. Honda, Improvement of PCA-based Approximate Nearest Neighbor Search Using Distance Statistics, Journal of Advanced Computational Intelligence and Intelligent Informatics, 査読有, Vol.18, pp. 1-7 (2014)

〔学会発表〕（計　１８件）

① H. Ichihashi, K. Honda, A. Notsu, Comparison of Scaling Behavior Between Fuzzy c-Means Based Classifier with Many Parameters and LibSVM, IEEE International Conference on Fuzzy System, 2011 年 6 月 28 日, Taipei, Taiwan

② H. Ichihashi, K. Honda, A. Notsu, Fuzzy c-Means Based Classifier - Comparison with LibSVM in Terms of Training Time for Parameter Optimization, 12th International Symposium on Advanced Intelligent Systems, 2011年9月30日, Suwon, Korea

③ 小林卓矢，市橋秀友，本多克宏，野津亮，ファジィ c-平均識別器の MEX と Visual-C での計算時間の改善，ファジィシステムシンポジウム，2012 年 09 月 12 日～2012 年 09 月 14 日，名古屋

④ T. Kobayashi, H. Ichihashi, K. Honda, A. Notsu, Mixed Usage of MATLAB and Visual C for Improving Classification Time and Training Time of FCM Classifier , International Conference on Soft Computing and Intelligent Systems and 13th International Symposium on Advanced Intelligent Systems, 2012 年 11 月 20 日～2012 年 11 月 24 日, Kobe, Japan

〔その他〕ホームページ:
http://ichihashi.jimdo.com/

６．研究組織

(1)研究代表者
市橋　秀友（ICHIHASHI Hidetomo）
大阪経済法科大学・経済学部・教授
研究者番号：３０１５１４７６

(2)研究分担者
本多　克宏（HONDA Katsuhiro）
大阪府立大学・大学院工学研究科・教授
研究者番号：　８０３３２９６４

野津　亮（NOTSU Akira）
大阪府立大学・大学院工学研究科・准教授
研究者番号：　４０４０５３４５