様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２９　年　　６　月　１２　日現在

機関番号：　２５４０６
研究種目：　基盤研究(C)（一般）
研究期間：　2014～2016
課題番号：　２６３３０２５１

研究課題名（和文）実時間カーネル行列学習によるストリームデータ分類

研究課題名（英文）Stream data classification with real time learning of kernel matrix

研究代表者
岡部　正幸（Okabe, Masayuki）

県立広島大学・経営情報学部・講師

研究者番号：５０３６２３３０

交付決定額（研究期間全体）：（直接経費）　　3,500,000円

研究成果の概要（和文）：本研究では，機械学習による高精度なデータ分類器の生成に欠かせない技術であるカーネル行列学習を，大規模かつ流動的な性質を持つストリームデータの分類タスクにおいて利用可能とすることを目的とし，それに対応する逐次・高速処理可能なカーネル行列学習アルゴリズムの構築と制約付与データ対の能動的選択アルゴリズムの構築を行った．また，ストリームデータ分類の応用例として，ネットワークトラフィックデータからの異常検知システムを構築し，実環境における運用を通じて構築したアルゴリズムの実用性について検証した．

研究成果の概要（英文）：Kernel matrix learning is an indispensable technique for machine learning to make of high accuracy. In this research, we developed an algorithm of kernel matrix learning that can be applied to incremental stream data classification and then proposed an active learning method that selects candidate data pairs to be labeled as constraints. We verified the utility of our developed algorithms through the experiments of outlier detection from network traffic.

研究分野：　知能情報学

キーワード：　カーネル行列学習　アンサンブル学習　制約付きK-means

１．研究開始当初の背景

機械学習において，データ間の類似尺度は分類・クラスタリング精度を決定する極めて重要な要素である．カーネル行列学習は，いくつかのデータ対に目標となるカーネル値（＝類似度）を制約（＝訓練データ）として与えることにより，それらの制約に応じた全データ対のカーネル値を学習する技術であり，既存のカーネル関数ではうまく分離できないデータ集合に対しても適切な類似尺度を提供することができる．

カーネル行列学習は種々の分類・クラスタリングアルゴリズムに組み込むことができる汎用性の高い要素技術であり，データマイニング・情報検索・画像認識・統計的言語処理など多岐に渡る応用分野での利用が見込まれる．しかしながら一方で，従来のカーネル行列学習方法の多くは，半正定値計画などの計算量の高い最適化問題をベースとしておりデータ数に対する拡張性に難がある．また，逐次学習に対応していないためバッチ処理を前提としたタスクに限られるなど，実用上解決すべき問題点を抱えている．

カーネル行列学習を実用的な分類システムにおいて幅広く利用するには，大規模かつ流動的なデータへ適用可能なアルゴリズムの開発が必要不可欠である．このようなデータはストリームデータと呼ばれ，通信パケット，データベーストランザクション，オンラインニュースなど実世界に幅広く存在する．ストリームデータは性質の変動を伴いながら絶え間なく発生するため，継続的な学習によって分類器を更新していく必要がある．

ストリームデータ分類では，一般に蓄積可能なデータ量は限られており，学習対象となるデータ集合の更新が頻繁に発生する．カーネル行列学習はこれに対応し，逐次的な再学習，更には制限時間内に必ず結果を返す実時間処理に対応することが必要となる．これには，従来のように精度のみを重視するのではなく，計算コストとのトレードオフを考慮した新たなアルゴリズム設計が必要となる．また，ストリームデータ分類では新規データの継続的な分類処理が行われるが，データの性質が変動した場合に備え，制約集合の更新を能動的に行う必要がある．学習効果を高める制約の能動的選択により，再学習による分類器の性能維持・向上を効率的に行うため，通常の能動学習よりも計算コストを抑えた期待効用計算のためのアプローチが必要となる．

２．研究の目的

本研究では，ストリームデータ分類に対応するカーネル行列学習アルゴリズムの構築とその実環境での評価を目的とする．具体的に行うのは以下の 3 点である．

(1) 実時間高速カーネル行列学習アルゴリズムの構築

ストリームデータ分類では，大規模かつ流動的なデータに対し継続的な学習による分類器の更新が必要となる．これに対応する逐次・高速な分類処理を実時間で行うことのできるアルゴリズムの構築を目指す．

(2) 制約付与データ対の能動的選択アルゴリズムの構築

ストリームデータ分類における性質の変動を伴うデータの継続的な分類処理に対応するため，新たに制約を付与すべきデータ対を能動的に選択するためのアルゴリズムを開発する．ランダム選択と異なり，能動的な選択では学習の費用対効果を高めるため，制約として利用した場合に得られる効用の期待値（期待効用）を各候補について計算し，その最も高いものを選択する．この期待効用の計算は通常は全候補について計算する必要があるが，ストリームデータ分類においてこの計算を実時間で行うのは困難である．よって，期待効用の計算対象を合理的に限定し，計算コストを削減するためのアルゴリズムの構築を目指す．

(3) 実環境ストリームデータ分類システムへの応用展開

上記 1，2 で構築したアルゴリズムの実用性について検証するため，ストリームデータ分類の応用例として，ネットワークトラフィックデータからの異常検知システムを構築する．このシステムは，本研究で構築したアルゴリズムを基に作成する分類エンジンを核とし，データ可視化機能・制約付与のためのインタフェースなどを備えたものとする．また，本システムの実環境における運用を通して構築アルゴリズムの実用性について検証する．

３．研究の方法

(1) 実時間高速カーネル行列学習アルゴリズムの構築

ストリームデータ分類に対応する実時間処理を達成するには，学習対象となるデータ集合が更新された時に効率的な再学習を行うことが必須となる．ベースとなるカーネル行列学習アルゴリズムには，高速処理可能なものであること，前回の学習時の途中経過などを再利用しやすいものであることなどの条件が求められる．

(2) 制約付与データ対の能動的選択アルゴリズムの構築

本研究では，能動学習の基本的なアプローチに基づき，制約を与えるべきデータ対，つまり，制約として利用した場合に得られる効用の期待値（期待効用）の最も高いデータ対の選択アルゴリズムを構築する．ただし，本研究ではストリームデータを対象としているため，通常の能動学習のように選択の可能性のある全候補の期待効用について計算する余裕はない．特にカーネル行列学習では，制約を与える対象はデータ対であり，その候補数は全データ数の 2 乗のオーダになる．そ

のため各候補の期待効用の計算を実時間で処理することは困難となることが予想される．よって能動学習の対象とする候補集合を合理的に縮小するための方法論が必要であり，本研究ではその合理性に関する説明と具体的なアルゴリズムの構築を行う．

(3) ネットワークトラフィックからの異常検知システムの構築

検知対象は，主に学内におけるファイル交換ソフトの利用とする．組織内におけるファイル交換ソフトの使用は著作権法違反をはじめとするコンプライアンス違反につながる可能性があり，その検知システムは実用上の有用性も高い．

学習対象となるデータはネットワークトラフィックパケットであり，それらがファイル交換ソフトによるものであるかどうかを判定する分類器の生成を目的とする．トラフィックデータは，学内ネットワークからタップ機器を介したミラーリングにより，tcpdump，tcptrace などのツールを利用して取得する．分類器には k-NN，SVM など既存の高速分類器に本研究で構築したカーネル行列学習を組み込んだものを使う．また，ファイル交換ソフトによる通信パケットの判定にはソフトに依存した個別の専門知識を必要とするため，システム評価のための正解データの作成には，専門のファイアウォール機器による判定結果を用いる．

４．研究成果

(1) アンサンブル学習を利用した制約付き k-means によるカーネル行列学習

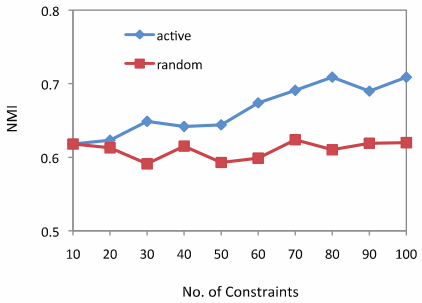
本研究では，計算コストの低い制約付きクラスタリングアルゴリズムの一つである COP-Kmeans に着目し，そのクラスタリング性能を補うため，アンサンブル学習の原理を利用し複数のクラスタリング結果をカーネル行列に変換・統合することでカーネル行列学習を行う方法を提案した．提案方法は，優先度に基づいて制約充足を行う制約付き K-means アルゴリズムをデータ対の半教師付き 2 値分類を行う弱学習器として利用し，AdaBoost をベースとしたブースティングによってその分類精度を向上させる．また，データ対の分類結果を要素とした行列を統合することで，カーネル行列学習を実現している．実験では，提案手法と従来手法のクラスタリング性能と計算時間について 12 種類のデータセットにおいて比較し，提案手法が多くのデータセットにおいて計算時間が少なくかつ同等以上の性能を持っていることを示した（図 1）．また，提案手法ではクラスタリング性能がブースティングラウンド数の増加とともに向上することを示し，ブースティングラウンド数を適切に設定することで計算時間を必要最小限に抑えることが可能であることがわかった．

	Iris (UCI)		
手法	1%	5%	10%
BCKM	0.02	0.27	0.49
KBST	134.71	172.94	122.15
BSTC	0.68	1.22	1.73
ITML	0.23	2.01	2.21
PCP	0.59	1.38	2.45
CKM	0.01	0.01	0.02

図 1 実験結果の例

(2) 能動的制約サンプリングによるカーネル行列学習

本研究では，カーネル行列学習を効率的に行うための，能動的な制約サンプリング方法の提案を行った．提案手法は，制約付き K-means を弱クラスタリングアルゴリズムとしたバギング（Bagging）に基づくカーネル行列学習をベースとしている．このアルゴリズムでは，バギングの過程において制約付き K-means によるクラスタリングが繰り返され，各クラスタリング結果において任意のデータペアは同一クラスタに属する・属さないかのどちらか 2 つの状態になる．能動的サンプリングは，この 2 つの状態のバギング過程における推移に基づき行う．この方法は教師あり分類学習における不確実性サンプリング（Uncertainty Sampling）に基づいている．つまり，あるデータペアについて同一クラスタに属するまたは属さないかのどちらか一方の状態が多ければ，そのデータペアの推定は容易でサンプリングの効果は期待できないとし，逆にあるデータペアについて同一クラスタに属するまたは属さないかの状態が拮抗していれば，そのデータペアの推定は困難でサンプリングによる制約としてのラベル付けの効果が期待できるとの仮定に基づいた方法である．6 つのデータセットを用いて，ランダムサンプリングによる方法と提案手法を比較した結果，提案手法の方がより少ない制約数で同じ性能を実現できるとする効果が確認できた（図 2）．



(a) glass

図 2 実験結果の例

(3) 外れ値尺度の可視化によるネットワークトラフィックログからの異常検知システム

本研究では，ネットワークトラフィックデータからの異常検知を対話的に行うシステムを提案した(図 3)．このシステムの特徴は，2 種類の外れ値尺度に基づいてデータを可視化する点にある．一つ目の尺度は相対的な外れ度で，特定の 2 軸を用いた散布図により外れ値を可視化する．特定の 2 軸の選択は，対話的に高速に行えるようになっており，研究成果(1)のアルゴリズムを用いて選択の優先順位を付与することも可能である．二つ目の尺度は時系列変化における外れ度で，あるデータ点の散布図上におけるプロット点の時系列変化をアニメーションにより可視化する．ある時点の散布図上において外れ値の可能性が高くても時系列変化がなければ，定常処理を行っている場合も考えられ異常トラフィックである可能性は下がる．逆に時系列変化が大きくある時点の散布図で突然外れ値になるトラフィックは異常である可能性が高い．このように，本システムを用いることで，まず多く通信ホストの中から異常性の高いトラフィックをもつホストを絞り込み，次にそのホストが行っているトラフィックの時系列変化を見ることで異常かどうかの判定を行うという 2 段階の作業によって効率的に異常を検知することができる．本システムを用いた予備実験では，P2P 通信を行っているホストなどを見つけることができた．

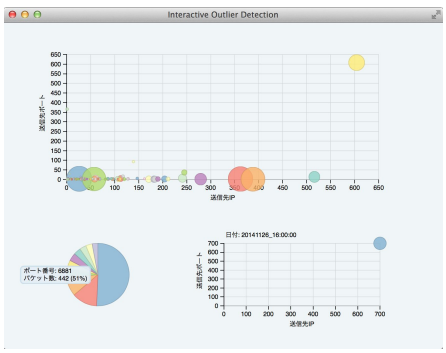


図 3 異常検知システム

5．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 1 件）

① 森國泰平，吉田光男，岡部正幸，梅村恭司，“ツイート投稿位置推定のための単語フィルタリング手法”，情報処理学会論文誌：データベース，Vol.8，No.4，pp.16-26（2015）（査読有）

〔学会発表〕（計 4 件）

① 岡部正幸，山田誠二，“外れ値検出に基づく対話的ファイアウォールログ分析”，第 28 回人工知能学会全国大会，3B3-OS-10a-4（2014）

② 福永度宗，山田誠二，岡部正幸，“非明示的フィードバックにより訓練データ選択を支援するインタラクションデザイン”，第 28 回人工知能学会全国大会，3B3-OS-10a-2（2014）

③ 中野翔平，菊地真人，吉田光男，岡部正幸，梅村恭司，“信頼区間の下限値による確率推定を用いた企業名抽出”，第 8 回データ工学と情報マネジメントに関するフォーラム，E8-1（2016）

④ 井本博之，岡部正幸，高間康史，“従属クラスタ動的生成機構を導入した Must-Link 制約付き K-means の提案”，第 9 回 WI2 研究会，pp.63-64（2016）

6．研究組織

(1)研究代表者

岡部　正幸（Okabe，Masayuki）
県立広島大学・経営情報学部・講師
研究者番号：50362330