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　３０　年　　６　月　１２　日現在

機関番号：　３４３１０
研究種目：　基盤研究(C)（一般）
研究期間：　2015～2017
課題番号：　１５Ｋ００３２３

研究課題名（和文）不均衡データ分類器の開発と応用

研究課題名（英文）Development and Application of an Imbalanced Data Classifier

研究代表者
大崎　美穂（Ohsaki, Miho）

同志社大学・理工学部・教授

研究者番号：３０３１３９２７

交付決定額（研究期間全体）：（直接経費）　　2,800,000円

研究成果の概要（和文）：がんの診断，交通事故の予測など多岐に渡る分野で，少数の危機的事例(少数クラス)と多数の通常的事例(多数クラス)の分類が望まれる．しかし，2クラス間の不均衡性は少数クラスの見落しを引き起こす．また，従来の解決策は対象に特化し，クラス間の性能バランスを制御し難い問題があった．そこで我々は，汎用的で，性能のバランス制御と向上を達成する不均衡データ分類器の開発を試みた．提案手法はカーネルロジスティック回帰，最小分類誤り学習・一般化確率的勾配法，混同行列の融合により実現した．実験の結果，提案手法は従来手法よりも高い有効性を持つと示された．そして，研究成果をまとめた学術論文の出版に至った．

研究成果の概要（英文）：In a wide range of domains such as cancer diagnosis, vehicle accident prediction, etc., there is a high demand for the classification of a small number of emergent instances (minority class) and a large number of ordinary instances (majority class). However, the imbalance of the two classes causes overlooking minorities. Conventional solutions for this were domain-specific and difficult to control the balance of performance between the classes. We therefore aim at the development of an imbalanced data classifier which is of high versatility and achieves the balance control and the improvement of performances. The proposed method is based on kernel logistic regression, minimum classification error and generalized probabilistic descent, and confusion matrix. The superiority of the proposed method to the conventional ones was confirmed by the evaluation experiments. We finally published an academic journal paper to report all this research results.

研究分野：　知能情報学

キーワード：　不均衡データ分類　混同行列　カーネルロジスティック回帰　最小分類誤り学習　一般化確率的勾配法

１．研究開始当初の背景
少数の危機的事例と多数の通常的事例から成る不均衡データの分類は，様々な分野(がんの診断，交通事故の予測，ネットワーク不正侵入の検知等)に必要な技術である．しかし，大半を占める多数クラスの影響により，少数クラスの見落としやクラス間で分類性能の差が生じる問題があった．

この問題に対する従来手法としては，事例数を補正するサンプリング法，クラスごとに経験的なコストを割り当てるコスト法，性能分散を低減するアンサンブル法が挙げられる．これらは一定の有効性を持つ反面，経験則的・タスク依存的なものが多く，クラス間の分類性能バランスを取りながら各クラスの分類性能を高めることも困難であった．

２．研究の目的
本研究では上述の背景を踏まえ，経験則やタスクに依存しない枠組みのもとに，分類性能のバランス制御と向上を可能にする不均衡データ分類を目的とした．具体的には，カーネルロジスティック回帰(KLOGR)，最小分類誤り学習・一般化確率的勾配法(MCE/GPD)，混同行列(CM)を融合した分類器(CM-KLOGR)を提案し，定式化，ソフトウェア開発・動作確認，検証実験を経て，提案手法の確立を目指した．

３．研究の方法
我々が提案する CM-KLOGR は，KLOGR のモデル構造を持ち，交差エントロピーを目的関数とする事前学習と，MCE/GPD と CM に基づきクラス間の分類性能バランスと各クラスの分類性能を総合した目的関数による再学習を行う．

KLOGR は，式(1)の上側のように入力変数を代入したカーネル関数の加重和を求め，式(2)の下側に示すソフトマックス関数に代入することで，クラスの事後確率を推定する手法である．KLOGR には，非線形なクラス境界を表現できる，結果の確信度としてクラスの事後確率を導出できる，不均衡データのように規模が小さいデータに適する等の利点があり，CM-KLOGR はこれらを引き継ぐ．

$$y_k(\mathbf{x}) = \boldsymbol{\alpha}_k^T \boldsymbol{\kappa}(\mathbf{x}) + b_k$$

$$\Pr(C_k|\mathbf{x}) = \frac{\exp(y_k(\mathbf{x}))}{\sum_{l=1}^{K} \exp(y_l(\mathbf{x}))}$$

・・・式(1)

CM-KLOGR の事前学習では，式(2)に示す目的関数を最小化することで KLOGR と同等以上の分類性能を確保する．この式において，第 1 項は KLOGR の目的関数と同じもの，すなわち，クラスの事後確率の推定性能を表す交差エントロピーである．第 2 項は過学習を抑制するパラメータの L2 ノルム正則化項，λは 2 つの項の影響力を調整するハイパーパラメータである．

$$J = -\sum_{n=1}^{N}\sum_{k=1}^{K} \delta_{k_n,k} \ln \Pr(C_k|\mathbf{x}_n) + \frac{\lambda}{2}\sum_{k=1}^{K} \boldsymbol{\alpha}_k^T \mathcal{K} \boldsymbol{\alpha}_k$$

・・・式(2)

再学習では，事前学習の結果を初期値として，少数クラスの分類性能，多数クラスの分類性能，および，これらのバランスを総合的に制御向上させる目的関数を導入する．このために，まずは MCE/GPD に基づき誤分類に対する平滑化 0-1 損失を式(3)のように定義する．

$$d_{k_n}(\mathbf{x}_n) = -\Pr(C_{k_n}|\mathbf{x}_n) + \left[\frac{1}{K-1}\sum_{j,j\neq k_n} \Pr(C_j|\mathbf{x}_n)^{\eta}\right]^{\frac{1}{\eta}}$$

$$l(d_{k_n}(\mathbf{x}_n)) = \frac{1}{1+\exp(-\epsilon d_{k_n}(\mathbf{x}_n))} \quad (\epsilon > 0)$$

・・・式(3)

次に平滑化 0-1 損失を用いて，CM の各要素に対応する，言い換えると少数クラスと多数クラスの正誤パターンに対応する，真陽性，偽陽性，真陰性，偽陰性を式(4)のように近似する．

$$N_{\mathrm{TP}} \approx \sum_{n=1}^{N} (1 - l(d_{k_n}(\mathbf{x}_n)))\delta_{k_n,2}$$

$$N_{\mathrm{FP}} \approx \sum_{n=1}^{N} l(d_{k_n}(\mathbf{x}_n))\delta_{k_n,1}$$

$$N_{\mathrm{TN}} \approx \sum_{n=1}^{N} (1 - l(d_{k_n}(\mathbf{x}_n)))\delta_{k_n,1}$$

$$N_{\mathrm{FN}} \approx \sum_{n=1}^{N} l(d_{k_n}(\mathbf{x}_n))\delta_{k_n,2}$$

・・・式(4)

さらに，これらを組み合わせることで，式(5)に示す Sens（少数クラスの見落としがないか），Spec（多数クラスの見落としがないか），PPV（少数クラスの予測が正確か），NPV（多数クラスの予測が正確か）が求まる．

$$\mathrm{Sens} = \frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}} + N_{\mathrm{FN}}} \qquad \mathrm{Spec} = \frac{N_{\mathrm{TN}}}{N_{\mathrm{TN}} + N_{\mathrm{FP}}}$$

$$\mathrm{PPV} = \frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}} + N_{\mathrm{FP}}} \qquad \mathrm{NPV} = \frac{N_{\mathrm{TN}}}{N_{\mathrm{TN}} + N_{\mathrm{FN}}}$$

・・・式(5)

最終的に，これらの評価基準 f_i (Sens, Spec, PPV, NPV を指す) の調和平均を取れば，全評価基準のバランスと値の大きさを定式化した式(6)の第 1 項が求まる．第 2 項とλは式(1)と同様に過学習防止のために加えた．式(6)による再学習を行えば，KLOGR の性能を越えるように，少数クラスの分類性能，多数クラスの分類性能，および，これらのバランスを総合的に制御向上できる．

$$J = -\left[\frac{1}{S_\gamma}\left(\sum_{i=1}^{N_{\mathrm{ec}}}\frac{\gamma_i}{f_i}\right)\right]^{-1} + \frac{\lambda}{2}\sum_{k=1}^{K}\boldsymbol{\alpha}_k^T\mathcal{K}\boldsymbol{\alpha}_k$$

・・・式(6)

４．研究成果

データセットの種類や，不均衡性比率，評価基準の重みを様々にして，CM-KLOGR の有効性を検証する実験を行った．

比較対象には，CM-KLOGR の事前学習に相当する KLOGR，最も普及したカーネル法であるサポートベクターマシン(SVM)，不均衡データ分類用の前処理手法であるサンプリング法を KLOGR，もしくは，SVM に組み込んだものを選定した．ただし，サンプリング法には多数事例を減らすアンダーサンプリング(US)，少数事例を増やすオーバーサンプリング(OS)の 2 種類を用いた．

表 1

Breast					
	Ideal Test Performance [%]				
Classifiers	Sens	Spec	PPV	NPV	HM
CM-KLOGR	95.83	100.00	100.00	97.87	**98.40**
KLOGR	100.00	95.65	92.31	100.00	96.88
SVM	100.00	95.65	92.31	100.00	96.88
Haberman					
	Ideal Test Performance [%]				
Classifiers	Sens	Spec	PPV	NPV	HM
CM-KLOGR	87.50	82.61	63.64	95.00	**80.36**
KLOGR	50.00	100.00	100.00	85.19	77.31
SVM	50.00	100.00	100.00	85.19	77.31
Ecoli-pp					
	Ideal Test Performance [%]				
Classifiers	Sens	Spec	PPV	NPV	HM
CM-KLOGR	100.00	100.00	100.00	100.00	**100.00**
KLOGR	100.00	96.55	83.33	100.00	94.43
SVM	100.00	100.00	100.00	100.00	**100.00**
Ecoli-imu					
	Ideal Test Performance [%]				
Classifiers	Sens	Spec	PPV	NPV	HM
CM-KLOGR	50.00	100.00	100.00	93.75	**78.95**
KLOGR	50.00	96.67	66.67	93.55	71.38
SVM	50.00	100.00	100.00	93.75	**78.95**
Pop_failures					
	Ideal Test Performance [%]				
Classifiers	Sens	Spec	PPV	NPV	HM
CM-KLOGR	100.00	95.92	71.43	100.00	90.04
KLOGR	80.00	97.96	80.00	97.96	88.07
SVM	80.00	100.00	100.00	98.00	**93.67**
Yeast-1_vs_7					
	Ideal Test Performance [%]				
Classifiers	Sens	Spec	PPV	NPV	HM
CM-KLOGR	100.00	100.00	100.00	100.00	**100.00**
KLOGR	66.67	100.00	100.00	97.73	88.43
SVM	100.00	51.16	12.50	100.00	33.46

データセットは，乳がんによる生死，蛋白質の細胞内局在場所，気候シミュレーションの成否など 6 種類であり，UCI リポジトリと KEEL リポジトリから入手した．これらの不均衡性比率(多数クラスが少数クラスの何倍の規模であるか)は 1.90～14.30 の範囲内で分散している．評価基準の重みに関しては，Sens，Spec，PPV，NPV に同じ重みを与える，Sens と PPV のみに，もしくは Sens と Spec のみに重みを与えるという 3 種類を考えた．

不均衡データは危機的な状況を想定するため，規模が比較的小さい．適正な汎化性能の見積りには，データセットを訓練用，検証用，試験用に分割して用いるべきであるが，規模の問題で正確性が低下する恐れがある．そこで，分割したサブセットを訓練，検証，試験に個別に用いる場合と，訓練と検証には同じものを用い，試験には個別のものを用いる場合の 2 種類を検討した．後者は検証におけるハイパーパラメータ設定が最適である理想状態を模擬する．

表 2

Breast					
	Ideal Test Performance [%]				
Classifiers	Sens	Spec	PPV	NPV	HM
CM-KLOGR	95.83	100.00	100.00	97.87	**98.40**
KLOGR-US	100.00	95.65	92.31	100.00	96.88
KLOGR-OS	100.00	95.65	92.31	100.00	96.88
SVM-US	100.00	95.65	92.31	100.00	96.88
SVM-OS	95.83	100.00	100.00	97.87	**98.40**
Haberman					
	Ideal Test Performance [%]				
Classifiers	Sens	Spec	PPV	NPV	HM
CM-KLOGR	87.50	82.61	63.64	95.00	**80.36**
KLOGR-US	87.50	73.91	53.85	94.44	73.91
KLOGR-OS	87.50	78.26	58.33	94.74	77.06
SVM-US	50.00	100.00	100.00	85.19	77.31
SVM-OS	50.00	100.00	100.00	85.19	77.31
Ecoli-pp					
	Ideal Test Performance [%]				
Classifiers	Sens	Spec	PPV	NPV	HM
CM-KLOGR	100.00	100.00	100.00	100.00	**100.00**
KLOGR-US	100.00	96.55	83.33	100.00	94.43
KLOGR-OS	100.00	93.10	71.43	100.00	89.40
SVM-US	100.00	96.55	83.33	100.00	94.43
SVM-OS	100.00	96.55	83.33	100.00	94.43
Ecoli-imu					
	Ideal Test Performance [%]				
Classifiers	Sens	Spec	PPV	NPV	HM
CM-KLOGR	50.00	100.00	100.00	93.75	**78.95**
KLOGR-US	50.00	100.00	100.00	93.75	**78.95**
KLOGR-OS	50.00	96.67	66.67	93.55	71.38
SVM-US	50.00	96.67	66.67	93.55	71.38
SVM-OS	50.00	100.00	100.00	93.75	**78.95**
Pop_failures					
	Ideal Test Performance [%]				
Classifiers	Sens	Spec	PPV	NPV	HM
CM-KLOGR	100.00	95.92	71.43	100.00	**90.04**
KLOGR-US	100.00	00.00	09.26	50.00	00.00
KLOGR-OS	100.00	93.88	62.50	100.00	85.74
SVM-US	00.00	100.00	50.00	90.74	00.01
SVM-OS	80.00	97.96	80.00	97.96	88.07
Yeast-1_vs_7					
	Ideal Test Performance [%]				
Classifiers	Sens	Spec	PPV	NPV	HM
CM-KLOGR	100.00	100.00	100.00	100.00	**100.00**
KLOGR-US	66.67	95.35	50.00	97.62	71.77
KLOGR-OS	100.00	97.67	75.00	100.00	91.80
SVM-US	100.00	88.37	37.50	100.00	68.99
SVM-OS	100.00	88.37	37.50	100.00	68.99

実験結果のうち，全評価基準に同じ重みを与え，理想的なハイパーパラメータ設定で得られた結果を抜粋して表 1，2 に示す．表中，HM は Sens，Spec，PPV，NPV の調和平均であり，これら全ての性能がバランス良く向上したかを意味する．

表 1 では,ほとんどのデータセットについて，CM-KLOGR が KLOGR や SVM よりも高い HM を達成している．表 2 では，全データセットについて，CM-KLOGR がサンプリング法を組み込んだ KLOGR-US，KLOGR-OS，SVM-US，SVM-OS よりも HM が高い．他の条件(評価基準の間で重みが異なる場合，サブセットを訓練，検証，試験に個別使用した場合)でも，同様の傾向が見られた．

以上より，CM-KLOGR の有効性が明らかになった．そして，本課題の全研究成果を学術論文として執筆出版し，公表することができた．

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 2 件）
(1) Miho Ohsaki, Peng Wang, Kenji Matsuda, Shigeru Katagiri, Hideyuki Watanabe, Anca Ralescu, Confusion-matrix-based Kernel Logistic Regression for Imbalanced Data Classification, IEEE Transactions on Knowledge and Data Engineering, vol.29, no.9, pp.1806-1819 (2017)，査読有．

(2) 大崎美穂,松田健司,ワンペン,片桐滋，横井英人，高林克日己，カーネルロジスティック回帰を用いたＣ型慢性肝炎の肝線維化ステージ推定，情報処理学会論文誌，vol.56, no.11，pp.2117-2130 (2015)，査読有．

〔学会発表〕（計 5 件）
(1) David Ha, Juliette Maes, Yuya Tomotoshi, Charles Melle, Hideyuki Watanabe, Shigeru Katagiri, Miho Ohsaki, A Class Boundary Selection Criterion for Classification, 情報処理学会関西支部大会，G-11 (2017 年 9 月)，査読無．

(2) 谷陵真，渡辺秀行，片桐滋，大崎美穂，最小分類誤り基準に基づくサポートベクター再学習による小規模カーネル分類器，信学技報，vol.116，no.461，PRMU2016-159，pp.41-46 (2017 年 2 月)，査読無．

(3) 谷陵真，渡辺秀行，片桐滋，大崎美穂，モデルサイズから見たカーネル最小分類誤り学習法の有用性の検証，情報処理学会関西支部大会，7p (2016 年 9 月)，査読無．

(4) Miho Ohsaki, Kenji Matsuda, Peng Wang, Shigeru Katagiri, Hideyuki Watanabe, Formulation of the Kernel Logistic Regression based on the Confusion Matrix, IEEE Congress on Evolutionary Computation CEC-2015, pp.2327-2334 (May 2015)，査読有．

(5) Peng Wang, Miho Ohsaki, Kenji Matsuda, Shigeru Katagiri, Hideyuki Watanabe, Kernel Logistic Regression based on the Confusion Matrix for Imbalanced Data Classification, 情報処理学会研究報告，vol. 2015-BIO-42, no.55, pp.1-2 (2015 年 6 月)，査読無．

６．研究組織
(1)研究代表者
大崎 美穂 (OHSAKI, Miho)
同志社大学・理工学部・教授
研究者番号：30313927