DBSJ Journal, Vol.11, No.1
June 2012

大域的なデータ分布を考慮したカーネル密度推定による外れ値検出

Outlier Detection using Kernel Density Estimation accounting for Global Data Distribution

立間 淳司 ♥　青野 雅樹 ♦
Atsushi TATSUMA　Masaki AONO

近年，大量にあるデータから，有益な情報や傾向を，解析・発掘するデータマイニング技術が注目を集めている．しかし，大量にあるデータ中には，全体の傾向から逸脱した外れ値 (Outlier) が含まれることがあり，解析結果に悪影響を与える．本論文では，カーネル密度推定による新しい外れ値検出手法 Local Kernel Density Outlier Factor (LKDOF) を提案する．まず，カーネル密度推定により局所的な密度を表す分布スコアを求める．そして，分布スコアの中央値を，正常値のものと仮定し，正常値との分布スコアの比を，外れ値スコア LKDOF とする．分布スコアの中央値を，正常値のものと仮定することで，データ全体の分布の傾向を考慮した外れ値スコアを算出する．実験から，Local Outlier Factor などの従来手法よりも，高い検出精度を得た．

Data mining technologies have getting more and more attention these days. Primary objectives for data mining include extracting valuable piece of information and predicting latent tendencies from massive data. However, most of the data mining technologies have been strongly affected by the existence of outliers within the data. In this paper, we propose a new quantitative score for outliers, which we call Local Kernel Density Outlier Factor (LKDOF). (1) We compute a distribution score indicating the density of local distribution using kernel density estimation. (2) We compute LKDOF as a ratio of distribution scores of each data to the median of distribution scores. We incorporate the distribution of the whole data into an outlier score by assuming the median of a distribution score to be a distribution score of normal data. From comparative experiments, we show that LKDOF exhibits higher detection precision than conventional techniques.

♥ 正会員　豊橋技術科学大学大学院電子・情報工学専攻　atsushi@kde.cs.tut.ac.jp
♦ 正会員　豊橋技術科学大学情報・知能工学系　aono@tut.jp

1. はじめに

インターネットの普及により，文書・画像・動画などのあらゆるデータが，爆発的に増加している．近年，これら大量にあるデータから，有益な情報や傾向を，解析・発掘するデータマイニング技術が注目を集めている．大量にあるデータ中には，全体の傾向から逸脱した外れ値 (Outlier) が含まれることがある．データマイニングでは，一般的に，クラスタリングや次元削減といった，機械学習アルゴリズムが用いられるが，機械学習アルゴリズムには，外れ値に対して頑健ではないものも多い．そのため，外れ値の存在により，誤った解析結果となってしまう場合がある．外れ値を検出する技術は，データマイニングの分野において，重要な研究テーマの一つとなっている [8, 11].

外れ値検出手法は，大きく分類して，教師あり手法と教師なし手法とに分けられる．教師あり手法 [5, 6, 13, 15] は，正常値・外れ値のラベル付データ用いて，Support Vector Machine などによる分類器から，外れ値であるかを判定する．教師あり手法は，ラベル付きデータを必要とすること，一般的に，外れ値のデータは数が少なく，十分な学習ができないことなどの問題がある．一方，教師なし手法 [1, 3, 7, 9, 16] では，直接，与えられたラベル無しデータから外れ値を検出するため，データに関する問題はなく，適用できる分野も広い．

外れ値検出手法は，また，外れ値か否かのラベル付けを行う手法 [1, 7, 9, 16] と，どの程度外れ値であるかのスコアを算出する手法 [3, 10, 12, 19] とにも分けられる．外れ値の検出結果として，ラベルを付与する手法は，出力が，正常値・外れ値の二値であるのに対して，スコアを算出する手法は，連続値となる．そのため，スコアを算出する手法では，スコアでの並び替えや，スコアが大きい上位 n 件を取り出すなどのデータ操作ができる．さらに，スコアを任意の閾値で区切ることで，正常値・外れ値のラベルを生成することもできる．外れ値の判定結果をスコアで表す手法は，ラベルで表す手法と比べて，応用範囲が広い．以上から，本論文では，外れ値スコアによる教師なし外れ値検出手法を対象とする．

教師なし外れ値スコアの代表的なものに Local Outlier Factor (LOF) [3] がある．LOF は，外れ値は分布が疎な箇所に位置するとし，どの程度外れ値であるかを量的に表す．局所的な密度は，各データにおいて，近傍との最大距離であらわされる到達範囲に，近傍データが，どの程度含まれているかで表す．LOF は，データ全体の分布の傾向を考慮せず，局所的な密度が疎であれば，高い外れ値スコアとなる．そのため，データ全体の分布から見れば，外れ値であると判定できないデータであっても，外れ値スコアが高くなってしまう場合がある．その他，教師なし外れ値スコアには，外れ値スコアを確率で表現する Local Outlier Probabilities [10]，カーネル密度推定により局所的な密度を計る Local Density Factor [12]，近傍との距離の比から外れ値スコアを求める Local Distance-based Outlier Factor [19] などがある．

本論文では，新しい教師なし外れ値スコア Local Kernel Den-

sity Outlier Factor (LKDOF) を提案する．提案手法では，まず，各データで，カーネル密度推定により局所的な密度を表す分布スコアを求める．そして，分布スコアの中央値を，正常値のものと仮定し，正常値との分布スコアの比を，外れ値スコア LKDOF とする．分布スコアの中央値を，正常値のものと仮定することで，データ全体の分布の傾向を考慮した外れ値スコアを算出する．比較実験から，LOF などの従来手法と比較して，高い外れ値検出精度を得た．

以下に，本論文の構成を示す．第 2 節では，教師なし外れ値スコアの関連研究について述べる．ついで，第 3 節では，カーネル密度推定による外れ値スコアの算出アルゴリズムと特長について述べる．そして，第 4 節で，提案手法の有効性を比較実験により示す．最後，第 5 節では，まとめと今後の課題について述べる．

2. 関連研究

本論文では，ラベル付データを必要とせず，並び替えなどデータ操作が可能である，教師なしかつ外れ値スコアによる，外れ値検出手法を対象とする．教師なし外れ値スコアには，Local Outlier Factor (LOF) [3]，Local Outlier Probabilities (LoOP) [10]，Local Density Factor (LDF) [12]，Local Distance-based Outlier Factor (LDOF) [19] などがある．教師なし外れ値スコアの算出アルゴリズムでは，外れ値は，分布が疎な箇所に位置するとし，各データの局所的な密度から，どの程度外れ値であるかのスコアを算出する．

Breunig らが提案した LOF は，局所的な密度による教師なし外れ値スコアの代表的な手法である．LOF は，各データ $\mathbf{x}$ と，その k-近傍にあるデータ $\mathcal{N}(\mathbf{x})$ により，各データ周辺の局所的な密度を評価する．LOF では，まず，データ $\mathbf{x}$ の k 番目の近傍にあるデータとの距離を $d_k(\mathbf{x})$，データ $\mathbf{y}$ との距離を $d(\mathbf{x},\mathbf{y})$ とすると，データ $\mathbf{x}$ とデータ $\mathbf{y}$ との Reachability Distance (RD) を次の様に定義する．

$$\mathrm{RD}(\mathbf{x},\mathbf{y}) = \max(d(\mathbf{x},\mathbf{y}), d_k(\mathbf{y}))$$

ついで，RD から，データ $\mathbf{x}$ 周辺の局所的な密度を表す値 Local Reachability Density (LRD) を計算する．

$$\mathrm{LRD}(\mathbf{x}) = \frac{k}{\sum_{\mathbf{y}\in\mathcal{N}(\mathbf{x})} \mathrm{RD}(\mathbf{x},\mathbf{y})}$$

そして，LOF は，データ $\mathbf{x}$ の LRD と，その近傍の LRD との比の平均で定義される．

$$\mathrm{LOF}(\mathbf{x}) = \frac{1}{k}\sum_{\mathbf{y}\in\mathcal{N}(\mathbf{x})} \frac{\mathrm{LRD}(\mathbf{y})}{\mathrm{LRD}(\mathbf{x})}$$

LOF は，局所的な密度による外れ値スコアの代表的な手法であり，距離に基づく手法 [9] が検出できない外れ値を検出できることや，外れ値を量的に表せることなどの利点から，その後，いくつかの改良手法が考案された．

Kriegel らが提案した LoOP は，LOF と同様に，局所的な密度による外れ値スコアであるが，外れ値スコアが，どの程度外れ値であるかの確率で定義されており，異なるデータセット間での外れ値スコアの比較が，容易であるなどの特長を持つ [10]．また，Zhang らが提案した LDOF は，各データで，k-近傍との平均距離と，k-近傍同士の平均距離との比で定義されるシンプルな手法であるが，近傍数 k を十分に大きくすると，外れ値スコアの下界が 0.5 となるなど，正常値・外れ値を区切る閾値の設定を容易にする特長を持つ [19]．

Latecki らによる LDF は，提案手法である LKDOF と同様に，カーネル密度推定を用いて，外れ値スコアを算出する [12]．LDF では，まず，LOF と同様に，各データの k-近傍との距離で定義される Reachability Distance から，局所的な密度を表す値 Local Density Estimate (LDE) を，カーネル密度推定を用いて計算する．

$$\mathrm{LDE}(\mathbf{x}) = \frac{1}{k}\sum_{\mathbf{y}\in\mathcal{N}(\mathbf{x})} \frac{\exp\left(-\frac{\mathrm{RD}(\mathbf{x},\mathbf{y})^2}{2(\sigma\cdot d_k(\mathbf{y}))^2}\right)}{(2\pi)^{D/2}(\sigma\cdot d_k(\mathbf{y}))^D}$$

ここで，D はデータの次元数であり，σ はカーネル幅である．LDF は，各データの LDE と，その k-近傍の LDE の平均との比により定義される．

$$\mathrm{LDF}(\mathbf{x}) = \frac{\frac{1}{k}\sum_{\mathbf{y}\in\mathcal{N}(\mathbf{x})} \mathrm{LDE}(\mathbf{y})}{\mathrm{LDE}(\mathbf{x}) + \frac{c}{k}\sum_{\mathbf{y}\in\mathcal{N}(\mathbf{x})} \mathrm{LDE}(\mathbf{y})}$$

ここで，c は任意の定数であり，Latecki らは $c = 0.1$ とした．さらに，Latecki らは，局所的なデータの分布を捉えるため，マハラノビス距離による LDE を再定義している [12]．

LDF は，局所的な密度を，カーネル密度推定を用いて計算することで，LOF を改良した手法と言える．しかし，LDE の式中に $(\sigma\cdot d_k(\mathbf{y}))^D$ を含むため，与えられたデータの次元数や距離によっては，LDF の値が正しく計算できない可能性がある．

LDF の他に，提案手法である LKDOF と同様に，カーネル密度推定を利用した外れ値検出手法は，いくつか提案されており [2, 4, 17, 18]，近傍の密度をカーネル密度推定により求め，その偏差から外れ値であるかを判定する手法 [17]，カーネル密度推定による確率密度関数を用いて，ベイズ決定則から外れ値を検出する手法 [18] などがある．

LOF など，多くの局所的な密度による外れ値スコアの算出アルゴリズムは，データ全体の分布を考慮せず，局所的な密度が疎であれば，外れ値スコアを付与する．そのため，データ全体の分布から見れば，外れ値であると判定できないデータに対しても，高い外れ値スコアを付与してしまう場合がある．本論文で提案する外れ値スコア LKDOF の算出アルゴリズムは，カーネル密度推定による確率密度関数から，データ周辺の局所的な密度を表す分布スコアを求め，全データの分布スコアの中央値を，正常値のものとして，正常値との分布スコアの比を，外れ値スコア LKDOF とする．分布スコアの中央値を，正常値の分布スコアとすることで，データ全体の分布の傾向を考慮した外れ値スコアを算出する．

3. カーネル密度推定による外れ値スコア

本論文で提案する，新しい教師なし外れ値スコア Local Kernel Density Outlier Factor (LKDOF) の算出アルゴリズムについて，以下に詳述する．LKDOF の算出アルゴリズムでは，外れ値は，データの分布が疎な箇所に位置するとし，まず，カーネル密度推定を用いて，各データ周辺の局所的な密度を表す分布スコアを求める．そして，分布スコアの中央値を，正常値のものと仮定し，正常値との分布スコアの比を外れ値スコア LKDOF とする．

データ $\mathbf{x}$ の k-近傍にあるデータを $\mathcal{N}(\mathbf{x})$ と表すと，局所的な確率密度関数は，ガウスカーネルを用いたカーネル密度推定から，以下のように求められる．

$$p(\mathbf{x}) = \frac{1}{k(2\pi\sigma^2)^{D/2}} \sum_{\mathbf{y}\in\mathcal{N}(\mathbf{x})} \exp\left(-\frac{d(\mathbf{x},\mathbf{y})^2}{2\sigma^2}\right)$$

ここで，D は次元数であり，σ はカーネル幅である．カーネル幅は，各データでの，k-近傍との平均距離の最小値とする．分布が疎な箇所に位置する外れ値は，k-近傍との距離が大きくなることから，正常値周辺の分布を基準にするため，最小値をとった．

$$\sigma = \min_i \frac{1}{k} \sum_{\mathbf{y}\in\mathcal{N}(\mathbf{x}_i)} d(\mathbf{x}_i,\mathbf{y})$$

さらに，推定された各データの局所的な確率密度関数から，各データ周辺の局所的な密度を表す分布スコアを，以下のように定義する．

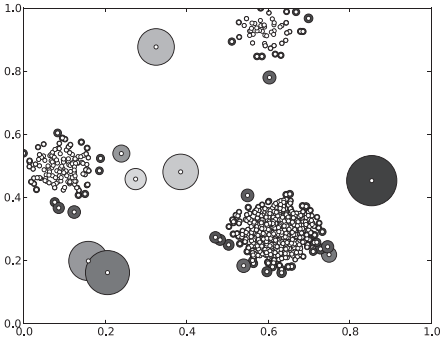
$$\mathrm{DS}(\mathbf{x}) = \frac{1}{k}\sum_{\mathbf{y}\in\mathcal{N}(\mathbf{x})}\frac{p(\mathbf{y})}{p(\mathbf{x})} = \frac{1}{k\cdot p(\mathbf{x})}\sum_{\mathbf{y}\in\mathcal{N}(\mathbf{x})} p(\mathbf{y})$$
$$= \frac{\sum_{\mathbf{y}\in\mathcal{N}(\mathbf{x})}\sum_{\mathbf{z}\in\mathcal{N}(\mathbf{y})}\exp\left(-\frac{d(\mathbf{y},\mathbf{z})^2}{2\sigma^2}\right)}{k\cdot\sum_{\mathbf{y}\in\mathcal{N}(\mathbf{x})}\exp\left(-\frac{d(\mathbf{x},\mathbf{y})^2}{2\sigma^2}\right)}$$

分布スコアは，データとその k-近傍との確率密度比の平均である．データとその k-近傍が，同程度の確率密度である場合，分布スコアの値は 1 に近づく．また，データが疎な分布中にあり，k-近傍が密な分布中にある場合，分布スコアの値は 1 より大きくなり，逆に，データが密な分布中にあり，k-近傍が疎な分布中にある場合は，分布スコアの値は 1 より小さくなる．つまり，分布スコアの値が 1 より遙かに大きな値であれば，データは疎な箇所に位置しており，外れ値である可能性が高くなる．

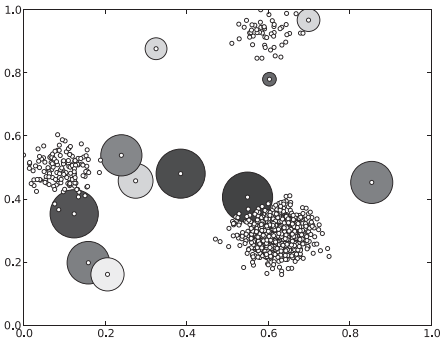
分布スコアは，データ周辺の局所的な分布の疎密を表す値である．外れ値は，データの分布が疎な箇所に位置するが，どの程度疎な分布であれば，外れ値であると言えるかは，全データの分布の傾向から決まる．そこで，全データの分布スコアの中央値 $\overline{\mathrm{DS}}$ を，正常値の分布スコアと仮定して，外れ値スコア LKDOF を以下のように定義する．

$$\mathrm{LKDOF}(\mathbf{x}) = \mathrm{DS}(\mathbf{x})/\overline{\mathrm{DS}}$$

図 1 は，人工的に作成した二次元のデータに対して，外れ値スコア LOF と LKDOF を，それぞれ算出したものである．人工データは，密度の異なる三つのデータ群と，一様乱数によるランダムなデータからなる．色が青から赤に近づくほど，円の半径が大きくなるほど，外れ値スコアが高くなる．外れ値スコアに LOF を用いた図 1(a) を見ると，データ点が密集したところから，わずかに離れた正常値であると判断できるデータ点に対しても， 外れ値スコアがついている．これは，LOF の算出アルゴリズムでは，データ全体の分布の傾向は考慮せず，局所的な密度が疎であれば，外れ値スコアを付与するためである．一方，LKDOF を用いた図 1(b) を見ると，全体的に見て，分布が疎な箇所に位置するデータ点に対して，外れ値スコアがついている．これは，LKDOF の算出アルゴリズムは，分布スコアの中央値を正常値のものとすることで，データ全体の分布の傾向を考慮するためである．LKDOF の算出アルゴリズムが，データ分布を捉えた，妥当な外れ値スコアを付与できることがわかる．



(a)LOF



(b)LKDOF

図 1　人工データに対して各手法で外れ値スコアを算出した例．
Fig. 1　Example of outlier scores on synthetic data set.

4. 実験

提案手法である Local Kernel Density Outlier Factor の有効性を確認するため，実際のデータを用いて，Local Outlier Factor など，代表的な従来手法と比較実験を行った．

4.1 実験内容

カーネル密度推定による新しい外れ値スコア Local Kernel Density Outlier Factor (LKDOF) と，従来手法とを比較することで，その有効性を確認する．

従来手法には，Local Outlier Factor (LOF) [3]，Local Distance-based Outlier Factor (LDOF) [19]，Local Outlier Probabilities (LoOP) [10]，Local Density Factor (LDF) [12] を用いた．このうち，LoOP はパラメータ λ を持つが，著者らの知見に従い $\lambda = 3$ とした．

データセットには，Wisconsin Breast Cancer Diagnosis (WBCD)，Vowel Recognition Data Set (VRDS)，Optical Recognition of Handwritten Digits (ORHD) [1] を用いた．

WBCD には，31 次元のデータが，benign クラスと malignant クラスに分かれており，benign には 357 個のデータが，maglignant には 212 個のデータが含まれている．実験では，benign クラスの全データを正常値とし，maglignant クラスから，ランダムに選び出した 10 個のデータを外れ値とした．VRDS には，462 個の 10 次元のデータが，11 のクラスに分かれている．実験では，この内 hid クラスに属する 42 個のデータから，ランダムに 10 個選び出したものを外れ値とし，残り 420 個のデータを正常値とした．ORHD には，64 次元の 0 から 9 までの数字画像データが 1,797 個含まれている．実験では，数字の 2 に該当するデータから，ランダムに 10 個選び出したものを外れ値とし，その他の数字に該当するデータを正常値とした．

評価尺度には，外れ値検出精度 (R-Precision) [19] を用いる．外れ値スコアは，どの程度外れ値であるかを量的に表すものであり，正常値・外れ値のラベルではない．そこで，各データを外れ値スコアでランク付けし，ランク上位に，外れ値データがどのように表れているかを評価する．

実験に使用するデータセットは，いずれも，外れ値の数は 10 個である．R-Precision は，外れ値の数と同じ，上位 10 件に，外れ値がいくつ含まれているかの割合である．上位 10 件が，全て外れ値となれば，R-Precision の値は 1.0 となる．また，R-Precision の値は，外れ値 10 個のうち，いくつがランク上位に含まれているかの割合 (Recall) とも同値となる．R-Precision を評価尺度に用いることで，各外れ値スコアの正確性と網羅性を評価する．

実験では，近傍数 k を変化させて，R-Precision を計算した．各データセットで，外れ値とするデータは，ランダムに選び出すため，評価尺度の値は，実験を 15 回行った上での平均値を用いた．

4.2 Wisconsin Breast Cancer Diagnosis (WBCD)

Wisconsin Breast Cancer Diagnosis (WBCD) は，569 個のデータが，benign クラスと malignant クラスに分かれている．実験では，maglignant クラスから，ランダムに 10 個選び出したものを外れ値とし，benign クラスの全データを正常値とした．

図 2 は，WBCD における各手法の外れ値検出精度である．

[1] http://archive.ics.uci.edu/ml/

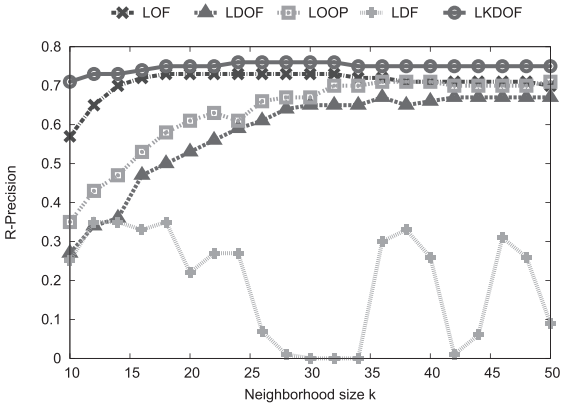


図 2　WBCD における各手法の外れ値検出精度．
Fig. 2　R-Precision of different outlier scores on WBCD.

横軸が近傍数 k，縦軸が外れ値検出精度 R-Precision である．図 2 を見ると，提案手法である LKDOF が，すべての近傍数 k で，もっとも高い検出精度となっていることがわかる．また，LKDOF と同様に，カーネル密度推定から，局所的な密度を推定する LDF よりも，高い検出精度となった．LKDOF の特長である，局所的な密度だけでなく，データ全体の分布の傾向を考慮することが，有効であることがわかる．

図 3 は，WBCD データセットを，主成分分析により二次元に次元削減し，可視化したものである．水色の丸印が正常値であり，赤色のバツ印が外れ値である．図 3 を見ると，正常値は，正常値の重心から外側に広がるように分布しており，わずかに疎である程度では，外れ値であるとは言えないことがわかる．LOF は，局所的な密度のみを基準とした外れ値スコアであるため，データ全体の分布の傾向を捉えることができない．これに対して，LKDOF は，局所的な密度だけでなく，分布スコアの中央値を正常値のものとすることで，データ全体の分布の傾向も捉える．このため，LKDOF が，従来手法と比較して，高い外れ値検出精度を得ることができたと考える．

4.3 Vowel Recognition Data Set (VRDS)

Vowel Recognition Data Set (VRDS) には，11 のクラスに 42 個ずつデータが存在する．実験では，この内 hid クラスに属するデータから，ランダムに 10 個選び出したものを外れ値とし，残りのデータを正常値とした．

図 4 は，VRDS おける各手法の外れ値検出精度である．横軸が近傍数 k，縦軸が外れ値検出精度 R-Precision である．図 4 を見ると，近傍数 $k \leq 30$ で，提案手法である LKDOF が，もっとも高い検出精度となっている．近傍数 $k > 30$ では，LDOF と同等程度の検出精度となっている．また，LKDOF と同様に，カーネル密度推定により局所的な密度を算出する LDF は，近傍数 k の値が大きくなるにつれ，検出精度が高くなっている．VRDS のデータは，10 次元と低次元なデータである．低次元かつ近傍数 k の値が大きい場合に，LDF は高い検出精度が得られる．一方，LKDOF は，次元数や近傍数に関わりなく，LDF よりも高

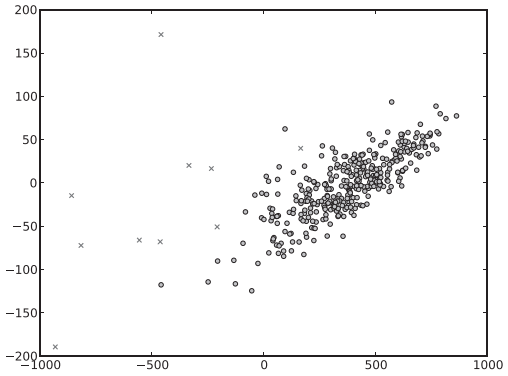


図 3　主成分分析により二次元に次元削減した WBCD.
Fig. 3　Visualization of WBCD by Principal Component Analysis.

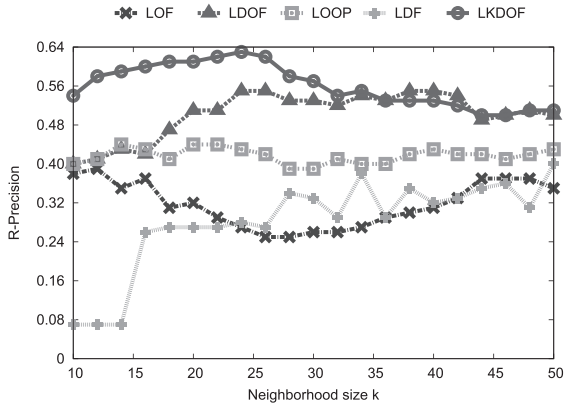


図 4　VRDS における各手法の外れ値検出精度.
Fig. 4　R-Precision of different outlier scores on VRDS.

い検出精度を得ている．LDF よりも LKDOF の方が，適応できるデータの種類が多いと考える．

4.4 Optical Recognition of Handwritten Digits (ORHD)

Optical Recognition of Handwritten Digits (ORHD) は，0 から 9 までの数字画像データからなる．実験では，数字の 2 に該当するデータから，ランダムに 10 個選び出したものを外れ値とし，その他の数字に該当するデータを正常値とした．

図 5 は，ORHD における各手法の外れ値検出精度である．横軸が近傍数 k，縦軸が外れ値検出精度 R-Precision である．図 5 を見ると，提案手法である LKDOF が，すべての近傍数 k で，もっとも高い検出精度となっていることがわかる．数字画像データでも，LKDOF が，従来手法と比較して，データ分布の傾向を捉えた外れ値スコアを算出できたと考える．また，どのデータセットでも，LKDOF と同様に，カーネル密度推定から局所的

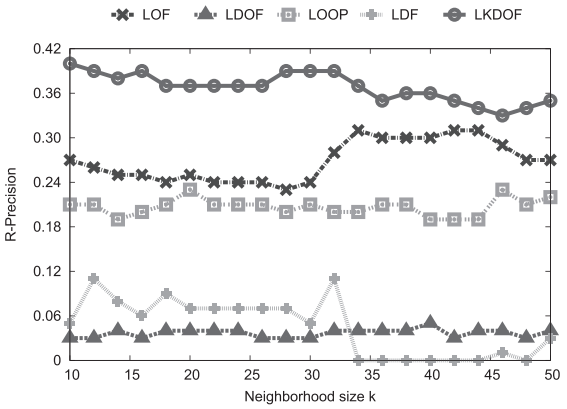


図 5　ORHD における各手法の適合率.
Fig. 5　R-Precision of different outlier scores on ORHD.

な密度を算出する LDF に比べて，近傍数 k の値に対する検出精度の変化が少ない．近傍数 k の値は，局所的な密度の推定に影響を与える．LKDOF は，局所的な密度だけでなく，データ全体の分布を考慮するため，LDF と比べて，近傍数 k の値の影響を抑えられたと考える．

5. おわりに

本論文では，教師なし外れ値スコア Local Kernel Density Outlier Factor (LKDOF) を提案した．提案手法では，まず，各データで，カーネル密度推定により局所的な密度を表す分布スコアを求める．そして，分布スコアの中央値を，正常値のものと仮定し，正常値との分布スコアの比を，外れ値スコア LKDOF とする．分布スコアの中央値を，正常値のものと仮定することで，データ全体の分布の傾向を考慮した外れ値スコアを算出する．比較実験から，Local Outlier Factor (LOF) などの従来手法よりも，高い外れ値検出精度を得た．

今後の課題は，近傍数 k の選択方法の考案である．局所的なデータ分布を手がかりに学習を行う多様体学習の分野でも，同様に，近傍数 k の選択が問題となっており，適応的に近傍数 k を決定する手法が，いくつか提案されている [14, 20]．これらアルゴリズムを，LKDOF の近傍数選択に応用することが考えられる．また，LKDOF による外れ値スコアから，正常値・外れ値のラベルを生成する際の，閾値の設定も問題となる．外れ値であるとする外れ値スコアの閾値を，適応的に求めることができれば，外れ値検出アルゴリズムとして，より有用なものとなる．さらに，スパムブログの検出や機械学習アルゴリズムのロバスト化など，LKDOF の応用も考えていく必要がある．

【謝辞】

本研究は日本学術振興会科学研究費補助金 (特別研究員奨励費) の助成，ならびに科学研究費基盤研究 (C) 課題番号 23500119 を受けて行われた．

【文献】

[1] C. C. Aggarwal and P. S. Yu, *Outlier Detection for High Dimensional Data*, In Proc. of the 2001 ACM SIGMOD International Conference on Management of Data, 30 (2001), pp. 37–46.

[2] T. Ahmed, *Online Anomaly Detection using KDE*, In Proc. of the 28th IEEE Confernce on Global Telecommunications, (2009), pp. 1009–1016.

[3] M. M. Breunig, H. -P. Kriegel, R. Ng and J. Sander, *LOF: Identifying Density-Based Local Outliers*, In Proc. of the 2000 ACM SIGMOD International Conference on Management of Data, 29 (2000), pp. 93–104.

[4] H. C. M. Bossers, J. L. Hurink, G. J. M. Smit, *Online Bivariate Outlier Detection in Final Test Using Kernel Density Estimation*, In Proc. of the IEEE International Workshop on Defect and Adaptive Test Analysis, (2011).

[5] K. Das and J. Schneider, *Detecting Anomalous Records in Categorical Datasets*, In Proc. of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining , (2007), pp. 220–229.

[6] A. K. Ghosh and A. Schwartzbard, *A Study in Using Neural Networks for Anomaly and Misuse Detection*, In Proc. of the 8th Conference on USENIX Security Symposium, 8 (1999), pp. 141–152.

[7] V. Hautamaki, I. Karkkainen, and P. Franti, *Outlier Detection Using k-Nearest Neighbour Graph*, In Proc. of the 17th International Conference on Pattern Recognition, 3 (2004), pp. 430–433.

[8] V. Hodge and J. Austin, *A Survey of Outlier Detection Methodologies*, Artificial Intelligence Review, 22 (2004), pp. 85–126.

[9] E. M. Knorr and R. T. Ng, *Algorithms for Mining Distance-Based Outliers in Large Datasets*, In Proc. of the 24rd International Conference on Very Large Data Bases, (1998), pp. 392–403.

[10] H-P. Kriegel, P. Kröger, E. Schubert and A. Zimek, *LoOP: Local Outlier Probabilities*, In Proc. of the 18th ACM Conference on Information and Knowledge Management, (2009), pp. 1649–1652.

[11] H-P. Kriegel, P. Kröger and A. Zimek, *Outlier Detection Techniques*, Tutorial on 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, (2010).

[12] L. J. Latecki, A. Lazarevic, and D. Pokrajac, *Outlier Detection with Kernel Density Functions*, In Proc. of the 5th International Conference on Machine Learning and Data Mining in Pattern Recognition, (2007), pp. 61–75.

[13] W. Lee and S. J. Stolfo, *Data Mining Approaches for Intrusion Detection*, In Proc. of the 7th Conference on USENIX Security Symposium, 7 (1998), pp. 79–94.

[14] N. Mekuz and J. K. Tsotsos, *Parameterless Isomap with Adaptive Neighborhood Selection*, In Proc. of the 28th Conference on Pattern Recognition, (2006), pp. 364–373.

[15] S. Mukkamala, G. Janoski, and A. Sung, *Intrusion Detection Using Neural Networks and Support Vector Machines*, In Proc. of the 2002 International Joint Conference on Neural Networks, 2 (2002), pp. 1702–1707.

[16] S. Papadimitriou, H. Kitagawa, P. B. Gibbons, and C. Faloutsos, *LOCI: Fast Outlier Detection Using the Local Correlation Integral*, In Proc. of the 19th International Conference on Data Engineering, (2003), pp. 315–326.

[17] H. K. Verma and S.K. Samparthi, *Outlier Detection of Data in Wireless Sensor Networks Using Kernel Density Estimation*, International Journal of Computer Applications, 5 (2010), pp. 28–32.

[18] D-Y. Yeung and C. Chow, *Parzen-Window Network Intrusion Detectors*, In Proc. of the 16th International Conference on Pattern Recognition, 4 (2002), pp. 385–388.

[19] K. Zhang, M. Huter and H. Jin, *A New Local Distance-Based Outlier Detection Approach for Scattered Real-World Data*, In Proc. of the 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, (2009). pp. 813–822.

[20] Z. Zhang, J. Wang, and H. Zha, *Adaptive Manifold Learning*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 34 (2012), pp. 253–265.

立間 淳司 Atsushi TATSUMA

平成 18 年豊橋技術科学大学情報工学課程卒．平成 20 年豊橋技術科学大学大学院修士課程情報工学専攻修了．同年ヤフー (株) 入社．平成 22 年より豊橋技術科学大学大学院博士後期課程電子・情報工学専攻，現在に至る．情報検索，データマイニングなどの研究に従事．日本学術振興会特別研究員．電子情報通信学会，電気学会，日本データベース学会各会員．

青野 雅樹 Masaki AONO

昭和 56 年東京大学理学部情報科学科卒．昭和 59 年東京大学大学院理学系研究科情報科学専攻修士課程修了．同年日本アイビーエム (株) 入社．平成 6 年, 米国レンセラー工科大学コンピュータサイエンス学科 Ph.D. 課程修了．平成 15 年より豊橋技術科学大学情報工学系教授, 現在に至る．情報検索，データマイニング, 情報抽出などの研究に従事．著書に『Java で学ぶコンピュータグラフィックス』など．電子情報通信学会, 人工知能学会，情報処理学会，言語処理学会，日本データベース学会，ACM, IEEE 各会員．