ISSN 2186-5647

P-18

多様体学習を統合したSOINNによる多変量密度推定法

日大生産工（学部）　○香山隆太郎　日大生産工　山内 ゆかり

1.まえがき

近年、ビッグデータを解析及び活用することに関心が高まっている。しかし、ビッグデータを解析することについて困難な点があることも確かである。まず、ビッグデータは膨大な数のセンサーなどからリアルタイムで生成され続けるので、学習が間に合わなくなる点である。次に、ビッグデータというものがパラメータに一切制限を求めない(ノンパラメトリック)ものであるという点である。最後に、データを実環境から取り入れるためノイズを多く含むと考えられ、高いロバスト性(ノイズ耐性)を必要とする点である。

この問題に対して中村らは、カーネル密度推定法とSOINNを拡張した、ロバスト高速オンライン多変量ノンパラメトリック密度推定法[1]を提案した。この手法は、まずSOINN[2]を用いてクラスタリングを行い、その後カーネル密度推定法を使用して確率密度を推定する。この時使用するSOINNは、今までユーグリット距離を使用していた部分にマハラノビス距離を利用することで、確率分布を求める時不要なノードを省きやすくしている。

そのほかに、定期的なノード削除の後、K近傍を求め双方向にK近傍だった時にエッジを張ることで、高次元データやサンプル数が少ないなどの理由でエッジが張られず密度を表せないということへの対策をしている。

その結果、学習時間、ロバスト性に対して既存手法より高い性能を示し、推定精度に対しても同等またはそれ以上の性能を示した。しかし、データが高次元になるにつれ推定精度が低下していくという問題点も判明した。

本研究では従来研究に、多様体学習の一つであるLocally Linear Embedding(LLE)[3]の特徴の一つである、高次元空間でのデータの局所的な位置関係の保持を利用する。この位置関係をマハラノビス距離の代わりに使用することで、データ空間における位相を考慮した不要なノードの削除が行えると考える。

また、K近傍を求め双方向にエッジを張る場合にもこの位置関係を使用することで、従来研究よりも正確にエッジを張ることが出来ると考える。そして、高次元データでの位置関係を保持しつつ正確に密度を表すことが出来れば、データが高次元になるにつれ推定精度が低下していくという問題点も解決できると考える。

2.従来研究

中村らは、SOINN の第一勝者ノードの更新が平均のオンライン更新と一致することに目を付け、ノードの周辺ネットワークは周辺サンプルの広がりとその方向を示していると考えた。

このサンプルの分布の情報を用いて、密度推定を行う。関数は以下のように定義される

$$\hat{p}(x) = \frac{1}{T_N}\sum_{n\in N} t_n K_{C_i}(x - w_n) \quad (1)$$

ここで$K_{C_i}(x - w_n)$はガウスカーネルである。各ノードにガウスカーネルを配置して、その線形和として確率密度を推定する。

また、勝者ノード間が離れていても、密度を表すエッジが張られる可能性を防ぐために、従来研究では、式（２）に示すマハラノビス距離を使用して、閾値の値がネットワークに沿うようにしていた。

$$\Theta_i = (\xi - w_i)^T M_i^{-1}(\xi - w_i) \quad (2)$$

この式のM_i^{-1}の転置前M_iには、局所ネットワーク共分散行列C_iを用いている。

$$C_i = \frac{1}{T_{p_i}}\sum_{p\in P_i} t_p (w_p - w_i)(w_p - w_i)^T \qquad T_{p_i} = \sum_{p\in P_i} t_p \quad (3)$$

ここで、t_pはpの勝者回数である。

局所ネットワーク共分散行列は通常の共分散行列と異なり、中心が平均ではなく対象ノードiの位置ベクトルで作られている。そして、勝者回数tで重み付けられている。これは、iのノードの周辺のサンプルの広がりを近似するためである。これを用いて、M_iは以下のようになる

$$M_i = C_i + \rho\gamma_i I \quad (4)$$

Multivariate Density Estimation Method Using SOINN Integrating Manifold Learning

Ryutaro KAYAMA and Yukari YAMAUCHI

$$\gamma_i = min_{p \in P_i} \| w_p - w_i \| \quad (P_i \neq \emptyset)$$

$$\gamma_i = mix_{p \in N \setminus i} \| w_p - w_i \| \quad \text{(otherwise)}$$

ここで、 ρ はハイパーパラメータ、I は単位行列である。この改良型 SOINN では、サンプル数が少ないと閾値範囲内にエッジが作られない。これを補うために、ノード集合に対して k 近傍グラフを作り、そのグラフで双方向にエッジがあるノードのペアに対してエッジを追加する機能を加えている。これにより、ノードが密に分布する場所では双方向にエッジが張られるので密度を表すエッジを残すことが可能になる。

3.提案手法

本研究では、従来研究のデータが高次元になるにつれ推定精度が低下していくという問題点を解決するために、LLEを使用する。

従来研究から変化している点は

1. ノードの周辺サンプルの広がりを意味する共分散行列の代わりにLLEの高次元データの局所的な位置関係を使用する。

2. K近傍を求める時、ユーグリット距離ではなく、LLEの局所的な位置関係を使用して近傍を求める。

3. 従来研究の密度推定法では、バンド幅行列に共分散行列C_iを用いていたが、提案手法ではLLEの高次元データの局所的な位置関係を使用する。

点である。

4.実験環境

本実験では、UCL Machine Learning Repository[4]のデータセットを用いて実験を行う。ハイパーパラメータを$\lambda = 1000, \rho = 1.0, k = 2d, age_{max} = 100$とし、各データセットからランダムに 75%を抽出しそれを学習データとする。残りをテストデータとし推定密度を測定する。これを 20 回繰り返しデータセットごとに評価する。

表1:実験に使用するUCLデータセット

	サンプル数	次元数	クラス数
Iris	150	4	3
Wine	178	13	3
Pima	768	8	2
Wine_Red	1,599	11	6
Wine_White	4,808	11	7
MAGIC	19,020	10	2
Skin	245,057	3	2

評価尺度は、負の対数尺度を使用した。

現在、従来研究が完成しておらず精度を求めることが出来ていない。そして、提案手法は、まだ正しく動作しておらず実験できなかった。

5.まとめ

本研究では、高次元データに対して、従来研究では出来なかった精度低下の克服そして、計算速度の改良を目的として LLE を使った密度推定法を提案した。

正常に動作した場合、高次元データ解析時に従来研究よりも、精度が高くなると考える。

また、逆行列を求めなくて済むため計算速度が速くなると考えられる。

今後の課題は、まず従来研究の完成と提案手法を稼働させることである。

6. 参考文献

1) 中村圭宏,長谷川修,ロバスト高速オンライン多変量ノンパラメトリック密度推定法. 電子情報通信学会(2014-8)

2) 山崎和博,巻渕有哉,申富饒,長谷川修,自己増殖型ニューラルネットワーク SOINN とその実践,日本神経回路学会誌. The Brain & neural networks, 17-4, (2010-12)

3) K Saul, Lawrence & T Roweis, Sam, An introduction to locally linear embedding. Journal of Machine Learning Research. 7. (2001).

4) UCL Machine Learning Repository, http://archive.ics.uci.edu/ml/index.php（アクセス日時 2017 年 10 月 19 日 14：05）