F-027

コピュラを用いた多変量ゼロ過剰データの分布推定・異常検知
Density Estimation and Anomaly Detection of Multivaraite Zero-inflated Data Using Copula

濱本 敬大†
Keita Hamamoto

1. はじめに

医療や金融など様々な分野において、統計的手法を用いた意思決定へのデータ活用が進んでいる。データの従う分布関数を推定する分布推定は、データの特徴把握や統計的パターン認識、および異常データの検出などのタスクにおいて、その根幹をなす基礎技術である。データ活用業務の拡大に伴って様々な特性を持ったデータが分析の対象となっており、データの各種特性に合わせた分布推定手法の整備が求められる。

業務データの特徴の一つに、ゼロ過剰データと呼ばれる、多くのレコードがゼロ値をとるデータがある[1]。例えば、「製造現場における不良品数」や「商品の購入個数」といった、離散的な値をとるデータに対しては、分布推定手法としてゼロ過剰ポアソン分布などを用いたモデル化がよく知られている。一方で、「昨日の薬剤投与量」や「配偶者の年収」といった、多くのゼロ値を含み有限かつ連続的な正値をとるデータの分布は、離散型と連続型の確率変数の混合で表現されるという特異な性質を有する。こうした変数を含む多変量データにおいては、後述する複雑な部分空間構造と確率密度関数の発散などの問題のため、従来の分布推定手法では取り扱いが困難である。

本研究では多変量のゼロ過剰連続変数に対する分布推定の困難点をまとめ、コピュラを用いた新しい分布推定手法を提案し、教師なし異常検知タスクへと適用した結果を報告する。

2. ゼロ過剰多変量連続変数の分布推定の困難点

一変数のゼロ過剰連続変数xの確率密度関数は、離散型と連続型の確率変数の混合である mixed random variable [2] の枠組みを用いて

$$q\delta(x) + (1-q)\,\tilde{f}(x)$$

とあらわすことができる。混合比qはゼロ値データの比率を表し、$\tilde{f}(\cdot)$は正値データの 1 次元連続確率密度関数であり、正の実軸上にサポートを持つ。原点$x = 0$に位置する確率質量はディラックのデルタ関数$\delta(\cdot)$を用いて、ゼロ値に局在した発散する確率密度として表現される。変数が D 個の場合の確率密度関数は、ゼロ値でないと仮定した変数の組み合わせに対応して異なる次元を持つ2^D項の和で表現され、それぞれ異なる部分空間上にサポートを持つ。

こうした確率密度関数の発散や異なる次元性を持つ大量の部分空間の存在により、従来の手法では正しい分布推定が困難である。それは、部分空間ごとに分布推定を繰り返す従来方針では、指数的に計算時間が増大するうえ、高次元の部分空間において正しく推定するのに十分なデータが得られないためである。また、図 1 に示すように、多峰性の分布を扱う手法として例えば混合ガウスモデル(GMM: gaussian mixture model)を用いると、低次元の部分空間に沿って無限小幅のガウシアンを配置してしまう。これによって尤度が無限大に発散するうえ、すべての部分空間を覆うために必要なガウシアンの数が指数的に増大し、精度や計算量の点で困難となる。カーネル密度推定(KDE: kernel density estimation)を用いた場合には、データの広がりの大きさや方向が部分空間ごとに大きく異なるため適切なカーネル幅の選択が困難である。結果として図 1 に示すようにゼロ値の周りに大きく広がった推定結果が得られ、負値や微小の正値データにも大きな尤度を与えてしまう。

このように、多変量ゼロ過剰連続データを単なる多峰性の分布として扱おうとすると、発散や部分空間構造に起因して様々な問題が発生する。そこで本研究では多変量ゼロ過剰連続データのもつ部分空間構造を考慮に入れた分布のモデル化を行う。

3. 提案手法：コピュラを用いた分布推定モデル

我々のモデルでは D 変量の同時確率密度関数を以下の形で表す。

$$f(x_1, \dots, x_D) = \left[\prod_{i=1}^{D}\left[q_i\delta(x_i) + (1-q_i)\tilde{f}_i(x_i)\right]\right] \times c$$

括弧内の積は、先に示した mixed random variable を用いた単変量の場合の確率密度関数を掛け合わせたものであり、変数間の依存関係を無視した場合の確率密度関数に相当する。括弧を展開することで、各変数がゼロ値か正値のいずれかを取る組み合わせに対応した2^D項の和が現れ、それぞれが異なる次元の部分空間と一対一に対応する。これによって、多変量の場合に指数的に増大する次元の異なる部分空間構造を一挙に表現することができる。

末尾のcはコピュラ密度と呼ばれる、変数間の依存性を表現する因子である。一般にコピュラを用いた分布モデリングでは、同時分布を周辺分布の積とコピュラ密度の積として記述する[3]。周辺分布とコピュラ密度のそれぞれを独立にモデル化することができるため、柔軟かつ表現力の高い手法として様々な分野で活用されている。本提案ではコピュラとしてガウシアンコピュラを用いる。ガウシアンコピュラのコピュラ密度は

$$c = \left.\frac{\phi_D(\omega_1, \dots, \omega_D|\Sigma)}{\phi(\omega_1) \times \cdots \times \phi(\omega_D)}\right|_{\omega_i = \Phi^{-1} \circ F_i(x_i)}$$

と表される。式中の$\phi(\cdot)$およびΦは標準正規分布の確率密度関数および累積分布関数、$\phi_D(\cdot\,|\Sigma)$はΣを共分散パラメータ、ゼロベクトルを平均ベクトルとして持つ D 変量正規分布の確率密度関数、F_iは変数x_iの累積分布関数である。非線形な変換$\omega_i = \Phi^{-1} \circ F_i(x_i)$によって各変数を標準正規分布に従う変数へと変換することで、変数間の相関を多変量の正規分布で表現することができる。

モデルのパラメータは学習データから以下のように推定できる。周辺ゼロデータ比率であるq_iは各変数x_iのゼロ値

† （株）日立製作所 研究開発グループ
Research and Development Group, Hitachi Ltd.

発生頻度から推定する。正値データの周辺分布$\tilde{f}_i(\cdot)$は、x_iの値が正であるデータのみを用いて従来の 1 変数分布推定によって推定することができる。本研究では KDE を用いて推定する。ガウシアンコピュラの相関パラメータであるΣは従来のガウシアンコピュラでよく行われているように[3]、非線形変換を施した後の変数である$\omega_i = \Phi^{-1} \circ F_i(x_i)$の共分散行列によって推定する。

我々のモデルは確率密度関数にデルタ関数を含むため発散箇所があり、そのままでは尤度として使用できない。そこで一変数の mixed random variable における尤度関数の構成法[2]に従い、同時尤度算出の際にはデルタ関数$\delta(x_i)$を指示関数$I(x_i = 0)$に置き換える。これによっていかなる入力データに対しても有限の尤度を出力することが可能となる。

本提案モデルでは指数的に増加する部分空間を周辺分布の積によって一括に表現しており、部分空間構造に起因した高次元での精度低下は生じない。またパラメータの推定および尤度の算出はともに入力次元 D の多項式時間で実行可能であり、高次元での計算量爆発も存在しない。

4. 数値実験

提案手法の有効性を示すため、台湾での融資サービスにおいて取得されたオープンデータ[4]を用いた数値実験を行った。各サービス利用者の n か月前の請求額と支払額を表す”BILL_AMT{n}”と”PAY_AMT{n}” (n=1,…,6)は、6~24%程度のゼロ値比率を持つゼロ過剰連続変数であり、相互に概ね正の相関を持ち右に歪んだ分布を持つ変数である。これら 12 変数に対して外れ値除去と対数変換を行ったのち、各手法で分布推定を実行し、その性能を測定した。分布推定の性能指標として通常用いられる対数尤度値は、例えば GMM によって無限に獲得されてしまうため適切な指標とならない。そのため、二値分類の精度指標を用いて性能を評価した。

指標の算出においては、まず 30000 件のデータを 21000 件の学習データと 9000 件の正常テストデータに分割した。次に 9000 件の正常テストデータを複製し、値の書き換えによって同数の異常テストデータを作成した。書き換えにおいては、各データに対し正値をとる変数x_iの値を一様乱数からのサンプル値で置き換えた。一様分布の下限と上限には、学習データにおける変数x_iの 1 および 99 パーセンタイル値を用いた。性能を評価するため、学習データ 21000 件で学習した GMM、KDE、提案手法の三種の分布推定モデルを用い、計 18000 件のテストデータに対して対数尤度値を算出し、正常/異常フラグとの ROC-AUC 値を性能指標値とした。データの分割および一様乱数のシード値を変えた 15 回の実験における平均値及び標準偏差を表 1 に示す。

表 1 によると提案手法による異常検知性能が最も高く、多変量ゼロ過剰連続変数の分布の特徴を最もよく捉えていることがわかる。また、GMM における最適なガウシアン成分数は 155 に及んでおり、より高次元では計算時間に困難が生ずると考えられる。加えて KDE においては次元の呪いのため、より高次元ではさらに性能が低下すると考えられる。一方、我々の提案手法ではより高次元においても著しい計算量の増加や精度の低下は見られず、優位性が大きいと考えられる。

表 1　ROC-AUC 値

GMM	KDE	提案手法
0.9058(0.0039)	0.9200(0.0017)	0.9684(0.0014)

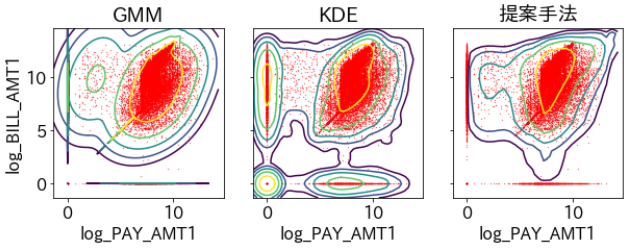

図 1　従来手法及び提案手法による分布推定結果

次に、各手法による分布推定で得られた確率密度関数を、等高線によって表したものを図 1 に示す。図示のため”BILL_AMT1”と”PAY_AMT1”の二変数のみを対象に 2 次元の分布推定を行っている。提案手法ではゼロ値データを指示関数$I(x_i = 0)$によって判定するため、等高線プロットが困難である点には注意が必要である。このことは、ゼロ値に無限の尤度を与えてしまったり確率密度関数が不自然に広がったりしていた従来手法の欠点が取り除かれていることを意味する。一方で、変数間の強い依存関係を的確には捉え切れていない部分が見て取れる。これは依存関係を単なるガウシアンコピュラのみで表現したためで、ゼロ過剰データの特性に合わせた適切なコピュラの選択など改善の余地が残されている。

5. おわりに

本研究では多変量のゼロ過剰連続変数において、確率密度関数の発散や異なる次元を持つ部分空間構造などの特異な性質によって従来手法による分布推定が困難であることを示し、これらに対応する新たな分布推定モデルを提案した。提案手法では周辺分布を mixed random variables を用いて記述し、変数間の相関にガウシアンコピュラを導入することで多数の部分空間を一挙に表現する。これにより、パラメータの推定や尤度の算出は入力次元 D の多項式時間で実行可能となり、高次元においても精度低下や計算量爆発のおそれがなくなった。提案手法を、オープンデータを用いた数値実験によって従来手法と比較することで優位性を示し、ゼロ値の扱い方がより適切であるとの見込みを得た。

一方で変数間の相関はガウシアンコピュラのみでしか表現できていないことから、例えば異なる変数間のゼロ値発生の相関などを記述できていないため、適切なコピュラを選択することでより柔軟で表現力のある統計モデルとなると期待できる。

参考文献

[1] L. Liu, et. al., “Statistical analysis of zero-inflated nonnegative continuous data: a review,” Statistical Science, vol. 34, no. 2, pp. 253–279, 2019.
[2] H. Pishro-Nik, "Introduction to probability, statistics, and random processes", available at https://www.probabilitycourse.com, Kappa Research LLC, 2014.
[3] E. Bouyé, et. al., “Copulas for finance-a reading guide and some applications,” Available at SSRN 1032533, 2000.
[4] I.-C. Yeh and C.-h. Lien, “The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients,” Expert systems with applications, vol. 36, no. 2, pp. 2473–2480, 2009.