異常検出モデルを用いた多次元時系列データの特徴量選択手法に関する研究

情報数理応用研究　5218C002-1　新井浩健
指導教員　後藤正幸

A Study on Feature Selection Method of Multidimensional Time Series Data Using Anomaly Detection Model

ARAI Hirotake

1. 研究背景と目的

近年，IoT 技術の発展により，多くの機器にネットワークが接続されることが可能になり，企業には大量な情報を収集可能の環境が整いつつある．それらの大量なデータを用いて顧客分析を行い，マーケティングに活用することが求められている．このとき，企業での実務的な分析では，大量のデータをいかに効率よく扱うかという課題がある．そのため，分析のコストや理解の平易さが重要であり，複雑なモデル構築による高性能な手法よりも，現場レベルでのマーケティングに活用可能な結果が必要とされている．特に時系列情報を考慮した，顧客の離脱予測や行動が変化する変化点検出といった分析が望まれている．

本研究では，某プリンタメーカーが提供するネットワークプリンタから取得される，プリンタ製品の使用履歴データを対象とする．その企業では，顧客のプリンタの使用履歴に関するデータが日々蓄積されている．プリンタに関する特徴量は印刷枚数やトナー交換回数などの一般的なプリンタ利用状況だけでなく，印刷ジャム回数や通電時間など，多様な顧客の使用状況を表す特徴量も取得されている．また，印刷枚数についても，トータルの印刷枚数に加えて，用紙サイズや出力・入力の方式など多種の特徴量が蓄積されている．これらの特徴量はプリンタ毎に時系列データとして観測され，かつプリンタ台数も非常に多い．そのため，これらの特徴量を全て用いて顧客の利用状況をモニタリングしたり，その変化を分析することは，分析コストや解釈性の観点から望ましくない．そこで，本研究ではこれらの特徴量から顧客の特性を分析するために重要となる，比較的少数の特徴量を選択することを目的とする．これらの特徴量は，時系列のデータであると共にそれぞれスケールが異なり，単純に特徴量を比較する分析が困難である．この問題に対して，Zhang et al.[1] は，特徴量の平均値などを用いるよりも，異常検知手法の 1 つである One-Class SVM[2] を用いて時系列データを顧客全体に対する各顧客の外れ度という指標に変換することにより，顧客の性質に基づく分類が可能になることを示している．

そこで本研究においても時系列データに対し，One-Class SVM により，外れ度を算出し，これらを用いて顧客が優良，非優良かを予測する上で重要となる特徴量を選択する分析モデルを構築する．このとき，特徴量選択においては，各特徴量をネットワーク分析によりコミュニティに分割することにより，顧客の性質を決定づける重要な特徴量を特定する手法を提案する．本手法を某プリンタメーカーから提供された顧客のプリンタの使用履歴データに適用し，有効性を示す．

2. 準備

2.1. ビジネスデータにおける特徴量選択

ビジネスの現場において，単にデータを処理するだけではなく現場レベルで活用できる，解釈性の高い分析が望まれている．特徴量を解釈する手法としては，主成分分析などを用いて特徴量を次元圧縮する方法と重要な特徴量を選択する方法が考えられる．主成分分析などは次元圧縮手法であるが入力に全ての特徴量が必要となる．そのため理解の平易さは問題となる．そこで本研究は，重要となる特徴量を絞り込むための分析モデルを考察する．Zhang et al.[1] の研究では，プリンタの時系列データに対し，異常検知手法の 1 つである One-Class SVM を用いて求められる外れ度指標を用い，特徴量を変換することで，顧客の時系列の特徴を表現することが有用であることを示している．

2.2. 異常検知と One-Class SVM

異常検知とは，データセット内の大多数のパターンから異なったパターンを検出する手法であり，クレジットカードの不正利用の検出や，システムの故障検知などの分野で用いられる．異常検知技術は大きく分けて教師あり学習，半教師あり学習，教師なし学習の 3 つに大別される．本研究で対象とするプリンタのデータはユーザが予め優良か非優良かのようなラベルが存在しないため，教師なし機械学習の枠組みで考える必要がある．One-Class Support Vector Machine (One-Class SVM) は教師なし機械学習手法の 1 つであり，全てのデータが 1 つのクラスに所属していると仮定し，そのクラスからの外れ度を算出し，異常か否かを識別する手法である．観測された学習データを $\boldsymbol{x_i}(i = 1, 2, ..., n)$ とする．$\boldsymbol{w}$ を識別境界の法線ベクトルとする．一般に線形分離不可であるため，スラック変数 ξ_i を導入する．One-Class SVM の目的関数を L_p とすると，それを最小化するための最適化問題は以下のように与えられる．ここで ρ は識別関数の切片で，$\varphi(\cdot)$ はカーネル関数とする．

$$\min L_p(\boldsymbol{w}, \xi) = \min \left(\frac{1}{2} \boldsymbol{w}^T \boldsymbol{w} - \rho + \frac{1}{vN} \sum_{i=1}^{N} \xi_i \right), \quad (1)$$
$$\text{s.t. } \boldsymbol{w}^T \varphi(\boldsymbol{x_i}) - \rho + \xi_i \geq 0, \xi_i \geq 0.$$

この問題に対してラグランジュの未定乗数 α_i, α_j を導入し，対応する双対問題の目的関数を L_d とすると，以下のように

表すことができる．

$$\max L_d(\boldsymbol{\alpha}) = \min\left(-\frac{1}{2}\sum_{i=1}^{N}\sum_{i=1}^{N}\alpha_i\alpha_j\varphi^T(\boldsymbol{x_i})\varphi(\boldsymbol{x_j})\right), \quad (2)$$
$$\text{s.t. } 0 \le \alpha_i \le \frac{1}{vN}, \sum_{i=1}^{N}\alpha_i = 1.$$

ここで，$\rho^T(x_i)\rho(x_j)$ の部分を式 (3) のようなガウシアンカーネル $K(\boldsymbol{x_i}, \boldsymbol{x_j})$ で置き換える．$\|\cdot\|$ はユークリッド距離を表す．

$$K(\boldsymbol{x_i}, \boldsymbol{x_j}) = \exp\left(-\frac{1}{2\sigma^2}\| \boldsymbol{x_i} - \boldsymbol{x_j} \|^2\right) \quad (3)$$

この時データの外れ度合いは式 (4) で表される．

$$d = \sum_{i=1}^{N}\alpha_i K(\boldsymbol{x_i}, \boldsymbol{x_j}) - \rho \quad (4)$$

2.3. ネットワーク分析

ネットワーク分析は，ネットワークの構造が仮定される問題に対し，その構造を元に分析する手法であり，コミュニケーションネットワークなど，社会的ネットワークの分析などに用いられる．本研究で対象とするデータの様に予めネットワークが定義されていない場合，ノードの接続方法は大きく分けて 2 つある．最近傍の k 個のノードと接続する k-NN 法と閾値 ε 以下の類似度をもつノード同士を接続する ε-NN 法が存在する．k-NN 法では接続するノード数を指定するため，k を大きくすればするほど密なネットワークが構築される．一方，ε-NN 法では類似度を元にネットワークを作成するため，ε を小さくすればするほどネットワークの構造が複雑になり，密なネットワークが構築される．また，ε-NN 法は接続するノード数に指定がないため，多くのノードと相関の持つ変数に対して多くのネットワークを構築し，コミュニティを検出することができる [4]．本研究では特徴量のネットワークを作成することで，似たような性質を持つ特徴量から構成されるコミュニティを検出し，分析に重要となる特徴量の選択を試みる．

3. 提案手法

3.1. 概要

本研究では某プリンタメーカーが提供するプリンタの使用履歴に関する時系列ログデータを用いる．100 以上の特徴量がある中で，顧客の行動を予測する上で重要な特徴量を選択する手法を提案する．本研究で対象とするデータは，特徴量数が多く月次ごとに蓄積された多次元の時系列データであること及び特徴量間の相関が想定されるといったが特徴である．そこで本研究では上記の 2 つの特徴を考慮した特徴量選択を試みる．まず，特徴量が多く時系列データであり，また，印刷枚数とトナー交換回数など，スケールの異なる特徴量が存在する．そこで Zhang et al.[1] と同様に異常検知手法の 1 つである One-Class SVM を用いて，そこから求められる特徴量の外れ度を用いることで，時系列を考慮した指標を分析に取り入れることを考える．その上で，100 以上ある特徴量の中から分析コストを削減するため，本研究では顧客を優良，非優良に分類する上で重要な特徴量を選択・発見することを考える．そのための分類器としては分類精度が高い機械学習モデルとして知られる，Random Forest[3] を用いる．Random Forest を用いた分類手法は，回帰分析のように変数を逐次的に追加することができず，変数をあらかじめ選択する必要がある．そのため，変数の組み合わせごとに実験が必要になるが，全ての組み合わせを計算することは現実的に困難である．一方で，Random Forest の重要度による特徴量選択では，トナー回数に関する類似の特徴量が上位に来るため，精度が高い一方で実務上扱うには適さない．そこで，ネットワークの作成により検出された各特徴量コミュニティの中から特徴量を 1 つずつ選択することで少数かつ分析に重要である特徴量の選択を行う．

以下にその具体的な手順を示す．

3.2. One-Class SVM による特徴量の外れ度算出

本研究で対象とするデータは，プリンタの時系列データである．本研究で対象とするプリンタの使用状況の変動に関しては，顧客が離脱するか否かがマーケティングの観点から重要であり，時系列における顧客の使用傾向の変化を考慮することが重要である．まず，各顧客のある特徴量の時系列データに対し，One-Class SVM を適用し，ある特徴量における顧客ごとの外れ度を算出する．この作業を n 個全ての特徴量に対して行うことで，顧客の特徴量に関する外れ度ベクトル $\boldsymbol{w} = (x_1, x_2, ..., x_n)$ を作成することができる．

3.3. ネットワーク分析によるコミュニティの抽出

One-Class SVM から求めた特徴量の外れ度行列を元に，特徴量間のコサイン類似度を算出し，各特徴量に対する類似度行列を作成する．そして，その類似度を元にネットワーク図を作成する．本研究では特徴量間の外れ度の類似度を元に，相関のある特徴量で構成されるコミュニティを検出する．本研究では特徴量を選択するために，似た異常傾向を持つ特徴量同士をコミュニティに構成することを考える．そこで全特徴量に対し，ε-NN 法でネットワークを作成し，コミュニティ検出を行う．

3.4. 重要度と相関の考慮した特徴量選択

本研究での特徴量選択は比較的少数の特徴量かつ分類精度もある程度高くなるような特徴量の探索を行う．ε-NN 法を元に作成したネットワークから，コミュニティに所属する特徴量数の多い順に M 個のコミュニティに着目する．同一コミュニティに所属する特徴量は似たような異常度を持つため，コミュニティ内の代表の特徴量のみを選択すれば良いと考えられる．そこで，各コミュニティから 1 つずつ特徴量を選択し，M 個の特徴量を元に分類モデルを構築してその精度を比較することで，分析に重要となる特徴量の選択が可能になると考えられる．各コミュニティに所属する特徴量全てを用いて全数探索することも考えられるが，その組み合わせ数は各コミュニティに所属する特徴量数の積となるため，コミュニティ数を多くすると組合せ数が爆発する．本研究で

は適切なコミュニティ数も事前に不明であるため，貪欲的探索を行う．所属する特徴量数の多いコミュニティから順に貪欲的に精度の高い特徴量を選択する．そのことにより，組合せ数を大幅に抑えることができる．従来の特徴量選択においては，特徴量の選択は経験と勘，もしくはすべての組合せを網羅的に探索する必要があったが，本研究のようにネットワークからコミュニティを検出し，貪欲的に探索することにより，計算量を抑えつつ，重要な特徴量を抽出することが可能になる．

3.5. 顧客分類モデルの構築

プリンタの使用履歴の時系列から，将来的に顧客が優良顧客か，非優良顧客のどちらに推移するのか予測できることは企業にとって非常に有用であると考えられる．そこで，顧客の時系列をもとに顧客分類モデルを構築することを考える．まず，過去のデータとして得られる時系列をもとに顧客を優良顧客と非優良顧客をラベル付けし，これらを学習して分類器を構築する．まず，One-Class SVM を各特徴量の前半 w 月の時系列に適用し，外れ度 $d_w^{'(p)} = (d_w^{'(1)}, d_w^{'(2)}, ..., d_w^{'(p)})$ を求める．それらを Random Forest で学習し，分類モデルを構築する．そして，後半 w 月のデータをテストデータとし，モデルの予測性能を評価する．また，この予測性能を用いて前述の通り特徴量を選択し，モデルを構築することで重要な特徴量を特定する．

4. 実データ分析

4.1. データセットと実験条件

本研究で対象とするデータは 2018 年 4 月から 2019 年 3 月までに使用され，全ての月においてデータが正確に取得されたプリンタ 53,780 件を対象とする．データは月次で保管され，各月の月末のある地点での，その月の合計値が蓄積されたデータを用いる．特徴量は 112 の量的変数を用いる．程よい数と大きさのコミュニティを検出するという観点からネットワークを作成する際の ε-NN 法の ε は 0.5 とした．また，優良顧客は，半年間の間でのトナー交換回数が 4 回以上の顧客と定義する．

4.2. 分析結果とその考察

4.2.1. ネットワークの作成とその考察

図 1 に外れ度を用いた指標においてできたコミュニティを示す．全特徴量 112 のうち，どの特徴量とも接続しなかった特徴量を除く，65 の特徴量では合計 12 個のコミュニティが作成された．一番大きなコミュニティはトナー交換回数やローラー回転数などに関する特徴量であり，二番目に大きなコミュニティは印刷枚数に関する特徴量，三番目に大きなコミュニティは特殊な印刷方法に関する特徴量など，各コミュニティごとに所属する特徴量の外れ度の特性が表れている．

そして，一番大きいコミュニティは 24 もの特徴量から構成され，トナー交換に関する外れ方は多くの特徴量の外れ方と性質が似ていることがわかる．

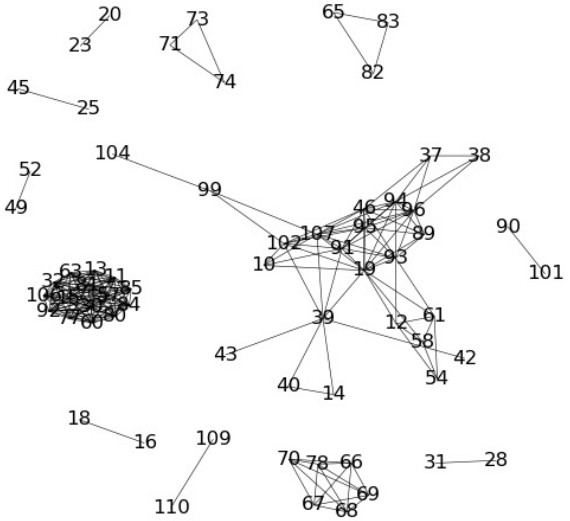

図 1: 検出された 12 個の特徴量コミュニティ

4.2.2. 分類モデルを用いた重要度による特徴量選択

表 1 に全特徴量を入力とした分類器の重要度の上位の特徴量を示す．

表 1: 分類モデルにおける重要な特徴量上位 7 件

重要度	特徴量名	特徴量 No
0.183	トナー交換回数①	99
0.138	トナー交換回数②	107
0.081	トナー交換回数③	102
0.065	ローラー回転数①	91
0.048	印刷ジャム回数①	37
0.047	電圧に関する変数	19
0.037	印刷ジャム回数②	39
0.035	トナー交換回数④	104
0.033	ネット接続ユーザ数	10
0.024	印刷ジャム回数③	43

表 1 より，分類に重要な特徴量はトナー交換回数の他に，印刷枚数やジャム回数，ローラー回転数などが挙げられる．しかし，これらの多くは，図 1 の最大コミュニティに属し，相関の強い似たような特徴量がである．そのため実務の観点からこれらの類似した特徴量をモニタリングしたり，分析対象とすることは，費用対効果の面で合理的ではない．

4.2.3. 特徴量選択とその精度

ネットワーク分析によって 12 の特徴量コミュニティの大きいコミュニティから順に貪欲的に探索的し特徴量の選択を行う．得られた分類精度の結果を図 2 に示す．

選択する特徴量を 2 から 12 まで変化させたとき，特徴量数が 2 から 6 までの間は高い精度を示し，特徴量数が 6 のとき最も精度が高くなる．一方，それ以上増やすと精度が落

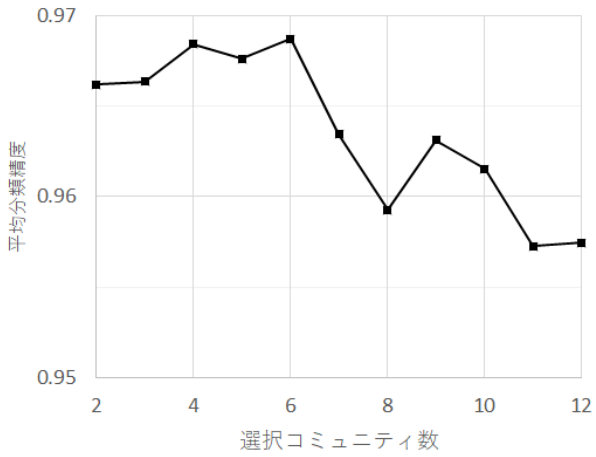

図 2: コミュニティ数と分類精度

表 2: コミュニティ数 6 で選択された特徴量

重要度順位	特徴量	特徴量 No
1	トナー交換回数①	99
2	印刷枚数①	85
3	特殊な印刷枚数①	66
4	特殊な印刷枚数②	83
5	特殊な印刷枚数③	71
6	特殊な印刷枚数④	52

ちることがわかる．したがって顧客が優良顧客かどうかを識別する上で，重要となる特徴量が 6 つほどであることが考えられる．また最も精度の高い 6 つの特徴量を表 2 に示す．これらは異なった解釈ができる特徴量となっている．したがってこのような代表的な特徴量のみを考慮することで幅広い特徴量が選択される．

4.2.4. 全数探索と貪欲的探索の比較結果

提案手法では計算量の観点から貪欲的探索を行った．しかしこれは局所最適解である可能性がある．そこで，全数探索と貪欲的探索のコミュニティ数が 3 の場合の比較結果を表 3 に示す．

表 3: コミュニティ数 3 における比較結果

比較手法	全数探索	貪欲的探索
組合せ数	2160	45
MAX 精度	0.966	0.966
最適な組合せ	99,85,68	99,85,66

表 3 のように，貪欲的探索は全数探索と比較して組合せ数が非常に少なくなっているのにも関わらず精度がほとんど落ちていないことがわかる．分類精度を見るとほとんど誤差の範囲で収まる．しかし，最適な特徴量の組合せは変化する．この入れ替わる 2 つの特徴量は類似しているため代替が可能である．したがって，貪欲的探索を行うことで，全数探索と同等の結果が得られていることが分かる．

5. 考察

本研究ではデータ分析のしやすさの観点から特徴量を選択する手法を提案した．外れ度をもとにネットワークを作成し，コミュニティ検出を行った結果，一見重要でない特徴量が，様々な特徴量の外れ度と類似することで，顧客分析にとって重要な特徴量になりうることがわかった．実応用上の観点から，今回示された重要な特徴量のみに着目し，より詳細な時系列を分析することで，計算コストを削減した上で顧客の使用パターンの変更や，離脱といった特徴的な行動を予測できることが期待される．

本研究では優良顧客をトナー交換回数で定義したが，印刷枚数が時系列とともに増加する顧客を優良と定義することもできる．問題設定により優良顧客の定義を変更することで問題に適した幅広い分析が可能になる．

また，分類に寄与する特徴量の選択に対し，コミュニティごとに特徴量を貪欲的に探索することで選択する特徴量の組合せ数の爆発的増加を回避し，全数探索とほとんど変わらない精度で特徴量を選択することが可能であることを示した．コミュニティ内の特徴量が多く，どの程度コミュニティを考慮するか事前に予測できない状況下ではこのような貪欲的探索も有効であることが考えられる．

6. 結論と今後の課題

本研究では，プリンタを使用する顧客の利用履歴データに対し，外れ度を用いることにより特徴量の時系列を考慮し，かつ相関の想定される特徴量の中から顧客の行動傾向を予測する上で重要かつ解釈性の高い特徴量の選択手法を提案した．そして，実データを用いて提案手法が顧客の分類精度を保ったまま，使用する特徴量数を削減することが可能であることを示した．今後の課題としては，選択した特徴量のみを用いた，時系列データに対する詳細な分析を行い，顧客の使用状況が変化するタイミングの予測とその考察などがあげられる．

参考文献

[1] Zhang.X, Yamashita.H, Kumoi.G, and Goto.M, “A Model to Detect Unique Customers Based on Time-series Log Data,” *The 19th Asia Pacific Industrial Engineering and Management Systems*, ID–207, 2018.

[2] Schölkopf.B, Platt.J.C., Shawe-Taylor.J, Smola.A.J, and Williamson.R.C, “Estimating the support of a high-dimensional distribution,” *Neural computation*, 13(7), pp. 1443–1471, 2001.

[3] Breiman.L, “Random Forests,” *Machine learning*, vol.45(1), pp.5–32, 2001.

[4] Leonardo.N.Ferreira and Liang.Zhao, “Time series clustering via community detection in networks,” *Information Sciences*, vol.326, pp. 227–242, 2016.

[5] Chandora,V., Banerjee,A., Kumar,V., “Anomaly detection: A survey,” *ACM computing surveys(CSUR)*,Vol.51,No.11,Article.15,2009.