F-026

マハラノビス-タグチ法を用いた不均衡データ処理手法の提案
A Study on the Imbalanced Data Processing with Maharanobis Taguchi System

榑松理樹†
Masaki Kurematsu

1. はじめに

予兆検知とは、工場内などでの機器が故障する前に「いつもと違う」状態、異常状態を故障や停止する前に検知することを指す。故障や停止する前に予兆を検知することで、重大事故を未然に回避することができるほか、機器の安定した利用に繋がり、社会的にも重要なタスクである。異常検知や変化検知、外れ値検知なども広い意味で同様なタスクといえる。

予兆検知に用いられる技術としては、機械学習があげられる。近年、ソフトウェア、ハードウェアの双方において利用環境整備が進み使いやすくなったこともあり、多くの研究が試みられており、一定の成果を得ている[1]。

一方で、ニューラルネットワークや SVM に代表されるこれら機械学習の多くでは、予兆検知モデルを大量のデータからモデルを学習する。良い検知モデルを構築するためには、質・量ともに十分なデータが必要となる。しかし、予兆検知におけるデータは、正常状態が大多数をしめ、異常状態とのデータ量に大きな差がある不均衡データとなる。このようなデータから構築した学習したモデルは正常な状態に偏る傾向がある。そのため、良い検知モデルを構築するためには、不均衡データ処理が必要[1][2]である。

不均衡データ処理としては、データの重み付け・水増し・間引きが初等的手法[2]として用いられている。しかし、重みの決定方法、間引かれたデータの扱い、水増ししたデータの信頼性など各手法とも課題が残されており、新たな手法を検討する必要がある。

一方、予兆検知は、高性能、高品質を目指す技術について扱う品質工学においても取り組まれている。品質工学においては、マハラノビス・タグチシステム（MTS）[3]が代表的な手法である。MTS は、正常なインスタンスの範囲を示す単位空間を作り、その中心からの距離に基づいて新しいインスタンスを評価する。MTS は小規模なデータに適用可能であり、計算量も比較的小さいという特徴を持つ。しかし、正常・異常の境界値については、利用者の経験や勘によるところが大きく、その明確な定義方法が必要である。また MT 法は実例に適用した研究が多く，手法自体への研究は少ない[4]。

機械学習、MTS については、同様のタスクに利用できるものであり、比較も行われている。しかし研究領域の違いからか両者を融合した研究例は少ない。融合した研究としては，MTS の前処理として，主成分分析やカーネルトリックを用いた手法[5]がある。機械学習と MTS は共通点も多く、またそれぞれが得意としていることが異なる。よって、これらを融合することで新たな手法が構築できる可能性は高い。

†岩手県立大　Iwate Prefectural University

以上の背景から、本論文では、機械学習、MTS のそれぞれの特徴に着目し、MTS を用いた不均衡データ処理手法を提案する。本手法では、MTS を用いて学習データを正常・異常の境界データ、すなわち、不均衡データの中でも均衡がとれる分を抽出する。抽出した境界データに対し、機械学習を適用し、検知モデルを構築する。予兆検知においては、初めに MTS で判定を行い、境界データに入る場合は機械学習による検知モデルで判定を行う。本手法により、機械学習に対しては、不均衡データ処理の新しい手法を、MTS には境界データの処理を実現する。

以降、第 2 章では、背景となる不均衡データ処理および MTS について述べる。続く第 3 章において、提案手法を説明した後、実験に基づく評価結果を第 4 章で示す。最後に、第 5 章で結論および将来の課題を述べる。

2. 関連研究

2.1 不均衡データ処理

水増し法は、不均衡データ処理の初等的手法の 1 つであり、多数クラスと少数クラスの間の不均衡を解消するために新しいデータを作成する。代表的な手法としては、Chawala ら[6]が提案した Synthetic Minority Over-sampling Technique (SMOTE)がある。SMOTE は、クラス間のバランスを保つために、少数クラスからランダムに K 個の最近傍サンプルを選択し、式(1)に基づいて選択したインスタンスの間に新しいインスタンスを作成し、少数インスタンスの領域を拡大する。式(1)において、x_i は元のサンプル、x_j はランダムに選択されたサンプル、x_{n+i} は新しいサンプルをそれぞれ示す。また r は 0 から 1 の間の乱数である。

$$x_{n+i} - x_i + r(x_j - x_i), (0 \le r \le 1) \qquad (1)$$

また Han H ら [7]による Borderline-SMOTE、 Haibo He ら [8]による Adaptive Synthetic Sampling (ADASYN) などの SMOTE の派生手法が提案されている。

また間引き法では、多数クラスと少数クラス間の不均衡を解消するためにデータからインスタンスの削除を行う。間引くインスタンスの選択方法として、既存のインスタンスを選択する方法と、近傍のインスタンスから新たに生成するインスタンスを利用する方法が主流である。前者の方法としては、ランダムに選択する方法 RUS（Random Under-Sampling algorithm）[9]がある。しかし、RUS によって変更されたデータセットには偏りがある可能性があるため、いくつかのアプローチが提案されている。1 つはクラスタベースの間引き方法であり、主要なクラスを選択するクラスタを作成し、各クラスタから順に間引いていく。一方、後者においては、複数のインスタンスを統合したインスタンスをもとのインスタンスの代わりに利用する方法である。

このように、不均衡データを処理する手法はいくつか提案されており、さらに改良が試みられている。

2.2 MTS

MTS は、田口玄一らが提唱したマハラノビス距離（MD）を利用した統計手法である。MTS では、正常の集団（単位空間）を考え、判定対象データと単位空間の中心との MD を基に、正常か異常かを判定する。判定対象データから MD を計算するのが主体であることから、判定対象データのみかつ小規模データに適用できる特徴がある。MTS は単位空間内に含まれるインスタンスの発見を行うことから、パターン認識手法としての役割ももち、判別、予測、推定、診断のタスクに活用され、バイオテクノロジー、マーケティング、医療などの領域で応用されている。さらに、MTS は改良が加えられ、いくつかのバリエーションが提案されている。その中の 1 つである Recognition Taguchi Method（RT）は判別タスクに重点を置き、データを高速に処理することができる。

一方、MTS には，正常異常判定の閾値決定方法が不明瞭という課題[10]がある。実際の利用においては、経験や学習データに基づいて定義される傾向がある。また MTS の別の課題としては「単位空間が多重正規分布しない場合の判定方法」がある[10]。この課題に対しては、主成分分析やカーネルトリックを適用したデータに MTS が提案[10]されている。これは機械学習と MTS とを融合する点本研究と類似するが，融合方法は異なっている。

3. 提案手法

3.1 概要

本研究では、MTS を用いて不均衡データを処理する手法を提案する。本手法の概要を図 1 に示す。本手法は、学習フェーズと検知フェーズに分割される。

学習フェーズは、次の①から④のステップで検知モデルを構築する。

① 学習データの少数クラスに MTS を適用し、単位空間を構築する。この単位空間は、検知フェーズでも用いる。
② 全学習データに対し、この単位空間の中心からの MD を求める、
③ 境界データ（機械学習用データ）を以下の方法で設定する。
  (ア) 上記②で求めた MD の平均値±標準偏差の範囲を取り出す。
  (イ) 上記の範囲に入る正常・異常データを取り出す。
  (ウ) 式(2)を用いて不均衡度を求める。式(2)において、N^{minor} は少数クラスのインスタンス数、N^{major} は多数クラスのインスタンス数を意味する。

$$不均衡度 = \frac{N^{minor}}{N^{major}+N^{minor}} \qquad (2)$$

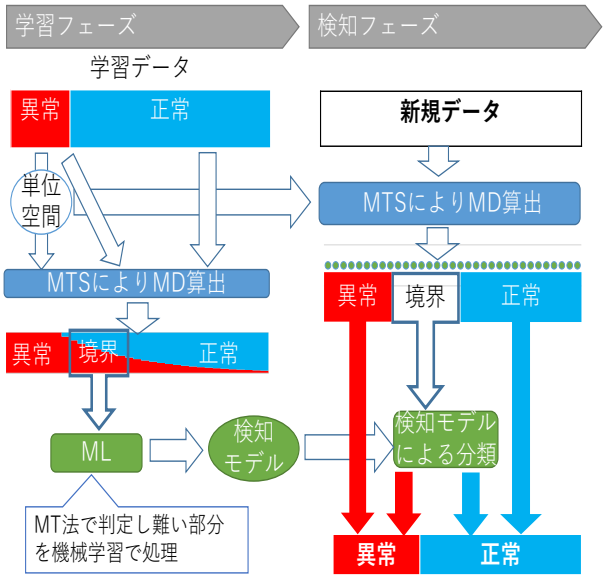

図 1　提案手法の概要

(エ) 得られた不均衡度が一定範囲以内であれば、今得られた範囲を境界データの範囲とする。一定範囲外であれば、標準偏差の範囲を変更し、一定範囲内に入るまで調整する。
④ 境界データに機械学習を適用し検知モデルを構築する。

本手法は、機械学習から見れば、学習データの一部を取り出すという点から、間引き法の一種と捉えることができる。一方で、間引いたデータは、MTS として予測においても利用する点が、大きく異なっている。また間引いたデータに MTS を適用することで，それらデータの有効活用，検知精度向上が期待できる。

検知フェーズでは、次のステップで新規データから予兆検知を行う。

⑤ 新規データに対し、学習データの①で構築した単位空間の中心からの MD を求める、
⑥ 上記⑤で得た MD が、学習フェーズの③で得た MD の範囲外であれば、MTS の判定結果を検知結果とする。範囲内であれば、機械学習で用いた検知モデルを適用し、判定する。

4. 評価

4.1 実験概要

本手法の有用性を評価するために Python を用いた実験を行った。実験の概要を図 2 に示す。

評価実験においては、本手法を、不均衡データ処理を行わない場合、SMOTE、RUS との比較を行う。SMOTE、RUS については Imbalanced-learning ライブラリ[1]を用いる。また機械学習モデルとしては、Random Forest (RF), Logistic Regulation (LR), 線形 SVM, SVM, K-Nearest Neighbors (KNN), Multi-Layer Perceptron (MLP)を使用する。これらについて

[1] https://imbalanced-learn.org/stable/

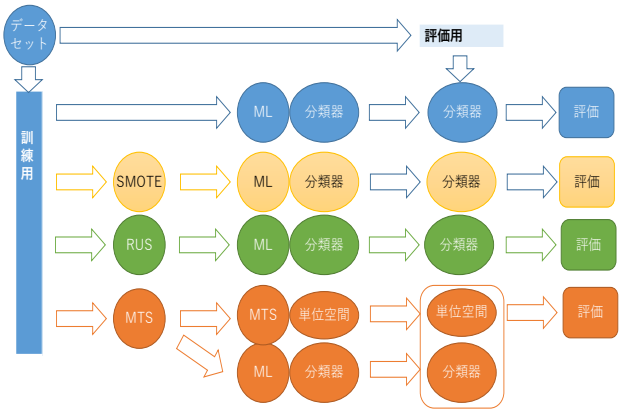

図 2　評価実験概要

表 1　各機械学習モデルの設定

モデル	パラメータ
RF	citerion=gini,estimators=100,min_sample_split=2
LR	penalty=12, C=1.0
SVM	C = 1, kernel=rbf
線形 SVM	penalty=12, loss-function=squeard hinge, C=1.0
KNN	k-nearests = 5
MLP	中間層 100 ノード 1 層、activation 関数：relu.

は、Python 機械学習用のライブラリとして著名な scikit-learn ライブラリ[1]を用いる。これらのモデルのパラメータを表 1 に示す。

実験データとしては scikit-learn ライブラリの関数{Thinkit make_classification}を用いて機械的に作成したものを用いる。本関数はガウス分布に基づいたデータセットを作成する。この関数で生成されたデータセットの特徴は、あるレベルで制御可能である。よって、この人工データセットを用いて、提案手法と特徴量との関係を分析する。本実験では、この関数を用いて、10 個の特徴量を持つ 5000 個のインスタンスを作成した。また多数クラスと少数クラスの割合については、式(2)で得られる不均衡度が 0.1 になるように設定する。作成したデータのうち、70%を学習フェーズで利用し、残りのインスタンスを検知フェーズで利用する。実験においては、5-Fold Cross Validation を実施し、その評価の平均を用いる。

評価基準には、式(3)で表される Sensitivity、式(4)で表される Specificity、式(5)で表される G-mean（Sensitivity と Specificity の幾何平均）を使用する。

$$Sensitivity = \frac{TP}{TP + FN} \quad (3),$$

$$Specificity = \frac{TN}{FP + TN} \quad (4)$$

$$G - means = \sqrt{Sensitivity \times Specificity} \quad (5)$$

式(3)(4)において、TP は P クラスを正しく P クラスと識別したデータ数、FN は N クラスを誤って別のクラスと識別したデータ数、FP は、P クラスを誤って別のクラスと識別したデータ数、TN は N クラスを正しく N クラスと識別したデータ数を示す。本実験においては、P クラスが多数クラス、N クラスが少数クラスになる。よって、Sensitivity は多数クラスを正しく識別した割合を、Specificity は少数クラスを正しく識別した割合を示す。

4.2 実験結果

表 2 に作成したデータの概要を示す。SMOTE においては、少数クラスを多数クラスと同数になるように水増しし、RUS では、多数クラスを少数クラスと同数になるように間引きしている。また提案手法における境界データの不均衡度の範囲は、0.4 以上 0.6 未満としている。

表 2　作成した実験データの概要

	元データ	SMOTE	RUS	提案
データ数	3500	6274	726	661.4
不均衡度	0.103	0.500	0.500	0.431

実験結果を表 3 に示す。表 3 の各値は 5 回分の平均値である。Sensitivity は多数クラスを正しく識別した割合を、Specificity は少数クラスを正しく識別した割合を示す。どちらの値も高いことが望ましいが、特に不均衡データで問題となる Specificity が重要となる。

表 3　実験結果

RF	元データ	SOMTE	RUS	提案手法
Sensitivity	0.99	0.97	0.94	0.98
Specificity	**0.81**	**0.88**	**0.93**	**0.83**
G-means	0.90	0.92	0.94	0.90
LR	元データ	SOMTE	RUS	提案手法
Sensitivity	0.99	0.95	0.94	0.98
Specificity	**0.82**	**0.92**	**0.93**	**0.83**
G-means	0.90	0.93	0.93	0.91
線形 SVM	元データ	SOMTE	RUS	提案手法
Sensitivity	0.99	0.95	0.94	0.98
Specificity	**0.82**	**0.92**	**0.92**	**0.84**
G-means	0.90	0.93	0.93	0.91
SVM	元データ	SOMTE	RUS	提案手法
Sensitivity	0.99	0.97	0.94	0.98
Specificity	**0.81**	**0.88**	**0.93**	**0.83**
G-means	0.90	0.92	0.94	0.90
KNN	元データ	SOMTE	RUS	提案手法
Sensitivity	0.99	0.93	0.91	0.97
Specificity	**0.68**	**0.88**	**0.93**	**0.74**
G-means	0.82	0.90	0.92	0.84
MLP	元データ	SOMTE	RUS	提案手法
Sensitivity	0.99	0.97	0.92	0.98
Specificity	**0.80**	**0.82**	**0.92**	**0.82**
G-means	0.89	0.90	0.92	0.90

不均衡データ処理を加えない場合（表中の元データ）では、機械学習手法に関係なく 0.99 と高い値となった。今回の元データは表 2 で示した通り不均衡度が 0.1 である。す

[1] https://scikit-learn.org/stable/

べての識別結果を多数クラスとしても、Sensitivity が 0.9 になることから、この値が高くなることは妥当である。一方、Specificity は、0.68 から 0.82 の値となり、Sensitivity よりも低い値となっている。この要因は不均衡データが影響していると考えられる。特に値が低い KNN は、アルゴリズムの特性上から、その影響が大きい。

SOMTE および RUS は、Sensitivity は、高い値を得ているが、元データと比較すれば、やや劣る結果となっている。一方で、Specificity は改善が見られ、本実験では、RUS が有用に働いている。結果、G-means も向上している。これらのことから、従来の不均衡データ処理は有用に働いていることが見て取れる。

一方、提案手法の値は、機械学習手法による違いもあるが、Sensitivity が 0.97 から 0.98、Specificity が 0.74 から 0.83 となった。Sensitivity は、SMOTE や RUS を上回るが、元データよりも低く、Specificity は逆に、SMOTE や RUS を下回り、元データよりも高い結果となっている。また G-means は元データより上回る場合もあるが、SMOTE や RUS よりも低い。

4.3 考察

今回の実験結果では、Specificity は、処理を加えなかった場合より、2 から 6 ポイントの向上が見られたが、G-means は 0 から 1 ポイントの向上に留まっている。G-means については、SMOTE や RUS よりも低い。これらのことから、本提案手法の優位性は確認できなかった。このような結果となる理由としては、次のことが考えられる。

第一に、境界クラスの抽出方法の問題が挙げられる。提案手法においては、多数データ・少数データともに、MD の範囲が同じものを取り出している。これは、同じ MD に偏るデータ間の識別率を上げることにより、全体の識別率をあげること、また、この部分が判断の難しい箇所であり、この点に明確な基準を設けることが必要と考えたからである。しかし、結果としては識別率向上にはつながっていない。これは、MD の範囲が同じため、KNN での値が低いことから予想されるように、データが近すぎ、識別が困難になったと考えられる。

第二に、学習データと検証データの関係がある。提案手法においては、学習データと検証データが同じ傾向にあると仮定している。仮にこれらのデータの特徴が異なる場合、提案手法では正しく分類することが困難である。この点は、他の手法でも同様の問題があるが、本提案手法は特に敏感といえる。

以上の点を踏まえ、今後の課題としては、データの抽出方法の再検討が挙げられる。現在同じ範囲としているが、識別率を上げるために、それぞれの平均±標準偏差の範囲から取り出すことを試みる。また、この際、不均衡度を考慮し、多数クラスの方を間引くことを行う。この点では、RUS などに近い形となる。

また、本実験では、1 つのパラメータ設定に基づくデータセットを作成したが、今後は様々なパラメータ設定に基づくデータセットを用いて提案手法を評価検証する。その結果をもとに、本提案手法を適用するためのデータセットの特徴を明らかにするとともに、適切なデータの抽出方法を検討する。さらに、実際の不均衡データを収集し、それに適用することで、本手法の有用性を評価する共に、特性を捉える。

一方、本提案手法では、各インスタンスが持つ特徴を、ベータ値や SN 値、MD など少数の特徴量に圧縮しているため、次元圧縮を行うことも可能である。MTS は、パターン認識の特性変数選択に SNR（信号対雑音比）実験計画を適用したものである。直交表と SNR ゲインを用いて特性変数の妥当性を検証し、影響力のある特性変数を選択することで、特性変数の最適化・簡略化を実現する。そのため、認識品質を落とすことなく、システム内の元特徴の数を最小化することを実現することが期待できる。Ting Mao ら[12]の研究において、その可能性が示されている。この点を踏まえて、改めて本提案手法を検討する。

5. おわりに

本論文では、MTS を用いた不均衡データ処理を提案した。本提案手法では、MTS によって求める MD を元に抽出したデータに対して、教師あり機械学習アルゴリズムを適用し、検知モデルを生成する。このことから、本手法は、機械学習アルゴリズムの観点からは、学習データセットを間引き手法の一種とみることができる。一方、間引き法では、学習データに含まれる全てのインスタンスを利用しないが、本提案では MTS において全インスタンスを利用する。

人工的な不均衡データを対象とした評価実験においては、本提案手法の優位性を示すまでには至らなかった。

今後の課題としては、様々なパラメータ設定に基づくデータセットを用いて評価実験の結果分析を通し、本提案手法の特徴を把握し、適切なデータ抽出方法を決定する、MTS を次元圧縮として利用することを検討することが挙げられる。これらを通し、MTS と教師あり機械学習の効果的な融合方法についての検討を進める。

参考文献

[1] N.Japkowicz,“Learning from Imbalanced Data Sets: A Comparison of Various Strategies”， AAAI Technical Report WS-00-05 ,(2000).
[2] 井手,“入門 機械学習による異常検知”, コロナ社 (2015)
[3] 田口,“タグチメソッドわが発想法”，経済界（1999）
[4] 立花,“品質工学（タグチメソッド）とは何か”,第 9 回横幹連合コンファレンス，B-1-1(2018)
[5] 佐野,黒木,“カーネル MT 法とその応用”,品質,Vol.42，No.1, pp.127-138(2012)
[6] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P.,“ SMOTE: synthetic minority over-sampling technique”, Journal of artificial intelligence research, 16:321-357. (2002)
[7] Han, H., Wang, W. Y., & Mao, B. H.,“ Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning ”, International conference on intelligent computing, p.878-887(2005)
[8] He, H., Bai, Y., Garcia, E. A., & Li, S. ADASYN:,“Adaptive synthetic sampling approach for imbalanced learning ”, IEEE International Joint Conference on Neural Networks, p.1322-1328(2008).
[9] 藤原,“スモールデータ解析と機械学習”,オーム社(2022)
[10] 永田,“ MT システムの諸性質と改良手法”, 応用統計学, 42 巻, 3 号, p. 93-119,(2013)
[11] Ting Mao, Lanting Yu, Yueyi Zhang, Li Zhou,“Modified Mahalanobis-Taguchi System based on proper orthogonal decomposition for high-dimensional-small-sample-size data classification”, Mathematical Biosciences and Engineering, 18(1):426-444(2021).