マハラノビス・タグチ法と最適化手法による
データ分析に関する研究

提出先　　大阪大学大学院情報科学研究科

提出年月　2023 年 1 月

村 田　真 一

発表論文リスト

学術論文

Shinichi Murata and Hiroshi Morita, “Feature Analysis Using Mahalanobis-Taguchi Method and Genetic Algorithm for Recorded TV Data”, International Journal of Innovative Computing, Information and Control, Vol. 18, No. 1, pp. 173-181, 2022.

Shinichi Murata and Hiroshi Morita, “Feature Analysis applying Clustering and Optimization Methods to Mahalanobis-Taguchi Method”, International Journal of Data Science, Accepted.

査読付き国際会議

Shinichi Murata and Hiroshi Morita, “Feature Analysis Using MT Method and Genetic Algorithm for Recorded TV Data”, 15th International Conference on Innovative Computing, Information and Control, Online, September 15-16, 2021.

口頭発表

村田真一, 森田浩, ”MT 法および最適化手法の適用による特徴分析”, 日本オペレーションズ・リサーチ学会 2021 年春季研究発表会＆シンポジウム, 2-B-1, 東京工業大学（オンライン）, 2021 年 3 月 1-3 日.

目次

第1章 序論

ICT 技術の進展により，IoT・AI・ビッグデータの利用環境が益々身近になり，データ分析において Amazon や Microsoft 等が提供するデータ分析サービスの活用が企業でも試行されている．また，ユーザーの操作履歴や使用履歴の分析による新たな製品開発の実施，購買データや Web 閲覧データの分析によるマーケティング，実需データの分析による SCM の最適化，商品やサービスの様々なメトリクスデータの取得，取得したデータによる品質管理の強化等，IoT・AI・ビッグデータは企業における様々な活動へ活用されはじめている．そして，そのスピード感や経営からの要請は更に強くなっていくと考えられる．加えて，データは今や経営の意思決定へ影響を及ぼす重要な要素の一つとなっており，企業は様々な方法で収集したビッグデータを AI やデータ分析の技術を用いて意味あるものにし，経営に役立てることが求められている[1].

このように多くの企業では IoT・AI・ビッグデータの利活用が試行されているが，一方で，上手く利活用できていないケースがある．そこには大きく 3 つの課題が存在している．

一つはデータの整備状況である．2 種類の正解データから学習を行い，その学習結果に沿って未知のデータに対して回帰や分類するという事が求められるが，企業の実際の現場にあるデータは製品本体性能の制約やデータ収集仕様の制約等により，データ分析で半ば前提となっている 2 種類の正解データが十分に揃わないことが多い．その結果，2 種類の正解データが不足している，もしくは準備できないといった，不完全なデータ整備状態の中で分析を実施することになる．すぐに仕組みを改善できればよいが，データ分析のために商品やサービスをスピーディにアップデートすることは

費用対効果の面から難しい．そのため，現場の実態は改善されず，分析に必要なデータの欠如，サンプル数の不足，データの不備が継続し発生している．実際にはこのような状況が多く見受けられ，Amazon の AWS や Microsoft の Azure 等が提供している一般的なデータ分析サービスを適用するには，1 種類の正解しかないデータを 2 種類の正解データへ変換する等，データ準備に多くの労力が必要になる．

二つ目の課題は，データが複雑化・高度化しており，特徴量の数も増大している点である．企業ではデータの取得を優先して実施してきた経緯がある．本来は何らかの目的を持った上で，それを達成するため，検証するためのデータを設計する．しかしながらデータ取得そのものを優先したため，特徴量が非常に多く，その中身についても非常に複雑になっている．そのため，該当データを適切に理解するには，データに関する深い業務知識を保有している必要がある．また，特徴選択へも影響がでている．特徴選択は有効かつ効率的なデータ分析を実施するために，データが保有する全ての特徴量を利用するのではなく，分析への寄与度が高い特徴量を上手く選択した上でデータ分析を実施していく方法論である．特徴量の増大により，データに関連する業務知識や該当データに対する知見やある程度のデータ分析スキルを有していないと，データが保有する特徴量の中からどの特徴量を利用し分析を実施するかの判断が難しくなっている．加えて，データ分析を適切に実施するには，適切なデータのグループを見極めた上で分析することも重要である．具体的には，動物というデータを，動物全体でまとめて分析するか，哺乳類などの種別に分割して分析するか，それとも更に細かい単位である犬等の種類へ分割し分析するのが適切かという判断である．データが複雑化・高度化していることにより，その意思決定が非常に難しくなってきている．多くの企業の現場では，データの中からどの特徴量を利用してデータ分析を実施するのか，データ分析を行うデータグループをどの単位にするのかという意思決定は，スキルや業務知識が異なる企業の業務担当者に任されている．そのため，同じ分析を行う場合においても，分析を行う担当者により，利用する特徴量や分析を行うデータグループが違うため，出力されるアウトプットが異なる．その結果，分析の品質は大き

く変動する可能性があり，経営者が意思決定に使用することが難しくなる．

最後の課題は，データ分析技術者の不足である． AI，機械学習，データ分析は，企業の R&D 部門からマーケティング部門，生産管理部門をはじめとする様々な部門や職能で使用されるテクノロジーになっている．しかしながら，多くの企業においては各部門・職能にデータ分析専門の担当者が配置されていることは少ない．一方でデータ分析業務においては，目的に応じ，データ分析に使うデータの単位や範囲の決定，データセットの中から分析に利用する特徴量の選択，目的に応じた分析アルゴリズムの選択等を適切に行うことが求められる．そして，これらを適切に実行するには高度なスキル，知識が要求されるため，誰もが簡単に実施できる業務ではない．また，企業においてデータ分析を実行する場合には，単純にサービサーが提供している AI を利用してブラックボックス的に結果をだすだけでなく，結果に対する説明責任を果たす必要がある．これらのことから，データ分析専門のスキルを備えた人材がいない場合，組織としてデータ分析を安定的かつ継続的に実施していくことが困難な状況になっている．

このように，IoT・AI・ビッグデータを用いて経営への貢献が要請される一方で，実際の現場ではデータの欠如，データの複雑化，データ分析技術者の不足等の複合的な要因のため，データ分析およびデータの利活用が進んでいない．そこで本論文ではこれらの企業のデータ分析課題の解消を目的とする．

本論文は以下の内容で構成されている．

まず，第1章では研究の背景および目的を述べる．第2章では提案の概要を述べる．提案するデータ分析手法の概要を説明し，提案手法の企業活動への適用方法を述べる．そして，提案手法や企業のデータ分析課題に対する関連研究を中心に説明し本論文の位置づけを明らかにする．第 3 章では提案手法を述べる．手法の詳細を記載するとともに，提案手法に利用する MT 法，最適化手法およびクラスタリング手法の概要と提案手法への適用について説明する． 第 4 章では提案手法を TV 録画データへ適用し検

証した結果を記載する．提案手法を用いた数値検証結果に加えて，他のデータ分析手法との比較も実施することにより提案手法の有効性を示す．第 5 章は本論文のまとめであり，各章のまとめと，本論文で得られた知見を整理するとともに，今後の課題および展開について述べる．

第2章 提案の概要

2.1 提案概要

本論文では，データ分析を行うための教師データの欠如，データの高度化・複雑化，データ分析人材の不足等の要因により，データ分析を業務や経営に活かせないという企業の課題解消を目的とする．そのため，以下の要件を具備したデータ分析手法を提案する．

【提案手法の概要】

- 企業が現在保有しているデータをそのまま活用しデータ分析を行う事ができる．
- 企業の現場すなわち専門家でもない一般的なスキル・知識しかもたない担当者でも簡単に実施できる．
- 人による属人性を排除し，業務品質を標準化，高位平準化する．

また，提案手法の特徴は以下の通りであり，より企業・ビジネスの現場へ適用しやすい形を目指す．

【提案手法の特徴】

- 分析目的に応じたデータ分析アルゴリズムの選択，アルゴリズム毎に求められるパラメーター設定，分析目的に応じたデータ分析グループの判断，効果的・効率的な分析に必要となる特徴選択といった，データ分析における一連の作業に求められる高度な専門知識や作業を極力排除．

- 企業が簡単には入手できない場合も多い 2 種類の正解データだけを用いるのではなく，主に 1 種類の正解データを利用し分析を行うことが可能.
- どの特徴量を利用したか確認可能であり，他者へ伝える際にある程度の説明責任を果たすことが可能.

企業では様々な経営活動が行われているが，その中の重要な取り組みの一つに優良会員の増加がある．企業には顧客が存在しており，企業ではその顧客を会員として管理している．会員の中にも区分を設けており，企業の様々な企画やイベントへの参加，有料プランへの加入等を行ってくれる優良会員という区分が存在している．企業としては優良会員を増やしていくことが，事業を永続させ，価値を届け続けるために重要な事の一つである．そこで，本論文ではこの優良会員を増やすために，提案手法を用いてアプローチを行う.

提案概要は次の通りである.

【用語の定義】

- 優良会員が保持するデータ群(行動履歴，購買履歴等)を優良会員データとする.
- 通常会員が保持するデータ群(行動履歴，購買履歴等)を通常会員データとする.

【提案概要】

1. 優良会員データを分析し特徴を抽出.
2. 通常会員データを分析し特徴を抽出.
3. 算出された優良会員の特徴と通常会員の特徴を比較.
4. 通常会員の中から優良会員が保持する特徴と同じ特徴を保有する会員を抽出する.

上記の一連の流れを，提案手法を用いて実施する．

企業では通常会員は以下の 2 つの要素で成り立っていると捉えており，いずれであるのかを判断する事が非常に重要である．

【通常会員の特徴】

- 明確に優良会員へなる意思がないために通常会員である．
- 優良会員へのきっかけがないだけで優良会員へなる意思がないわけではない．

提案概要の具体的イメージと実際の企業での活用例を記載する．本論文では提案手法を TV 録画データへ適用し検証を実施している（ 優良会員および通常会員データとして TV 録画データを利用)．TV 録画データの例であれば，まず，優良会員の TV 録画データを分析し特徴を抽出する．次に通常会員の TV 録画データを分析し特徴を抽出する．その結果をマッチングさせることで，通常会員グループの中から，優良会員に属している人と同じ属性を持った人を抽出することが可能となる．そして，抽出された通常会員に対して，イベントやキャンペーンへの参加依頼や有料プランへの加入依頼を行う事で，より効果的・効率的なマーケティング活動を実施することが可能となり，優良会員の増加に繋げることができる．

2.2 関連研究との関係と位置づけ

ここでは提案手法に関連する研究や企業が抱える問題・課題に対する研究を述べ，本論文の位置づけを明らかにする．

顧客の行動，特徴，属性を抽出する研究や抽出した情報に基づき予測をする研究は過去から多く実施されている[2, 3, 4]．レコメンデーションやパーソナライゼーション技術としてビジネスへの応用も多く行われている．Amazon の過去の購買履歴や閲覧履歴から商品をレコメンデーションする技術が代表的である．これらは情報推薦・情報フィルタリングの研究として行われていた．その後，協調フィルタリングの考え方が提案された．また，情報フィルタリングを一種の文書分類問題と捉えることによる，様々な機会学習アルゴリズムの適用が行われるようになった．そして，精度向上等，より実用的な技術として研究が進みビジネスの現場でも広く利用される技術となった．

昨今，顧客の行動，特徴，属性の抽出，レコメンデーション，データ分析等の幅広い分野に深く関係しているのが機械学習に関する研究である．機械学習は大きく 3 つに分類されている．一つ目は教師あり学習，二つ目は教師なし学習，三つ目は強化学習である．

一つ目の教師あり学習については，教師データすなわち正しい答えが事前に与えられる．教師あり学習ではこの教師データを学習に利用し，未知の情報に対して判断することができる回帰モデルや分類モデルを構築する．

回帰モデルは与えられた教師データを利用し，主に値を予測する目的で利用される．主な手法としては線形回帰，リッジ回帰[5]，ラッソ回帰[6]が挙げられる．線形回帰は説明変数（独立変数）を用い、連続的な従属変数の予測を行うことであり，それを可能とする最適な直線を見つける事が求められる．主な用途としては予測が挙げられる．売上高に対する来店客数，世帯数，広告費等との関連を明らかにすることで，企業として宣伝費をどの程度増やせば売上がどの程度上がるという売上を予測すること

が可能となる．線形回帰では説明変数が増えると過学習を起こす可能性が高くなる．その過学習を抑えるために正則化項の概念を取り入れたのがリッジ回帰，ラッソ回帰である．いずれの場合においても回帰モデルは数値予測が主な用途になっている．企業の活動における数値予測を行う場合には非常に有用な手法群である．一方，提案概要で述べた様な通常会員の中から優良会員と同じ特徴を有したデータを抽出するといった用途には向かない．

分類モデルは回帰とは異なり種類やカテゴリが異なる複数のデータを区分できる境界線を求める事が目的で利用される．与えられた教師データを利用し学習することで 2 項分類や多項分類を行う．多数の方法が提案されており，主な手法としてはロジスティックス回帰(LT)[7]，決定木(DT)[8]，サポートベクター―マシン(SVM)[9]，ニューラルネットワーク(NN)[10, 11]等が存在している．分類を目的とするこれらの手法では予測する変数が所属カテゴリ等になる．具体的な用途としては，動物の画像を入力データとし猫か犬かを識別する，着信メールがスパムメールか正常メールかを判断する等に活用される．企業においては顧客の購買情報から該当顧客が新製品を購買するか否かを分類したり，顧客情報からその顧客の評価を分類したりすることが挙げられる．

これらの教師あり学習に共通している内容として，正解データが必要になるという点である．学習させるために，事前に猫という正解ラベルとその情報，犬という正解ラベルとその情報というように，最低 2 種類の正解データを用いて学習させる事が必要となる．正解データを潤沢に取得できる場合や準備できている場合においては非常に効果的である．一方で，企業の現場においては正解データを潤沢に準備できない場合も多く発生しており，その場合においては適用が困難である．

教師あり学習に対し，2 種類の正解データを必要としない教師なし学習という手法が存在している．教師なし学習では正解ラベルがない状態での分類になるため，何かしらの基準に基づきデータを分類もしくはグループ化していく事が主な目的となる．主にクラスター分析と次元の削減にわけられる．

クラスター分析は大量のデータから，特徴が近いデータを集め，それらを集団に分類する分析手法である．これにより，データの特性や共通項の把握が可能となり，大量のデータを扱いやすくすることができる．企業においては多くの顧客を分類し，マーケティング施策を検討する際に利用される．主な手法として k-means 法[12]，混合ガウス分布[13]等が挙げられる．

次元の削減は，大量の説明変数から，それらの変数を組み合わせた変数やより少ない指標へ要約する手法である．代表的な方法に主成分分析[14]が挙げられる．主成分分析は大量の変数を要約し主成分と呼ばれる合成変数へ次元を削減する．これによって，元のデータそのままでは非常に多くの変数が存在しており理解しにくい情報を，データの持つ情報をできる限り損なわず可視化し，理解しやすくする事が可能である．

クラスター分析および主成分分析等の次元圧縮は分類を実施することは可能であるが，分類されたそれぞれのグループが持つ意味までを学習により明らかにすることはできない．そのため，企業等で利活用する場合においては，学習により分類された結果に対して，人が意味づけを実施する必要がある．企業で意味づけを適切に行うには該当業務に関する専門的な知識やスキルを有している必要があるが，そのようなデータ分析にも業務にも精通している人材が少ないのが現状である．そのため適切な意味づけをすることは実際には難しく，これらの手法をそのまま適用することはハードルが高い．

一方で教師なし学習は人の目では判断できない大量のデータやグループに対し分類を行う事が可能であり，異常なデータの発見等に役立てる事ができる．異常なデータの発見は主に異常検知の分野において多く研究されている． 2 種類の正解データ(教師データ)が揃っていることが少ない状況でデータ分析を実施しないといけない異常検知においては，1 クラスのみのデータセットを含む少数のデータセットのみを使用してデータを分析する方法を用いる事が多い．製造業を例に見ても不良品の異常検知を行う場合，不良品はそもそも高い確率で発生するものではないため，異常データつまり不良品データを十分に収集すること自体が難しい状況である．そのため教師なし異常

検知はそのような状況に対して非常に重要な手法のひとつである．

異常検知は蓄積された大多数のデータと比較し，振る舞いが異なるデータを検知する技術である．この技術はビジネスへの応用も広がっており，クレジットカードの不正利用，システムダウン，異常行動検知等様々な分野で実装が試行されている．

企業のデータ分析においても，2 種類の正解データが揃っている事は少なく，限られたデータを用いて分析することが求められるため，これらの異常検知等で用いられているデータ分析手法を上手く活用することは有効であると考える．

具体的な手法としては，k 最近傍法（k-Nearest Neighbor：k-NN）[15]，局所外れ値因子（Local Outlier Factor：LOF）[16]および 1 クラス SVM（One-Class Support Vector Machine：OCSVM）[17]がある．

k-NN は回帰および分類が可能であり，特徴量空間にデータをプロットしたときに，プロットと距離が近い学習データを近いものから順に k 個選び出し，それらの学習データのクラスを多数決することでデータのクラスを決定する分類モデルである．最近傍法の場合は，他クラスの鋳型に囲まれる孤立点ができるが，k-NN 法では，多数決型により決定するため，異常値を許容し孤立点が少なくなる．k-NN の特徴として計算量が多いことが挙げられ，低減のための方法もいくつか提案されている．一つは k-NN で誤って識別されたデータを削除する方法である．削除後はデータセットが変化するため，再帰的に繰り返していくことも可能である．二つ目は識別境界から遠いデータは重要度が低いと考えられるため分析対象から消していくという方法である．計算量が多いという特徴からトレーニングデータ数や特徴量が多い場合には予測が遅くなるため，一般的にはデータの中に大量の特徴量を有する高次元データには向かない．加えて，k の値によりパフォーマンスが変わるため，最適な k の値をデータ分析者が探し設定する必要がある．

LOF は空間におけるデータの密度に着目する方法である．近傍 k 個の点といかに密かであるかを表す局所密度（Local density）を推定する．対象データの局所密度と近傍点の局所密度が等しいほど正常データであり，その差が大きいほど異常データ（外

れ値）である可能性が高いと解釈する．これを定式化し，ある点が外れ値である可能性を表す外れ値スコアを算出し，該当のデータが正常データなのか異常データなのかを判別する． この外れ値スコアは，大きい値をとるほど外れ値である可能性が高いということを表しており，異常検知では，このような外れ値スコアに閾値を設け，それを超えるものを異常値と判断する．ただし，実際に LOF による異常検知を行う場合にはこのスコアに対する閾値設定を慎重に行う必要があり，難しさの一つである．

OCSVM は SVM を領域推定問題に応用した手法であり，異常値検出やデータ密度を推定することが可能である．データ密度により，モデルの適用範囲・適用領域を設定することもできる．従来の SVM では分析に 2 つのクラスのデータ（2 種類の正解データ）が必要であったが，OCSVM では 1 クラス（1 種類の正解データ）だけのデータで分析が可能である．また，SVM は教師あり学習なのに対して，OCSVM は教師なし学習に分類されるため，教師データは不要である．OCSVM は正常値データをグループ 1，原点のみをグループ 2 とし，SVM と同様にカーネル法で高次元の空間に変換する．このとき，カーネル法により都合の良い空間に変換されており，グループ 1 は原点（グループ 2）から離れたところに位置する．そして，SVM のマージン最大化と同様に，原点からの距離が最大となる境界を求める．そこに異常値データを入力すると，そのデータは変換された空間内で原点に近いところに集まる．これにより原点に近い（境界の原点側にある）データを異常値データとみなすことができる．ただし，カーネルに何を用いるか（デフォルトは RBF（Radial Basis Function：放射基底関数））やデータセットに占める外れ値の割合の上限等をパラメーターとして閾値設定をする必要があるため，データ分析に精通した人材でないと難しい．また，数多くの実験を実施しないと適切な値を見極めるのが困難である．

LOF と OCSVM は，複雑な分布を持つデータの外れ値を検出するのに特に効果的である．ただし，企業において実際に適用する際には個々の分析に応じた高度で正確な閾値とパラメーターの設定が必要になる．この閾値設定やパラメーター設定は簡単に設定できるものではなく，幾度も試行を重ね最適解を導き出していく必要があり，企

業の様々な部門において簡単に導入できるものではない.

その他の手法に品質工学の分野で利用されているパターン認識・予測の手法であるマハラノビス・タグチ法(MT 法) [18]がある．日本の工学者である田口玄一氏によって提唱された手法であり，品質工学会を中心に現在も発展を続けており，異常検知や健康診断等の分野へ応用されている．判別分析を考える際，多くの手法が母集団に対等な二群を想定している． MT 法では正常と異常という概念で判別し，正常はある一つの群をなすが、異常は群をなさないという考え方を行う．そのため，正常群のみのデータを利用してモデルを構築する．そして，正常群を基準とし、基準からのパターンの相違を距離として表現する。したがって，１クラス（1 種類の正解データ）だけのデータで分析(モデルの構築)が可能である．ただし，MT 法を用いる場合も，２クラスのデータを保有している場合においては，２クラスのデータを利用し，構築したモデルの妥当性検証や特徴選択を実施することが望ましい．MT 法におけるパラメーター設定については複雑性が低い．正常か異常かを判定するために基準からの距離の閾値を設定する必要があるが， 4 という数字が品質工学等の分野で一般的であり多くの実績がある.

企業が保有する 1 クラスのデータの利活用や一般的なスキル・知識しかもたない担当者でも簡単に実施できることを目指す場合，MT 法は企業の実態に即した非常に適用しやすい手法である.

最後の強化学習はシステムが試行錯誤を繰り返して、適切な制御方法を学習していく技術である．教師データとしてデータと正解をセットで準備し与えるのではなく，コンピュータが目的として設定された報酬を最大化するための行動を学習していくものである．具体的な手法としては Q-Learning[19]や DQN[20]が存在している．また，実際の活用例として、囲碁AI，将棋AI等のゲーム関連，自動運転制御やロボット制御等が挙げられる。

企業においてはある明確な目的が定まっており，それを達成するために状況に応じて対応を最適化していく場合には適するが，他者への説明のしやすさや試行錯誤しな

がらデータの分析を進める場合には向いていない.

データが高度化・複雑化しているという課題に対しては，様々なアプローチが考えられる．上述した教師なし学習の一つであるクラスター分析を活用するのが一つの方法である．データ分析をする際，適切な対象グループを見極める方法について，データセットを適切な空間（グループ）に分割するクラスタリング手法が応用できる．クラスタリングはデータの類似度に基づき，データを分類する手法である．クラスタリングにより作成された類似度が高いデータ同士の集団をクラスターと呼ぶ．クラスタリングは，データの分類の仕方により大きく階層的クラスタリングと非階層的クラスタリングが存在している.

階層的クラスタリングは，データ間の類似度が近いものからまとめていく（凝集型階層的クラスタリング)，あるいは遠いものから離していく（分割型階層的クラスタリング）が存在している．あらかじめクラスター数を与える必要がないが，クラスタリングの対象数が多い場合には複雑になり計算や結果の解釈ができない可能性がある.

一方，非階層的クラスタリングとは，分類の良さを評価する関数を定義し，反復的に計算することで，その関数が最適となる分類を探索するものである．クラスタリングの対象数が多い場合においても活用可能である特徴がある．一方であらかじめクラスター数を与える必要がある.

階層的クラスタリングのための k-means 法などは，分析に適したデータグループを形成する手法として研究されており，学術研究と実社会の両方に適用されている．しかしながら，k-means 法を使用する場合，クラスター分割数を指定するという難しさが存在し，この部分については人間の判断に委ねられている．そのため，企業へ適用する際には，データ分析を実施する担当者の知識と能力に依存することになり，業務品質が均一化されない問題がある.

しかし，近年では，クラスターの数をデータ分析者が指定する必要がなく，分割数を自動的に決定できるアプローチが研究されている．具体的な方法としては，x-means 法[21]および g-means 法[22]がある．分割数を自動判定できるクラスタリング

手法の開発により，データ分析者はどのデータグループで分析を行うかを自身のスキルや経験によって決める必要がないため，一定レベルの品質を維持することができる．

データの高度化・複雑化やデータ分析人材の不足といった課題の中で特にデータの深い知識を必要とする特徴選択に対するアプローチとしては，主に機械学習，深層学習，AI の分野で研究が行われてきている．データ分析をする際の有益な特徴量を判断するための手法としては主に以下が存在している．

- Filter Method

  統計手法を用いて各特徴量に対してデータ分析に対する寄与度をスコア化するものである．算出されたスコアで各特徴量にランクを付与し，データ分析に利用するか否かを決定していく手法である．スコアの決定方法として特徴量と分析対象の関係性を考慮して決定する方法や特徴量だけに着目し統計的に決定する方法も存在する．この手法の特徴として複数の特徴量による寄与度は考慮されないことが挙げられる．主な手法としてはカイ二乗検定（Chi-Square）や ANOVA（Analytics of Variance）が挙げられる．

- Wrapper Method

  複数の特徴量を組み合わせながらデータ分析・予測精度の検証を行い，その精度が最も高くなるような特徴量の組み合わせを見つける手法である．様々な特徴量を使って予測を繰り返し実施し，精度が高くなる特徴量へ絞り込みを実施していくアプローチである．そのため，この手法では一般的に計算の負荷がかなり高くなる傾向になる．分析対象データが少ない場合には問題にならないが，超大量データを分析する際などはこの手法を用いると特徴量を選択するだけで膨大な時間が必要となる．主な計算方法は前進法（Forward Search）と後退法（Backward Elimination）である．両者の違いは特徴量をすべて除いた状態から特徴量を加えていくか，すべての特徴量を使用した状態から特徴量を除いていくかという部分になる．

■ Embedded Method

主に機械学習モデルを適用する手法であり，機械学習モデルにより学習の一部として特徴量の選択を実施するものである．学習の一部として特徴選択も含まれているため，学習と特徴選択を同時に実施できることが特徴である．利用される主なアルゴリズムとしてはラッソ回帰や DT などが挙げられる．

これまでに述べてきた研究に加え，各手法を組み合わせた研究も多く行われている．各研究手法を単独で用いるのではなく，手法を組み合わせて利用することで課題を解決しようとするアプローチである．

一つ目は特徴選択に関する組み合わせ研究である．最適化手法を用いて特徴選択を実施するというアプローチは過去から行われており，最適化手法の一つである遺伝的アルゴリズム（Genetic Algorithms： GA）を用いた特徴選択手法[23]等がある．近年ではデータ分類手法である SVM および k-NN の特徴選択に加えてパフォーマンスを最適化手法により評価することで，より少ない特徴量でより高い分類精度を実現する試みがなされている [24, 25]．その他，最適化手法の Particle Swarm Optimization(PSO)[26]を利用し多目的最適化問題を定義し，利用する特徴量の数を最小化するとともに分類性能の向上を実現させている事例もある[27]．また，進化的計算手法を用いて，今までにはない新たな特徴選択アルゴリズムを提案しているものもある[28]．他にはクラスタリング手法を応用し特徴量の抽出を行う方法もある[29]．

MT 法に関連する組み合わせ研究も行われている．MT 法における特徴選択には通常，直交表を用いるが，直交表ではなく主成分分析を用いた特徴選択を実施する方法が提案されている[30]．また，ノイズを伴ったデータに対し新たな解析プロセスを追加した上で MT 法を適用する方法も議論されている[31]．その他にも，MT 法に機械学習を用いて分析の精度向上を目指すアプローチ[32]，多様な認識性能を持つ単位空間を作成し，特徴量の設計を変更することで，精度向上を目指す方法[33]，MT 法を超高次元データへも対応可能とした手法[34]等がある．

二つ目は各手法を実行する際のパラメーター設定に関するものである．パラメーター設定は有効な分析を行う際に非常に重要な要素である一方で，実際の現場においてはスキル，知識，時間的な制約によって最適なパラメーター値を見つけることが難しい．これらの課題を解消するためにパラメーター設定に関する様々な方法が研究されている．具体的にはOCSVMのパラメーターをPSOによって最適化し性能を向上させる方法[35]．流出モデルのパラメーターの最適化をPSOで行うもの[36]．適切なパラメーター設定を得るために機械学習の分野で研究されているアンサンブル法を適用した方法もある[37]．特徴選択と分析手法のパラメーターの両方を効果的なものにするハイパーパラメーター設定を人口蜂コロニーアルゴリズムで解決しようとするアプローチ[38]．パラメーターを最適化するのではなく，パラメーター自体を不要にするという方法も考えられている[39, 40]．また，すべてを自動で行うのではなく1回の実験データを用いてパラメーターチューニングを最適化していく方法もある[41]．

このように各論の課題に対し様々な研究・アプローチが行われ成果があがっている．また，手法を組み合わせたアプローチも多く提案されている．

しかしながら，そのほとんどが，データ分析の精度や効率向上を目的としているものである．もしくは，特徴選択の最適化や，データ分析手法に対するパラメーターの最適化等，データ分析における一部分の処理を対象にしているものである．そのため，様々な企業の現場で誰もが簡単に利用できること，誰が実施しても同じ業務品質を担保できること，データ分析業務におけるより多くの処理を自動化することを目的としていない．結果として，企業にて実際にデータ分析を業務として実施するには，特定の目的やニーズに応じて，上述した各手法の中から，データ分析アルゴリズムの選択，適切なパラメーターの調整，最適な特徴量の選択を適切に実施する必要がある．組み合わせによる提案を採用する場合も，一つの組み合わせ手法だけでなく，複数の組み合わせ手法を用いて補完する必要がある．

このような状況を踏まえた上で，本論文の目的である，「企業が保有するデータをそのまま活用する」，「一般的なスキル・知識しかもたない担当者でも簡単に実施できる」，

「人による属人性を排除し業務品質を標準化する」に合う手法を考え，具体的には以下とした．

データ分析の方法は MT 法を用いる．理由としては 1 つの正解データで分析を実施することができ，複雑なパラメーター設定を必要としないためである．2 種類の正解データが必要な手法や複雑パラメーター設定が必要な手法は現場への適用や長期的な運用が困難である．

データ分析における特徴選択においては最適化手法を用いる．特徴選択を自動化でき，属人性の排除と業務品質の標準化を実現できるため，最適化手法を適用することが一番望ましいと考えた．

データ分析に利用するデータグループの判別にはデータ分割数をアルゴリズム側で判断するクラスタリング手法を用いる．理由はデータグループを自動的に分割することで業務の標準化が可能になるからである．一般的なクラスタリング手法ではデータ分析者がデータ分割数を指定する必要があり，人のスキルや知識に依存する．

ＭＴ法，最適化手法，クラスタリング手法を組み合わせることで，本論文では企業のデータ分析業務における，分析目的に応じたデータ分析アルゴリズムの選択，アルゴリズム毎のパラメーター設定，データ分析グループの判断，特徴選択といった，データ分析における一連の作業に求められる高度な専門知識や作業を極力排除する．加えて，企業が保有している 1 クラスデータによる分析を可能にし，誰が実施しても同じ結果となる業務の標準化・高位平準化を目指す．

第3章 提案手法の概要

3.1 MT 法と最適化手法を用いた分析

本論文ではデータ分析に MT 法[42, 43]を用いる．MT 法では，判断の基準となる単位空間データを 1 つのクラスのデータ（データの中に正解ラベルが 1 つしかない）で作成する．そのため企業の現場において 2 クラスのデータを揃える事が困難な場合や２クラスのデータが非常に少ない場合においても適用する事が可能である．また，複雑なパラメーター設定の必要がないデータ分析を実行できる．

基本的な MT 法では，特徴選択に直交表を利用する．しかしながら，この手順は特徴量が増えると難しい．また，システム上のオンラインではなくシステム外のオフラインで作業する場合もあるため手間もかかる．そして，これらの作業にはある程度の専門知識が必要になる．そこで特徴選択の複雑さを最小限に抑えるために，機械学習における特徴選択等で利用される最適化手法を MT 法へ適用する．これにより，特徴量の自動選択を実現するとともに，どの特徴量を利用したかを可視化することで，ある程度の説明を可能にする方法を提案する．

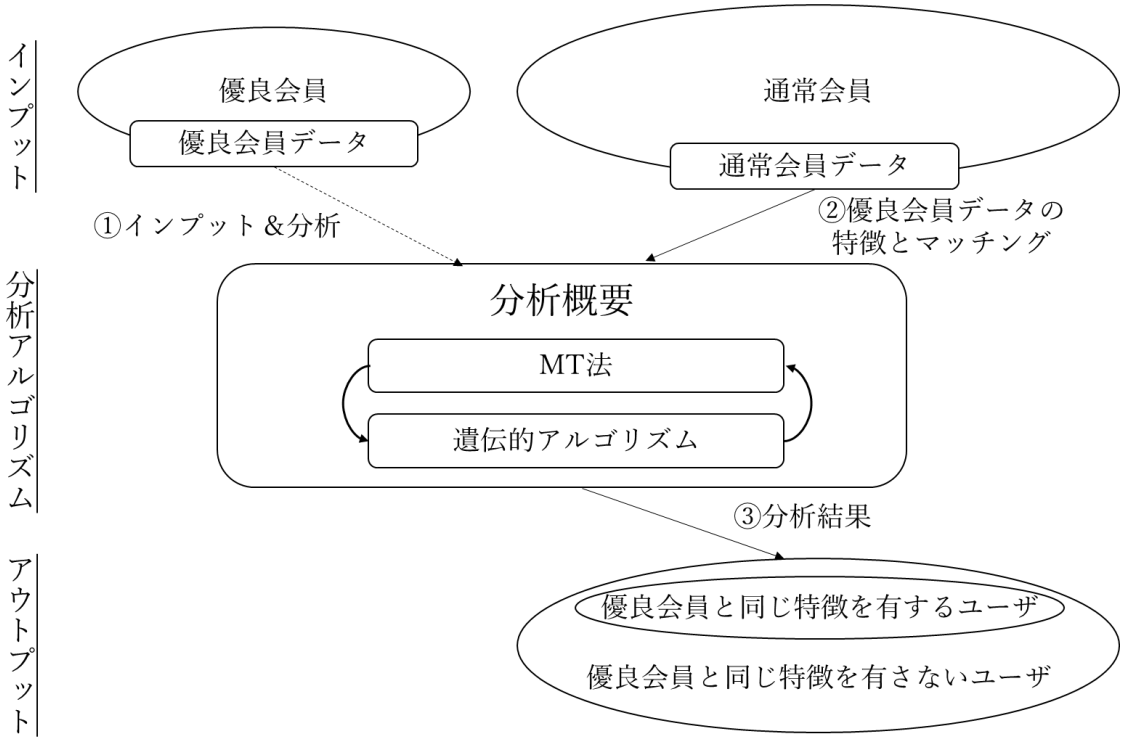

図 1　MT 法と最適化手法を用いた分析の概要

提案手法の概要を図 1 に示す．最初にインプットデータである優良会員データに対し最適化手法の一つである GA[44, 45]を用いてデータ分析に利用する特徴量を自動的に選択する．

次に選択された特徴量に対し MT 法を適用しマハラノビス距離[46]による判定の基準となる単位空間を形成しモデルを構築する．その後，このモデルの評価を行う．モデルの評価には検定データを与え，その検定データに対する優良会員か通常会員かの正解率で評価する．そして，この特徴選択～単位空間作成～モデル評価のサイクルを繰り返し実施し，評価結果に基づき最終的に利用する特徴量を決定する．次に，通常会員データに対して MT 法によりマハラノビス距離を算出し，優良会員データからなる単位空間との距離を判定する事で通常会員の中から優良会員と同じ特徴を有しているデータを抽出する仕組みを構築する．

具体的な処理フローを図 2 に示す．

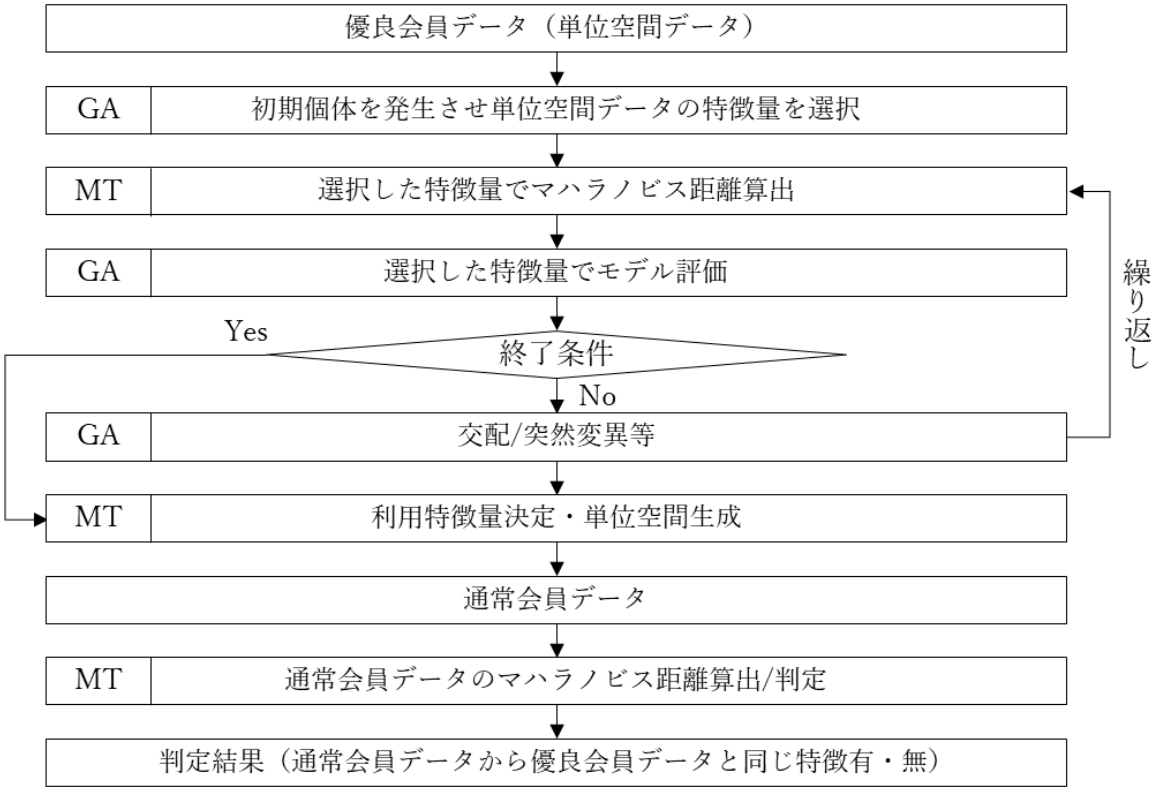

図 2　MT 法と最適化手法を用いた分析の流れ

3.1.1 MT 法の概要と提案手法への適用

MT 法[47, 48, 49]は，マハラノビス距離に基づいて異常値を決定する方法である．この方法は田口源一博士によって開発された調査データから個体を判定する手法であり，診断，予測，パターン認識，検査における判定等，幅広い用途に応用が可能である．これまで主に品質工学[50]および異常検出[51, 52]において，MT 関連の広範な研究が行われている．

異常検知分野では，ある個体（状態）があるとき，その個体が正常な個体の集団の中心からどの程度ずれているかで判定する方法，あるいは，正常の境界の中にいるか外にいるかで判定する方法をとる．いずれにおいても過去のデータ等から経験的・統

計的に判定する場合が多く，異常検知の分野では大きく 3 つの課題が存在している．

一つ目は正常と異常の判断を行うにあたり，どのような数値を用いるのか，二つ目は正常と異常の境界線をどのように定めるのか，三つ目は正常と異常を判断する境界線における閾値等の値をどのように設定するかという課題である．

一つ目のどのような数値を用いるのかという課題については，世の中のデータの特徴量が増えてきており，単体の指標の数値のみを用いて異常検知等を実施することは事実上困難な状況になっているのが現状である．二つ目，三つ目の正常と異常の境界線や判断に用いる閾値についても，データの特徴量が増えるだけでなく，複雑化している実態を鑑みると，過去から蓄積されたノウハウや閾値を適切にチューニングできる高度なスキルが必要である．そのため広く一般に利用する方法としては難易度が高いと考えられる．MT 法はこのような課題をある程度解消し，実際の異常検知の現場で使える技術である．

MT 法ではマハラノビス距離を用いるが，マハラノビス距離はインドの数学者により考案された，いくつかの特性間で相関関係を持つ多次元（多変量）空間中の距離を測定する方法である．MT 法はこのマハラノビス距離を使用して異常検知を実施する手法である． MT 法においては，正常・異常を判定する問題であれば，正常な集団のデータのみを用いて相関係数行列を求め，マハラノビス距離を算出している．このように MT 法ではデータ分析の基礎データとして一つの集団のデータで実施することが可能である点が大きなポイントである． 2 種類の正解データが取得できていない場合や異常データの取得が非常に難しい場合，取得できたとしても潤沢なデータ数が揃っていない・揃わない場合において非常に有効な手法である．

また，MT 法はデータ分析における重要なポイントである単位空間の作り方（適切なデータ分析対象グループ）や特徴選択（項目選択）の方法やその評価についても包含した手法である．

特徴選択はデータの特徴量をすべて用いて分析を実施するのではなく，結果に影響しない特徴量や，結果への影響が小さい特徴量，また結果に悪影響を与える可能性の

ある特徴量を特定して除外しながら，分析に有効な特徴量を特定していく技術である．MT 法においては異常な個体はその異常の程度が大きいほど正常な集団から求めたマハラノビス距離平均との距離Dが大きくなる．そこで，横軸に異常の程度をとり，縦軸にD^2/kをとれば，図 3 のようになる．

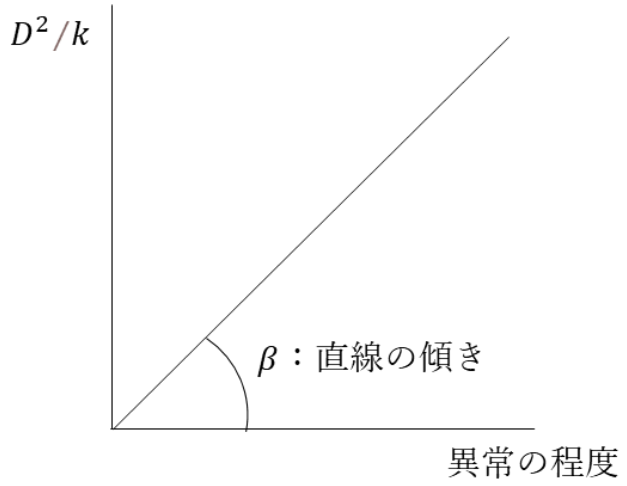

図 3　マハラノビス距離と異常の程度

母集団のマハラノビス距離の 2 乗であるΔ^2の期待値$E(\Delta^2)$が式(1)のようにデータの特徴量であるkに依存するため，kで割り正常集団D^2の平均を 1 に正規化する．

$$E(\Delta^2) = k \tag{1}$$

ここで異常なデータが入力されたと仮定すると，算出される結果はD^2/kである．そこで，式(2)のように，異常の程度に対するD^2/kのＳＮ比を考える．

$$\text{SN 比 } \eta\text{（デシベル）} = 10\log\frac{\beta^2}{\sigma^2} \tag{2}$$

ここで，βは図 3 直線の傾き，σ^2は各点の直線からの残差の分散である．ここでのＳＮ比はノイズに対する強さではなく，直線性の良さと感度の大きさを意味する．

つまり，D^2/kのＳＮ比を大きくできれば異常の判定精度が上がるため，このＳＮ比を大きくする特徴量を見つけるために，MT 法では直交表を用いて分析に用いる特徴

量の有効性を検証する．これは MT 法で用いる各特徴量を 2 水準系の直交表へ割り付け，分析を実施し SN 比に対する要因効果図を作成して，各特徴量が分析に与える影響度を測定するものである．例えば，2 水準系の直交表へ割り付けについては，直交表の第 1 水準は、「この特徴量を単位空間作成に使用する」、第 2 水準は「この特徴量を単位空間作成に使用しない」とする．

ここで，その特徴量を使うあるいは使わないの意味は，単にマハラノビス距離を計算するときにそのデータを除外するのではなく，相関係数行列の計算の前から除外するという意味である．表 1 に，特徴量が 11 個あるときに，L_{12}という直交表に特徴量を割り付ける場合を示す．

表 1　直交表$\boldsymbol{L_{12}}$を使用した特徴選択

	x_1 1	x_2 2	x_3 3	x_4 4	x_5 5	x_6 6	x_7 7	x_8 8	x_9 9	x_{10} 10	x_{11} 11	判定に使用する特徴量
1	1	1	1	1	1	1	1	1	1	1	1	すべての特徴量を使う
2	1	1	1	1	1	2	2	2	2	2	2	x_1，x_2，x_3，x_4，x_5
3	1	1	3	3	3	1	1	1	2	2	2	x_1，x_2，x_6，x_7，x_8
4	1	2	2	2	2	1	2	2	1	1	2	x_1，x_6，x_9，x_{10}
5	1	2	1	1	2	2	1	2	1	2	1	x_1，x_3，x_4，x_7，x_9，x_{11}
6	1	2	2	2	1	2	2	1	2	1	1	x_1，x_5，x_8，x_{10}，x_{11}
7	2	1	2	2	1	1	2	2	1	2	1	x_2，x_5，x_6，x_9，x_{11}
8	2	1	1	1	2	2	2	1	1	1	2	x_2，x_3，x_4，x_8，x_9，x_{10}
9	2	1	2	2	2	2	1	2	2	1	1	x_2，x_7，x_{10}，x_{11}
10	2	2	1	1	1	1	1	2	2	1	2	x_3，x_4，x_5，x_6，x_7，x_{10}
11	2	2	2	2	1	2	1	1	1	2	2	x_5，x_7，x_8，x_9
12	2	2	1	1	2	1	2	1	2	2	1	x_3，x_4，x_6，x_8，x_{11}

直交表で計画した内容に沿って，順に単位空間を作成し，作成した単位空間に対して異常なデータを用いて異常データのマハラノビス距離を求める．算出されたマハラノビス距離を用いて，SN 比を算出し，第 1 水準（その特徴量を使う）の方が S N比が高い特徴量を利用してマハラノビス距離を求めるのがよいと考えられる．一方でSN比の算出では，異常データを用いてマハラノビス距離を算出するため，異常データの取得が難しい場合には実施することが難しいという点が存在している．

このように MT 法における特徴選択は効果的かつ詳細な手法が確立されている．しかし，この作業を実行するには，ある程度の知識・スキル・専門性がないと実際の現

場で実行するのは難しい内容となっており，一般的な知識やスキルしか保有していない担当者が同じ業務品質で簡単に実施できる方法ではないのが現状である．

次に MT 法全体の解析手順[53]と図 4 に MT 法による判定のイメージを示す．

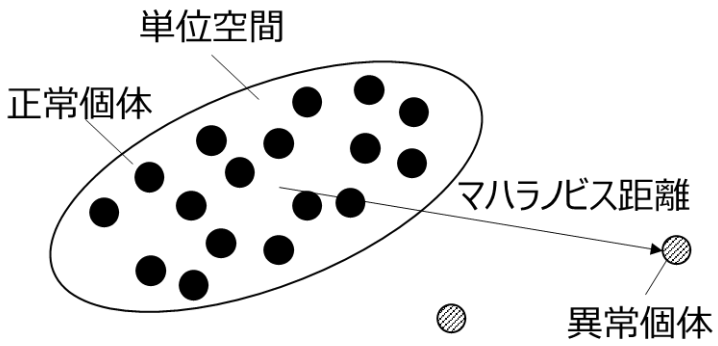

図 4　MT 法による判定イメージ

1. 単位空間（正常）データを決め，単位空間データに対して基準化を行う．
2. 基準化されたデータで相関係数行列，逆行列を算出する．
3. 基準化されたデータでマハラノビス距離を算出し閾値を決める．一般的な閾値として 4 が用いられる．
4. 異常なデータから直交表を用いて特徴選択を実施する．
5. 選択した特徴量を用いて新たなデータについてマハラノビス距離を算出し閾値に基づき判定を実施する．

具体的な計算方法は，単位空間に属するm次元のn個のデータ$\mathrm{X} = \{X_{m1}, X_{m2}, \cdots, X_{mn}\}$の平均$\mu = \{\mu_1, \mu_2, \cdots, \mu_m\}$と標準偏差$\sigma = \{\sigma_1, \sigma_2, \cdots, \sigma_m\}$を計算し，式(3)のようにXの標準化を行う．

$$\hat{X}_{ij} = \frac{X_{ij} - \mu_i}{\sigma_i} \tag{3}$$

ここで，相関係数行列 A の要素$a_{i1,i2}$を式（4）とし，A の逆行列の各要素を$\bar{a}_{i1,i2}$とおく．

$$a_{i1,i2} = \frac{1}{n}\sum_{j=1}^{n} X_{i1j}X_{i2j} \tag{4}$$

m次元の観測データYのマハラノビス汎距離MD_Yの二乗は式(5)で与えられる．このMD_Y^2に閾値を設定し，観測データが正常か異常かを判別する．

$$MD_Y^2 = \frac{1}{m}\sum_{ij} \bar{a}_{ij}\left(\frac{Y_i - \mu_i}{\sigma_i}\right)\left(\frac{Y_j - \mu_j}{\sigma_j}\right) \tag{5}$$

次に，l個の異常なデータ$\mathrm{Z} = \{Z_1, Z_2, \cdots, Z_l\}$を，2 水準系の直交表に従い，計算に使用する特徴量を選択する．そして，再計算したマハラノビス距離をMD_{Zi}とおく．SN 比 η[db]を式(6)と定義する．

$$\eta = -10\log\frac{1}{l}\left(\frac{1}{MD_{z1}^2} + \frac{1}{MD_{z2}^2} + \cdots + \frac{1}{MD_{zl}^2}\right) \tag{6}$$

2 水準系の直交表に従い，ある特徴量を使用する場合の SN 比の平均を$\bar{\eta}_a$，使用しない場合の SN 比の平均を$\bar{\eta}_b$とし，その特徴量の有意度e_iを式(7)で求める．

$$e_i = \bar{\eta}_a - \bar{\eta}_b \tag{7}$$

有意度e_iにも閾値を設定することで，利用する特徴量を決定することができる．

本論文ではデータ分析にＭＴ法を適用し，MT 法における特徴選択・評価について，直交表を用いるのではなく最適化手法を用いる．直交表における評価は精緻に設計すると，少ない計算量で特徴量の関係や結果に大きな影響を及ぼしている特徴量を特定できるという利点がある．その反面，オフラインによる手作業が含まれる事が多い．そして，データの特徴量が多い場合は直交表による評価の設計や設計に沿った評価に非常に時間がかかる課題がある．また，直交表の利用には知識と経験が必要であり，誰もが同じ品質で作業を実施できないという難しさもある．それとともに，直交表を用いた特徴選択は正常なデータではなく，異常なデータを用いて実施する．しかしながら，既に述べてきた通り，実際の企業においては正常なデータ，異常なデータが全て事前に揃っていないことも多くあり，適用が難しい場合も多い．

特徴選択に最適化手法を用いる事で直交表を利用した特徴選択の課題をある解消することが可能になる．

【特徴選択に最適化手法を用いる利点】

- データが保有する特徴量が非常に多い場合においても特徴選択を自動で実行することが可能となる．
- データに関する深い知識がなくても特徴選択の設計にバラツキが発生せず，一定の精度・品質を担保できる．
- 事前に準備できるデータが限定されている場合においても実施可能.

最適化手法を用いた特徴選択の考え方を図 5 に示す.

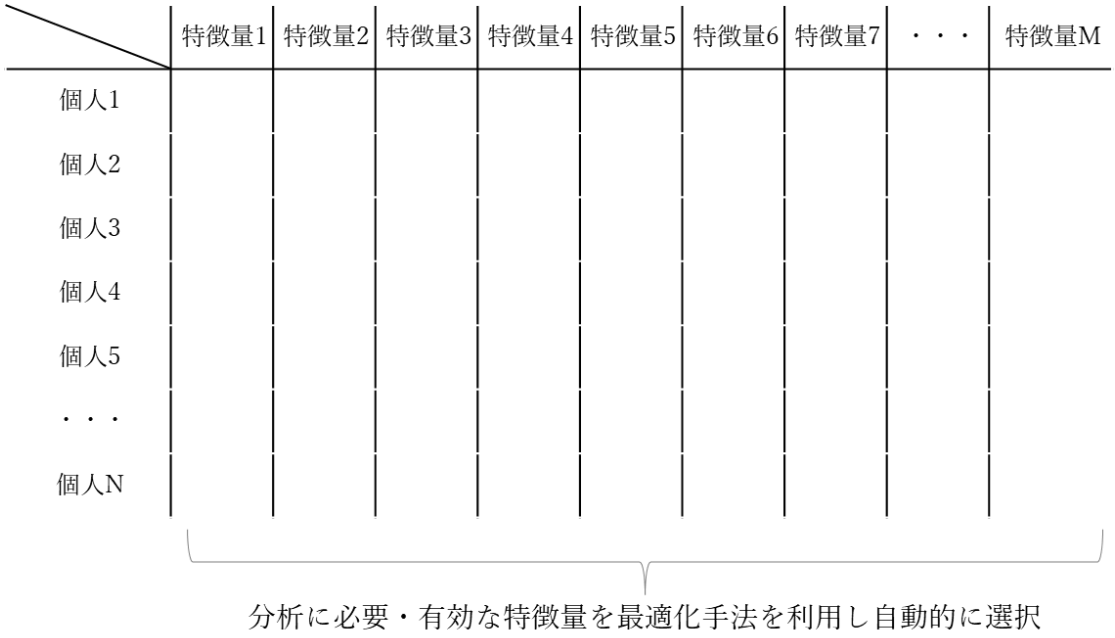

図 5　最適化手法を用いた特徴選択

3.1.2 遺伝的アルゴリズムの概要と提案手法への適用

最適化手法の中でも進化論的な手法[54]に基づく計算は，生物における遺伝子の複製や選択淘汰のプロセスを模倣した考え方である．進化論的な手法はこの考え方に基づいて，データを操作し問題や目的に応じた最適解の探索，学習、推論を行う方法である．

進化論的計算手法で扱う情報の一つはGTYPE（genotype）と呼ばれる，これは遺伝物質の構成である遺伝子型のアナロジーである．この GTYPE に対して，進化論的計算手法による操作が行われる．もう一つの PTYPE（phenotype）は表現型（発現型）であり，このPTYPEで表現した解に応じて，進化論的計算手法の判断に用いられる適合度（fitness value）が算出され・評価される．

進化論的手法では，複数の個体が集団を構成する．各個体は各々GTYPEとして遺伝子コードを有し，それらが発現したPTYPEに応じて適合度が決まっている．そして，

これらの個体群は生物における生殖活動を行い，次の世代の子孫を作り出す．この次の子孫を作り出す際にPTYPEにより算出された適合度が良いものを優れた個体と見做し，より次の世代へ残りやすいようにし，適合度が悪いものは死滅しやすいようにする．このようにして生物における進化プロセスを模倣し最適解を算出する．

進化プロセスを模倣するため，生殖の際には単に前世代の内容がコピーされるのではない．生物を模した，交叉，逆位，突然変異等を起こすことで遺伝子を変容させ，バリエーションのある次世代を発生させることで，問題に対する解の探索効率を高めている． 交叉には一点交叉や複数点交叉や一様交叉等のバリエーションが存在している．また，交叉や突然変異についてはどの程度発生させるかという発生率を決める必要がある．

進化計算では，適者生存（Survival of the fittest）の原則を実現して，適合度がよいものほど，より多産で生き残りやすいように集団内の個体を選択する必要がある．選択法としては通常，以下の3つの方式が使われる．

- ルーレット方式（Roulette）

  適合度に比例した割合で選択する方法である．一番単純な実現法は重み付けのルーレットによるものになる．これは適合度に比例した領域を持つルーレットを回し，ルーレットの球が入った領域の個体を選び出すことで行う．
- トーナメント方式（Tournament）

  これは集団の中からある個体数をランダムに選びだして，その中で一番良いものをトーナメント方式で選択する．この過程を集団数が得られるまで繰り返すというものである．
- エリート戦略

  ルーレットとトーナメント方式による選択では，親の候補はあくまで確率的に選択されるため，最良個体が次の世代に残らない可能性がある．加えて，最良個体が親の候補として残ったとしても，遺伝操作である交叉や突然変異が行われると遺伝子が変更されるため，最良個体が残らない場合がある．そのため，かならず最良個体を次世代に残す方法をエリート戦略という．この場合，次世代では前世代の結果が最低限保証される．

進化論的計算手法の代表的なものとして，GA[55]や遺伝的プログラミング（Genetic Programming：GP）がある．おもに GA はパラメーターの最適化を目的とするのに対し，GP は遺伝子型がプログラムなどの構造表現であり，人工知能の問題解決やプログラムの自動生成を目指しているものである．GA や GP は様々な実際的な問題に応用されその有効性が確認されている．これらの応用例には，工学的なパラメーター最適化，ロボットの学習，設計問題，画像処理，芸術分野へのデザイン問題などが含まれている．

GA は上述の通り，生物の自然選択の進化過程において，集団内で環境適応性の高い個々の生物は，生き残り，次世代をリードする可能性が高いという原則をモデル化している．常に最適な解を見つけるという保証はないが，通常，短時間で比較的高い精度で近似解を見つけることを可能とする手法である．

このように進化論的計算および GA は問題を解法する際に解となりうる選択肢・候補を遺伝子として表現し，それぞれの選択肢・候補に対して評価を実施する．そして，遺伝子を操作することで異なる遺伝子を生成し，より適合度の高い個体を探索するものである．分析に利用する分析対象データそのものや，分析アルゴリズムと連動することなく，遺伝子という変数のみを変化させることで最適解の探索が可能である．そのため，様々な問題に適用しやすく，インプットに利用するデータや分析に利用するアルゴリズムとの分離性も高い．データ分析手法を企業実務の現場へ適用するにはデータ分析システムや分析ツールといった仕組みを構築して運用することになる．GA であれば，そのシステムやツールの長期運用における機能としての独立性・疎結合を実現できる．また，企業における仕組みの保守・運用業務を簡素化することが容易である．

本論文では，どの特徴量を利用してデータ分析を実施するかの特徴選択を自動化し，担当者毎のデータ分析のバラツキを抑えることで，データ分析の高位平準化と標準化および効率化を実現することが必要である．そのための方法として，運用面におけるデータやデータ分析アルゴリズムとの高い独立性から GA を利用する．

本論文においては GTYPE として，分析対象のデータが保持している特徴量に応じ

て長さが変わる一次元のビット列を定義する．これをバイナリで表現したものをPTYPEとする．具体的には1として表現された場合は該当のデータの特徴量を利用するとし，0 として表現された場合は該当のデータの特徴量を分析に利用しないとしている．

GAの基本的な流れは以下になる[56]．GTYPEの集合$M(t) = \{g_t(m)\}$をある世代tにおける個体群とする．各$g_t(m)$の表現型$p_t(m)$に対して環境内における適合度$u_t(m)$が決定される．適合度の高い個体は，生き残り，次世代をリードする可能性が高いという原則をモデル化するため，適合度の良い GTYPE が選択され，その結果生成された新たなGTYPEは適合度の悪いGTYPEと置き換えられる．これにより，次の世代$(t+1)$の GTYPE の集合$M(t+1) = \{g_{t+1}(m)\}$が生成される．定義した収束条件に達するまで同様にしてこれらの過程は繰り返される．

Step 1　初期集団の生成

ランダムに初期世代の集団$M(0)$を生成する．

Step 2　各個体の適合度の計算

現在の集団$M(t)$内の各個体mに対して適合度$u(m)$を計算する．

Step 3　選択

$u(m)$に比例する確率分布を用いて，$M(t)$から個体mを選び出す．具体的には一部をエリート戦略で残りをルーレット選択で定める．エリート戦略では，適合度の高いものを無条件に残す．ルーレット選択では，ルーレットの板に占める各個体の割合を適合度に基づいて定め，ルーレットを回して当たった個体を残す．

Step 4　交叉

2つの個体で染色体を交叉させ，新しい個体を作る．親となる2つの個体はルーレット選択で選ぶ．親の染色体を一様交叉（各遺伝子ごとに乱数を取りどちらの親から継承するかを定める）させ，子として染色体（個体）を作る．

Step 5　突然変異

個体に対する遺伝子操作の1つで，遺伝子情報の一部をあらかじめ決めておいた確率で変化させる．これにより局所解から抜け出す機会ができる．

Step 6　集団の評価

一定世代変化がない場合，もしくは決められた世代数に達した場合終了する．この条件を満たさないときステップ2へ戻り，次の世代の集団$M(t+1)$を生成する．

3.2 クラスタリングによる MT 法と最適化手法を用いた分析の改良

MT 法と最適化手法を用いた分析では，企業が現状保有しているデータを利用して，簡単にデータ分析を実施することが可能となる．また，人による判断を極力排除しているため，一定のレベルの品質を確保したアウトプットを安定的に出力できる．そして，ある程度の説明責任を果たすことができるため，企業のあらゆる部門や部門内の担当者の業務へ適用し活用することが容易であると考えられる．

一方でデータが今後更に複雑化し高度化していくと，分析するデータセットの単位が適切であるのかの判断を，誰もが同じ考え方や品質で実施していくのが難しくなることが予想される．例えば，動物全体のデータを入力値とするのがよいか，犬，猫，その他等のデータに分割したデータを入力値とするのがよいか等である．

この対象・目的に応じた適切なデータのグループを判断して単位空間を作成する問題に対処するために，クラスタリング手法の一つである x-means 法を使用して，データを適切なグループへ分割する．

x-means 法は k-means 法等のクラスタリング手法と異なり，データ分析者がクラスター分割数を指定する必要がない．k-means 法ではクラスター分割数を指定する必要があるため，企業においてはデータ分析者のスキルや知識によって指定する分割数が異なり，データ分析の品質が均一化できない問題があった．x-means 法はアルゴリズム側でクラスター分割数を自動的に決定する．これにより，業務知識やスキルを保有していない企業の担当者でも適切な単位空間を作成できるようになる．また，企業側としても個人のスキルや知識に依存することなく，業務を標準化することができる．

具体的には，優良会員データに対して，最初に x-means 法を使用して適切なグループに分割する．この際にクラスター分割数はデータ分析者が試行錯誤して設定するのではなく，x-means 法のアルゴリズムによりクラスターの分割数が自動的に決定され

る．分割された優良会員グループに対して，GA を適用し，分析・評価に利用するデータの特徴量を決定し単位空間を形成する．この各単位空間に対して MT 法を適用しモデルを構築する．その後，各モデルの評価を行う．モデルの評価には検定データを与え，その検定データに対する優良会員か通常会員かの正解率で評価する．この特徴選択～単位空間作成～モデル評価のサイクルを繰り返し実施し，評価結果に基づき最終的に利用する特徴量を各単位空間ごとに決定する．次に，通常会員データに対して各単位空間ごとに MT 法でマハラノビス距離を算出し，事前に算出した各単位空間との距離を判定する事で，通常会員の中から優良会員と同じ特徴を有しているデータを抽出する仕組みを構築する．

図 6 はクラスタリングによる MT 法と最適化手法を用いた分析の改良の概要を示している．

図 7 で具体的な処理フローを示す．

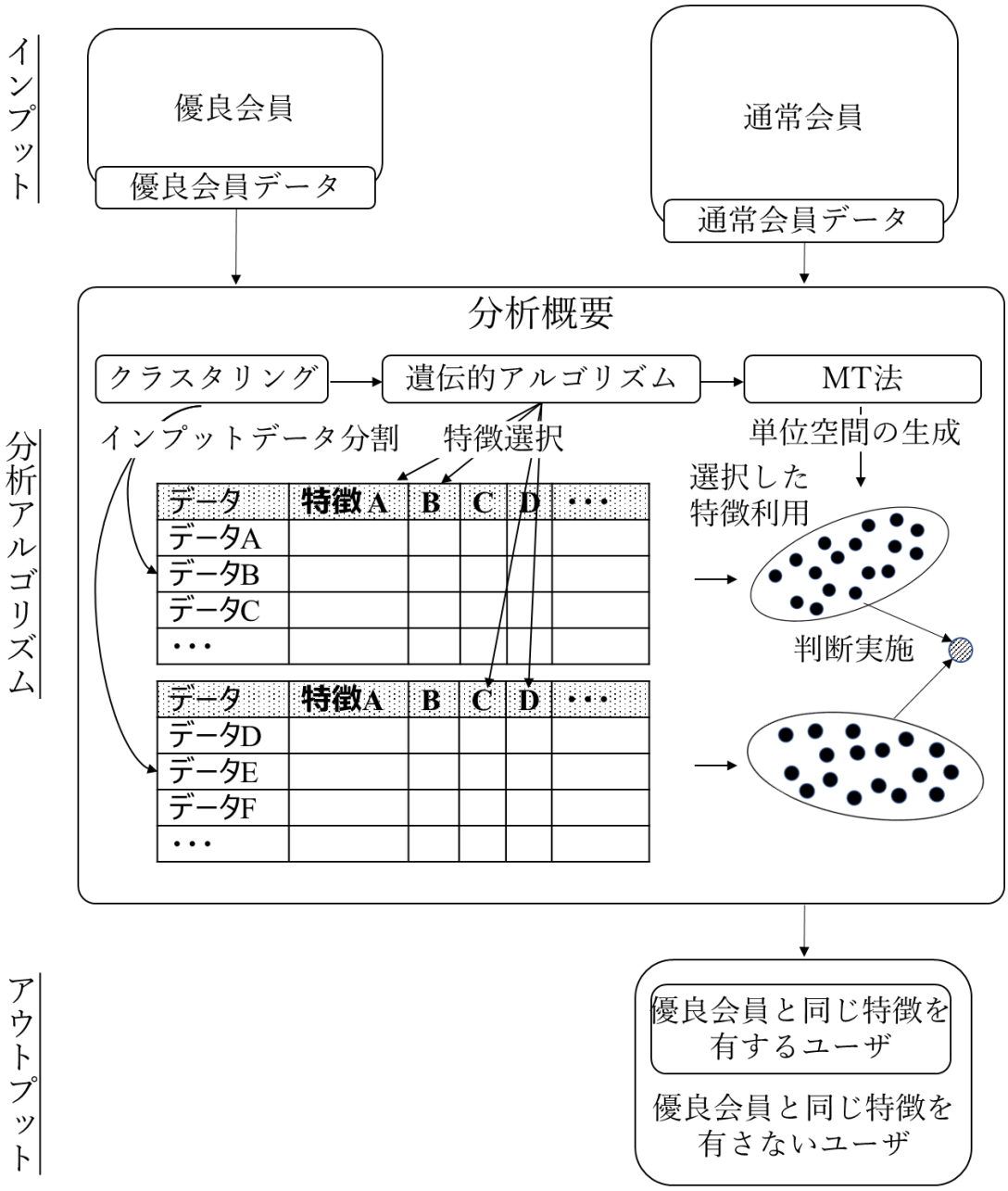

図 6　クラスタリングによる MT 法と最適化手法を用いた分析の改良の概要

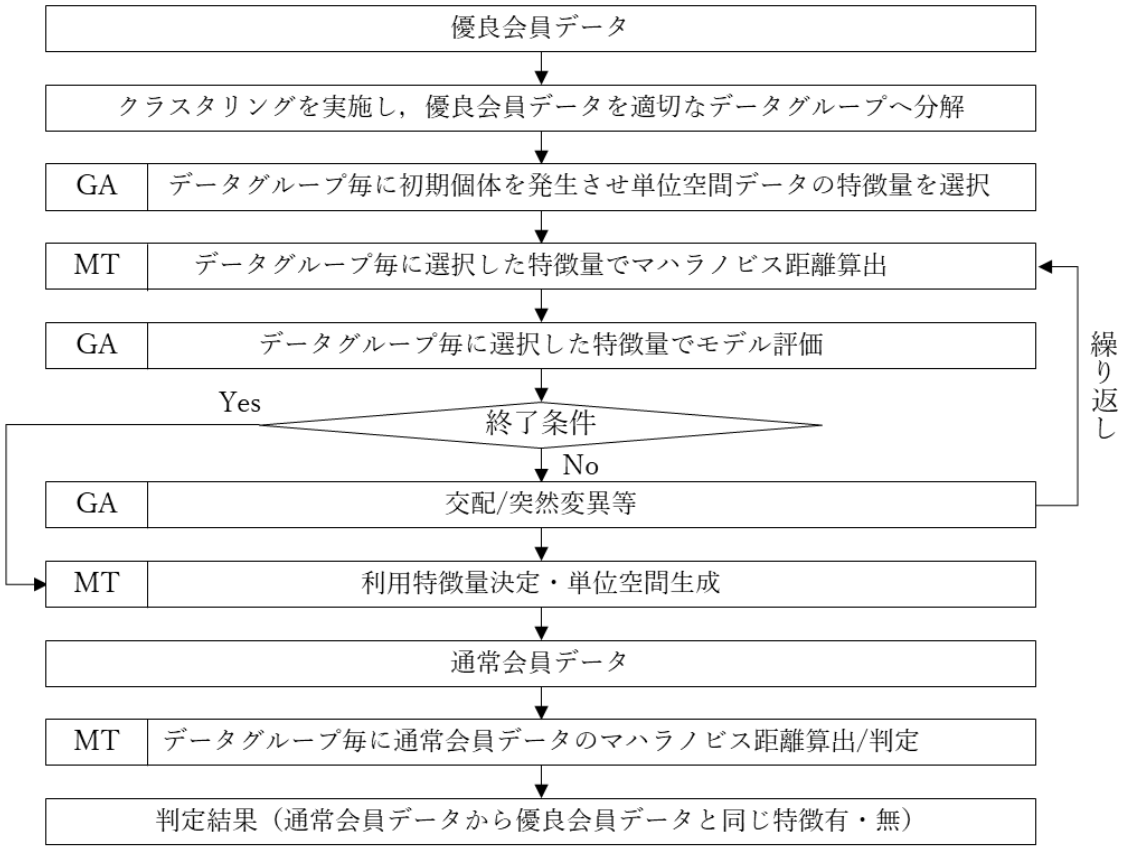

図 7　クラスタリングによる MT 法と最適化手法を用いた分析の改良の流れ

3.2.1 クラスタリング手法の概要と提案手法への適用

MT 法では，判断の基礎となるマハラノビス距離は，単位空間と呼ばれるデータグループから計算される．単位空間が適切なデータグループであることを確認することは，データ分析を実行するために重要である．今日，データは益々複雑になり，分析に適したグループ単位を人間が適切に判断することは困難になっている．クラスタリング手法を使用して適切な単位空間を機械的に作成することが可能となれば，データ分析の基礎となる単位空間の定義（これまで人間の判断に委ねられていた作業）が体系化・自動化され，誰もがより標準化された均一な品質のデータ分析を実行できるよ

うになる．そして，より正確なデータ分析システムの構築が可能になる．

クラスタリングには，データが階層的に分類される階層的アプローチと，データが特定の数のクラスターに分類される非階層的アプローチの両方が含まれる．階層的アプローチの典型に Ward の方法[57]がある．クラスタリング結果をツリー構造で出力し，結果取得後の分類状態からクラスター数を決定することができる．ただし，データ量が多い場合は，必要な計算にかなりの時間がかかる場合がある．一方，k-means 法は，典型的な非階層的アプローチを表している．固定数のクラスターに従って，類似した属性を持つデータをグループ化する．そのため，k-means 法においてはクラスターの数を事前に決定する必要がある．また，計算量の観点からみると，大量のデータがある場合でも比較的迅速に完了することができるという特徴を有している．ビジネスの分野では，扱うデータ量も年々増大しており，大量のデータを処理するシーンが非常に多くなっている．このような状況においては，クラスターの数を事前に固定するという問題があるものの，この部分への対処ができる限り，非階層的な方法を使用することが望ましい．

実際の企業内でデータ分析を実施する場合においては k-means 法の様な非階層的な方法を選択する必要がある．しかし，クラスター数を事前にデータ分析者が指定する必要があるため，この部分について企業内担当者のスキルや経験への依存が発生してしまい，データ分析結果の品質にバラツキが発生する．極力，担当者のスキルや経験に依存しないようにするため，本論文ではクラスター数を事前に決めなくても実施可能であるクラスタリング手法が必要となる．近年，クラスター数を事前指定せず，分析とともに分割数を自動的に決定する方法が研究されてきており，具体的な手法として，x-means 法が存在している．

本論文では，大量のデータの分析をサポートし，ユーザーの判断や経験に依存しない分析を可能にするために，x-means 法を使用する．

k-means 法とは異なり，x-means 法はクラスターの数を自動的に決定する． その x-means 法の手順を以下に示す[58].

1. クラスター数の初期値k_0を指定する(通常は 2).
2. $k = k_0$として k-means 法を適用する．分割後のクラスターを$C_1, C_2, \ldots, C_{k_0}$とする．
3. $i = 1, 2, \cdots, k_0$とし，手順 4〜9 を繰り返す.
4. クラスターC_iに対して$k = 2$として k-means 法を適用する．分割後のクラスターをC_i^1, C_i^2とする．
5. C_iに含まれるデータx_iに以下のp変量正規分布を仮定する．

$$f(\theta_i; x) = (2\pi)^{-\frac{p}{2}} |\mathrm{V}_i|^{-\frac{1}{2}} \exp\left[-\frac{(x-\mu_i)^t \mathrm{V}_i^{-1}(x-\mu_i)}{2}\right] \tag{8}$$

そのときのBICを以下により計算する．

$$\mathrm{BIC} = -2\log L\left(\widehat{\theta_i};\ x_i \in C_i\right) + q\log n_i \tag{9}$$

ここで$\widehat{\theta_i} = \left[\widehat{u_i}, \widehat{\mathrm{V}_i}\right]$は$p$変量正規分布最尤推定値とする．$\mu_i$は$p$次の平均値ベクトル$\mathrm{V}_i$は$p \times p$の共分散行列である．$q$はパラメーター空間の次元数であり，各パラメーターがそれぞれ独立であると仮定して共分散を無視すると$q = 2p$である．x_iはクラスターC_iに含まれるp次元データとし，n_iはC_iに含まれるデータ数とする．Lは尤度関数で$L(\cdot) = \prod f(\cdot)$である．
6. C_i^1, C_i^2に対して，それぞれパラメーターθ_i^1, θ_i^2をもつp変量正規分布を仮定し，2 分割モデルにおいてデータが従う確率密度を以下とおく．

$$g\left(\theta_i^1, \theta_i^2; x\right) = \alpha_i \left[f(\theta_i^1; x\right]^{\delta_i} \left[f(\theta_i^2; x\right]^{1-\delta_i} \tag{10}$$

ここで

$$\delta^i = \begin{cases} 1, & x_i \in C_i^1 \\ 0, & x_i \in C_i^2 \end{cases} \tag{11}$$

とする．また，α_iは確率密度とするための基準化定数であり，その近似として

$$\alpha_i = \frac{0.5}{K(\beta_i)} \tag{12}$$

により計算する．ここで$K(\cdot)$は標準正規分布の下側確率とする．

β_iは$f(\theta_i^1; x)$と$f(\theta_i^2; x)$の分離の程度を示す指標で以下に示すものとする．

$$\beta_i = \sqrt{\frac{\|\mu_i - \mu_2\|^2}{|\mathrm{V}_1| + |\mathrm{V}_2|}} \tag{13}$$

これらを用いて 2 分割モデルにおけるBICを以下により計算する．

$$\mathrm{BIC}' = -2\log L'(\widehat{\theta_i'}; x \in C_i) + q'\log n_i \tag{14}$$

ここで$\widehat{\theta_i'} = [\widehat{\theta_i^1}, \widehat{\theta_i^2}]$は$p$変量正規分布の最尤推定値とする．各パラメーターがそれぞれ独立と仮定して共分散を無視すると，各pに対し平均と分散の 2 つのパラメーターが存在するので，パラメーター空間の次元は$q' = 2 \times 2p = 4p$となる．

7. $\mathrm{BIC} > \mathrm{BIC}'$ならば 2 分割モデルをより好ましいと判断し，2 分割を継続するため$C_i \leftarrow C_i^1$とする．C_i^2についてはp次元データ，クラスターの重心，対数尤度とBICを保持し，これらをスタックに積み，手順 4 へ戻る．
8. $\mathrm{BIC} \leq \mathrm{BIC}'$ならば 2 分割しないモデルをより好ましいと判断し，C_i^1について 2 分割を停止する．手順 7 で作成されたスタックからデータを取り出して$C_i \leftarrow C_i^2$として，手順 4 へ戻る．スタックが空なら次の手順へ進む．
9. C_iにおける 2 分割が全て終了．手順 4～8 で作成された 2 分割のクラスターの番号をC_i内で一意になるように振り直す．
10. はじめにk_o分割したクラスター全てについて 2 分割が終了．全データに対してそれらの属するクラスター番号が一意になるようにデータの属するクラスター番号を振り直す．
11. 全データの属するクラスター番号，および各クラスターの重心，各クラスターに含まれるデータ数を出力し，全ての処理を終了とする．

第4章 提案手法への TV 録画データの適用と検証

4.1 検証概要

本論文では提案手法を検証するために， TV 録画データを用いた検証で提案手法の有効性を確かめる．本検証で利用する TV 録画データとは，録画した番組やその内容，放送日時やチャンネル情報，ジャンル情報等のメタ情報，視聴有無やどのようなモードで録画したか，どのようなデバイスへ録画したか，どのようなデバイスで視聴したか，どこまで視聴しているか，途中でやめたのか，やめなかったのか，どのような速度で視聴しているか等の情報で構成されている．人に関する各種情報（居住地，家族構成，年代，収入等のデモグラフィック情報）は含んでいない．本検証ではデータ件数 892 件，保有する特徴量 25 個のデータを利用した（表 2 参照）．

表 2　TV 録画データ例（一部抜粋）

項目	補足
録画機器	・本論文では 25 個の特徴量を選択して利用 ・892 機の録画機器に含まれる TV 録画データを対象
録画数	
チャンネル情報	
ジャンル情報	
モード情報	
・・・	

比較手法として，1 クラスデータ（1 つの正解データ）で分析が可能な異常検知の分野で多く利用されている OCSVM および LOF を用いた．また，2 クラスデータ（2 つの正解データ）で分析が可能な手法として， NN，DT，LT を用いた．提案手法は優良会員データのみを用いてデータを分析する．一方，NN，DT，LT は教師あり学習であり，教師データとして，優良会員データと通常会員データの 2 種類を用いて分析を行う．そのため，単純比較は困難であるが，参考として NN，DT，LT と比較してどの程度のパフォーマンスを示せるのかを確認した．

OCSVM は 1 クラス（1 種類の正解データ）だけのデータで分析が可能であり，教師なし学習に分類され，教師データが不要であるため，分析を行うために 2 種類の正解データを必要とせず，本論文で検証したい内容と合致した検証を実施することが可能である．OCSVM は SVM と同様にカーネル法で高次元の空間に変換する．そのため，データ分析を実施するにはカーネルに何を用いるか（デフォルトは RBF（Radial Basis Function：放射基底関数））やデータセットに占める外れ値の割合の上限等をパラメーターとして閾値設定をする必要がある．

LOF はデータの中から外れ値を見つける外れ値検知のアルゴリズムの一つであり，OCSVM と同様に 1 クラス（1 種類の正解データ）のデータのみで分析が可能であるため比較対象として選択した．近傍 k 個の点といかに密かであるかを表す局所密度

（Local density）という指標を定式化し，ある点が外れ値である可能性を表す外れ値スコアを算出し判別する．この際に最終的に正常値なのか異常値なのかを判別するために LOF では外れ値スコアの閾値設定が必要となる．

2 種類の正解データを学習に必要とする NN は人間の脳を抽象化し，それをコンピュータ上で表現するために考えられたモデルである．パターン認識のための一連のアルゴリズムであり，認識できるものとして画像、音声、テキスト、時系列など様々なものが存在している．NN は人間の脳から着想を得ており，大きく，入力層，中間層，出力層の 3 つから成り立っている．入力層は外部からデータを収集し，入力されたデータを次の層へ転送していく役割を担う．中間層はバックグラウンドでの計算の実行を担い，中間層を複数設けることも可能である．出力層は中間層で計算された結果を伝達する役割を担う．NN の特徴として，入力データをある目的をもって分析するのではなく，入力データに付与されている正解ラベルから，その正解ラベルへ判別するために必要になる入力データの特性を学習していく点にある．そのため，入力データから何を取得するのかというプログラムを記述する必要がない．このように与えられた正解がわかっているデータから学習していくという特徴から，精度を高め，より正確な結果を算出するには，入力データを可能な限り多くしていくことが求められる．企業等に 2 クラスデータが潤沢にある場合は非常に有効な学習を実施できる可能性が高まる．また，NN における学習とは、出力層で人間が望む結果（正しい答え）が出るよう、パラメーター（重みとバイアス）を調整する作業になる．そのパラメーターには，層の数、ニューロンの数、活性化関数の種類等，考慮すべきパラメーターは非常に多く，適切な結果を得るためにはそれぞれを適切にチューニングしていく必要がある．また，NN は原則として学習するために 2 種類の正解データが必要となる．そのため， 1 種類の正解データしかない場合や片方のデータサンプルが非常に少ない場合におけるデータ分析には適用しにくい．

DT は木構造（ツリー構造）を利用してデータを分類していく方法である．ツリー構造を用いて各特徴量に対して分岐を繰り返していく．どの特徴量での分割が有効で

あるかを算出し，有効度の高い特徴量での分割から適用し，繰り返し実施を行い，ツリーを構築し分類を実施する．企業等に適用する際の利点として最も大きいのは，計算・出力された結果に至る理由が非常にわかりやすく，組織内部の意思決定に用いる際の説明責任が果たしやすいことである．また，企業におけるデータの特徴量が複雑化している中で，特徴量の多いデータや質的変数・量的変数が混在しているデータでも扱いやすいという利点もある．ただし，性能面においては他の手法に劣る事が多く，過学習を起こしやすいといった難点もある．また，DT は 2 種類の正解データが必要なデータ分析手法であるため，1 種類の正解データしかない場合には適用しにくい手法である．

LT はインプットである入力データを元に，そのデータに含まれている特徴量を用いて 2 値の結果（答えが二つしかない値）が起こる確率を説明・予測する，もしくはデータが 2 値の結果のどちらに属するのかを予測・分類する分析手法である． 特徴としてアルゴリズム自体が単純でわかりやすいため，技術的に実際の現場へ実装しやすく適用しやすい点がある．一方で分析性能をみると，ある特徴量の集合が二次元平面上にマッピングされていると仮定した場合に，特徴量の集合によって，あるクラスのデータと別のクラスのデータを 1 本の直線で分割できる，線形分離が可能な場合のみに高い性能を発揮する点が挙げられる．他の手法と同様にパラメーター数やどのような値を設定すべきかの数値設定が多い．また，2 種類の正解データが必要なデータ分析手法であるため，1 種類の正解データしかない場合には適用しにくい手法である．

これらの手法を提案手法の比較手法として検証した．

4.2 MT 法と最適化手法を用いた分析の検証

4.2.1 検証方法

優良会員の TV 録画データをインプットとし，分析に利用する特徴量をランダムに自動的に選択する．そして選択された特徴量を用いて MT 法を適用し単位空間を生成する．次に選択した特徴量を用いたモデルの評価を行う．モデルの評価には検定データを与え，その検定データに対する優良会員か通常会員かの正解率で評価する．その後，GA を用いて特徴選択，単位空間生成，選択した特徴量を用いたモデル評価のサイクルを繰り返し実施し，分析に最適な特徴量を決定し単位空間を生成する．そして，通常会員の録画データをインプットとし，事前に決定した特徴量を用いて通常会員の録画データからマハラノビス距離を算出する．最後に，優良会員の録画データから得られた単位空間との距離を比較することで通常会員の TV 録画データの中から優良会員の録画データと同じ特徴を有しているデータを判断する．
処理ステップの詳細を以下に示す．

Step 1　優良会員の録画データをインプットとする．

Step 2　GA を適用し特徴選択を実施する．最初はランダムに分析に利用する特徴量を選択する．

Step 3　選択された特徴量を利用して MT 法を実施し，単位空間とマハラノビス距離を算出する．

Step 4　GA により選択された特徴量から算出した分析モデルに対して，検定データを与えてモデルの評価を実施する．モデルの評価は与えた検定データに対する正解

率で評価を実施する．

Step 5　GA による，エリート選択，ルーレット選択，突然変異を実施し，次世代に利用する特徴量を算出する．

Step 6　GA により新たに選択された特徴量で，マハラノビス距離を算出し，検定データを与えてモデル評価を実施する．

Step 7　5〜7 を繰り返し，分析に最適な特徴量を探索する．

Step 8　収束条件（今回は定義した世代数）に達したら，GA によるループを終了し分析に利用する特徴量を決定する．

Step 9　通常会員の録画データからマハラノビス距離を算出する．

Step 10 算出されたマハラノビス距離から，優良会員の録画データと同じ特徴を有しているか否かを判断する．

4.2.2 提案手法の検証

検証には単位空間を作成するための基準データとして優良会員の録画データ 200 件を用いた．検証データとして，通常会員データを用いるが，今回特別に企業にてアンケート等の追加調査を行うことにより，通常会員から優良会員になったデータと通常会員のままであったデータを特定している．結果として，検証データは優良会員データ 140 件，通常会員データ 752 件，合計 892 件のデータを用いた．

MT 法の判定の閾値となるマハラノビス距離は品質工学の分野等で多く用いられている数値である 4 とした．GA はエリート選択，ルーレット選択，突然変異を利用し，各パラメーターは個体数 100，世代数 10，エリート選択 0.1，突然変異 0.2 を用いた．

また，提案手法の TV 録画データへの有効性を確認するため，1 クラスのみのデータで分析が可能である，他の 2 つの方法 OCSVM および LOF のパフォーマンスと比較

した.

OCSVM，LOF ともにデータに対する特徴選択は実施せず，そのままの状態で分析を実施した．実装は Python 用の機械学習ライブラリ scikit-learn を用い，パラメーターは標準設定である.

企業の様々な業務や職能へ適用することを考えると，各実務担当者が適切な分析アルゴリズムを選択し，データ特性や業務特性に応じた特徴選択処理を行う事は難しい．また，様々な実験を行いパラメーターを調整しハイパーパラメータを見つける事も困難である．そのため，提供されるライブラリをそのまま利用する場合を仮定し比較した.

表 3 と図 8 に，TV 録画データへ適用し検証した結果を示す.

表 3　検証データに対する正解率

	合計	優良会員	通常会員
提案手法	870/892	136/140	734/752
OCSVM	752/892	10/140	742/752
LOF	753/892	21/140	732/752

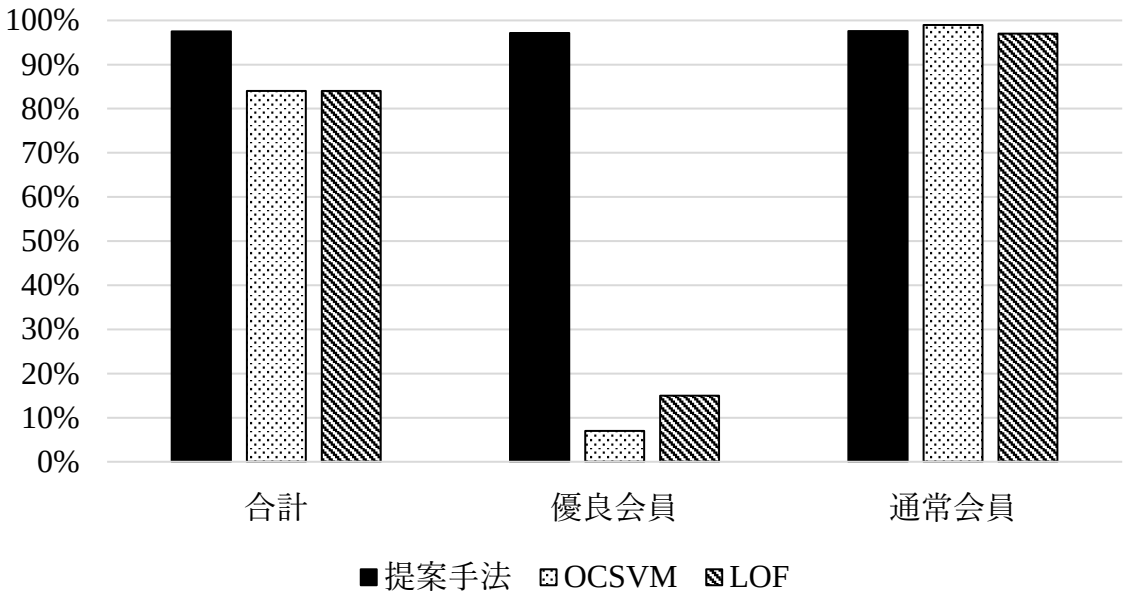

図 8　提案手法と各手法の比較（正解率）

検証データ全 892 件に対する正解数は，提案手法が 870 件正解である．これに対して比較手法である OCSVM では 752 件で，LOF では 753 件正解している．これらの事から検証データ全体に対する正解数では TV 録画データに対する提案手法の有効性が確認できた．

次に検証データに含まれる優良会員データを正しく判別できたかという観点で見ると，優良会員データ全 140 件に対する正解数は，提案手法が 136 件である．これに対して比較手法である OCSVM では 10 件，LOF では 21 件正解している．これらの事から優良会員データに対する正解数でも TV 録画データに対する提案手法の有効性が確認できた．OCSVM および LOF の正解率が低い点については 2 つの理由があると考えられる．一つ目はデータに対して特徴選択処理を実施していないためである．すべての特徴量を利用したため分析精度が低くなっていると考えられる．二つ目は OCSVM および LOF のパラメーター設定である．OCSVM，LOF ともに適切なパラメーター設定が求められるため，適切にパラメーター設定を行わないと分析精度が低くなってしまう．ただし，その適切なパラメーター設定は非常に難しく，適切なパラメーターを見つけるには実験と調整を繰り返し実施していく必要がある．しかし，ハイパーパラ

メーターを見つけることができれば大幅に正解率は向上すると考えられる．一方で提案手法における，特徴選択の有効性や複雑なパラメーター設定を必要としない分析手法の有効性も示すことができたと考えられる．

最後に検証データに含まれる通常会員データを正しく判別できたかという観点で見ると，通常会員データ全 752 件に対する正解数は，提案手法が 734 件である．これに対して比較手法である OCSVM では 742 件，LOF では 732 件正解している．これらの事から検証データから通常会員データを判別するという点においては他手法の方が高い正解数となった．この検証での内容を実際の企業の現場で活用する場合を考えると，通常会員データから優良会員データと同じ属性を有しているデータを抽出し，そのデータを保有している通常会員へ対してターゲティングやコミュニケーションを図り，企業とのタッチポイントを作っていくことが大切である．それには検証データ全体に対する正解数および優良会員データの正解数が重要なポイントになる．そのため，通常会員データ全 752 件に対する正解数単体で評価するよりも，全体に対する正解数に重点をおき評価することが大切になる．

提案手法は比較手法である OCSVM および LOF よりも検証データ全体に対する正解数および優良会員データの正解数で良い結果となっている．このことから，TV 録画データに対する提案手法の有効性が確認できた．

4.2.3 まとめと今後の課題

TV 録画データを用いた検証において提案手法は他手法よりも高い正解率を得ることができた．企業や職種によってはデータ分析の有識者人材が枯渇し，その現場では 2 クラスデータが不足している．このような場合においては，今あるデータで分析に着手でき，できるだけ簡単で品質のばらつきがない分析が行える事は重要である．提案手法はほとんど人の判断を必要とせず，高い結果を得ることができる．加えて，提案

手法は結果だけでなく，最適化手法によって選択された特徴量を確認することができるため，どのような特徴量を用いると有効な結果が得られたかという点を確認することが可能である．このような点から，企業の実際のデータ分析業務に適用しやすい手法となっている．

一方で今後，企業において様々なデータが取得されデータ量が膨大になり，データが大きくなっていくと，どのようなデータのグループでデータを分析するべきかの判断が難しくなる．この判断を適切に行わないまま，様々なデータが含まれた状態でデータ分析の基準となる単位空間を作成すると，単位空間の精度や粒度が荒くなり，分析性能が低下することが考えられる．

そこで，これらの課題について 4.3 にてこの課題を解消する手法の検証を実施した．

4.3 MT 法と最適化手法を用いた分析の改良手法の検証

4.3.1 検証方法

データ分析のインプットとなる，優良会員の録画データを x-means 法を用いて分割する．これにより対象データが大きい場合や特徴量が非常に複雑な場合においても適切なデータグループへ自動的に分割することができる．結果として，詳細な単位空間群を作成することができ，よりデータ特性に応じた細かな分析が理論上可能となる．その上で，分割されたデータグループごとにランダムに特徴量を選択し MT 法を適用し単位空間を生成する．次に選択した特徴量を用いたモデルの評価を行う．モデルの評価には検定データを与え，その検定データに対する優良会員か通常会員かの正解率で評価する．その後，分割されたデータグループごとに GA を用いて特徴選択，単位空間生成，選択した特徴量を用いたモデル評価のサイクルを繰り返し実施し，データグループごとに分析に最適な特徴量を決定し単位空間を生成する．そして，通常会員の録画データをインプットとし，事前に決定した特徴量を用いて通常会員の録画データからマハラノビス距離をそれぞれの単位空間ごとに算出する．最後に，優良会員の録画データから得られた単位空間との距離を比較することで通常会員の録画データの中から優良会員の録画データと同じ特徴を有しているデータを判断する．

処理ステップの詳細を以下に示す．

Step 1　優良会員の録画データをインプットとする．

Step 2　インプットデータに対して x-means 法によるクラスタリングを実施しデータグループへ分割する．

Step 3　分割されたデータグループ毎に GA を適用し特徴選択を実施する．最初はデータグループ毎にランダムに分析に利用する特徴量を選択する．

Step 4　選択された特徴量を利用して MT 法を実施し，単位空間とマハラノビス距離を算出する．

Step 5　GA により選択された特徴量から算出した単位空間毎の分析モデルに対して，検定データを与えてモデルの評価を実施する．モデルの評価は与えた検定データに対する正解率で評価を実施する．

Step 6　GA による，エリート選択，ルーレット選択，突然変異を実施し，次世代に利用する特徴量を単位空間毎に選択する．

Step 7　GA により新たに選択された特徴量で，マハラノビス距離を算出し，検定データを与えてモデル評価を実施する．

Step 8　6〜8 を繰り返し，分析に最適な特徴量を探索する．

Step 9　収束条件（今回は定義した世代数）に達したら，GA によるループを終了し分析に利用する特徴量を単位空間毎に決定する．

Step 10 通常会員の録画データからそれぞれの単位空間ごとにマハラノビス距離を算出する．

Step 11 単位空間毎に算出されたマハラノビス距離で，優良会員データと同じ特徴を有しているか，否かを判断する．

優良会員データと同じ特徴を有している通常会員データを特定することが目的であるため，通常会員データが優良会員データを分割し作成された各単位空間のどれかに属していれば優良会員データと同じ特徴を有していると判断した．

検証データは 4.2 での検証で用いたものと同じものである，データ件数 892 件，保有する特徴量 25 個の TV 録画データを利用した．

4.3.2 提案手法の検証

検証には単位空間を作成するための基準データとして優良会員の録画データ 200 件を用いた．検証データとして，優良会員データ 140 件，通常会員データ 752 件，合計 892 件のデータを用いた（基準データ，検証データともに 4.2 の検証で用いたものと同じ）．

MT 法の判定の閾値となるマハラノビス距離は品質工学の分野等で多く用いられている数値である 4 とした．GA はエリート選択，ルーレット選択，突然変異を利用し，各パラメーターは個体数 100，世代数 10，エリート選択 0,1，突然変異 0.2 を用いた．

また，TV 録画データへの提案手法の有効性を確認するため，他の 2 つの方法 OCSVM および LOF のパフォーマンスと比較した．4.2 での検証と同じ理由で OCSVM，LOF ともに録画データに対する特徴選択は実施せず，そのままの状態で分析を実施した．実装は Python 用の機械学習ライブラリ scikit-learn を用いて，パラメーターは標準設定を用いた．

表 4 と図 9 に，TV 録画データへ適用し検証した結果を示す．

表 4　検証データに対する正解率

	合計	優良会員	通常会員
提案手法	872/892	136/140	736/752
OCSVM	752/892	10/140	742/752
LOF	753/892	21/140	732/752

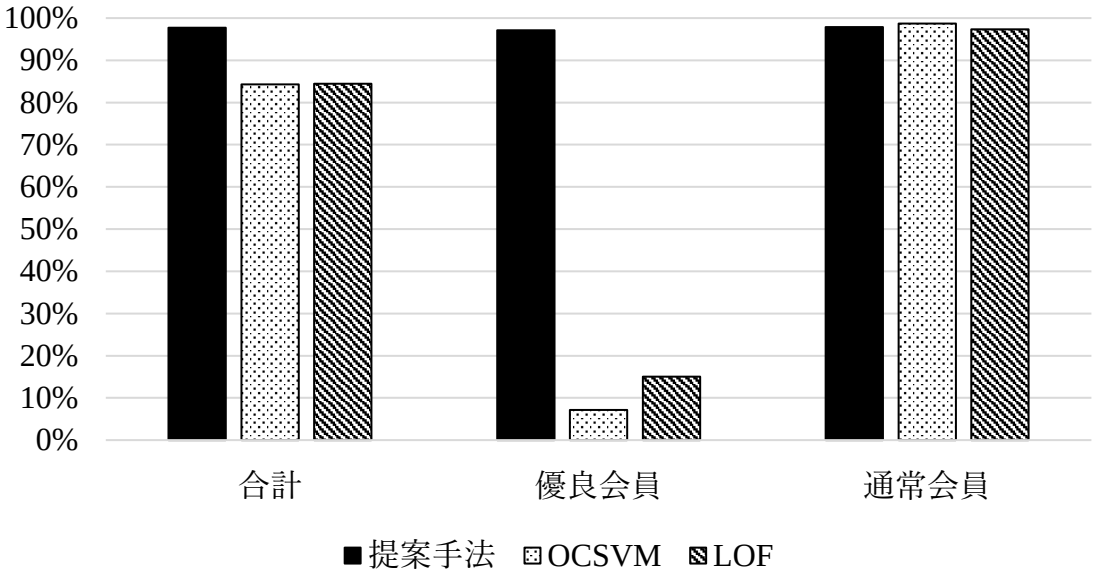

図 9　提案手法と各手法との比較（正解率）

検証データ全 892 件に対する正解数は，提案手法が 872 件正解している．これに対して比較手法である OCSVM では 752 件，LOF では 753 件正解している．これらの事から検証データ全体に対する正解数では TV 録画データに対する提案手法の有効性が確認できた．

次に検証データに含まれる優良会員データを正しく判別できたかという観点で見ると，優良会員データ全 140 件に対する正解数は，提案手法が 136 件正解している．これに対して比較手法である OCSVM では 10 件，LOF で 21 件正解している．これらの事から優良会員データに対する正解数でも TV 録画データに対する提案手法の有効性が確認できた．OCSVM と LOF の正解率が悪い理由は 4.2 での検証と同様で特徴選択処理を実施していないこと，パラメーターに標準設定を用いていることであると考えられる．企業の現場において，分析に最適な特徴選択を実施することや適切なパラメーターを見つけることは非常に難しい．しかし，ハイパーパラメーターを見つけることができれば大幅に正解率は向上すると考えられる．一方で 4.2 での検証と同様に提

案手法における，特徴選択の有効性や複雑なパラメーター設定を必要としない事の有効性も示すことができたと考えられる．

最後に検証データに含まれる通常会員データを正しく判別できたかという観点で見ると，通常会員データ全 752 件に対する正解数は，提案手法が 736 件正解している．これに対して比較手法である OCSVM では 742 件，LOF では 732 件正解している．これらの事から検証データから通常会員データを判別するという点においては他手法の方が高い正解数となった．4.2 での検証と同様に，企業の現場で活用する場合を考えると，検証データ全体に対する正解数および優良会員データの正解数が重要なポイントになる．

提案手法は比較手法である OCSVM および LOF よりも検証データ全体に対する正解数および優良会員データの正解数で良い結果となっている．このことから，TV 録画データに対する提案手法の有効性が確認できた．

4.3.3　2 クラスデータを用いる分析手法と提案手法の性能比較

1 クラスデータに対応したデータ分析手法との比較に加えて，データ分析の分野でよく実施される 2 クラスのデータ(2 種類の正解データ)を用いたデータ分析手法とのパフォーマンスの比較を実施した．

提案手法の検証方法は，4.3.2 の 1 クラスデータの検証と同じ要領で検証した．単位空間を作成するための基準データとして優良会員の録画データ 200 件を用いた．検証データとして，優良会員データ 140 件，通常会員データ 752 件，合計 892 件のデータを用いた．

MT 法の判定の閾値となるマハラノビス距離は品質工学の分野等で多く用いられて

いる数値である 4 とした．GA はエリート選択，ルーレット選択，突然変異を利用し，各パラメーターは個体数 100，世代数 10，エリート選択 0.1，突然変異 0.2 を用いた．

TV 録画データへの提案手法の性能を確認するために，2 クラスデータに対する分析手法として NN，DT，LT を比較対象の手法として使用した．

NN，DT，LT の検証には，録画データに対し優良会員ラベルと通常会員ラベルの 2 種類の正解ラベルが付与されている教師データ 200 件を準備した（通常会員から優良会員になったデータと通常会員のままであったデータを特別に特定し準備）．検証データは提案手法と同じである，優良会員データ 140 件，通常会員データ 752 件，合計 892 件のデータを用いた．

NN，DT，LT ともに 4.2 での検証と同じ理由で録画データに対する特徴選択は実施せず，そのままの状態で分析を実施した．実装は Python 用の機械学習ライブラリ scikit-learn を用いて，パラメーターは標準設定である．

表 5 と図 10 に，TV 録画データへ適用し検証した結果を示す．

表 5　検証データに対する正解率

	合計	優良会員	通常会員
提案手法	871/892	135/140	736/752
NN	645/892	71/140	574/752
DT	874/892	140/140	734/752
LT	861/892	140/140	721/752

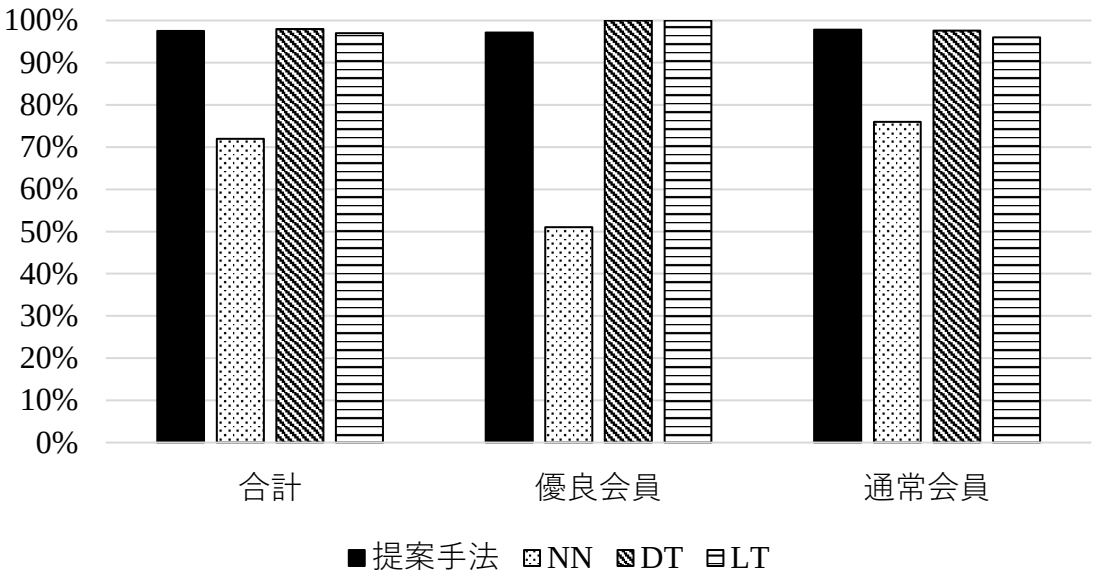

図 **10**　提案手法と各手法の比較（正解率）

検証データ全 892 件に対する正解数は，提案手法が 871 件である．これに対して比較手法である NN では 645 件，DT では 874 件，LT では 861 件である．これらの事から検証データ全体に対する正解数では DT が最も高い正解数となった．次に提案手法の正解数が高くなっている．ただ，提案手法と最もパフォーマンスの高かった DT の正解数の差は 3 件であり，提案手法も同等の性能を有していると考えられる．

次に検証データに含まれる優良会員データを正しく判別できたかという観点で見ると，優良会員データ全 140 件に対する正解数は，提案手法が 135 件である．これに対して比較手法である NN では 71 件，DT では 140 件，LT では 140 件である．優良会員データ対する正解数では 2 種類の正解データを利用して精緻な分析が行える DT および LT が非常に高いパフォーマンスを示した．ただし，提案手法も 140 件中の正解数 135 件と高いパフォーマンスを示している．

最後に検証データに含まれる通常会員データを正しく判別できたかという観点で見ると，通常会員データ全 752 件に対する正解数は，提案手法が 736 件である．これに対して比較手法である NN では 574 件，DT では 734 件，LT では 721 件である．これらの事から，検証データから通常会員データを判別するという点においては提案手法

が高い正解数となった.

この検証では本来であれば取得できていない 2 種類の正解データという環境を疑似的に作りだし，2 種類の正解データが必要な分析手法と提案手法とのパフォーマンスを比較した．この検証においては，1 種類の正解データのみで分析を行う提案手法に比べ，2 種類の正解データを用いる DT と LT の方が優良会員データの正解数が高い．また検証データ全体に対する正解数では DT が最も高くなっている．一方で，提案手法も 1 種類の正解データだけという条件の中でもある程度の性能を有している事が示された.

4.3.4 まとめと今後の課題

提案手法は 1 クラスデータを使った検証において他手法よりも高い正解率を得ることができた．また，2 クラスデータを使った検証では，2 クラスデータを用いて学習し分析を実施する他手法に数値面では劣るものの，同等の性能を発揮することができている.

企業では 1 クラスのデータしかない場合や 2 クラスデータが非常に少ない場合に，分析を諦めているケースも存在している．そのような場合においても，企業が保有するデータをそのまま用いて，データ分析を開始でき，ある程度の精度の結果をアウトプットできることは非常に意味があると考えられる．2 クラスのデータを用いる手法が高い性能を発揮することを考えると，まずは提案手法により，データ分析を実施する．そして，データ分析からうまれる施策を通じ教師データ(2 クラスの正解データ)を蓄積していく．その後，十分な教師データが確保できた段階で 2 クラスデータを用いる精度の高い分析手法に切り替え，データ分析活動を更に発展させていくということが考えられる.

今回，データ分析の対象データを x-means 法を用いて適切なデータグループへ自動

分割する方法を提案した．この手法の有効性という観点においては，4.2 で検証した x-means 法によるクラスタリングを適用しない方法と比較して，多少の正解率向上は実現しているものの，大幅な性能向上には至っていない．

x-means 法を用いる事で大きなデータ群を適切なデータのグループへ分解できる．そして，各グループごとに単位空間を作成することで，データの特性や偏りをより正確に表現できるという特徴がある．一方で，そのインプットとなる単位空間を作成するデータ群が小さく，複雑性も低い場合，x-means 法のクラスタリングによる効果が少なく，データ群を分割しない場合と比較して単位空間の形成に大きな違いがでない．今回検証に利用した TV 録画データについては複雑性もそこまで高くないものであるため，x-means 法のクラスタリングによる効果がでなかったと考えられる．

今回は提案手法を TV 録画データを用いて検証したが，今後の課題としては，複雑性の高いデータセット，特徴量が非常に多いデータセット等，様々なデータセットで検証することである．

第5章 結論と今後の課題

情報技術の飛躍的な進歩により，近年 IoT・AI・ビッグデータの重要度が増し，データの収集と保存が盛んに行われ，データ量自体は非常に大きくなっている．しかし，一番重要である蓄積したビッグデータを様々な企業活動へ利活用するための，データ活用が道半ばである．

多くの企業においては，データ分析技術者が足りていない．そのため，データ分析の業務知識や専門スキルを保有した人材ではなく，各業務担当者がデータ分析も担当する必要があることが多い．さらに，企業の現場においては 2 種類の正解データが揃い，十分な教師データが確保されているというデータ分析に万全な状態が確立されている事が少ない．このような状況では，少量の 2 クラスデータ（2 種類の正解）もしくは 1 クラスデータ（1 種類の正解）を上手く利活用する必要がある．また，実施者によるデータ分析の品質に大きなばらつきがなく，可能な限り簡単かつ自動的にデータ分析を実行できることが重要である．

提案手法は品質工学の分野で多く用いられている MT 法，クラスタリング手法である x-means 法，最適化手法の一つである GA をデータ分析へ適用している．各手法を組み合わせる事で，特別な知識やスキルがなくても自動でデータ分析を実施できる．かつ，人の判断を少なくすることで，分析結果のばらつきを低減させている．提案手法は，1 クラスのデータでも分析が可能であり，誰もが同じ品質のアウトプットを出力できる．これらの特徴から，企業の実際のデータ分析に容易に適用できると考えられる．

提案手法を TV 録画データへ適用し検証した結果では，1 クラスデータを用いた他の

データ分析手法に比べ有効である結果が得られた．また，2 クラスデータを用いる他のデータ分析手法との性能比較では 2 クラスデータを用いる分析手法に劣るものの，ある程度の性能を有することが示された．企業においては，提案手法を用いて現在保有する 1 クラスのデータで分析に着手し，データ分析を通じた施策により 2 クラスデータを取得・蓄積する．その後，2 クラスデータを用いる精度の高い分析手法へ切り替える．このようなデータ分析の取り組みをステップアップさせていくアプローチが考えられる．

今後，企業の各部門や幅広い業務へ適用させるために，複雑性が高いデータや非常に量の多いデータ等の様々なデータによる検証，他のデータ分析手法や最適化手法の適用，より汎用的なモデルの構築を実施していく．

IoT，AI，ビッグデータやデータ分析は技術の進化により近年急速に発展している．それに伴い多くの研究が行われ，多くの成果が創出されている．今後も更に技術の進化のスピードは速くなり，より多くの研究成果が世にでていくことが想定される．一方でアカデミックの研究成果と実際の企業における業務や活動は大きく離れているのが現状である．アカデミックの世界で研究が行われ様々な成果が創出されようとも，それが実際の企業やくらしの中で利用されなければ，人々のくらしの向上や社会課題の解決は実現しない．

特に日本においては企業の大半が中小企業で構成されている．大企業においては，その資金力や人材力を活用し自社での研究開発や大学等との共同研究を実施することができる．そして，実際の世界で実験を行い，研究成果を社会へ実装するという組織能力が備わっている．しかしながら，大部分を占める中小企業においては，そこまでの組織能力はないことが多い．

今後日本は，少子高齢化社会，地方の過疎化と都市部への人口集中，空き家問題，ダブルケア問題などの様々な課題に直面する．これらの課題を解消し，持続可能な国をつくっていくには，アカデミックの世界で創出された成果や技術を，大企業，中小企業のすべての企業で享受しなければならない．そして，その成果や技術を活かし，

新たな価値あるサービスを創出することで，人々のくらしをよりよくして，社会課題を解決していく必要があると考える．

本論文においては，データ活用という観点からアカデミックでの研究成果と企業の実務との間にある乖離を埋め，新たな研究技術を誰でも簡単に早く使えるデータ分析技法を提案した．今後はデータ活用を軸としながらも，アカデミックと企業実務の間にある溝を埋める方法を，様々な角度から提案していくことが研究課題の一つに挙げられる．

そして，あらゆる企業におけるデジタル技術のフル活用による，顧客，社会，地球への新たな価値創出を実現させていきたい．

謝辞

本研究を進めるにあたり，終始適切なご指導およびご助言を賜りました大阪大学大学院 情報科学研究科 森田 浩教授に厚く御礼申し上げます．論文審査の副査をお引き受け戴き，懇切なるご指導およびご指摘を戴きました，大阪大学大学院 情報科学研究科 谷田 純教授，沼尾 正行教授に謹んで感謝を申し上げます．森田研究室のみなさまには謹んで感謝申し上げます．またこの論文を直接または間接的に支援していただいた皆様に厚く御礼申し上げます．最後に，私の我が儘を叶えてくれた妻に心から感謝いたします．

参考文献

[1] 小川裕克, 永井義明, IoT 等の進展が与える 情報システムへの影響に関する研究, 産業経済研究所紀要, Vol. 27, pp.27-88, 2017.

[2] 土方嘉徳, 嗜好抽出と情報推薦技術, 情報処理学会論文誌, Vol.47, No.4, pp.1-10, 2006.

[3] 林眞司, 金盛克俊, 大和田勇人, エージェントモデルシミュレーションと機械学習を用いた顧客解約予測, 情報処理学会全国大会講演論文集, Vol.77, pp.2.303-2.304, 2015.

[4] 小野俊之, 吉川裕, 森田真弘, 薦田憲久, 定期預金残高増加見込み顧客予測方法, 電気学会論文誌 C, Vol.126, No.4, pp.556-562, 2006.

[5] Hoerl,A.E. and Kennard,R.W., Ridge Regression: Biased Estimation for Nonorthogonal Problems, *Technometrics*, Vol. 42, pp.55-67, 1970.

[6] R. Tibshirani, Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, Vol. 58, No. 1, pp.267-288, 1996.

[7] 中澤秀夫, ロジスティック回帰, 日本医科大学医学会雑誌, Vol. 10. No. 4, pp.186-191, 2014.

[8] 波部斉, ランダムフォレスト, 情報処理学会研究報告, Vol. 2012-CVIM-182, No. 31, 2012.

[9] 小野田崇, サポートベクターマシンの概要, オペレーションズ・リサーチ：経営の科学, Vol.46, No. 5, pp225-230, 2001.

[10] F. Rosenblatt, The perceptron: a probabilistic model for information storage and organization in the brain., *Psychological Review*, Vol. 65, pp386-408, 1958.

[11] D. E. Rumelhart, G. E. Hinton and R. J. Williams, Learning internal representations by error propagation , *Parallel distributed processing: explorations in the*

microstructure of cognition, Vol. 1, pp. 318-362, 1986.

[12] J. MacQueen, Some Methods for classification and Analysis of Multivariate Observations, *1. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press*, pp.281–297, 1967.

[13] David M. Blei and Michael I. Jordan, Variational Inference for Dirichlet Process Mixtures, *International Society for Bayesian Analysis*, Vol.1, No. 1, pp.121-144, 2006.

[14] 川崎能典, 多変量時系列に対する主成分・因子分析, 統計数理, Vol. 49, No. 1, pp.109-131, 2001.

[15] T. Cover and P. Hart, Nearest neighbor pattern classification, *IEEE Transactions on Information Theory*, Vol.13, No.1, pp.21-27, 1967.

[16] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng and Jörg Sander, LOF: Identifying Density-Based Local Outliers, *SIGMOD '00: Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pp.93-104, 2000.

[17] Yunqiang Chen, Xiang Zhou and S. Huang Thomas, One-Class SVM for Learning in Image Retrieval, *Proceeding IEEE Int'1 conference on Image processing*, 2001.

[18] 永田靖, MT システムの諸性質と改良手法, 応用統計学, Vol. 42, No.3, pp.93-119, 2013.

[19] Christopher Watkins, Learning From Delayed Rewards, Thesis (Ph. D.) King's College, Cambridge, 1989.

[20] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra and Martin Riedmiller, Playing Atari with Deep Reinforcement Learning, *NIPS*, 2013.

[21] D. Pelleg and A. Moore, X-means: extending k-means with efficient estimation of the number of clusters, *International Conference on Machine Learning (ICML)*, pp.727-734, 2000.

[22] Hamerly, Greg and Charles Elkan, Learning the k in k-means, *Advances in neural information processing systems*, Vol.16, pp.281-288, 2003.

[23] 木村 義政, 鈴木 章, 小高 和己, 遺伝的アルゴリズムを用いた類似文字識別のための特徴選択, 画像電子学会誌, Vol. 40, No. 1, pp200-207, 2011.

[24] Jacob Sakhnini, Hadis Karimipour and Ali Dehghantanha, Smart Grid Cyber Attacks

Detection Using Supervised Learning and Heuristic Feature Selection, *2019 IEEE 7th International Conference on Smart Energy Grid Engineering (SEGE)*, pp.108-112, 2019.

[25] Tohari Ahmad and Mohammad Nasrul Aziz, Data Preprocessing and Feature Selection for Machine Learning Intrusion Detection Systems, *ICIC Express Letters*, Vol.13, No.2, pp.93-101, 2019.

[26] Kennedy, J. and Eberhart, R., Particle Swarm Optimization, *in Proc. of The 1995 IEEE International Conference on Neural Networks*, Vol. 4, pp.1942–1948, 1995.

[27] B. Xue, M. Zhang, and W. N. Browne, Particle swarm optimization for feature selection in classification: A multi-objective approach, *IEEE Transactions on Cybernetics*, Vol. 43, No. 6, pp.1656-1671, 2013.

[28] 川村 篤志, Basabi Chakraborty, 進化的計算手法を用いた特徴選択アルゴリズムの提案, 第 31 回人工知能学会全国大会論文集, 3A2-2, 2017.

[29] 宇野光平, クラスタリングと特徴抽出の融合, 日本知能情報ファジィ学会誌, Vol.33, No.2, pp.57–63, 2021.

[30] 小林靖之, MT 法の原因診断における主成分分析適用の提案：マハラノビス距離の主成分分解による項目診断の高速化, 日本機械学会関東支部ブロック合同講演会講演論文集, pp.127-128, 2011.

[31] 大久保豪人, 永田靖, スパース・モデリングを応用したマハラノビス・タグチ法による異常検知, 情報処理学会第 79 回全国大会講演論文集, pp.2-51-2-52, 2017.

[32] 新村 諭, 浜 勉, MT 法と機械学習を用いたプレス加工音による金型摩耗の検知, 長野県工 業技術総合センター研究報告, Vol.13, pp.134-137, 2018.

[33] 石田秀一, 田原竜夫, 岩崎渉, 宮本弘之, 薄型 AE センサと MT システムの適用によるワイヤボンディングの接合状態評価, 第 30 回エレクトロニクス実装学会春季講演大会, pp187-190, 2016.

[34] 大久保豪人, マハラノビス・タグチ・システムにおける高次元データ解析法の展開, 横幹, Vol. 13, No. 2, pp.117-122, 2019

[35] Zhongfeng Wang, Yatong Fu, Chunhe Song and Peng Zeng, Power System Anomaly Detection Based on OCSVM Optimized by Improved Particle Swarm Optimization, *IEEE*, Vol. 7, pp.181580-181588, 2019.

[36] 多田毅, PSO アルゴリズムによる流出モデルパラメータの最適化, 水文・水資源学会誌, Vol.20, No.5, pp.450-461, 2007.

[37] 錦織 護直, 下川 朝有, 宮岡 悦良, アンサンブル法に基づく異常値検知について, 計算機統計学, Vol. 33, No. 2, pp.77-90, 2020.

[38] 近藤久, 浅沼由馬, 人工蜂コロニーアルゴリズムによるランダムフォレストとサポートベクトルマシンのハイパーパラメータの最適化と特徴選択, 人工知能学会論文誌, Vol. 34, No. 2, pp.G-I36_1-11, 2019.

[39] 木津左千夫, 澤井秀文, 遠藤哲郎, パラメータ不要の遺伝的アルゴリズム, 電子情報通信学会論文誌, Vol. J81-D-Ⅱ, No.2, pp.450-452, 1998.

[40] 近藤史孝, 藪下良樹, 渡邊俊彦, 分散化したパラメータ不要の遺伝的アルゴリズム, 第 26 回ファジィシステムシンポジウム論文集, Vol.26, pp.479-481, 2010.

[41] 相馬将太郎, 金子修, 藤井隆雄, 一回の実験データに基づく制御器パラメータチューニングの新しいアプローチ, システム制御情報学会論文誌, Vol.17, No.12, pp. 528-536, 2004.

[42] G. Taguchi and R. Jugulum, The Mahalanobis-Taguchi Strategy: A Pattern Technology System, *John Wiley and Sons*, 2002.

[43] 田口玄一, MT システムにおける技術開発, 日本規格協会, 2002.

[44] D. Goldberg, Genetic Algorithm in Search, Optimization, and Machine Learning, *Addison-Wesley*, 1989.

[45] J. H. Holland, Escaping Brittleness: The Possibilities of General- Purpose Learning Algorithms Applied to Parallel Rule-Based Systems, *Machine learning*, pp.593-623, 1986.

[46] 救仁郷誠, 解説 マハラノビスの距離 入門, 品質工学, Vol.9, No.1, pp.13-21, 2001.

[47] 立林和夫, 入門タグチメソッド, 日経 BP マーケティング, 2009.

[48] 鴨下隆志, 高田圭, 高橋和仁, おはなし MT(マハラノビス・タグチ)システム―予測・推測の可能性を広げる品質工学手法, 日本規格協会, 2004.

[49] 永田 靖, MT システムの研究, 関西学院大学国際学研究, Vol.6, No.2, pp.29-36, 2017.

[50] 間ケ部明, 高田圭, 矢野宏, はんだ自動外観検査へのマハラノビスの距離の適用, 品質工学, Vol.6, No.6, pp.66-73, 1998.

[51] 高濱正幸, 三上尚高, ガスタービンプラントの異常予兆検知, 品質工学, Vol.20, No.4,

pp.437-443, 2012.

[52] M. Ohkubo and Y. Nagata, Anomaly detection in high-dimensional data with the Mahalanobis-Taguchi system, *Total Quality Management & Business Excellence*, Vol.29, No.9-10, pp.1213-1227, 2018.

[53] 佐々木友弥, 國島丈生, MT 法を用いたトランプゲームのイカサマ行為防止のための異常検知システム, 第 21 回 IEEE 広島支部学生シンポジウム論文集, pp.136-141, 2019.

[54] 伊庭斉志, 遺伝的アルゴリズムと進化のメカニズム, 岩波書店, 2002.

[55] 山村雅幸, 小野貴久, 小林重信, 形質の遺伝を重視した遺伝的アルゴリズムに基づく巡回セールスマン問題の解法, 人工知能学会誌, Vol.7, No.6, pp.1049-1059, 1992.

[56] 野口博範, 大森健児, 遺伝的アルゴリズムによる重み付きグラフの多分割について, 情報処理学会論文誌, Vol.44, No.9, pp.2383-2389, 2003.

[57] Joe H. Ward, Jr. , Hierarchical Grouping to Optimize an Objective Function, *Journal of the American Statistical Association*, Vol. 58, Issue. 301, p. 236-244, 1963.

[58] 久保尚輝, 今村幸祐, 橋本秀雄, x-means クラスタリングによるクラスタ数を用いた動オブジェクト抽出, 電子情報通信学会技術研究報告, Vol.110, No.189, pp.23-28, 2010.