機械学習を用いた会計データのリスク定量化による内部監査業務の効率化

太田 雄也，藤井 徹

近年、不適切会計の開示企業数は増加傾向にあり、内部監査の実施によるリスク低減が重要である。そのため、監査人による会計データのリスク確認が必要だが、膨大な会計データに対し網羅的に人手で確認することは現実的ではない。

本稿では、BRF（Balanced Random Forest）をベースに、摘要欄の活用と監査人のリスク分析ノウハウを考慮したモデル構造の採用により内部監査特有の技術課題を解決した異常検知手法を提案する。これにより、会計データのリスクを異常スコアとして定量化でき、網羅的かつ効率的なリスク分析が可能となることを示す。

自社の会計データに提案手法を適用し、異常スコア上位 2%を閾値とした場合の再現率が 72.1%に達することを確認した。さらに、実際の内部監査業務に活用し、従来の人手でのリスク分析と比較して 1 会計単位あたりの分析時間を 43%削減することができた。

Risk Prediction in Accounting Data in Internal Audit using Machine Learning

OTA Yuya and FUJII Toru

In recent years, the number of companies disclosing inappropriate accounting has been increasing, and it is important to conduct internal audits to reduce risks. Therefore, it is necessary for auditors to inspect the risks of accounting data. However, it is difficult to manually inspect all the large amount of accounting data accumulated daily.

In this paper, we propose an anomaly detection method that addresses the unique technical challenges of internal auditing by leveraging the Balanced Random Forest (BRF) approach, utilizing descriptions, and incorporating auditors' risk analysis processes into the model structure. This method allows for the quantification of accounting data risks as anomaly scores, enabling comprehensive and efficient risk analysis.

We applied the proposed method to our company's accounting data and confirmed that the recall rate reached 72.1% when the top 2% of anomaly scores were used as the threshold. Furthermore, by applying this method to actual internal audit operations, we were able to reduce the analysis time per accounting unit by 43% compared to traditional manual risk analysis.

1. まえがき

1.1 背景

近年国内における不適切会計の開示件数は増加傾向にある[1)]。そのため、内部監査の実施を通して社内の内部統制意識を向上させることで、不正行為は当然のこと、会計処理におけるミスやムダの発生に対するリスクを低減させることが重要である。具体的には、会計および内部統制に関する知見・ノウハウを有する社内の担当者が、監査人として会計データを確認してミス・ムダ・不正につながるレコードを抽出するリスク分析を行う必要がある。しかし、膨大な会計データに対して人手で網羅的な分析を行うことは困難である。

この問題の解決策の一つとして、会計データを効率的に分析するためのCAAT（Computer Assisted Audit Techniques, コンピュータ利用監査技法）ツールが商用化されてい

Contact : *OTA Yuya* yuya.ota@omron.com

る[2]。しかし、高リスクなレコードを抽出するための仮説の洗い出しや、具体的な分析手法の選定・構築は監査人自身で行う必要がある[3]。そのため、リスク分析の効率化に対する効果は限定的、かつリスク分析の質は監査人の知見・ノウハウに依存する。もう一つの解決策として、外部監査においては機械学習を用いて会計データの不正のリスクを異常スコアとして定量化し異常検知を実現することで、監査業務を効率化する試みがなされている[4]。一方で内部監査を対象とした先行研究は公開されていない。これは、外部監査における異常検知手法を内部監査に適用して効率化を実現するには課題が存在するためである。それらの課題は、外部監査と内部監査の目的と業務内容の違いから説明できる。

1.2 内部監査におけるリスク分析の特徴

外部監査は企業の財務情報の信頼性を担保することが目的であり、監査法人や公認会計士などの第三者機関によって行われる監査である。一方、内部監査は企業の経営目標に対する経営および業務の適切性を確認し改善することが目的である。そのため、業務の不正や正確性（ミス）だけでなく、効率性（ムダ）も対象とすることが特徴である[5]。内部監査におけるリスク分析を対象とした異常検知手法では、機械学習の観点から以下の特徴を考慮する必要がある。

（1）正常データに対して異常データが著しく少ない（不均衡データ）
（2）入力に数値データ・カテゴリデータに加えテキストデータも含まれる
（3）真値に対してラベル付きデータは一部にのみ存在する（半教師あり）

（1）は外部監査・内部監査に共通する特徴であり、（2）（3）は内部監査において重視される特徴である。（1）は、膨大な会計データと比較するとミス・ムダ・不正の疑いがあるレコードがごく僅かであることによる。このようなデータに対し対策なく学習を行うと、すべてのデータに対して正常であると予測する無意味なモデルを構築してしまう可能性がある。

（2）は、会計処理の背景にある内部統制意識を確認するために摘要欄も重視することによる。摘要欄とは取引内容をわかりやすくするために記入されるフリーテキストの項目であり、取引の詳細や用途、特記事項などが記載される。外部監査では、財務情報の信頼性担保の観点から金額や日付、勘定科目といった数値データやカテゴリデータを主な分析対象とする。一方で、内部監査ではそれらに加えてテキストデータである摘要欄も考慮する必要がある。

（3）は、ミス・ムダ・不正の疑いがあるレコードすべてにラベル付けすることは現実的ではないことによる。外部監査では不適切会計として公開された情報を教師ラベルとして扱うことができる。一方、内部監査の対象であるミス・ムダ・不正の疑いがあるレコードは軽微なものまで含めると膨大な件数となる。また、取引実態の調査には工数を要するためすべての対象レコードを抽出することは現実的ではない。このようなデータで学習を行い、新たなレコードに対して異常スコアを予測すると、真値では高リスクとなるレコードであっても異常スコアが低く予測される可能性がある。

本稿の貢献は、内部監査の特徴（2）（3）を考慮した教師あり機械学習手法を構築し、内部監査におけるリスク分析に有効な異常検知を実現することにある。本稿では、特徴（1）のみ考慮した外部監査で活用実績のある機械学習手法をベースに、まず特徴（2）への対応として摘要欄を対象とした監査人のリスク分析の知見・ノウハウに基づいた特徴量の抽出を行った。さらに、特徴（3）への対応として監査人のリスク分析プロセスを考慮したモデル構造を採用した。以上により内部監査を対象とした異常検知を提案する。

本稿の構成は次の通りである。2 章では特徴（1）への対応として外部監査での機械学習手法を述べたうえで、特徴（2）（3）に対する関連研究を述べる。3 章では提案手法を述べる。4 章では自社の会計データを用いて提案手法の有効性を示す。最後に 5 章で本稿のまとめを示すとともに、他の業務プロセスへの展開による事業貢献の展望を述べる。

2. 先行研究

2.1 外部監査における異常検知

外部監査を対象とした異常検知の先行事例として文献 6）がある。文献 6）では、まず会計データから抽出した特徴量を不正リスクの種別毎に分類する。そして、各種別に対して、不均衡データに対応可能な教師あり機械学習手法の Balanced Random Forest（BRF）を適用し、サブモデルを構築する。最後にサブモデルを統合して判定結果を得る。

監査人は、蓄積された経験知に基づいて会計データから多岐にわたる特徴を捉えたうえで、様々な不正の手口を想定して当該レコードのリスクを判断する。そのため、説明変数が高次元であることと、識別モデルが高い非線形性を有することが前提となる。BRF は決定木を用いた識別モデル Random Forest（RF）をベースとしていることから、これらの前提にも適した手法である。これらの前提は内部監査でも共通する。

2.1.1 Balanced Random Forest（BRF）

BRF は、RF に対し不均衡データ適用時の課題解決策を施した機械学習手法である。RF では、復元抽出によるサンプリング（ブートストラップサンプル）を複数回行ったうえで各ブートストラップサンプルから決定木を構築し、それらの多数決などにより識別を行う。一方、BRF では各ブートストラップサンプルから決定木を構築する前に、各クラスラベルのサンプル数が同数になるように多数派のクラスラベルのサンプル数を減らすことでデータの不均衡を解消する手法である（図 1）。

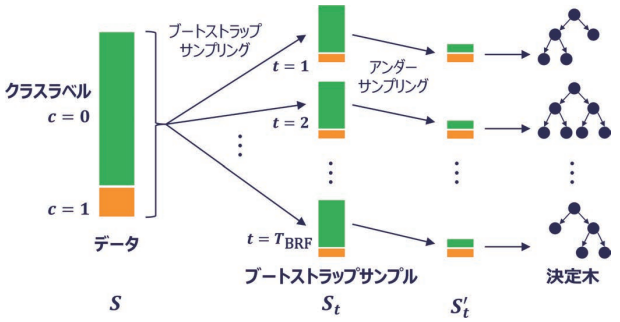

図 1　BRF の概要

まず、学習時ではデータ S から T_{BRF} 個の決定木を構築する。S から抽出した t 番目 $(t=1,\ldots,T_{\mathrm{BRF}})$ のブートストラップサンプル S_t において、$c=0$ のデータ数 $n_{c=0}$ と $c=1$ のデータ数 $n_{c=1}$ を比較し、$n_{c=0}=n_{c=1}$ となるように、多数派のクラスラベルのデータからランダムに間引き（アンダーサンプリング）を行うことでサンプル S'_t を得る。S'_t に対して決定木アルゴリズムを適用することで 1 つの決定木が構築される。すべてのブートストラップサンプルに対して同様の処理を行うことで T_{BRF} 個の決定木が構築される。t 番目の決定木のある葉ノードに $x \in S$ が属するとき、一つの決定木の葉ノードにおけるクラスラベルの予測確率 $p_t(c \mid x)$ は、葉ノードにおける各クラスラベルのデータの割合で表すことができる。

次に、テストデータ $x_i(i=1,\ldots,N)$ の各クラスラベル $c \in \{0,1\}$ の予測確率 $p(c \mid x_i)$ を求める。まず、t 番目の決定木に注目すると、x_i は決定木の条件分岐に従ってある葉ノードに属することになる。従って、x_i に対する t 番目の決定木の葉ノードにおけるクラスラベルの予測確率 $p_t(c \mid x_i)$ が推定結果となる。

以上より、最終的な予測確率 $p(c \mid x_i)$ は、T_{BRF} 個の決定木における予測確率の相加平均として得られ、次式で表される。

$$p(c|x_i)=\frac{1}{T_{\mathrm{BRF}}}\sum_{t=1}^{T_{\mathrm{BRF}}}p_t(c|x_i) \tag{1}$$

2.2 関連研究

特徴（2）に対応する関連研究を 2.2.1 と 2.2.2 で、特徴（3）に対応する関連研究を 2.2.3 で述べる。

2.2.1 テキストデータの数値ベクトル化

特徴（2）に関し、テキストデータに対する異常検知手法として、数値ベクトル化と外れ値検知を組み合わせる手法が提案されている[7]。具体的には、テキストデータを数値ベクトル化したのち、得られた数値ベクトルに対して外れ値検知を行うことで異常なテキストデータを検知する方法である。

数値ベクトル化はテキストを単語単位に分割したうえで、各単語を何らかの数値に変換することで、テキストを数値ベクトルに変換する手法である。一般的な手法として、BoW（Bag of Words）や TF-IDF（term frequency - inverse document frequency）がある。

・BoW

BoW は文書における各単語の出現回数である。N_{doc} 個の文書が N_{word} 個の単語で構成されるとする。また、i 番目の文書 $d_i(i=1,\ldots N_{\mathrm{doc}})$ の単語 $w_j(j=1,\ldots,N_{\mathrm{word}})$ の出現回数を n_{ij} とする。このとき、文書 d_i に対する BoW は次式を要素とする N_{word} 次元のベクトルとなる。

$$\mathrm{bow}(w_j \mid d_i)=n_{ij} \tag{2}$$

・TF-IDF

TF-IDF は、BoW に対して全文書における各単語の珍しさを重視した指標である。文書 d_i に対する TF-IDF は次式を要素とする N_{word} 次元のベクトルとなる。ここで、$|\cdot|$ は要素数を表す。

$$\text{tf-idf}\,(w_j \mid d_i)=\mathrm{tf}_{ij}\cdot\mathrm{idf}_j \tag{3}$$

$$\mathrm{tf}_{ij}=\frac{n_{ij}}{\sum_{k=1}^{N_{\mathrm{word}}}n_{ik}} \tag{4}$$

$$\mathrm{idf}_j=\log\frac{N_{\mathrm{doc}}}{\left|\{d_i : w_j \in d_i\}\right|+1} \tag{5}$$

TF は BoW を各文書の単語数で除した値であり、文書における各単語の出現割合を示している。IDF は全文書に対して、ある単語が出現する文書数の割合の逆数であり、各単語の全文書における珍しさを示している。これらを掛け合わせることで TF-IDF が得られる。

2.2.2 外れ値検知

外れ値検知は、正常と異常を識別する境界を直接推定する識別モデルベース、推定した正常の確率分布に基づいて生起確率の低いデータを異常と見なす確率分布ベース、異常と正常は距離的に離れている前提のもとデータ間の距離に基づいて異常を判別する距離ベースの 3 つに分類され

る[8]。摘要欄は部門や勘定科目によって記載内容の傾向が大きく異なるため、正常データの確率分布の仮定を要する確率分布ベースの手法は適さない。従って本節では、提案手法として採用した識別モデルベースの OC-SVM（One Class-Support Vector Machine）および、距離ベースの LOF（Local Outlier Factor）と IF（Isolation Forest）について述べる。ここでは、数値ベクトル化により文書 d_i から得られる N_{word} 次元の数値ベクトルを $v_i=(v_{i,1},\ldots,v_{i,N_{\text{word}}})$ とする。

・OC-SVM

OC-SVM[9] は、教師ありの識別タスクで用いられる機械学習手法 SVM（Support Vector Machine）を教師なしの外れ値検知に改良した手法である。図 2 に OC-SVM の概念図を示す。

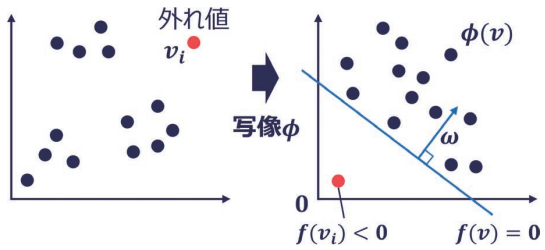

図 2　OC-SVM の概要

図内の赤色のデータ点が検知すべき外れ値とする。まず、異常なデータほど原点に近くなるような写像 ϕ によって特徴空間にデータ点群を変換する。その後、異常データの割合 $\rho\in(0,1]$ をハイパーパラメータとして、特徴空間において正常データと異常データを識別する超平面 $f=0$ を作成する。すると、数値ベクトル v_i に対して関数 f は次式で表される。ここで、ω は特徴空間において超平面と直交し、超平面から原点方向を負とするベクトルである。

$$f(v_i\,|\,\rho)=\omega\cdot\phi(v_i)-\rho \tag{6}$$

式 6 より、外れ値を判別する閾値を 0 として、$f(v_i)$ が正であれば正常データ、負であれば異常データと予測される。従って、$f(v_i)$ の値を用いて異常スコアが定義できる。

・LOF

LOF[10] は各データ点に対する周囲のデータ点群までの距離にその密集度合いを加味して、どの程度乖離しているかを異常スコアとして算出する教師なしの外れ値検知手法である。LOF の概念図を図 3 に示す。

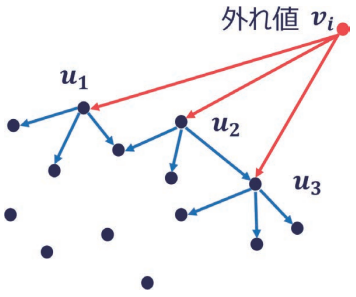

図 3　LOF の概要

数値ベクトル v_i に対する LOF による異常スコアは次式で定義される。

$$\mathrm{lof}(v_i\,|\,k)=\frac{1}{k}\sum_{u\in N_k(v_i)}\frac{\mathrm{dist}_k(v_i)}{\mathrm{dist}_k(u)} \tag{7}$$

ここで、$N_k(v_i)$ は v_i の k 近傍であり、v_i からのユークリッド距離が近い上位 k 個のデータ集合を表す。また、$\mathrm{dist}_k(v_i)$ は v_i の k 近傍内の各データ点 u から v_i への近傍有効距離の平均である。近傍有効距離 $l_k(u\to v_i)$ は次式で定義される。ここで、$\varepsilon_k(u)$ は u の k 近傍のデータ点をすべて含む u を中心とする最小球の半径である。

$$\mathrm{dist}_k(v_i)=\frac{1}{k}\sum_{u\in N_k(v_i)}l_k(u\to v_i) \tag{8}$$

$$l_k(u\to v_i)=\begin{cases}\varepsilon_k(u), & v_i\in N_k(u)\wedge u\in N_k(v_i)\\ \mathrm{dist}(u,v_i), & \text{otherwise}\end{cases} \tag{9}$$

LOF の異常スコアは式 8 の定義より正常なデータであれば 1 に近くなり、異常なデータほど異常スコアが高くなることから、一般に外れ値を判別する閾値は 1.5 とされる。

・IF

IF[11] は対象のデータ点を二分木によって孤立させるときに必要な二分木の分割回数（二分木の深さ）に基づいて異常スコアを算出する手法である。IF の概念図を図 4 に示す。

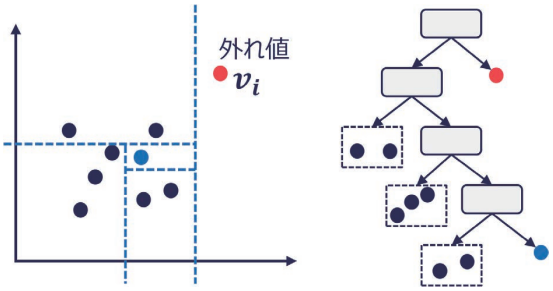

図 4　IF の概要

同図より、疎な領域に存在する外れ値のデータ点は必要な分割回数が少なく、密な領域に存在する正常なデータ点は分割回数が多くなるため、二分木の深さ（浅さ）が異常

スコアとして利用できることが分かる。数値ベクトル v_i に対する IF による異常スコアは次式で定義される。

$$s(v_i \mid \psi) = 2^{-\frac{E[h(v_i)]}{c(\psi)}} \tag{10}$$

$$c(\psi) = H(\psi - 1) - \big(2(\psi - 1)\big)/\psi \tag{11}$$

$$H(i) = \log i + \gamma \tag{12}$$

ここで、ψ は文書数 N_{doc} からの T_{IF} 回繰り返すサブサンプリングサイズで、$E\big[h(v_i)\big]$ は数値ベクトル v_i に対する T_{IF} 個の二分木の深さの平均である。$c(\psi)$ は各データ点を孤立させる（≒二分木探索に失敗する）ときの二分木の深さの平均であり、γ はオイラー定数（≈ 0.5772156649）である。式 10 より異常スコアは（0,1］の値をとり、正常なデータの異常スコアは 0 に近く、異常なデータの異常スコアは 1 に近くなり、外れ値を判別する閾値は 0.5 となる。

2.2.3 ラベル付きデータ不足への対応

特徴（3）への対応方法は大きく 2 つに分類される。一つは、機械学習分野において半教師あり学習と呼ばれる手法である。既存のラベル付きデータの内挿または近傍のデータは同種のクラスラベルを持つという前提のもと、ラベルなしデータのクラスラベルを推定し、擬似的なラベル（疑似ラベル）を付与するアプローチである[12)]。しかし、内部監査におけるリスク分析では、多岐にわたる分析観点の組合せによってリスクが判断されている。そのため、監査人が知見・ノウハウとして保有する分析観点の組合せ全体から想定される真値のドメインに対して、ラベル付きデータはカバーしていない可能性が高い。この場合、半教師あり学習では真値のドメインの一部にしか疑似ラベルが生成できない。

もう一つは先見知識を活用することで、少ないラベル付きデータからでも汎用的なモデルが得られるようにモデル構造を設計する方法である。具体的な方法は適用対象によって異なる。最も簡易な例として、適用対象の非線形性が強くなく線形性を仮定できることが分かっている場合に、深層学習などの非線形性が強い手法ではなく、線形回帰を採用することなどが挙げられる。

3. 提案手法

3.1 全体像

提案手法は 3 つのステップで構成される。ステップ 1 は特徴抽出で、会計データから監査人のリスク分析の知見・ノウハウに基づいて特徴量を抽出するステップであり、摘要欄からの特徴抽出も本ステップで行う。ステップ 2 はサブモデルの構築で、ステップ 1 で抽出した特徴量に対して監査人の分析観点毎に BRF を適用することで分析観点毎のモデル（サブモデル）を構築する。ステップ 3 はモデル統合で、ステップ 2 で構築したサブモデルを統合し最終的なモデルを得る。以下では、各ステップの内容を示すとともに、特徴（2）（3）への対応方法の詳細を示す。

3.2 ステップ 1：特徴抽出

3.2.1 概要

監査人は会計や内部統制の知識および実務経験に基づいた様々な分析観点を持っている。分析観点とは会計データからミス・ムダ・不正の兆候を判断する観点であり、例えば、「取引額が他の取引と比べて異常に高い」「仕訳日がある特定日に近い」「摘要欄が不記載」などのように金額や仕訳日、摘要欄といった会計データの項目とその見方から構成される。特徴量は分析観点を定量的に表現するために会計データから算出されるものであり、例えば「月次取引額の前年度同月比」「仕訳日から特定日までの日数」などが特徴量として考えられる。

監査人へのヒアリングを通じ「金額の異常」などの分析観点のカテゴリ（分析観点カテゴリ）を整理したうえで、各分析観点カテゴリに属する特徴量をデータサイエンティストが設計・実装した。抽出した 10 種の分析観点カテゴリ計 147 種の特徴量を表 1 に示す（グループについては 3.3 節で述べる）。ある会計単位のある勘定科目の会計データに N 件のレコードが存在する場合、i 番目のレコードから抽出した特徴量 $x_i(i = 1,\ldots,N)$ は、147 次元のベクトルとなる。

表 1　分析観点カテゴリの一覧

グループ	分析観点カテゴリ	会計データ項目	特徴量数
主要	金額の異常	取引額	85
	摘要欄の珍しさ	摘要欄	19
層別	国固有のリスク	国	1
	勘定科目固有のリスク	勘定科目	11
調整	取引額が基準値超え	取引額	1
	仕訳作成プロセスの異常	仕訳作成プロセス種別	14
	仕訳日の異常	仕訳日	2
	取引の継続性の異常	取引額	2
	取引通貨の珍しさ	取引額通貨	3
	摘要欄内の単語固有のリスク	摘要欄	9

本稿では全特徴量の内、特徴（2）に関する特徴量「摘要欄の珍しさ」に絞って詳細な抽出方法を述べる。

3.2.2 特徴量「摘要欄の珍しさ」

「摘要欄の珍しさ」は、特定の期間・会計単位・勘定科目の会計データに含まれる摘要欄を俯瞰した際の珍しさを定量化した特徴量である。頻出する単語が使われているレコードは頻繁に発生する通常の取引である。一方、珍しい単語が使われているレコードは滅多に発生しない取引であり、リスク分析上注視すべきという考えに基づいている。

まず、摘要欄の記載内容を単語に分割する。摘要欄の記載内容は、「①」と「1」など等価な文字を統一する Unicode 正規化や、数値はすべて 0 に置換するなどの前処理が施されているものとする。単語の分割には多言語対応が容易なオープンソース自然言語処理ライブラリ spaCy[13] の言語毎（日本語, 英語, 中国語）の中規模モデル（表2）を用いた。

表 2　自然言語処理モデル

言語	モデル
日本語	ja_core_news_md
英語	en_core_web_md
中国語	zh_core_web_md

次に分割した単語を、BoW および TF-IDF により数値ベクトル化する。i 番目のレコードの摘要欄 $d_i (i = 1,\ldots N)$ に単語 $w_j (j = 1,\ldots, N_{\text{word}})$ が出現したとすると、得られる数値ベクトルは N_{word} 次元のベクトル $v_i = (v_{i,1}, \ldots, v_{i,N_{\text{word}}})$ となる。BoW および TF-IDF による数値ベクトルの要素はそれぞれ式 2 と式 3 より次式の通りとなる。

$$v_{i,j} = \text{bow}(w_j \mid d_i) \tag{13}$$

$$v_{i,j} = \text{tf-idf}(w_j \mid d_i) \tag{14}$$

続いて、数値ベクトル v_i に対して外れ値検知手法を適用することで、各レコードの摘要欄の異常スコア $a_{\text{desc}}(v_i)$ を得る。適用する外れ値検知手法は 2.2 節で述べた OC-SVM、LOF、IF の 3 種類である。

OC-SVM を適用する場合、式 6 に示した通り $f(v_i)$ が異常スコアとして利用できる。具体的には、異常データの異常スコアを正の値として取得するため、$f(v_i)$ の正負を入れ替える。すなわち、数値ベクトル v_i に対する OC-SVM による異常スコアは次式で得られる。

$$a_{\text{desc}}(v_i) = -f(v_i \mid \rho) \tag{15}$$

LOF、IF を適用すると、式 7 と式 10 よりそれぞれ異常スコアは次式で得られる。

$$a_{\text{desc}}(v_i) = \text{lof}(v_i \mid k) \tag{16}$$

$$a_{\text{desc}}(v_i) = s(v_i \mid \psi) \tag{17}$$

さらに、会計単位や勘定科目によって異常スコアのレンジや分布形状が異なるため、各異常スコアに対する閾値の判別結果（カテゴリデータ）と、異常スコアを降順で並べたときの順位（順序データ）も特徴量として算出した。このとき用いた閾値は 2.2 節で示した各手法の一般的な閾値とした。加えて、TF-IDF はそれ自体が異常スコアとしての性質を持つため、数値ベクトル v_i に対して次式で定義される TF-IDF の平均値も特徴量として算出した。

$$a_{\text{desc}}(v_i) = \frac{1}{N_{\text{word}}} \sum_{j=1}^{N_{\text{word}}} \text{tf-idf}(d_i, w_j) \tag{18}$$

以上より、数値ベクトル化手法 2 種と外れ値検知手法 3 種の組合せ 6 種それぞれに対して、異常スコアと異常スコアに対する判別結果と順位の 3 種の特徴量（18＝6×3）を抽出した。さらに、式 18 の TF-IDF の平均値を加えて、表 1 の「摘要欄の珍しさ」特徴量計 19 種を抽出した。

3.3 ステップ 2：BRF によるサブモデル構築

ステップ 1 で抽出した特徴量から主要な分析観点カテゴリ毎に BRF を適用しサブモデルを構築することで、分析観点カテゴリ毎の異常スコアを算出する。

まず、i 番目のレコードから抽出した特徴量 $x_i = (x_{i,1}, \ldots, x_{i,M})$ として、対応する分析観点カテゴリをリスク分析での使われ方に応じて 3 つのグループ（主要 G_m、層別 G_s、調整 G_a）に分類する。ここで、M は特徴量数 147 のことである。特徴量と各グループの対応を図 5 に示す。主要グループは各レコードのベースとなるリスクの程度を判別する特徴量であり、「金額の異常」や「摘要欄の珍しさ」などが該当する。層別グループは対象の部門や勘定科目に応じてリスクの考え方を変更する特徴量であり、「国固有のリスク」などが該当する。調整グループは主要グループに対して付加的にリスクを考慮する特徴量であり、「仕訳日の異常」などが該当する。各グループと特徴量 $x_{i,j}$ の対応関係を図 5 に示す。

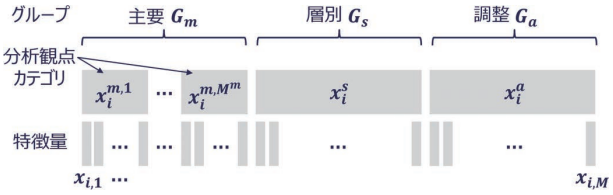

図 5　特徴量と分析観点カテゴリの対応

主要グループ G_m の l 番目の分析観点カテゴリに対応する特徴量を $x_i^{m,l}(l=1,\ldots,M^m)$、層別グループ G_s 全体に対応する特徴量を x_i^s、調整グループ G_a 全体に対応する特徴量を x_i^a とする。サブモデルは主要グループの各分析観点カテゴリに適用される。つまり、主要グループの l 番目の分析観点カテゴリ特徴量 $x_i^{m,l}$ と特徴量 x_i^s を説明変数、監査人のリスク判別結果を教師ラベル（目的変数）として BRF で学習しサブモデルを構築する。従って、M^m 個のサブモデルが構築される。なお、層別グループの特徴量 x_i^s は共通して用いられる分析観点であるという考えから、各サブモデル共通の説明変数として用いる。

得られたサブモデルを用いて各レコードの特徴量 $x=x_i$ を入力として、監査人が高リスク（$c=1$）と判別する予測確率（異常スコア）を出力する。主要グループの l 番目の分析観点カテゴリにおいて、BRF による予測確率は式 1 に基づいて次式で算出される。

$$\mathrm{score}(x\,|\,l)=p\left(c=1\,|\,(x^{m,l},x^s)\right) \tag{19}$$

3.4 ステップ 3：モデル統合

ステップ 2 で算出した各サブモデルの異常スコアを統合し、1 つの異常スコアにする。特徴（3）に対応するため、監査人の分析ノウハウに適した統合方法とする必要がある。提案手法におけるモデル構造の全体像を図 6 に示す。監査人はまず、主要な分析観点を用いて会計データを概観し大まかな絞込みを行う。そのうえで、主要な分析観点以外の分析観点に基づいて優先順位付けすることで高リスクなレコードを抽出する傾向がある。このとき、主要グループの分析観点による絞込みでは、特定の分析観点でリスクが低い場合でも抽出候補からは除外せず、いずれかの分析観点でリスクが高いと判断されれば抽出候補として残す傾向があることが分かった。

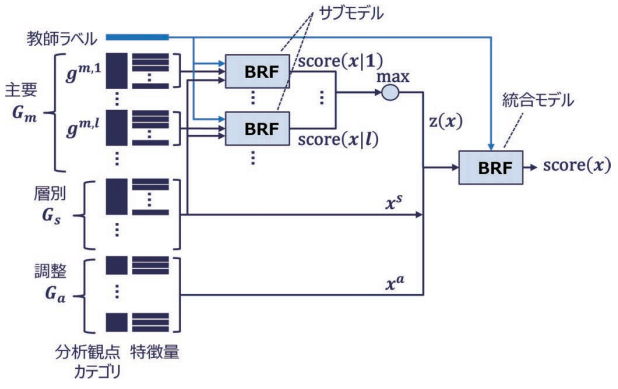

図 6　モデル構造の全体像

このような監査人の分析ノウハウを踏まえ、主要グループに対応するサブモデルの異常スコアを最大値により統合する。すなわち、次式で得られる値を新たな特徴量 z とする。

$$z(x)=\max_{l=1,\ldots,M^m}\{\mathrm{score}(x\,|\,l)\} \tag{20}$$

最大値を用いることで、いずれかのサブモデルの異常スコアが高ければ統合後の異常スコアが高くなるため、監査人の分析ノウハウに適した異常スコアを算出することが可能となる。最後に、新たな特徴量 z と層別グループの特徴量 x^a と調整グループの特徴量 x^s を説明変数として、BRF によるモデルを構築し（統合モデル）、1 つの異常スコアに統合する。このとき、層別グループの特徴量 x^a は調整グループの分析観点の使い分けにも用いるため、統合モデルでも説明変数とする。以上より、会計データの各レコードに対する特徴量 x を入力として、対象レコードの異常スコアは次式により定量化される。

$$\mathrm{score}(x)=p\left(c=1|(z(x),x^a,x^s)\right) \tag{21}$$

4. 効果検証

4.1 検証方法

実際の内部監査で用いる自社の会計データに提案手法を適用し、監査人のリスク分析の結果を教師ラベルとして量・質両面から有効性を検証する。ここで教師ラベルは高リスクか低リスクであり、監査人が実際のレコードを確認した結果である。なお、監査人が確認していないレコード（ラベルなしデータ）には低リスクのラベルを付与した。

まず量の観点として、監査人が異常スコアの高いレコードから優先的に抽出候補とする場合を考え、学習データにおける異常スコア上位 2%に基づいて閾値を設定する。ここで 2%という値は、監査人が目視でレコードを確認する作業を想定した時に業務上許容されるレコード数の上限の目安として設定した。そして質の観点として、テストデータにおける高リスクのレコードの内、異常スコアが閾値以上のレコード数（検出件数）の割合（再現率）を評価する。

検証に用いた会計データの内訳を表 3 に示す。事前検証として、学習データの期間を半年・1 年・2 年と変えた場合の、テストデータでの再現率を比較した。最も再現率が高い結果より、学習データの期間はテストデータの期間の直近 1 年間とした。

表 3　データの内訳

学習／テスト	年度	期	レコード件数	
			全体	高リスク
学習	2021	上期	1,850,264	1,259
	2021	下期	1,938,135	1,976
テスト	2022	上期	1,987,539	1,991

また、学習およびテストデータの期間中に新設・廃止された部門や子会社は検証対象から除外し、137 の会計単位を対象とした。会計単位とは、企業において個別に管理すべき会計の範囲のことである。対象の勘定科目は監査人が通常リスク分析の対象とする 140 種に限定した。

なお、実行環境は OS：Windows10、CPU：インテル Core i7-8700、動作周波数：3.2 GHz、RAM：16 GB の PC を用い、python 3.8、spaCy 3.6.0、scikit-learn 1.1.1、imbalanced-learn 0.9.1 にて実装した。imbalanced-learn[14] はアンダーサンプリングなど不均衡データに関する機能を提供するライブラリである。

提案手法の有効性を示すため、特徴（2）（3）への対応有無を変更した 3 つの手法を検証対象とする。1 つ目は、すべての特徴量のうち摘要欄から抽出される特徴量を除外し、サブモデルの異常スコアをそのまま統合モデルの入力とする手法（Desc(－)Max(－)）。2 つ目は、摘要欄から抽出される特徴量も含めたすべての特徴量を用いるが、サブモデルの統合には最大値を用いず、サブモデルの異常スコアをそのまま統合モデルの入力とした手法（Desc(＋)Max(－)）。3 つ目は、すべての特徴量を用いたうえで、最大値による統合を行う手法（Desc(＋)Max(＋)）である。なお、本検証では表 1 に示した通り「金額の異常」と「摘要欄の珍しさ」を主要グループとした。

また、OC-SVM、LOF、IF の各種ハイパーパラメータは会計単位・勘定科目毎に設定することが困難なため、一般的な初期値で固定した。BRF における決定木アルゴリズムの処理を規定する各種ハイパーパラメータはクロスバリデーション[15]（分割数 3）とベイズ最適化[16] により得た最適値を設定した。最適値探索における評価関数は F値[15] とした。

4.2 検証結果と考察

テストデータにおける各手法の再現率と検出件数を表 4 に示す。Desc(＋)Max(＋) の再現率が 72.1% と最も高く、Desc(－)Max(－) と比較して 3.8 ポイント高かった。

表 4　検証結果（再現率）

手法	再現率	検出件数
Desc(－)Max(－)	68.3%	1,360
Desc(＋)Max(－)	66.9%	1,332
Desc(＋)Max(＋)	72.1%	1,436

Desc(－)Max(－) と Desc(＋)Max(－) では検出できなかったが、Desc(＋)Max(＋) では検出できた事例をもとに、提案手法のポイントである「摘要欄の珍しさ」特徴量と最大値によるモデル統合の効果を確認する。対象とする事例の部門と勘定科目における各手法の異常スコアの違いを図 7 に示す。同図内横軸は取引額が高いレコードから順に並べた時の順位である。

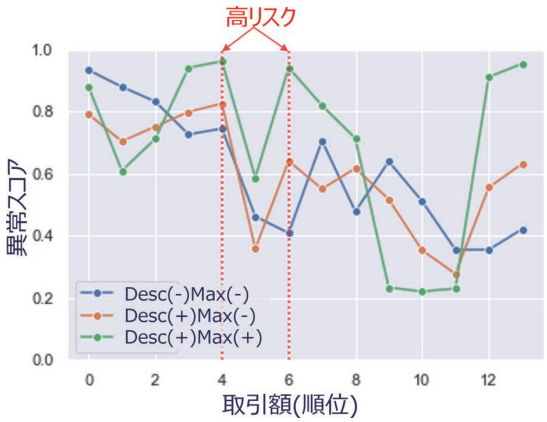

図 7　各手法の異常スコアの比較

Desc(－)Max(－) と Desc(＋)Max(－) では異常スコアが低いレコードの一部が、Desc(＋)Max(＋) では異常スコアが高いことが確認できる。実際、当該レコードでは摘要欄以外の金額などの分析観点では高リスクと判断できるような情報が乏しいが、摘要欄の内容から、滅多に発生しない珍しい用途の経費利用であることが分かった。これは通常発生しない取引であり、高リスクなレコードとして検出すべきである。このようなレコードに対し、Desc(－)Max(－) は摘要欄を考慮していないため検出が困難である。また、Desc(＋)Max(－) でも「摘要欄の珍しさ」以外の「金額の異常」の異常スコアが低いことに引っ張られ、最終的な異常スコアが低くなったと考えられる。

同様に複数レコードについて、監査人 2 名により妥当な異常スコアが算出されていることを確認した。さらに、検出できなかったレコードは、高リスクなレコードの中でも相対的に優先度が低く、監査人によって抽出要否が分かれるレコードであることを確認した。これは、リスク分析に異常スコアを用いることで監査人の属人性の低減にも効果があることを示唆している。

4.3 リスク分析工数の削減効果

Desc(＋)Max(＋) を実際のリスク分析業務に適用し、工数削減に対する効果を検証した。監査人には、各レコードの異常スコアが降順で併記された会計データを用いて、異常スコアが高いレコードから順に目視確認し、高リスクと判断したレコードを抽出してもらった。この時、実際に調査が行えるレコード件数は限られることから、抽出件数は従来と同程度とし、結果 1,176 件であった。なお、検証対象の会計データは、表 3 とは異なる期間、かつ監査人がリスク分析未実施の会計データとした。リスク分析実施済みのデータセットでは、既知であることがバイアスとなって分析時間が短くなる可能性が高いためである。

高リスクなレコードを抽出し終えるまでの分析時間を計測し、従来のすべて人手で行っていた場合の分析時間と比較した。なお、モデル学習と異常スコア推定にかかる処理

時間は分析時間に含めない。その結果、従来の1会計単位あたりの平均分析時間2時間40分に対し、Desc(＋)Max(＋)を用いた場合は1時間30分となり、43%削減となることを確認した（図8）。

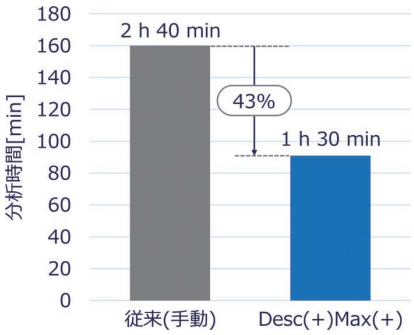

図8　分析時間の比較

従来の人手によるリスク分析では8.4人月/年の監査人の工数を要している。本検証より、監査人は異常スコアを用いたリスク分析を行うことで、従来の人手によるリスク分析での質を維持しつつ、3.6人月／年の工数削減が期待できる。

5. むすび

本稿では、内部監査において膨大な会計データの網羅的なリスク分析が人手では現実的に困難であることに対して、異常検知によるリスク分析業務の効率化を試みた。具体的には、BRFをベースとした機械学習モデルにより、ミス・ムダ・不正のリスクを定量化する異常検知手法を提案した。提案手法においては、テキストデータに対する異常検知手法を用いて摘要欄の珍しさを定量化することと、分析観点に応じてサブモデルを構築したうえで、監査人の分析ノウハウに適した最大値によるサブモデルの統合方法が、内部監査特有の技術課題解決のポイントであることを示した。

実際に、提案手法を自社の会計データに適用することで、前述の技術課題の解決策により監査人のリスク分析結果に対する再現率が72.1%に達することを確認した。さらに、監査人に異常スコアを参照しながらリスク分析を行ってもらうことで、従来の人手による分析方法と比較して、分析時間が大幅に削減されることを確認した。

本稿では内部監査におけるリスク分析を対象に検証を行った。リスク分析が対象とする業務を一般化すると、外部環境の変化が発生する中で、リスクに関する判断材料を収集し、リスクを見通し、対応を意思決定・アクションし、その結果を評価するサイクルを継続的に行う業務と捉えることができる。本稿で提案した手法は「リスクの見通し」における判断基準が属人化する課題に対して、形式知化を実現する解決策の一つである。一方で「判断材料の収集」「意思決定・アクション」においては、それぞれ外部環境の変化への追従や、意思決定・アクションの妥当性評価が課題として挙げられる。前者に対しては入力データ分布やラベルのコンテクストの変化を検知し対策するコンセプトドリフト適応が、後者に対しては予測の不確かさを定量的に評価する技術、例えばベイズ推定やConformal Predictionの活用が有効と考えられる。これらの技術確立によりリスク分析を核とする業務の工数削減と再現性・質向上の両立を目指す。具体的には、SCMでの需要予測に基づく在庫・生産管理や、製造設備の保全履歴や稼働データに基づく予兆保全などへの展開可能性がある。

参考文献

1）株式会社東京商工リサーチ. “2023年上半期の「不適切会計」開示　過去2番目の35社（36件）、最多はサービス業の9社.” 東京商工リサーチ. https://www.tsr-net.co.jp/data/detail/1197827_1527.html（Accessed: Jan. 17, 2024）.

2）伊集院大助. “「CAATツール」利用状況についてのアンケート結果 報告書.” 一般社団法人日本内部監査協会. https://www.iiajapan.com/pdf/kenkyu/c0041510.pdf（Accessed: Feb. 19, 2026）.

3）中村元彦, “会計監査におけるCAAT活用の影響と課題,” *現代監査*, vol. 2015, no. 25, pp. 162–170, 2015.

4）島田裕次, “監査業務へのAIの導入可能性 AI導入における課題,” *日本情報経営学会第84回全国大会*, pp. 59–60, 2022.

5）久保恵一, *図解 一番はじめに読む内部監査の本*, 第2版. 東洋経済新報社, 2010, pp. 44–45.

6）清水多賀雄, 深見英二, 宇宿哲平, “AI・機械学習を用いた不正リスク検知SUNモデル,” *KPMG Insight*, vol. 39, pp. 15–18, 2019.

7）S. Kalra et al., “Automatic classification of pathology reports using TF-IDF Features,” 2019.［Online］. Available: https://arxiv.org/abs/1903.07406.

8）吉澤亜耶, 橋本洋一, “異常検知技術の概要とその応用動向について,” *Intec Tech. J.*, vol. 17, pp. 42–47, 2016.

9）B. Schölkopf et al., “Support vector method for novelty detection,” in *NeurIPS 1999 Conf.*, 2019, pp. 582−588.

10）M. M Breunig et al., “LOF: Identifying density-based local outliers,” *ACM SIGMOD Rec.*, vol. 29, no. 2, pp. 93–104, 2000.

11）F. T. Liu et al., “Isolation-based anomaly detection,” *ACM Trans. Knowl. Discovery Data*（*TKDD*）, vol. 6, no. 1, pp. 1–39, 2012.

12）X. Zhu, “Semi-Supervised Learning Literature Survey,” 2007.［Online］. Available: https://pages.cs.wisc.edu/~jerryzhu/pub/ssl_survey_6_24_2007.pdf.

13）M. Honnibal et al. “spaCy: Industrial-Strength Natural Language Processing in Python.” spaCy. https://spacy.io（Accessed: Feb. 19, 2026）.

14）G. Lemaître et al. “GitHub - scikit-learn-contrib/imbalanced-learn: A Python Package to Tackle the Curse of Imbalanced Datasets in Machine Learning.” GitHub. https://github.com/scikit-learn-contrib/imbalanced-learn（Accessed: Feb. 19, 2026）.

15）井手剛, 杉山将, *異常検知と変化検知（機械学習プロフェッ*

ショナルシリーズ), 講談社, 2015, pp. 8–13.
16）J. Bergstra et al., “Algorithms for hyper-parameter optimization,” in *NeurIPS 2011 Conf.*, 2011, pp. 2546–2554.

執筆者

太田 雄也　OTA Yuya
ストラテジック R&D 本部
デジタルソリューションセンタ
専門：情報工学

藤井 徹　FUJII Toru
ストラテジック R&D 本部
デジタルソリューションセンタ
専門：情報工学、データ分析
所属学会：電子情報通信学会