研究報告

保険リスク予測のためのデータサイエンス
―クラスタリングとニューラルネットワーク―

一般社団法人　ＪＡ共済総合研究所
調査研究部　研究員
山名一史（やまな かずふみ）

アブストラクト

本稿では、保険リスクの予測にデータサイエンスを活用する方法として、先行研究の紹介とそこで用いられている機械学習・深層学習的手法の解説と考察を行う。

まず、保険市場のモデルを用いて、保険事故発生確率の正確な予測（リスク予測）の重要性を確認し、データサイエンスの活用可能性について議論した。次に、具体的なリスク予測手法として、クラスタリングとニューラルネットワークについて検討した。

データの分割がリスク予測に効果的な場合、標準的なクラスタリング手法であるK平均クラスタリングや階層的クラスタリングを用いることで、予測性能の改善が期待できる。また、複雑な関係を予測する場合、ニューラルネットワークを用いた方法が有力であり、とくに複数のネットワークを組み合わせた混合モデルを用いると、予測性能が高くなることが知られている。

データサイエンスは、大量のデータから有用な知識を探索、発見するために有効な方法で、高い予測能力が期待できる。予測性能には向上の余地があり、分析手法のさらなる改善が今後も期待される。

（キーワード）　K平均クラスタリング　階層的クラスタリング　ニューラルネットワーク

目　次

1．はじめに

保険契約は、当事者の一方に一定の事由が生じた場合、保険証券に記載された保険契約にかかる約款にしたがって財産上の給付（保険金）を相手方に行う義務を負い、その代わりに相手方から保険料を受け取る契約[1]であり、リスクに対するヘッジ手段として機能する金融商品である。保険商品を販売する側を保険者、保険証券の補償対象となっている側を被保険者と呼び、その保険がどのようなリスクに関連しているかによって、死亡保険、医療保険、自動車保険、火災保険など様々な種類の保険が存在する。

近年、様々な業界でよりよい意思決定を行うためにビッグデータや機械学習・深層学習などデータサイエンス手法の活用が試みられている。こうした取り組みは保険業界でも見られ、人工知能（ＡＩ）やビッグデータ、その他先端技術の活用はインシュアテック（InsurTech）と総称される。データサイエンスの応用先は幅広く、たとえばデータを活用した顧客の分析は、保険業界にとどまらず様々な業界で有用である。ただ、保険商品の特殊性を鑑みると、リスク評価と予測、不正請求の検知など業界特有の業務ニーズに対してデータサイエンスをどのように活用するかが、実務的により関心の高いところだろう。本稿では保険リスクの予測にデータサイエンスを活用する方法として、先行研究を紹介し、そこで用いられている分析手法の解説を行うことにする。

本稿の構成は以下の通りである。第２節では一般的な保険市場のモデルを使ってリスク予測の重要性を確認し、データサイエンスがどのように活用できるかを議論する。第３節ではリスク予測で使われる手法としてクラスタリング、具体的にはＫ平均クラスタリングと階層的クラスタリングの解説を行い、第４節ではニューラルネットワークについて解説する。

2．リスク予測の必要性

最初に、保険業界においてデータサイエンスを活用する目的を明確にするため、Einav, Finkelstein, Levin（2010）にしたがって一般的な保険市場のモデルを定式化してみよう。保険契約は保険者と保険契約者との間に結ばれる契約であり、これは保険の対象となる目的のベクトル$\boldsymbol{\phi}$と保険料のベクトル$\boldsymbol{p}$で表現される。一般に被保険者は契約者と同じとは限らないが、単純化のため、ここでは同一と仮定しよう。この契約者と被保険者との同一性の仮定は一般性を失うものではない。

被保険者はリスク態度や選好、学歴、所得など様々な性質のベクトル$\boldsymbol{\xi}$で表現できる。ここで、リスク態度や選好といった性質は直接観察できない一方、年齢や学歴などの性質は観察可能である。そこで、被保険者の性質のベクトル$\boldsymbol{\xi}$を観察できない性質$\boldsymbol{\xi}_u$と観察できる性質のベクトル$\boldsymbol{\xi}_o$とに分けておこう：

1　より正確な定義として、保険法第２条１号「保険契約、共済契約その他いかなる名称であるかを問わず、当事者の一方が一定の事由が生じたことを条件として財産上の給付（生命保険契約及び傷害疾病定額保険契約にあっては、金銭の支払に限る。以下「保険給付」という。）を行うことを約し、相手方がこれに対して当該一定の事由の発生の可能性に応じたものとして保険料（共済掛金を含む。以下同じ。）を支払うことを約する契約をいう。

$$\boldsymbol{\xi} = \{\boldsymbol{\xi}_u, \boldsymbol{\xi}_o\}.$$

被保険者（$\boldsymbol{\xi}$）が保険$(\boldsymbol{\phi}, \boldsymbol{p})$にどれだけの価値を見出すかは、保険期間内にどのような保険事故がどれくらいの確率で生じるかに依存して決まる。そこで、保険期間内に被保険者が取り得る行動の集合をA、生じうる保険事故の集合をMとする。自動車保険を例にとると、被保険者が自動車の運転中にどれくらい注意して運転するかの度合いが$a \in A$、運転中に交通事故の当事者になるかどうかが$m \in M$となる。被保険者が当事者になるかどうかは被保険者の性質や運転中の注意の度合いに依存して決まるので、保険事故mが生じる確率は$\pi(m \mid a, \xi)$と表現することができる。また、被保険者の効用はどのような保険事故が生じ、その際にどのような保険に加入しているかに依存して決まるので$u(m, a, \boldsymbol{\xi}, \boldsymbol{\phi}, \boldsymbol{p})$と表現できる。以上の議論から、標準的な期待効用仮説に基づいて被保険者が保険に見出す価値は、

$$v(\boldsymbol{\phi}, \boldsymbol{p}, \boldsymbol{\xi}) = \max_{a \in A} \sum_{m \in M} \pi(m|a, \boldsymbol{\xi}) u(m, a, \boldsymbol{\xi}, \boldsymbol{\phi}, \boldsymbol{p})$$

と表すことができる。ここで、被保険者は合理的なので、期待効用を最大化するように最適な行動$a^*(\boldsymbol{\phi}, \boldsymbol{p}, \boldsymbol{\xi})$を選択し、被保険者が最適な行動を選択したときに保険事故mが生じる確率は$\pi^*(m \mid \boldsymbol{\phi}, \boldsymbol{\xi}) \equiv \pi^*(m \mid a^*(\boldsymbol{\phi}, \boldsymbol{p}, \boldsymbol{\xi}), \boldsymbol{\xi})$となる。被保険者の性質が観察できるものと観察できないものとに分けられることから、$\pi^*(m \mid \boldsymbol{\phi}, \boldsymbol{\xi}) = \pi^*(m \mid \boldsymbol{\phi}, \boldsymbol{\xi}_u, \boldsymbol{\xi}_o)$と書き換えられることも併せて確認しておこう。

次に保険会社の保険金支払いについて考えよう。保険会社は生じた保険事故mと保険目的$\boldsymbol{\phi}$に基づいて保険金を支払う。この保険金支払い額を$\tau(m, \boldsymbol{\phi})$とすると、保険会社が支払う保険金の期待値は、

$$c(\boldsymbol{\phi}, \boldsymbol{\xi}) = \sum_{m \in M} \pi^*(m|\boldsymbol{\phi}, \boldsymbol{\xi}_u, \boldsymbol{\xi}_o) \tau(m, \boldsymbol{\phi})$$

となり、保険市場が完全競争的な場合は保険料と等しくなる（給付・反対給付均等の原則が成立する）[2]。この式から、保険金は（そして保険料も）被保険者の数だけでなく、被保険者の構成（つまり$\boldsymbol{\xi}$）にも依存することがわかる。

以上の議論から、保険会社がデータサイエンスを活用する目的を明確にできる。保険会社は収益を確保するため、保険金請求を賄え、かつ競争力のある保険料水準を設定する必要がある。適切な保険料水準を設定するには、保険金請求額の正確な見積もりが不可欠であり、保険事故が生じる確率$\pi^*(m \mid \boldsymbol{\phi}, \boldsymbol{\xi}_u, \boldsymbol{\xi}_o)$の正確な予測が求められる。観察可能な被保険者の性質ξ_oを使って、より正確な予測値$\hat{\pi}^*(m \mid \boldsymbol{\phi}, \hat{\boldsymbol{\xi}}_u, \boldsymbol{\xi}_o)$を得ることが、機械学習・深層学習などデータサイエンスの手法を活用する目的となる。

データサイエンスを用いたリスク予測の必要性を確認したところで、具体例として、将来支払うこととなる保険金等に対して保険会社が確保する積立について考えよう。積立額の見積もりには、決定論的（非確率的）な方法や確率的な方法が用いられてきた。決定論

2　単純化のために無視したが、実際には保険の解約が存在するので、保険会社は、保険料の変化が解約率に与える影響も考慮して最適な保険料の水準を決定する。解約率の予測はリスク予測と同様にデータサイエンスの主要な対象となっている。

的な方法は計算が簡単であるものの、支払いの見積もりに不可欠なリスクを定量化していないという欠点がある。リスクを定量化するためには$\pi^*(m \mid \boldsymbol{\phi}, \xi_w, \xi_o)$を考慮した確率的なモデルを用いる必要があり、典型的には一般化線形モデル（GLM）やそれを拡張したモデルが挙げられる[3]。ただ、GLMのような単純なモデルは線形性の仮定が強いため、予測精度を最大化するという観点からは必ずしも最適な選択肢でない。Smyth and Jørgensen（2002）は、Tweedie分布をベースにした一般化線形モデルを使って保険金の平均と分散を同時にモデル化し、予測モデルの柔軟性を高めることで予測精度の向上を図っている。

しかし、こうしたモデルの拡張を行ったとしても、保険データにおいて典型的に見られるような、多数のカテゴリカルな説明変数からなる複雑な関係をモデル化するには不十分なことが多い。保険契約の大部分から保険金が請求されず、データに切断（truncation）が生じている点も、分析を困難にしている要因である[4]。近年、こうした複雑な分析対象を扱う方法として注目されているのが、機械学習・深層学習的手法を用いたモデル化である（Christmann, 2004）。

3．リスク予測のためのデータサイエンス

観察可能な被保険者の性質ベクトルから、保険事故の発生確率や保険金の額を予測する問題は複雑であり、問題の複雑さに対処するには一般に大きく２種類の方法が考えられる。一つは、何らかの前処理を行うことで問題を単純化する方法、そしてもう一つは、より一般性のある関数形を使うことで、複雑さそのものをモデル化する方法である。この分野における単純化のための前処理としては、観察可能な性質ベクトルに基づいて被保険者を複数のリスクグループに分割する方法がよく用いられる。この方法の背景にあるのは、データを分割することで観察可能な性質と発生確率との関係を単純かつ同質的にできるのであれば、予測モデルは単純であっても、リスクグループ毎に行う予測精度は向上可能である、という仮説である。必ずしも適切なグループの分割方法が存在するという保証はないものの、問題の複雑さを回避する方法として試す価値は十分にあるだろう。

他方で、複雑さそのものをモデル化する場合、関数形を複雑にすれば、一般にそれだけモデル内パラメータの推定は難しくなる。そのため、推定難易度とモデルの複雑性とのトレードオフに十分な注意を払ったうえで、モデル選択を行う必要がある。以下では、前者の方法としてクラスタリング、後者の方法としてニューラルネットワークをそれぞれ検討しよう。

（1）クラスタリング

まず、複雑さを回避する方法として、データの分割を考える。データを分割するうえで重要なのは、サブサンプルのサンプルサイズ

3　たとえば、損害保険会社が保険契約準備金のうち、既経過責任部分を算出する決定論的なアルゴリズムとして、チェインラダー法やボーンヒュッター・ファーガソン法がある。GLMの拡張については、たとえばEngland and Verrall（2002）, Antonio and Beirlant（2007）, Bjökwallet al.（2011）などを参照のこと。
4　その他、保険金の分布の非対称性やファットテール性も分析を困難にしている要因である。

と同質性とのトレードオフである。いま、観察された性質と確率との間に以下のような真の関係を持つような被保険者のサンプルを仮定する[5]。

$$\pi_1^* = 0.9\xi_{o,1} + 2\xi_{o,2}$$
$$\pi_2^* = 2\xi_{o,1} + \xi_{o,2}$$
$$\pi_3^* = \xi_{o,1} + 2\xi_{o,2}$$
$$\pi_4^* = 1.8\xi_{o,1} + \xi_{o,2}$$
$$\pi_5^* = 0.6\xi_{o,1} + 1.9\xi_{o,2}$$

明らかに奇数番のサブサンプルと偶数番のサブサンプルとでは、観察された性質と発生確率の間に異なる関係が存在している。このようなケースでは、単一の関数形で予測を行うのではなく、異なるリスクグループに分けて予測を行ったほうが高い精度を期待できるだろう。他方で、サブサンプルへの過度な分割は、個々のサブサンプルのサンプルサイズが減少することによって推定精度の低下を招くので好ましくない。データを分割する際は、グループ内の同質性とグループ間の異質性のバランスに配慮する必要がある。

リスクグループは年齢や性別、自動車保険であれば初度登録（検査）年月や車両型式など、発生確率に影響を与えそうな（または、経験的に影響を与えることが知られている）少数の性質に応じて分割する方法がもっとも単純だろう。しかし、こうした恣意的な方法には改良の余地があるかもしれない。そこで、本稿では、Yeo et al.（2001）が用いているK平均クラスタリング、Williams and Huang（1997）が用いている階層的クラスタリングという２つの手法について検討してみよう。

クラスタリング（clustering）とは、似ているデータをクラスター（cluster）というグループに分ける方法全般のことである。先行研究で用いられているクラスタリング手法は、どちらもデータを互いに疎な（排他的な）クラスターに分ける手法である。このように重複を許さないクラスタリングをハードクラスタリング（hard clustering）と呼ぶのに対し、重複を許すクラスタリングはソフトクラスタリング（soft clustering）またはファジークラスタリング（fuzzy clustering）と呼ばれる。K平均クラスタリングと階層的クラスタリングの最大の違いは、前者がサンプルをあらかじめ定めたK個のクラスターに分けるのに対し、後者はクラスターの数を事前に定めずに分ける点である。

ここで、クラスタリングと混同しやすい手法である分類についても言及しておこう。分類とはそのデータがどのグループに属するのか、正答に基づいてグループ分けを行う方法である。クラスタリングは正答が無い状態でグループを分ける方法なので、分類は教師あり学習（supervised learning）と呼ばれるのに対し、クラスタリングは教師なし学習（unsupervised learning）と呼ばれる。教師あり学習は分類精度がどれくらいなのか、正答に基づいて検証することができる一方、教師なし学習には正答がないため、事後的に分割がどの程度正確だったのか、その精度を評価することが困難であり、教師あり学習を用いた場合とは異なる性能指標が求められることも併せて確認しておこう。

5　言うまでもなく、この真の確率は事前にも事後にも観察不可能である。

（1）－1　K平均クラスタリング

先ほど確認したように、グループ分けをする際は、グループ内の同質性とグループ間の異質性とのバランスに配慮する必要があるので、クラスタリングを行う際は同質性、つまり個々のデータがどの程度似ているのか、その類似度をどのように測るかが肝となる。K平均クラスタリングは、類似度を測るためクラスター内変動（within－cluster variation）と呼ばれる概念を導入する。いま、k番目のクラスターをC_kとすると、そのクラスター内変動は$W(C_k)$と書け、K平均クラスタリングは

$$\min_{C_1,\ldots,C_K} \sum_{k=1}^{K} W(C_k),$$

という総クラスター内変動の最小化問題として定義することができるようになる。総クラスター内変動はクラスター慣性（inertia）とも呼ばれ、ユークリッド2乗距離を類似度の指標として用いるので、

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{o=1}^{O} \left(\xi_{o,i} - \xi_{o',i'}\right)^2,$$

と表される。ここで、$|C_k|$はクラスター内の被保険者数、Oは観察できる被保険者の性質の数を表す。

最小化問題を数式で定義するのは簡単だが、N人の被保険者をK個のクラスタリングに分ける組み合わせはK^N通り存在するため、大域的最適解を得るのはそれほど容易ではない。そこで、大域解の代わりに局所的最適解を得るアルゴリズムを用いることが多い。具体的には、

1．K個の点を無作為に選ぶ。この点は重心（centroid）と呼ばれる。
2．各観測値を最も近い重心に割り当てる。近さはユークリッド距離に基づいて定義する。
3．各重心に割り当てられた観測値の中心を新しい重心として、重心の更新を行う。
4．2に戻って、重心の更新が行われなくなるまで、これを繰り返す（iteration）。

K平均クラスタリングの最大の問題点は、事前にクラスターの数を指定する必要があることである。可視化が可能な低次元データであれば、可視化した結果に基づいてKの数の当たりを付けることができる場合もあるだろう。しかし、被保険者の性質が多く含まれるなど高次元データの場合、可視化は非現実的であり、このような当て推量は難しい。また、事前にも事後にも真のクラスター数が分からない以上、分割精度の検証には限界があり、分析過程から恣意性を完全に排除するのは困難である。

客観的にクラスターの数を指定する方法としては、複数のKについて一通り分析を行ってみて、慣性が最小になるようなKを選ぶ方法がまず考えられる。具体的には、Kごとに慣性をプロットし、慣性の値が大きく変わるようなKの値（屈曲点）を探せばよい。慣性をプロットしたものがひじの形に見えることから、この方法はエルボー法（elbow method）と呼ばれる。

エルボー法は客観的にKの値を探す手段として有用だが、Kを選ぶ際にグループ間異質性が適切に考慮されていないという欠点がある。そこで、エルボー法の代わりに$(b-a)/\max(a,b)$で定義されるシルエット係数（silhouette coefficient）を用いる方法も提案

されている。aはクラスター内の同質性の指標として、ある観測値とその観測値が含まれるクラスターに含まれるそれ以外の観測値との距離の平均を計算することで得られる平均クラスター内距離を表し、凝集度とも呼ばれる。また、bはクラスター間の異質性の指標として、観測値とその観測値が含まれるクラスターを除いた中で最も近くにあるクラスターに属する観測値との距離の平均を計算することで得られる平均最近傍クラスター間距離を表し、乖離度とも呼ばれる。グループ内同質性が大きければ大きいほどaは小さくなり、グループ間異質性が大きければそれだけbは大きくなることから、シルエット係数が最大値である＋１に近ければ近いほど、そのクラスタリングは望ましいということになる。

シルエット係数を用いた実装方法はエルボー法とほぼ同様で、Kの値ごとにすべての観測値についてシルエット係数を計算し、その平均値であるシルエットスコアが＋１に近いKの値を探せばよい。ただし、この方法では一部の観測値のシルエット係数が異常に高い場合、その値にシルエットスコアが引っ張られる危険性がある。そこで、Kの値を選択する際は、クラスターごとのシルエット係数の値のばらつきも考慮したほうがよいだろう。

K平均クラスタリングは、空のクラスターが生じる可能性がゼロでない点も問題である。空のクラスターが生じた場合は、空のクラスターの重心から一番離れた観測値を代わりの重心として設定し、繰り返し計算に戻るといった対処法などが考えられる。

標準的なK平均法は重心の初期値を無作為に選ぶため、初期値の偏りによっては収束に時間がかかったり、非最適解に収束してクラスタリングに失敗したりする可能性がある。Arthur and Vassilvitskii（2007）は、K－means＋＋というアルゴリズムを用い、初期値を十分にばらつかせることで、より安定的なクラスタリング結果が得られることを示した。具体的には、

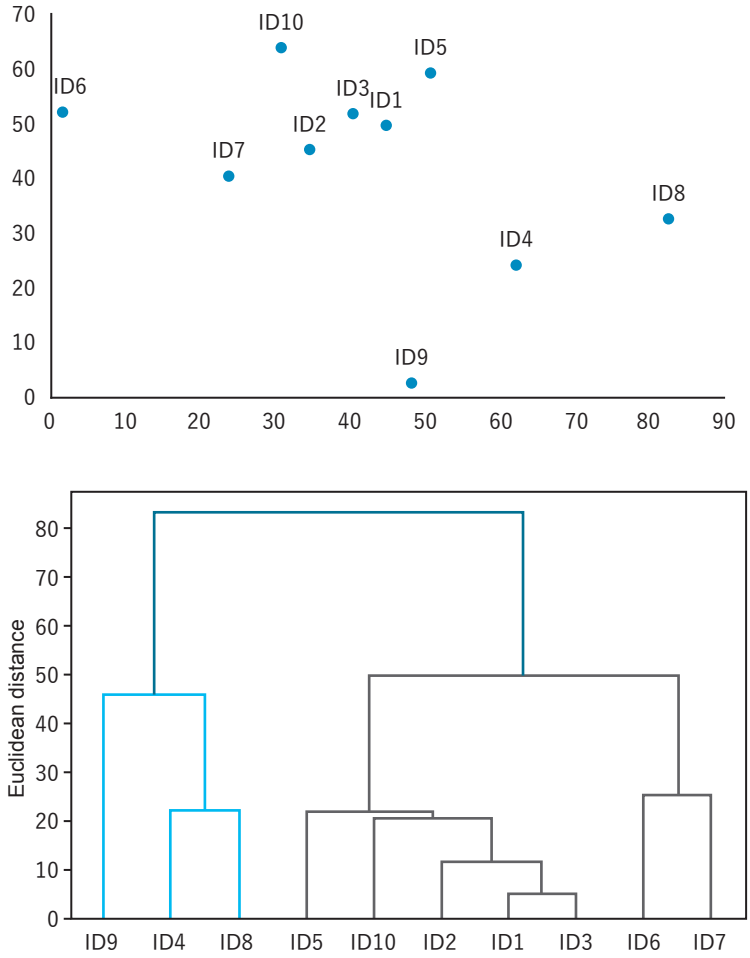
1．重心を１つ無作為に選ぶ。
2．ある観測値について、既に選ばれた重心の中で最も近い重心との距離を計算する。この距離をd_iとする。
3．$d_i^2 / \sum_{j=1}^{N} d_j^2$の確率で、観測値$i$を新たな重心とする。$d_i$が大きければ、それだけ$i$は新たな重心となる確率が高いので、既存の重心から離れた位置に重心が選ばれやすくなる。
4．2に戻って、重心をK個選ぶまで繰り返す。

（1）－2　階層的クラスタリング

K平均クラスタリングは有力なクラスタリングの手法であるが、あらかじめクラスター数を定めることで柔軟性が失われており、エルボー法やシルエットスコアを使ったとしてもこの点は改善されない。他方で、階層的クラスタリングはクラスター数を事前に指定する必要がないだけでなく、樹形図（dendrogram）でデータを分かりやすく可視化できるという利点がある。

階層的クラスタリングには凝集型（agglomerative）と分割型（divisive）という２つの方式がある。凝集型は個々の観測値をクラスターに振り分け、その後、クラスターを類似度に応じて統合し、最終的にクラス

（図１）ユークリッド距離を用いた完全連結法による階層的クラスタリングを行って得られた樹形図



ターが一つになるまでこれを続けるボトムアップ方式のクラスタリングである。逆に、分割型はすべての観測値をひとつのクラスターに振り分け、類似度に応じてクラスターを分割していくトップダウン方式のクラスタリングである。

具体例として、２つの属性が観測された被保険者10人のデータに対して、凝集型の階層的クラスタリングを行った結果を**図１**に示そう。初めはすべての被保険者がそれぞれ異なるクラスターに属しているが、まず１番目（ＩＤ１）と３番目（ＩＤ３）の被保険者が同じクラスターに統合され、されに２番目（ＩＤ２）、10番目（ＩＤ10）の被保険者がクラスターに統合されていき、最終的には一つのクラスターに統合されていく様子を確認することができるだろう。

個々のクラスターを統合する際にはクラスター間の類似度を参照するが、この参照方法には単連結法（single linkage）、完全連結法（complete linkage）、平均連結法（average linkage）がよく使われる。単連結法は対となるクラスターから最も距離の近い観測値のペアを取り出して、その距離が最小となるクラスターを結合する。逆に、完全連結法は最も距離の遠い観測値のペアを取り出して、その距離が最小となるクラスターを結合する。そして、平均連結法は対となるクラスターに

含まれるすべての観測値間の最小距離の平均に基づいてクラスターを結合する方法である。図１では完全連結法を用いているため、７番目の観測値が最初に統合されたのは２番目の観測値ではなく６番目の観測値である。なお、距離の計算にはユークリッド距離を用いているが、分析する対象によっては、（観測値の性質間の）相関に基づく距離を用いることが適切な場合もある。

樹形図が得られた後は、階層構造を確認したうえで切断する高さを決定する。低い位置で切断すればするほど多くのクラスターが得られ、逆に高い位置で切断すればするほどクラスターの数は少なくなる。切断する位置を決めることは、Ｋ平均クラスタリングにおいてクラスターの数Ｋを決めることに相当するが、樹形図によって各観測値の同質性に関する構造が得られたもとでクラスター数を決められる点は、階層的クラスタリングの強みと言えるだろう。ただ、構造が既知であったとして、最終的に分析者がクラスター数を選択する点はＫ平均クラスタリングと同様であり、分析過程から完全に恣意性を排除することは難しい。

（２）ニューラルネットワーク

前項では、リスクを予測する方法として、複雑な関係そのものをモデル化するのではなく、単純な関係が得られるようにデータを分割し、分割されたデータに基づいてモデル化および予測を行う方法として、クラスタリングの手法を解説した。実際に真の構造が企図した通りの形状をしていれば、そしてクラスタリングが適切に機能していれば、線形回帰モデルなど単純なモデルを使ったとしても、クラスターごとに予測を行うことで高い性能が期待できるだろう。しかし、データの構造がクラスタリングに馴染まない、もしくはデータの分割は必要だとしても、分割した後ですらクラスター内の被保険者の性質と保険事故の発生確率との関係が単純な形状で表せない場合、複雑な関係そのものをモデル化することは避けられないだろう[6]。

Chapados et al.（2001）は、自動車保険のデータベースを使って、どのような手法を用いると保険金を正確に予測できるか、アクチュアリーが伝統的に用いてきた方法に加えて一般化線形モデルや決定木、サポートベクターマシン回帰などと比較する実験を行い、ソフトプラス活性化関数を用いたニューラルネットワークの混合モデルの予測性能が最も優れていることを明らかにした。紙幅の都合により、比較されたすべてのモデルを説明することはできないので、本稿ではニューラルネットワークについて簡潔に説明する[7]。

ニューラルネットワークは、脳機能を計算機によって模倣することを目指した数学モデルで、単層のニューラルネットワークは入出力、重み、活性化関数で構成される。今回のケースでは入力値として被保険者の観察可能な性質を用い、重みは保険金を予測する際に個々の性質をどれくらい重視するか、を反映

6　もちろん、クラスタリングを行うことによって、求められる複雑さの程度が低下している可能性はある。
7　ニューラルネットワークに関するより初心者向けの内容としては、山名一史．2021.『データ活用の基本②－深層学習の基礎知識（ニューラルネットワーク）－』．共済総研レポート176. 56－62.

する値となる。

単純に観測値を重みづけし、活性化関数を通じて出力したものを最終的な出力とした場合、入力と出力との間に単純な関係しか考えることができない。しかし、出力を別のニューラルネットワークの入力値として受け渡し、重みづけと活性化関数による出力を繰り返せば、繰り返しの回数を増やすにしたがって、より複雑な関係を表現できそうである。このように、単層ニューラルネットワークを多層に重ね合わせ、複雑な関係の表現を試みたネットワークのことを多層ニューラルネットワークという。最初の入力部分を入力層、最終的な出力部分を出力層と呼ぶのに対し、中間部分は隠れ層と呼ばれる。

ニューラルネットワークを重ね合わせて多層にするということは、より複雑な関係が表現できるようになる一方で重みの訓練も難しくなることを意味する。しかし、現在はバックプロパゲーション（誤差逆伝播法）という効率的な訓練方法が提案されている。

バックプロパゲーションは出力層から入力層に向かって逆向きに誤差勾配の情報を伝播することで重み値を更新していくアルゴリズムである。バックプロパゲーションは勾配の情報を利用するため、ステップ関数のように微分不可能な関数を活性化関数として用いるのは望ましくない。そこで、多層ニューラルネットワークの活性化関数として用いられるようになったのがシグモイド（ロジスティック）関数である。シグモイド関数は0から1の値を取る関数で、連続で微分可能という特徴を持つ。

シグモイド関数は活性化関数として望ましい性質を持っているが、得られる勾配の最大値が小さいため、訓練に時間がかかるという欠点がある。そこで、訓練を高速化するため、連続かつ微分可能で、－1から1の値を取る双曲線正接（hyperbolic tangent）関数が用いられるようになった。ただ、双曲線正接関数を用いたとしても、ネットワークの層を増やすとともに訓練が行き詰まる、いわゆる勾配消失（vanishing gradient）問題は解消されなかった。これは、バックプロパゲーションが入力層に近づくにつれて勾配が小さくなり、0に近い領域では重みが更新されないために起こる問題である。

このような背景から、勾配消失を起こす可能性が低い活性化関数としてReLU（Rectified Linear Unit）関数が用いられるようになった。ReLU関数は$ReLU(z) = \max(0, z)$という形状をしており、連続関数であるものの$z = 0$で微分不可能なため、一見すると活性化関数には不適に見えるが、よく機能することが実証されている。Chapados et al. 2001で用いられているソフトプラス（softplus）活性化関数は、ReLU関数よりも滑らかなので訓練にかかる時間を高速化でき、ReLU関数と同様に出力が常に正になるような制約を加えることができる、$softplus(z) = \ln(1 + exp(z))$という形状の関数である。

残念ながらReLU関数も万能というわけではなく、訓練中に一部のニューロンが0以外の値を出力しなくなり、ネットワークが機能しなくなるdying ReLUという状態に陥ることがある。この問題を解決するため、$LeakyReLU(z) = \max(\alpha z, z)$で定義されるLeaky ReLU関数（ハイパーパラメータαは

固定）やLeaky ReLU関数のハイパーパラメータαをランダムにしたRandomized Leaky ReLU関数など、ReLU関数の亜種が提案された。さらに、2015年にはELU（Exponential Linear Unit）活性化関数が提案された。ELU活性化関数はReLUよりも計算時間がかかるが、$z<0$の領域で負の値をとるだけでなく勾配が非ゼロなので、勾配消失問題やdyingReLUの問題を回避できる。近年は、このELU活性化関数の亜種であるCELU（Continuously Differentiable ELU）やSELU（Scaled ELU）なども用いられている。

保険金の予測のように、回帰を目的としてニューラルネットワークを用いる場合、任意の値を出力するため、出力層に活性化関数を適用するのは望ましくない。一方で、出力の値を制限したい、たとえば確率の値を出力したい場合は出力層にシグモイド活性化関数を適用するとよい。同様に、正の値のみを出力したい場合は、ReLU関数やソフトプラス関数を適用することで、出力値を制限することができる。

ニューラルネットワークと一口に言っても、どの変数を入力に用いるか、また何層重ね合わせるかといった違いで、出力される予測値は大きく異なる。そして、単一のニューラルネットワークだけを使って予測を行うより、複数のネットワークを組み合わせて予測を行ったほうが、予測性能は高くなることが知られている。こうした方法はアンサンブル（ensemble）法と総称され、入力に応じてニューラルネットワークの重みを変えるアンサンブルの構成法の一つが、Chapados et al.（2001）で用いられているニューラルネットワークの混合モデルである。アンサンブルを構成するということは、入力変数のクラスタリングを行っているということなので、ニューラルネットワークの混合モデルを用いて分析を行ったということは、本稿で学んだクラスタリングとニューラルネットワークによる回帰を併せて分析を行ったものと理解できる。したがって、混合モデルの予測性能が最も優れていた、という結果は、保険金の予測がクラスタリングによる前処理だけでなく、ニューラルネットワークによるモデル化も必要とする複雑な問題であったと解釈できるだろう。

4．おわりに

データサイエンスは、大量のデータから有用な知識を探索、発見するために有効であり、経験や勘に基づく方法はもちろんのこと、統計学的であったとしても、過度に強い仮定に基づく数理モデルを用いた場合に比べて、高い予測能力が期待できる分析手法である。本稿では、一般的な保険市場のモデルを用いてリスク予測の重要性を確認し、機械学習・深層学習などデータサイエンス的な手法の有効性を議論した上で、リスク予測の具体的な手法として、先行研究で用いられているクラスタリングとニューラルネットワークの解説と考察を行った。リスク予測の性能には向上の余地があり、今後も分析手法のさらなる改善が期待される。

参考文献

- Antonio, Katrien and Beirlant, Jan, 2007. “Actuarial statistics with generalized linear mixed models,” Insurance：Mathematics and Economics, 40（1）, 58－76.
- Arthur, D. and Vassilvitskii, S. 2007. “k－means＋＋：the advantages of careful seeding.” SODA’07：Proceedings of the eighteenth annual ACM－SIAM symposium on Discrete algorithms 1027－1035.
- Björkwall, Susanna, Hössjer, Ola, Ohlsson, Esbjörn, and Verrall, Richard, 2011. “A generalized linear model with smoothing effects for claims reserving,” Insurance：Mathematics and Economics, 49（1）, 27－37.
- Chapados, N., Bengio, Y., Vincent, P., Ghosn, J., Dugas, C., Takeuchi, I., & Meng, L. 2002. “Estimating Car Insurance Premium：A Case Study in High－Dimensional Data Inference,” Advances in Neural Information Processing Systems：Vol. 14. Cambridge：MIT Press.
- Christmann, A. 2004. “An approach to model complex high－dimensional insurance data,” Allgemeines Statistisches Archiv 88, 375－396.
- England, P.D. and Verrall, R.J., 2002. “Stochastic claims reserving in general insurance,” British Actuarial Journal, 8（3）, 443－518.
- Einav, L., and Finkelstein, A., and Levin, J. 2010. “Beyond testing：empirical models of insurance markets,” Annual Review of Economics, 2（1）, 311－336.
- Géron, A. 2019. “Hands－On Machine Learning with Scikit－Learn, Keras, and TensorFlow, 2nd Edition,” O’Reilly Media.
- Patel, A.A. 2019. “Hands－On Unsupervised Learning Using Python：How to Build Applied Machine Learning Solutions from Unlabeled Data,” O’Reilly Media.
- Raschka, S., and Mirjalili, V. 2019. “Python Machine Learning：Machine Learning and Deep Learning with Python, scikit－learn, and TensorFlow 2, 3rd Edition,” Packt Publishing.
- Smyth, G., and Jørgensen, B. 2002. “Fitting Tweedie’s compound Poisson model to insurance claims data：dispersion modelling.” ASTIN Bulletin, 32（1）, 143－157.
- Williams GJ, Huang Z. 1997. “Mining the knowledge mine：the hot spots methodology for mining large real world databases.” Lecture Notes in Artificial Intelligence, 1342, 340－348.
- Yeo, A.C., Smith, K.A., Willis, R.J. and Brooks, M. 2001, “Clustering technique for risk classification and prediction of claim costs in the automobile insurance industry,” Intelligent Systems in Accounting, Finance and Management, 10：39－50.