1Q-04

２変数間に相関を持つアイテム集合を応用した分析方法の検討

小田 朱莉†　　嶋田 香†
†群馬大学 情報学部

1. はじめに

統計的な背景を持つアイテム集合群（ItemSB: Itemsets with Statistically Backgrounds ）を発見するための進化計算方法[1]が提案されている。しかし、発見された ItemSB 自体から何が見いだされるかについての提案はまだあまり行われていない。 本論文は、この方法を用いて発見されたアイテム集合からどのような知識や活用方法が見出されるかを検討するものである。

2. 材料と方法

今回用いたデータは UCI Machine Learning Repository に公開されている肝細胞癌患者に関する医療データセット[2]・[3]を[1]の方法によって ItemSB を抽出したデータである。[2]は肝細胞癌と診断された 165 人の診断された年齢、Clinical Practice Guideline に従った 49 の特徴などが含まれている。この 49 の特徴のうち年齢を除いた 48 個の特徴を 96 個の属性（アイテム）に離散化した（$A_1,\cdots,A_{96}$）。欠損データはデータセット全体の 10.22%を占めており、すべての特徴における完全なデータを持つ患者は 8 例のみである。一年生存についての情報も含まれており、68 例が死亡、102 例が生存している。

この医療データに[1]の方法で年齢と生死の間の相関係数が 0.75 以上、支持度 0.08 以上の ItemSB を発見した。また、負の相関を持つものもあるかを調べるために、2 変数間の相関係数が -0.75 以下の ItemSB も発見した。どちらも生死の人数がそれぞれ 3 人以上いる ItemSB だけに限定した。以下の図１は example(a),(b)は[1]の方法で得られた ItemSB の例である。検討項目としては年齢と生死の相関に注目した分析、構成する属性の分析、回帰直線とロジスティック回帰曲線の関係性の分析である。

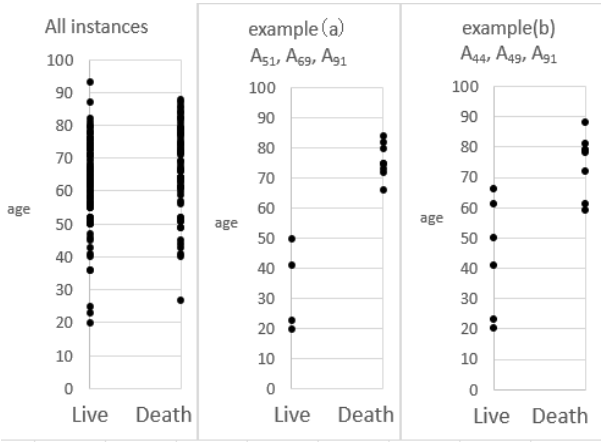

図１：ItemSB の例。example(a)はアイテム集合 A_{51}, A_{69}, A_{91} は生死と年齢に正の相関を持つ散布図のような統計的な背景を持つことが分かる。

3. 結果

3.1 年齢と生死の相関に注目した分析

まず、本論文においては[1]の方法を用いて抽出された年齢と生死に相関のある属性のセットをアイテム集合と定義する。年齢と生死に正の相関を持つアイテム集合は 14424 個、負の相関を持つアイテム集合は 135 個見つかった。次に、それぞれのアイテム集合に対して、該当する人たちを横軸に生を０、死を１として取り、縦軸に計測時の年齢をプロットしたグラフを考える。このグラフの回帰直線を作成し、回帰直線上の横軸の値が生と死の中間である 0.5 の点を境界点と呼ぶこととする。

この境界点は、発見した各 ItemSB を簡略化して利用するための数値として導入した。また、境界点における年齢を以下では、境界年齢と呼ぶこととする。境界年齢を求め、境界年齢が小さい順に並べ替えたところ 41.07 歳から 78.29 歳までの年代でアイテム集合が見つかった。このことにより各アイテム集合における年齢と生死に関する知見が得られるのではないかと考えた。また、各アイテム集合を構成する属性に年代ごとに差があるのではないかという予測が立った。

Analytical methods applying itemsets with correlations between two variables
Akari Oda †
Kaoru Shimada †
†Faculty of Infomatics, Gunma University

3.2 構成する属性の分析

次に、アイテム集合を構成する属性の年代ごとの出現頻度に差があるかを調べた。境界年齢の小さいほうから 5 段階の年代に分け、5 段階それぞれにどの属性がいくつ出現するかを調べた。これを正の相関と負の相関、両方で行った。この結果、正の相関においても負の相関においても年代ごとに出現する属性の差があることが認められた。図２は、正の相関における ItemSB が含む属性のうち総数が上位 3 個の属性の境界年齢ごとの積み上げグラフである。図２からも境界年齢によって出現する属性の個数に差があることが分かる。

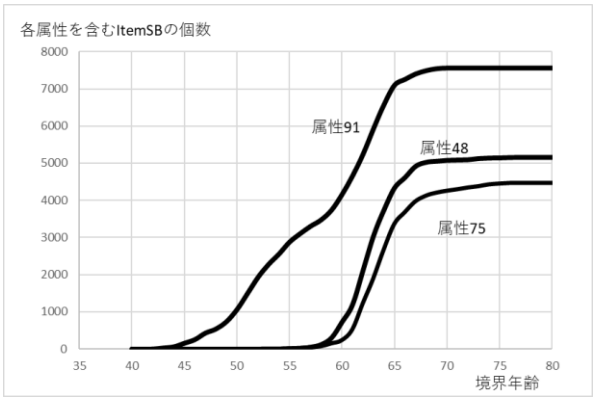

図２： 正の相関の ItemSB に含まれている総数が上位３個の属性の境界年齢における積み上げグラフ

3.3 境界点の妥当性について

本論文では、境界点を発見した各 ItemSB を簡略化して利用するための数値として扱ってきた。しかし、境界点で ItemSB の統計的な特徴を表せていると判断してもよいのかという疑問が生まれるだろう。そこで、ロジスティック回帰分析を各 ItemSB の年齢と生死に対して行い、ロジスティック曲線の対称点と境界点との関係性を考えていきたい。ここで、図 1 の example(a)のように生と死の年齢の間に明確な差がある ItemSB に関しては、ロジスティック回帰分析の収束条件を考慮する必要がある。そのため、今回の比較にはロジスティック回帰分析が比較的容易に行える図 1 の example(b)のような ItemSB のみを採用し、9 個の ItemSB について境界点と対称点の相関を調べた。

この実験を行った結果として 9 個の ItemSB に関しては、境界年齢と対称点における年齢には 0.98 という非常に高い相関係数が認められた。このことから、境界点は ItemSB を簡単に代表する値として妥当であると考えた。また、境界点と対称点に高い相関が認められることから、境界点から対称点を求める式が作れる可能性が高い。より多くの ItemSB に対して対称点と境界点についても相関を求め、境界点から対称点を求める補正式についての検討を進めている。

3.4 回帰直線とロジスティック回帰曲線の関係性

境界点と対称点に高い相関が認められ、境界点から対称点が求められると考えられる。このことから、ロジスティック回帰曲線の対称点における接線も回帰直線から求めることができるのではないかと考えた。これが可能であれば、ロジスティック回帰分析が行えなかった ItemSB についても簡易的な曲線を求めることができる。曲線を用いて各年齢における死亡確率を予測するために、ロジスティック回帰曲線と回帰直線の補正式の検討を進めている。

4. まとめ

今後の展望として、回帰直線から近似的なロジスティック曲線を求める方法の検討を進め、個人の各年齢における死亡確率を予測する手法を考えていく。また、予測を進める中で出現頻度の高かった属性は年齢と生死の２変数間の相関関係に影響を大きく与える要因として個人の持つ属性に着目した予測を行っていきたい。

参考文献

[1] Kaoru Shimada, Takaaki Arahira, Shogo Matsuno,ItemSB: Itemsets with Statistically Distinctive Backgrounds Discovered by Evolutionary Method, International Journal of Semantic Computing, 16 (3), 2022, 357-378
[2] UCI Machine Learning Repository, “HCC Servival” https://archive.ics.uci.edu/dataset/423/hcc+survival
[3] Miriam Seoane Santos, Pedro Henriques Abreu, Pedro J. García-Laencina, Adélia Simão, Armando Carvalho, A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients,Journal of Biomedical Informatics, Volume 58, December 2015, Pages 49-59