一般口演

一般口演10
データマイニング・テキストマイニング

2017年11月21日(火) 16:00 ～ 17:30 F会場 (10F 会議室1004-1005)

[2-F-3-OP10-2] 高額医療費の要因分析に適したデータマイニング手法に関する研究

村永 文学[1], 岩穴口 孝[1], 宇都 由美子[2], 熊本 一朗[2] （1.鹿児島大学病院 医療情報部, 2.鹿児島大学大学院医歯学総合研究科 医療システム情報学）

【目的】病院データウェアハウスに蓄積されたデータから、医療コストに影響を与える因子を発見・可視化する場合に最適なデータマイニング手法について比較調査する。
【方法】
2011年1月～2015年12月の間に院内がん登録で肝細胞癌の初回入院・初回治療として登録された患者を対象とし、当院の病院データウェアハウスから医療コスト情報、院内がん登録システムからがん登録情報、医事会計システムから会計情報、オーダリングシステムから検査結果情報を抽出し、全て匿名化してデータマートに格納した。
高コストとなる要因の分析について、クラスタ分析、アソシエーション分析、ベイジアンネットワーク分析、ニューラルネットワーク分析等の手法の有用性について評価を行った。
【結果及び考察】患者数は、365名であった。医療コストデータは約62万行、医療コスト費目名のバリエーションは3642種であった。まず、性別,在院日数,がんのステージ,年齢,総医療費及び一日当たりの医療費について、 k-means法によるクラスタ分析による特徴群の抽出を試みたが、明らかな高コスト要因は発見できなかった。次に一日当たりの医療費が大きい順に、高・中・低コスト群に3分割し、高コスト群を医療コスト費目名のみから判別可能であるか調査した。アソシエーション分析・及びベイジアンネットワークは、メモリオーバーフローの為分析不可能であった。
ニューラルネットワークでは、機械学習用274件と、評価用91件に症例を分割し、機械学習後、評価した。その結果、中間ニューロン数4の場合に約87%の高コスト群を判別できた。中間ニューロンが増えると処理時間は倍増するが、精度は上がらなかった。高コスト群と学習したニューロンの重みについて分析した結果、試行の度にニューロンの重みが関連する費目名が変化することが判明した。学習を分析を繰り返し、要因となる費目名を絞ることができた。

高額医療費の要因分析に適したデータマイニング手法に関する研究

村永 文学[*1]、岩穴口 孝[*1]、宇都 由美子[*2]、熊本 一朗[*2]
*1 鹿児島大学病院 医療情報部、*2 鹿児島大学大学院医歯学総合研究科 医療システム情報学

Research on optimal data mining method for factor analysis of high medical expenses

Fuminori Muranaga[*1], Takashi Iwaanakuchi[*1], Yumiko Uto[*2], Ichiro Kumamoto[*2]
*1 Medical Informatics, Kagoshima University Hospital ,
*2 Medical Information Science, Kagoshima University Graduate School of Medical and Dental Sciences

Abstract

[Background] The purpose of this study is to investigate the optimal data analysis method to analyze factors that increase medical cost, even though it is the same diagnosis. [Method] In this study, we investigate the causes of the medical costs of 365 patients diagnosed with hepatocellular carcinoma at Kagoshima University Hospital between 2011 and 2015, clusters by k-mean method, association analysis by apriori, Bayesian Network, and neural network. [Results and discussion] Among the four methods, it was possible to distinguish patients who were expensive only from machine learning using neural networks. However, the weight of the learned neural network changed greatly with each learning. From now on, we would like to continue our research on the method of analyzing the weights of neurons learned by neural networks.

Keywords: medical cost analysis, data mining, neural network.

1. 背景

少子高齢化が進行中の我が国では、厚生労働省によって、薬価改定や施設基準の見直し等、様々な医療コスト抑制策が施策されている。また、近年、検診や病院情報システムの情報、診療報酬のレセプト等から得られた膨大な医療データを分析し、費用対効果を可視化する等の研究が国内のみならず、海外でも盛んにおこなわれている。[1)]

我々は過去の研究で、DPC 別の医療コスト分析用データウェアハウスの開発を行い、高速に DPC 別医療コストを把握することが可能とした。[2-7)] また、医薬品の相互作用によって引き起こされる有害事象の検知システムも開発した。[8-13)] このシステムによって、データマイニング技法の1つである、アソシエーション分析技法（アプリオリアルゴリズム）を応用し、有害事象が発生した患者の薬歴から、医薬品相互作用の可能性がある薬剤を仮説として提唱することが可能となった。アソシエーション分析技法は一般的にマーケットバスケット分析に用いられている。消費者の「合わせ買い」の関連性を発見することが可能であり、既にオンラインショッピングモールでの商品紹介等で実用化されている。また我々は過去の研究で、糖尿病罹患患者の併存症発症の危険因子の関連性についてベイジアンネットワークを応用し、一定の成果を得た。[14,15)]

また、近年は人工知能に関する技術革新が進んでいる。例えば、多層化されたニューラルネットワークへの効率的な機械学習を行う「ディープラーニング」法の応用があげられる。

2. 目的

本研究の目的は、病院データウェアハウスから、医療コスト分析に最適なデータマイニング技法の複数のアルゴリズムを評価することである。

医療コストに影響を与える因子を可視化し、コストの適正化を支援するシステムを開発することが我々の研究の最終目的である。本システムの開発が成功すると、医療の質を担保しつつ医療コストの適正化が実現可能となる。

3. 方法

本研究では、まず、DPC 別医療コスト分析用データウェアハウスシステムを構築した。本システムには、既設の DPCBANK システムから DPCBANK 情報、院内がん登録システムから、がん登録情報、医事会計システムから会計情報、オーダリングシステムから検査結果情報を抽出し、全て匿名化して格納した。期間は 2011 年 1 月～2015 年 12 月までに蓄積されたデータとした。DPCBANK は、鹿児島大学で開発された DPC 別医療コスト分析専用のデータウェアハウスである。

本研究では、特に肝細胞癌の高コスト要因について、クラスタ分析、アソシエーション分析、ベイジアンネットワーク分析、ニューラルネットワーク分析等の手法を適用させ、評価を行った。

4. 結果及び考察

構築したシステムの性能を評価するために、本研究では肝細胞癌の初回入院・初回治療に関するコスト分析を行った。その理由は、がん治療は概ねプロトコル通りに施行され、コストのばらつきは原則的に発生しにくいと考えられる。したがってコストのばらつきの原因が特定しやすく、かつ、その結果はイレギュラーなものとして、改善の余地が期待できると考えたからである。

抽出期間は 2011 年 1 月～2015 年 12 月までのデータとし、病理診断にて HCC が確定したがん登録由来のデータベースから、入院日、退院日、在院日数、入院時年齢、性別、がんのステージ、手術の有無、化学療法の有無等を情報を抽出した。医療コストデータとしては、がん登録データに記録されていた初回入院、初回治療に該当する入院期間の、費目名と診療点数を費別に抽出した。

データ分析には統計解析ソフトウェア R 64bit 版 ver3.3.0 を用いた。利用したライブラリは、アソシエーション分析(arules ライブラリ)、ベイジアンネットワーク（deal ライブラリ）、ニューラル

ネットワーク(nnet ライブラリ)である。データ抽出とクレンジングには Perl スクリプトを作成し用いた。

分析用ワークステーションの CPU は、Intel Core i7 6900K@3.2GHz で、主記憶容量は 64GByte とした。

肝細胞癌患者数は、365 名であった。分析対象となった患者 ID の医療コストデータは、約 62 万行、費目名のバリエーションは 3642 種であった。

(1) クラスタリングによる分析

まずは、数値データに関する k-means 法によるクラスタ分析による特徴群の抽出を試みた。抽出したパラメータは、性別(sex),在院日数(hospitalization),がんのステージ(Stage),年齢(Age),総医療費(Point)とした。クラスタ分析の結果を図1に示す。

図 1 によると、明らかに関連性がみられるのは、在院日数(hospitalization)と総医療費(Point)のみであり、がんのステージ(Stage)や年齢(Age),性別(sex)との有意な関連性や、特別なクラスタは発見されなかった。

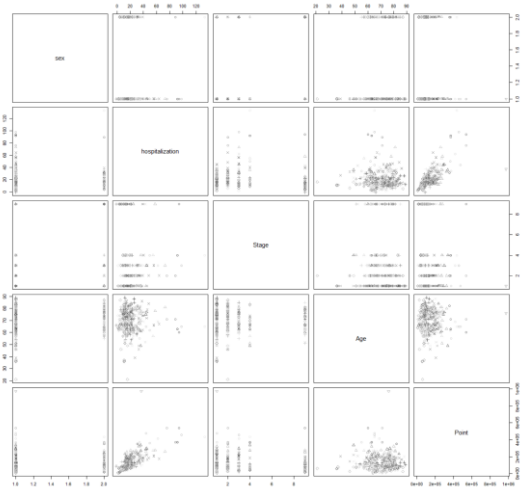

図 1 k-means 法による対散布図

この結果を元に、総医療費(Point)を在院日数(hospitalization)で除して、一日当たりの医療費(Point.per.day)を算出し、改めて k-means 法のクラスタリングを行った。その結果を図 2 に示す

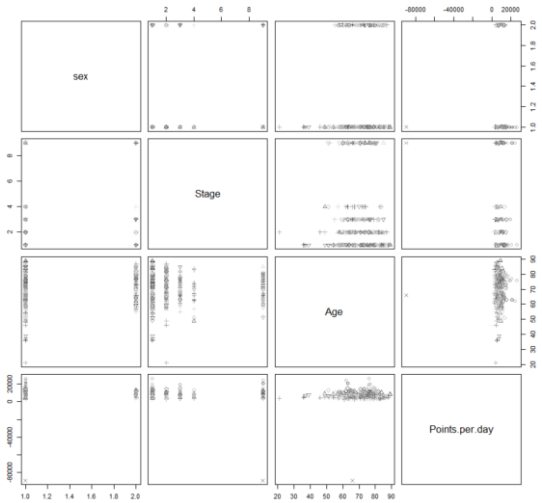

図 2 k-means 法による対散布図 その2

この結果からも、明確なクラスタは存在せず、一日当たりのコストと関連するパラメータは発見出来なかった。

(2) 高コスト群に特徴的な医療費目の発見に関する分析

クラスタ分析の結果では、明確なクラスタは存在しないことが判明したので、1日あたりの医療費が高い群に特徴的な医療費目の発見について、他の複数のデータマイニング手法を用いて分析した。365 件のデータの総医療費と在院日数から、一日あたりの医療費を算出し、一日当たりの医療費が大きい順に、高コスト群、中間的コスト群、低コスト群に3分割し、ラベルした。肝細胞癌の患者に行われた医療行為(薬剤、検査項目を含む)のバリエーションは 3642 種であった。

まず、アソシエーション分析による高コスト要因分析を試みたが、Item 数 3642 では、メモリオーバフローとなり解析不可能であった。分析可能な Item 数の上限について調査したところ、96 が上限であったので、本目的には利用不可能であると判断した。

次に、ベイジアンネットワークについて要因分析を試みたが、こちらもメモリーオーバフローにより分析不可能であった。分析可能な Item 数は 15 程度であった。R は 64bit 版を採用しており、分析用ワークステーションには 64GByte もの主記憶を準備していたため、これらのメモリーオーバーフローは、ハードウェアの問題ではなく、R システムのライブラリの実装上の問題であろうと推測した。

ニューラルネットワーク分析については、Item3642 種のままで問題なく分析実行できた。ニューラルネットワークでは、365 件のデータを機械学習用データ(トレーニングデータ)274 件と、評価用データ 91 件に均等に分割して分析を行った。

出力結果が高・中・低の3要素なので、中間ニューロン数を 4, 5と変化させて分析を行った。ニューラルネットワーク分析のイメージを図 3(中間ニューロン 4),図 4(中間ニューロン 5)に示す。なお、このイメージでは、Item 数が 3642 種のままでは作図困難であったため、Item 数を 10 に制限して作図した。

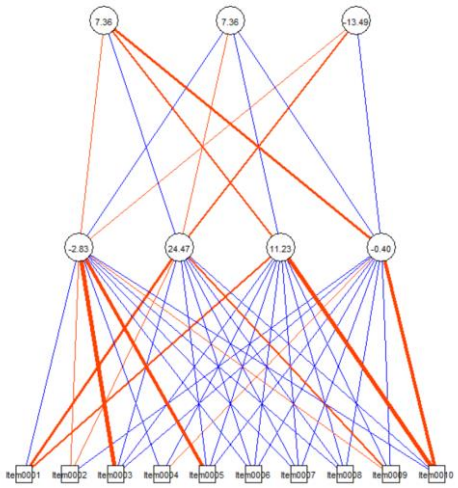

図 3　中間ニューロン 4 の場合の分析例
(Item 数を 10 に限定)

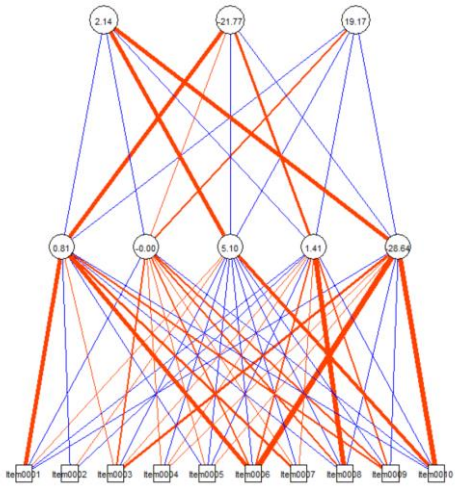

図 4　中間ニューロン 5 の場合の分析例
（Item 数を 10 に限定）

トレーニングデータで学習させたニューラルネットモデルで、評価データを評価させた結果を示す。

① 中間ニューロン数4の分析結果
・重み：14587
・処理時間：36 分 19 秒
・学習試行回数:1090 回

・予測結果

	High	Mid	Low	sum
High	25	5	0	30
Mid	5	17	9	31
Low	2	11	17	30
sum	32	33	26	91

横軸はニューラルネットワークが費目名のパターンから予測したコスト
縦軸は1日あたりの医療費から算出したコスト
・コスト判断一致率 0.6483516

この結果より、Mid と Low の判定が上手く処理できていないことが判明したので、High とそれ以外をまとめて集計した。

	High	Low	sum
High	25	5	30
Low	7	54	61
sum	32	59	91

横軸はニューラルネットワークが費目名のパターンから予測したコスト
縦軸は1日あたりの医療費から算出したコスト
・高コスト判定のみの一致率=0.868132

② 中間ニューロン数5の分析結果
・重み：18233
・処理時間：1時間 14 分 5 秒
・学習試行回数：1340

・予測結果

	High	Mid	Low	sum
High	25	5	0	30
Mid	9	11	11	31
Low	1	7	22	30
sum	35	23	33	91

横軸はニューラルネットワークが費目名のパターンから予測したコスト
縦軸は1日あたりの医療費から算出したコスト
・コスト判断一致率＝0.6373626

同様に、High とそれ以外をまとめて集計した結果

	High	Low	sum
High	25	5	30
Low	10	51	61
sum	35	56	91

横軸はニューラルネットワークが費目名のパターンから予測したコスト
縦軸は1日あたりの医療費から算出したコスト
・高コスト判定のみの一致率=0.835165

上記に示すように、機械学習したモデルを用いて、コスト費目から総医療費が高額になるか否かの判定は 80%以上の的中率で可能であった。中間ニューロンが増えると、処理時間が倍増するのみで、精度は上がらなかった。

次に、機械学習したニューラルネットワークの入力ニューロンの重みから、高コスト要因となったコスト費目を抽出した。
学習した重みの抽出には成功したが、中間ニューロン4、5共に、試行するたびにニューロンの重みは変化していた。一致する Item の割合は10%程度であった。このことより、ニューラルネットワークによる機械学習では、高コストに寄与する全

ての特徴を捉えているのではなく、高コストになる「断片的な特徴」を捉えていると推測された。試行毎に発見する特徴が変化する為、機械学習を複数回繰り返し、学習した重みから発見したItemの出現頻度を集計する必要があると思われた。1回の試行に30分以上の時間を要することから、試行の自動化等、更なるシステム開発が必要である。本研究期間では、残念ながらItem 頻度を自動計測する部分のシステム開発までは当初の計画にもなく、実現できなかった。今後の研究課題としたい。

5. まとめ

鹿児島大学病院で肝細胞癌の診断・治療を受けた症例データを用い、高コストとなる要因を分析するのに最適な分析手法についての評価を行った。ニューラルネットによる機械学習によって、患者に実施された医療費目から、該当患者が高コストになる可能性があることを80以上の確率で判別できた。しかし学習したニューロンの重みは、機械学習する都度、大きくから具体的な医療費目を特定するには、更なる解析が必要である。

6. 謝辞

本研究はJSPS科研費 JP26460868の助成を受けたものです。

参考文献

1) 森田正実:医療健康分野のビッグデータ活用の現状と課題:JPMA NEWS LETTER,No.167,Page1-9,2015. [http://www.jpma.or.jp/about/issue/gratis/newsletter/archive_after2014/67pc.pdf]
2) 熊本 一朗, 村永 文学:【DWH構築が果たす効果検証】総論 DWHの活用における今日的課題と将来展望 鹿児島大学病院の事例を踏まえて;新医療,43巻2号,24-27,2016.
3) 村永 文学, 熊本 一郎, 宇都 由美子:DPC別診療コストの算出を目的とした病院データウェアハウスの開発;医療経済研究(1340-895X)18巻2号,95-104,2006
4) 宇都 由美子, 熊本 一朗, 村永 文学, 宇宿 功市郎:包括評価と病院経営 重症度分類の差異によるDPCコスト分析と病院経営支援;医療情報学連合大会論文集,23回,125-126,2003.
5) 宇都 由美子, 熊本 一朗, 村永 文学:DPC別原価計算と看護ケア量の評価;病院管理,40巻Suppl.,176,2003.
6) 宇都 由美子, 村永 文学, 宇宿 功市郎, 熊本 一朗:品質管理・コスト管理のツールとして有効な患者別原価計算システムの開発 病院DWHを利用したDPCごとの患者別原価計算;医療情報学,23巻1号,23-31,2003.
7) 村永 文学, 宇都 由美子, 宇宿 功市郎, 熊本 一朗:【データウェアハウス・データマイニング】 病院情報システムに蓄積されたデータを活用した病院データウェアハウスの構築;BME,16巻4号,8-17,2002.
8) 村永 文学, 岩穴口 孝, 宇都 由美子, 熊本 一朗:医薬品相互作用による有害事象シグナル検知システムの分析手法の評価;医療情報学連合大会論文集,36回1号,446-449,2016.
9) 村永 文学, 岩穴口 孝, 宇都 由美子, 熊本 一朗:アソシエーション分析による医薬品相互作用検知手法の評価;医療情報学連合大会論文集,31回,1000-1002,2011.
10) Muranaga Fuminori, Uto Yumiko, Kumamoto Ichiro:Evaluation of a Data Mining Technique for Detection of Adverse Drug Events using a Pharmacoepidemiological Data Warehouse(英語);日本薬剤疫学会学術総会抄録集16回Abstracts,54,2010.
11) 村永 文学, 宇都 由美子, 熊本 一朗:薬剤疫学データウェアハウスを用いた医薬品有害事象検知手法の評価;医療情報学連合大会論文集,29回,772-773,2009.
12) 村永 文学, 熊本 一朗, 宇都 由美子, 宇宿 功市郎, 下堂薗 権洋:薬剤疫学データウェアハウスの医薬品副作用監視への応用;薬剤疫学,9巻Suppl., S38-S39,2004.
13) 村永 文学, 武藤 充, 松藤 凡, 岩穴口 孝, 宇都 由美子, 熊本 一朗:アソシエーション分析を用いた小児慢性特発性偽性腸閉塞症の診断基準案の策定;医療情報学連合大会論文集,33回,602-605,2013.
14) Nurjannah , Fuminori Muranaga, Takashi Iwaanakuchi, Yumiko Uto, Ichiro Kumamoto:The significance of a Bayesian Network in Type 2 Diabetes Mellitus(英語);医療情報学連合大会論文集,34回,352-355,2014.
15) Nurjannah , Fuminori Muranaga, Takashi Iwaanakuchi, Yumiko Uto, Ichiro Kumamoto:Discovering the Probability of relationship between Diabetes Mellitus Type 2 and Diabetic Complication Diseases with Bayesian Network(英語);医療情報学連合大会論文集,33回,598-601,2013.