様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　元 年　 ６ 月　 ５ 日現在

機関番号：１７７０１
研究種目：基盤研究(C)（一般）
研究期間：2016～2018
課題番号：１６Ｋ０９１７８

研究課題名（和文）人工知能を応用したデータマイニングによる糖尿病合併症危険因子発見に関する研究

研究課題名（英文）Research on diabetes complication risk factor detection by data mining applying artificial intelligence

研究代表者

村永　文学（Muranaga, Fuminori）

鹿児島大学・医歯学総合研究科・客員研究員

研究者番号：００３２５８１２

交付決定額（研究期間全体）：（直接経費）　　3,700,000円

研究成果の概要（和文）：【背景】糖尿病患者の合併を予防することは、日本でも重要な問題である。【目的】定型化されていない自由記載の診療記録等から、糖尿病患者の合併症の予知をする手法について研究する。【方法】GenSimライブラリのDoc2Vecを用いて、自由記載を含む患者データのコサイン類似度を算出し、糖尿病合併症を判別可能か検討した。研究対象は2011年～2012年に、鹿児島大学病院にて糖尿病の治療目的でインスリンを使用した患者とした。【結果及び考察】予備実験の結果より、Doc2Vecの最適パラメータを算出した。糖尿病合併症の無い患者と、糖尿病性網膜症または腎症を発症した患者の判別は有意に判別できた。

研究成果の学術的意義や社会的意義
今日の日本では非常に多くの糖尿病患者が存在している。さらに透析患者の37.6%が糖尿病腎症で占められている。糖尿病患者の合併を予防することは、日本でも重要な問題である。ただし、診療担当の医療従事者にテンプレート入力をさせるようなデータ収集を行うと、その負担の大きさから、入力漏れ等が発生する。データ精度の観点からも望ましくない。
本研究は、定型化されていない自由記載の診療記録等から、糖尿病患者の合併症の予知をする手法について研究することを目的とする。本研究の成果により、電子カルテ等に記録された診療記録から、糖尿病合併症を発症しつつある患者を自動的に判別し、警告することで、合併症予防に繋げる。

研究成果の概要（英文）：[Background]Preventing the merger of diabetic patients is an important issue in Japan.[Purpose]We study methods to predict the complications of patients with diabetes from non-stabilized free-handed medical records.
[Methods]Using the GenSim library Doc2Vec, we calculated cosine similarity of patient data including free description, and examined whether it was possible to distinguish diabetic complications.The subjects of the study were patients who used insulin for treating diabetes at Kagoshima University Hospital from 2011 to 2012.[Results and discussions]From the results of the preliminary experiment, the cosine similarity was calculated by setting the optimal parameters.Discrimination between a patient without diabetic complication and a patient who developed diabetic retinopathy or nephropathy was able to be discriminated significantly by discriminating cosine similarity with a threshold.

研究分野：医療情報学

キーワード：データマイニング　テキストマイニング　糖尿病合併症　GenSim　Doc2Vec

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

日本透析医学会の統計調査によると、2013 年末現在、国内の透析人口は 31 万 4,180 人であり、原因疾患は糖尿病腎症が 11 万 5,118 人で、透析患者全体の 37.6%と最も多い。厚生労働省の 2014 年患者調査では糖尿病の総患者数は 316 万 6,000 人であり、糖尿病予備群を含めると 1000 万人を越えると推計されている。糖尿病患者における合併症は合併症の割合は、神経障害がもっとも多く 11.8%、次いで腎症 11.1%、網膜症 10.6%、足壊疽 0.7%であった。糖尿病合併症は、患者の ADL・生活の質や寿命に著しい影響を与える。糖尿病患者が合併症を発症しやすくする因子を発見することは、わが国においても喫緊の課題と言える。学会等でも、テンプレート方式による患者情報の収集とデータ分析を行っているが、多項目のテンプレート入力は診療担当医の負担になるため、大学病院等の一部の医療機関以外では実施困難である。理想的には、定型化されていない診療記録やデータ等から、糖尿病合併症発症の予知を行い、警告するようなシステムが開発されることが望ましい。我々は過去にデータマイニングに関する様々な手法について研究を行ってきた。これらの研究成果を基礎とし、糖尿病合併症に関するデータマイニングに臨む。

２．研究の目的

本研究では、定型化されていない自由記載の診療記録等から、糖尿病合併症を発症する因子を発見するデータマイニング手法について研究を行う。最終的には糖尿病患者に合併症が発症しつつあることを警告するようなシステムを開発したいと考える。

３．研究の方法

フリーテキストの分析する手法として、TF-IDF 法(Term Frequency - Inverse Document Frequency)がある。しかし、TF-IDF 法は文書に対する単語の発生頻度（特異度）を特徴として算出する手法であり、例えば、症状の有無（腹痛がある、腹痛が無い、等）を区別することはできない。我々は、単語の前後関係についても分析できる文書ベクトル化法の Doc2Vec(word2vec)を応用し、糖尿病合併症患者の判別と、判別根拠となった要因の分析（機械学習内容の分析）が可能かどうかを検討した。Doc2Vec では、文書同士の類似度を「コサイン類似度」として算出することができる。GenSim ライブラリの Doc2Vec には、設定すべきパラメータとして、size（文書ベクトルの次元数）、window（単語の前後関係の幅）、min_count（破棄する単語の出現回数の閾値）が存在する。Doc2Vec などのニューラルネットワークを用いた解析では、計算処理の効率より乱数を用いて重みづけの調整が行われるため、得られるコサイン類似度は毎回の計測にて微妙に変化する。処理結果のバラつきを平均化するために、複数回、解析を行い、その平均値をとる必要があると考えた。適切な学習用データ数についても検討した。

まずは、Doc2Vec にて医療従事者の記録を分析するのに最適なパラメータの探索を行う予備実験を行った。診療記録分析用パラメータ最適化のために、誤嚥性肺炎に関するデータを用いた。対象は、2013 年から 2017 年の間に当院に入院し誤嚥性肺炎を発症した全ての患者を疾患症例群とし、2018 年 6 月 1 日～7 月 20 日の間に入院した患者から、誤嚥性肺炎症例群の 95%信頼区間に相当する年齢に該当した症例を、ランダムに抽出した症例群を対象症例群とした。解析対象とした看護観察記録は、記録時に 40 文字(80Byte)の文字数制限が設けられており、看護師が冗長な記録をせず、簡素な表現で患者の状態を逐次記録するよう設計されている。誤嚥性肺炎症例群は、発症日四日前～発症日前日の記録までを採用した。対照症例群については、対象期間中の入院日から４日分の記録を採用した。まず、全てのデータについて MeCab を用いて形態素解析を実施した。なお、形態素解析用辞書として、MeCab に付属されている ipadic 辞書、Comejisyo（ただし、１文字の単語は削除）に加え、我々が過去の研究にて開発した、鹿児島大学病院独自のユーザー辞書も使用した。ユーザー辞書の作成には、誤嚥性肺炎症例群・対照症例群のデータとは関係のない過去の看護観察記録を用いた。MeCab で形態素解析後に、適切に単語が分割されない用語（固有名詞等）を抽出し、ユーザー辞書として登録した。抽出された誤嚥性肺炎症例群を、学習用データ・テスト用データ・評価用データにランダムに 3 分割した。学習用データは全て誤嚥性肺炎症例となる。テスト用データ・評価用データには、誤嚥性肺炎症例に加え、対照症例群からランダムに抽出した症例を加える。テスト用データや評価用データの解析は、学習用データに対して各データ（テスト用データ・評価用データ）の症例を１例ずつ追加して処理した。過去の誤嚥性肺炎症例の記録である学習用データが、各データ（テスト用データ・評価用データ）と類似している程度を判断する。出力された学習用データのコサイン類似度の平均を算出し、本研究における各データ（テスト用データ・評価用データ）のコサイン類似度とした。実験手順としては、まず学習用データとテストデータを用いて探索パラメータ等を定める予備調査を実施後、適合性を評価する。その後、学習用データと評価用データを用いて最終的な適合性の評価を行った。最適な学習用データ数については、Mean Average Precision（MAP）を指標に、適切となるデータ数を決定した。データの上位 1～20 例と、データ全例で検証し、計算に用いる学習用データ数を決定した。

最適な解析繰り返し数については、学習用データに対してテスト用データの誤嚥性肺炎症例のコサイン類似度を複数回実施し、F 検定により分散が収束する試行数を求めた。

Doc2Vec パラメータの最適値の探索については、size=(10,20,30,40,50,60,70,80,90,100),

window=(4,5,6,7,8,9,10,11,12,13,14,15), min_count=(1,2,3)の範囲で、組み合わせたパラメータ設定を順次行い、Doc2Vec にてコサイン類似度を算出し、各パラメータ毎の Average Precision（AP）を基準に最適パラメータを決定した。これらの予備実験により、各パラメータの最適値を算出した。

次に、本研究の主題である、糖尿病患者のデータ解析を行った。対象は、我々が過去にベイジアンネットワークを用いて分析した患者と同じ、2011 年～2012 年に当院を継続受診し、糖尿病の診断名を有し、かつインスリン投与を行われている患者を対象とした。本研究では、自由記載の診療記録を一定量集める必要があったので、入院患者を対象とした。糖尿病合併症を発症していない患者、糖尿病性腎症の患者、糖尿病性網膜症の患者について、「患者の年齢（年代）、性別、入院時要約、検査結果項目（各項目の最新値）、投薬情報（薬剤名のみ）、基準日（入院日）から１０診療日分の診療記録データ」を収集し、これらの順でデータを結合し１文書とした。なお、本研究では合併症発症の検知が目的であるため、診断病名等は含めなかった。

「糖尿病合併症を発症している患者と、発症していない患者の文書データを、Doc2Vec のコサイン類似度で区別できるのか、区別できた場合に、Doc2Vec で機械学習した各文書を特徴づける単語フレーズを抽出可能か」について、試行した。なお、Doc2Vec 学習用データ数、繰り返し数、size, window, min_count の３つのパラメータについては、予備実験の誤嚥性肺炎患者データ解析で得られた結果を採用した。

なお、これら研究は、研究施設である鹿児島大学病院の疫学研究倫理委員会で審査され承認された。（糖尿病合併症=番号 412,誤嚥性肺炎=番号 411,697）。

４．研究成果

（１）誤嚥性肺炎症例データによる予備実験（最適パラメータ探索実験）の結果

対象となった誤嚥性肺炎症例は 98 例を用いた。対照症例は 228 例を用いた。看護観察記録の個々の記録量は、誤嚥性肺炎症例群は 2～16kbyte、対照症例群は 1～10kbyte であった。誤嚥性肺炎症例は学習用データとテスト用データにランダムに二分割を行い、対照症例は全例をテスト用データとした。適切となる学習用データ数（TopN）を検証した結果、MAP において Top9 が最大の適合性を示した。繰り返し施行階数毎の分散の変化について図１に示す。8 回と 9 回の繰り返し試行数間にて、F 検定によって同一群と判断した。(p = 0.241484009)。

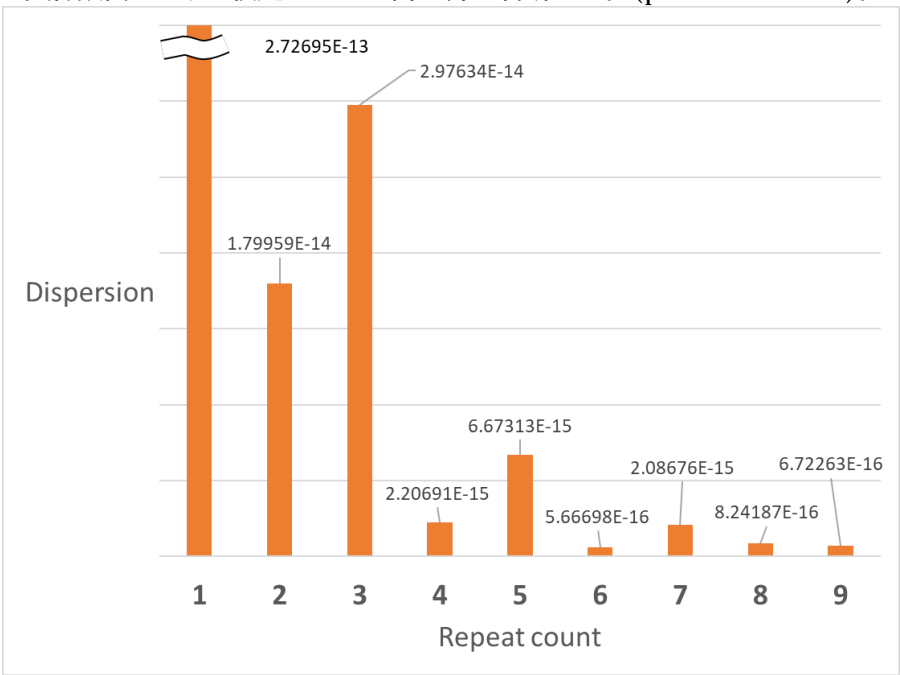

図１　繰り返し回数ごとの分散と前試行数との p 値
（検証用パラメータとして、size50、window10、min_count2 を使用）

以上の検証結果より、繰り返し試行数は 8 回、計算に用いる学習用データ数は上位 9 データと設定し、size=(10,20,30,40,50,60,70,80,90,100), window=(4,5,6,7,8,9,10,11,12,13,14,15), min_count=(1,2,3) の範囲で最適パラメータの探索を行った。最も高い AP を示したパラメータ値は size=40、window=12、min_count=1 であった。ROC 曲線を図２に示す。

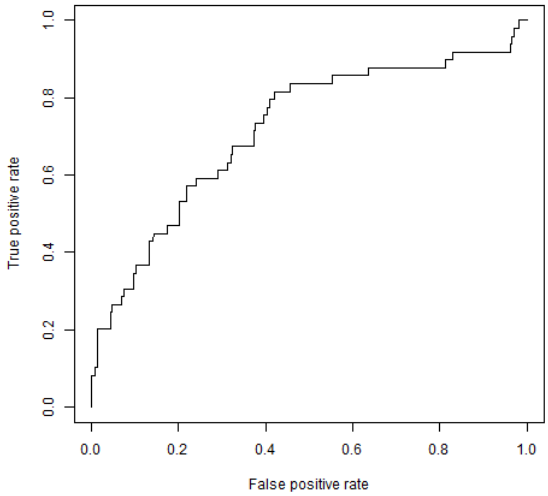

図２　size40、window12、min_count1 における ROC 曲線（テスト用データ）

AUC は 0.718 であった。感度・特異度の最大値は、感度=81.6%、特異度は=7.9%であった。

次に、これらの最適パラメータの正当性を検証するために、評価用データで検証を行った。評価用データの誤嚥性肺炎症例は、対象期間の症例からランダムに抽出した 33 例を用いた。また、対照症例は対象期間において抽出された 63 例を用いた。看護観察記録の個々の記録量は、誤嚥性肺炎症例群は 2～18kbyte、対照症例群は 1～12kbyte であった。繰り返し試行数、計算に用いる学習用データ数、探索パラメータ範囲はテスト用データの解析と同様の条件にて解析を行った。評価用データより、最も高い AP を示した最適パラメータ値は size=80、window=13、min_count=2 であった。ROC 曲線を図３に示す。

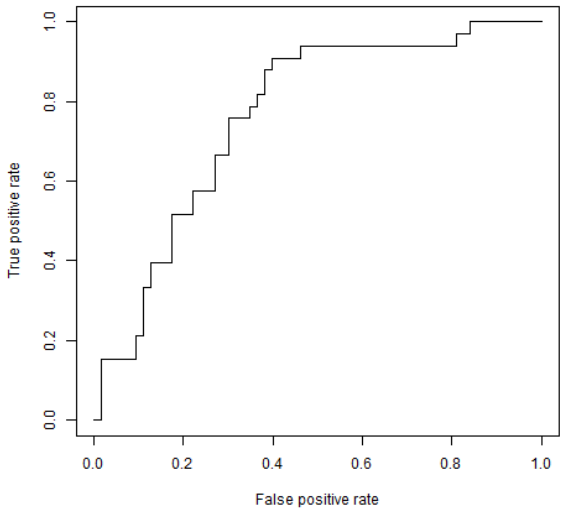

図３　評価用データ ROC 曲線

AUC は 0.763 であった。感度・特異度の最大値は、感度=90.9%、特異度=60.3%であった。

（２）糖尿病合併症の分析結果と考察

本研究では、2011～2012 年のインスリン治療中の糖尿病患者のうち、一時的な高血糖に対するインスリン治療ではなく、糖尿病の治療を目的として入院している患者から 277 例を抽出し、研究の対象とした。糖尿病合併症の無い患者は９４名（以下、N 群と呼ぶ）、糖尿病性網膜症を合併している患者が１３９名、糖尿病性腎症を合併している患者が１０７名、糖尿病性網膜症と糖尿病性腎症の両方を合併している患者は６３名(以下、RR 群)であった。したがって、糖尿病性網膜症のみ発症している患者は７６名（以下、Ret 群）、糖尿病性腎症のみ発症している患者は４４名(以下、RF 群)であった。NN 群、RR 群、Ret 群、RF 群より、ランダムに学習用データ２０例と、評価用データ２０例を抽出した。

例えば Ret 群の検出能力評価方法としては、NN 群の学習用データと Ret 群の学習用データを混在させ、NN 群の評価用データおよび Ret 群の評価用データから１例ずつ抽出し、学習用デー

タ群と評価用データのコサイン類似度を most_similar 関数で算出した。Doc2Vec 分析用の各パラメータは、予備調査の結果より、size=80、window=13、min_count=2 を設定した。なお、most_similer で算出されたコサイン類似度の上位 9 データを採用し、その平均を算出し、さらに同じ計算を８回繰り返し、平均を算出した。算出されたコサイン類似度を、ある閾値で分割し、Ret 群の評価データが、Ret 群の学習データに近いと判定された場合、および NN 群の評価データが、NN 群の学習データに近いと判定された場合を正答とし、そうでない場合をエラーとして２×２テーブルを作成した。その２×２テーブルに対して Fisher の正確確率検定で有意となるような閾値が存在した場合に、感度及び特異度が最大となるような閾値を求めた。

同様に、NN 群と Ret 群、NN 群と RF 群、NN 群と RR 群を組み合わせて試行を行った。NN 群と、各合併症群との比較では、コサイン類似度の閾値 0.9997 とした場合に、感度 90.2%、特異度 58.8%で判別可能であった。非常に大きなコサイン類似度の閾値であるが、各文書の類似点が多いために、閾値も大きな値になったと思われた。Ret 群と RR 群、および RF 群と RR 群の組み合わせでは、有意な閾値は算出できなかった。以上より、文書群のコサイン類似度により、合併症の有無については判別可能であると思われた。ただし、既に合併症を発症している群間を判別することはできなかった。記録の内容を目視で精査すると、例えば糖尿病性網膜症を発症した患者では、当然ながら他の合併症が発症していないかを同時に検査しており、記録内容に類似点が多くなっており、文書ベクトルとしても多きく類似していたためと思われた。

次に、具体的に学習された文書の特徴について、Doc2Vec の infer_vector 関数の出力データを分析可能か試みた。文書ベクトルを散布図として可視化して観察すると、その形状は試行の度に変化していた。Doc2Vec はニューラルネットワークを利用して文書ベクトルを算出するが、ニューラルネットワークは乱数を利用するために、同じ文書セットでも試行の度にベクトル構造が変化すると思われた。我々が過去に行った研究に於いても、ニューラルネットワークで機械学習されたニューロンの重みは試行の度に変化することが観察された [9]。コサイン類似度が試行の度に変化するのも同様の理由であると思われる。したがって、少なくとも同じ文書に対して８回の分析を繰り返し、文書ベクトルの類似点を抽出できないか試みた。残念ながら infer_vector から出力された文書ベクトルのデータ構造の解析に手間取り、合併症群について学習された特徴の詳細については、研究期間内に調べることはできなかった。これは今後の研究課題としたい。

５．主な発表論文等

〔雑誌論文〕（計 1 件）

1) 小牧祥太郎、村永文学、宇都由美子、岩穴口孝、熊本一朗: 誤嚥性肺炎予防の為の観察記録解析における文章ベクトル化技法の有用性の検討, 医療情報学,37 巻 Suppl., 380 - 383, 2017.（査読あり）

〔学会発表〕（計 1 件）

1) 小牧祥太郎、村永文学、宇都由美子、岩穴口孝、熊本一朗: 誤嚥性肺炎予防の為の観察記録解析における文章ベクトル化技法の有用性の検討, 第 37 回医療情報学連合大会,大阪市グランキューブ大阪（大阪国際会議場）, 2017.（査読あり）

〔図書〕（計　0 件）

〔産業財産権〕
○出願状況（計　0 件）
○取得状況（計　0 件）

〔その他〕
特記事項なし

６．研究組織

(1)研究分担者
研究分担者氏名：熊本　一朗
ローマ字氏名：Ichiro Kumamoto
所属研究機関名：鹿児島大学
部局名：医歯学領域医系
職名：教授
研究者番号（8 桁）：40225230

研究分担者氏名：宇都　由美子

ローマ字氏名：Yumiko Uto
所属研究機関名：鹿児島大学
部局名：医歯学領域医系
職名：准教授
研究者番号（8 桁）：50223582

研究分担者氏名：岩穴口　孝
ローマ字氏名：Takashi Iwaanakuchi
所属研究機関名：鹿児島大学
部局名：医歯学域医学部・歯学部附属病院
職名：助教
研究者番号（8 桁）：80619198

(2)**研究協力者**
研究協力者なし