原著論文 Original Paper

深層学習を用いた特徴抽出によるカルテ情報からのフェノタイプの同定

大北 剛 Tsuyoshi Okita
九州工業大学 Kyushu Institute of Technology
tsuyoshi.okita@gmail.com

山下貴範 Takanori Yamashita
九州大学病院 Kyushu University Hospital
t-yama@med.kyushu-u.jp

野原康伸 Yasunobu Nohara
（同　　上）
y-nohara@info.med.kyushu-u.ac.jp

井上 創造 Sozo Inoue
九州工業大学 Kyushu Institute of Technology
sozo@mns.kyutech.ac.jp

廣川佐千男 Sachio Hirokawa
九州大学 Kyushu University
hirokawa@cc.kyushu-u.ac.jp

中島直樹 Naoki Nakashima
九州大学病院 Kyushu University Hospital
nnaoki@infomed.kyushu-u.ac.jp

keywords: 深層学習, メモリネットワーク, 集合演算子, フェノタイピング, 電子カルテ.

Summary

大量のカルテ情報に記載された自然言語で書かれた内容から必要とする項目を信頼性をもって抽出して解析することにより, フェノタイプを同定する精度を上げることを目的とする. このため, メモリネットワークを拡張して, 語彙をベースとした探索を効率良く行なう集合演算子メモリネットワークを提案して, 半構造化データから特徴量を抽出するシステムを構築した. 80 万件 (3ヶ月分, 53,065 人分相当) のカルテ情報から集合演算子メモリネットワークを用いた方法により説明変数 8,398(× 53,065 人, 実質 3 万人分) の特徴を抽出して, これから 1 型糖尿病患者のフェノタイピングを行なった.

1. ま え が き

精密医療 (プレシジョンメディシン)[Fernald 11, Lu 14] は患者の病態（フェノタイプと呼ぶ）のみならず遺伝子レベル（「ジェノタイプ」）で解析して治療する。 これは、従来の万人に一様に適合する治療という方向を一転して, 個々人に応じた治療を可能とする画期的なものだが, 生活習慣や環境などの因子が個々のフェノタイプに大きく関わる疾患もある [Wenzel 12, Zhou 14, Delude 15, 河添 16]. ジェノタイプはこれまでの遺伝子解析研究により詳細まで急激にわかりつつあるのに対し, フェノタイプの同定やその記録方法などに関する研究は遅れている。 フェノタイプの方法論の確立が可能となれば, 明確にフェノタイプが分類された疾患 (もしくは疾患のサブタイプ) に対して正確にジェノタイプ情報を付加することで診断や治療の精度を向上させる効果が見込まれる. また, 現状でジェノタイプを用いて診察を行なっている病院は少ないため, 通常の診察や検査を用いて判定の難しい疾患の判定率を改善させるためにフェノタイプ手法を使えることになる [*1]. 本論文でのトライアルは 1 型糖尿病においている. 血液、尿検査などの通常の検体検査により判定しやすい疾患と判定が難しい疾患の存在 (例；急性腎不全は易、1 型糖尿病は中程度, うつ病は難など) が知られており, 1 型糖尿病は通常の検体検査により判定の難しさが中程度という理由により選択した.

上記したように, 疾患により個々の病態「フェノタイプ」には生活習慣や環境などの因子が大きく関わる. このため, フェノタイプの因子を分析する手掛かりの一つは, カルテ情報などに自然言語という形で隠された患者の生活習慣や環境など記述を抽出することである. さらに, 本論文の設定では検査データ, 投薬などのデータもカルテ情報から読み取らねばならないため, これらも同時に目的とする.

*1　近年, 特定の疾患の存在や進行度をバイオマーカーの濃度に反映させることに成功し, このことは, 病態「フェノタイプ」の特定に結び付ける客観的な測定評価項目を一つ加えた.

自由文から関心のある関係情報を抽出する技術の一つは関係抽出 (relation extraction)[Banko 07, Snow 05] の分野で機械学習の機構を用いて発展した. これは抽出する対象が自由文の中で頻度高く何度も登場する場合を想定している *2. 一方, 自然言語処理で行なわれる形態素解析, 構文解析は関係とする対象が一度登場すれば認識可能であるが, 本論文で行なうようなカルテ情報の解析とはドメインが非常に異なるため形態素解析器, 構文解析器をそのまま適用することは難しい. 通常, ドメイン適用 [DaumeIII 09] が行なわれるが, その場合には対応するペアの語のリストや言語モデルなどの妥当なトレーニング集合を用意する必要がある. 本論文では, メモリネットワーク [Weston 14, Kumar 15] を応用した形でこのタスクを遂行させる. メモリネットワークは自然言語理解を行なうべく開発された手法で, トレーニング集合の子集合である与えられた独立な文に対し, 自動的にさまざまな質問応答に答える. カルテ情報から得たい情報というのは 2 通りあり, 1 つ目は予め因子のペアの一方が判っている場合 (検査, 投薬など), 2 つ目は環境や生活習慣に絡む関係ペアというオープンエンドな場合である. 本論文では特に前者に注目した.

フェノタイプの同定に関する先行研究はいくつか存在する. 川添 et al.[河添 16] は癌のフェノタイプの同定を行う. この論文では, 登録病名, 投薬薬品名, 検査項目を患者に適応した結果を含まずして使用している. たとえば, ある患者が投薬薬品としてアマリール錠を服用した場合, 服用した事実があれば, それを特徴として加えるというものである. アマリール錠を 1mg 服用したのか 10mg 服用したかは関係なく, また, 服用後の患者の状況にも感知しない. この場合, われわれのような抽出法は必要とせず, 患者のレコードにアマリール錠があるかないかをチェックするだけでよい. この方法が示唆するように, 一般の質問応答のように質問解析により探索する内容を特定して探索するより, 最初に設定されている語彙にしたがって探索を開始するほうが効率的と思われる.

本論文の貢献は, カルテ情報解析に対して

- メモリネットワークを拡張した集合演算子メモリネットワークを開発したこと

である.

2. 1 型糖尿病の推定ロジック

1 型糖尿病を推定するフェノタイピングのロジックの原型を以下に示す. これは中島 [田ジマ 15] が開発した. 本研究ではこれを出発点として, この PPV を上げることを目的とする. この推定ロジックにおいて用いられたデータは, 対象期間を 2009 年 1 月 1 日から 2014 年 12 月 31 日の 6 年間とする.

(1) 以下, a, b,c,d,e を抽出する. a) 病名: 対象期間に外来または入院がある患者で,「1 型糖尿病」を含む病名を持つ患者. 病名は以下のように抽出する: 確定のみ (疑いを除く), 対象期間中にアクティブ (未転帰 または 転帰日 $\geq$ 2009/1/1 かつ 開始日 $\leq$ 2015/1/1) であること. b) インスリン処方が対象期間中にある. c) 血中 CPR 検査で陽性 (0.6ng/mL 未満) が期間中に 1 度でもある. d) 「ケトアシドーシス」を含む病名が対象期間中に登録されている. e) a の患者の全病名を抽出し,「膵移植」を含む病名を持つ患者を抽出する.

(2) A 集団を a AND (b OR c OR d OR e) または b AND c AND d とする.

(3) e で抽出した全病名から, 除外対象疾患を抽出する.「1 型糖尿病」が先行し,「2 型糖尿病」が先行し, 2 型糖尿病やその他の「除外対象疾患」が同日を含めて後に開始された場合を除外対象とする.「1 型糖尿病」,「除外対象疾患」が複数登録されている場合は, 各々の最終日付を採用する.

除外: 除外対象疾患. 開始日の最大 $\geq$ 1 型糖尿病. 開始日の最大. 除外しない: 除外対象疾患. 開始日の最大 $<$ 1 型糖尿病. 開始日の最大.

(4) 除外対象薬剤を「SU 剤」,「グリニド剤」,「DPP4 阻害剤」とし, 対象期間中の処方データを抽出する. 患者ごとの最終処方日を求め, 最終処方日より前に「1 型糖尿病」病名が登録されている場合を除外対象とする.「1 型糖尿病」が複数登録されている場合は, 初回の日付を採用する.

除外: 最終処方日 $>$ 1 型糖尿病. 開始日の最小. 除外しない: 最終処方日 $\leq$ 1 型糖尿病. 開始日の最小.

(5) 「1 型糖尿病」病名が死亡以外で転帰している場合, 除外対象とする.

(6) 血中 CPR 陽性が期間中に一度でもあれば, 除外の取消対象とする.

(7) e) で抽出した全病名に「膵移植」または「緩徐進行 1 型糖尿病」があれば, 除外の取消対象とする.

以上より,「1 型糖尿病」推測値= A - (B NOT C) として求める.

3. 集合演算子メモリネットワークによる特徴抽出

メモリネットワークはさまざまな質問タイプに応じた質問応答を行なうフレームワークである. 一般的な質問タイプに対応するため, 質問解析においては一般的な質問タイプに対応するための工夫がある. 一般的な語彙を用い, 質問タイプを一般化している. 一方, 本研究においては, 医学分野における特徴抽出を対象とするため, 語彙

*2 たとえば (会社, 創始者) のペアとして, マイクロソフトとビルゲイツなどを抽出する場合, これらの固有名詞が頻出する文章をコーパスとして用いる必要がある. 文書に一度しか登場しないペアは基本抽出するのは難しい.

は有限だが医学用語に偏り, 質問タイプも非常に偏りがある. こういう状況の違いに対処するため, われわれは語彙を有限であることを基本とする集合演算子メモリネットワークを導入する.

3·1 代表的な質問パターン

集合演算子メモリネットワークに対する質問は, 病態（フェノタイプ）の同定に役立つと思われる医学的なイベントを抽出する質問となる. HPO[Koehler 17] はフェノタイプオントロジを定義するが, この HPO は以下の 4 つのサブオントロジをもつ *3. (1) どのような病気の徴候 (フェノタイプ的な異常) が表われているのか, (2) 体のどの部位に表れているのか, (3) 頻度 (排他的に, 拘束的に, 頻繁に, 時折, ほんのたまに, 非常に頻繁に) はどのようなものか, (4) 医学的修飾子 (改善している, ーをきっかけにして, フェノタイプ的な変形, ひどいものか, 位置, 時系列パタン, 悪化している, どのような痛みか, 兆候は, 進行速度は) に関係する. (1) の徴候に関しては症状所見マスタ *4による項目を参考にした. また, 質問が有効か有効でないかの可否は分類作業を行なった際に重要度によりフィードバックをかけることを前提とすることにした. これを参考にわれわれの代表的検索のパターンを挙げると以下の 3 つとなる.

(1) 検査項目 CPR に対応する検査値を探索する
(2) 投薬薬品名インスリンに対する 1mg や 50mL といった投薬割合を探索する
(3) 観察用語の「しびれ」に対する is-a 関係 (しびれに対する「ない、残存、あり、消失」などの叙述) を抽出する

抽出プロセスの一段目は上述したように語彙を事前に登録する形である. たとえば, 患者 ID245 の患者が HbA1c の検査を行なったか否かは, 検査項目リストに HbA1c があるため, これらのリストにしたがってカルテ情報の中身がこれらのリストの用語を含むかどうかをチェックする. 抽出プロセスの二段目は, われわれのターゲットとしたものは事前に設定したクエリを投げかけることによりその属性を抽出する.

3·2 集合演算子と語彙

病院においては, 検査マスターなどとして, 検査、投薬、注射, 病名などの一覧表が得られる場合が多い. これらをベースとしてオリジナルの語彙とした. それぞれの統計量は以下のようになる. 出典は, medis のデータ (「病名」, 「診断名」, 「観察」) を用いた. 1 年生の解剖学辞典 *5(「部位」) のデータを用いた. それ以外のデータは九州大学病院のデータ (「薬」, 「検査」, 「注射」, 「手術」*6, 「処方」) を用いた.

	統計量	
	オリジナル (人)	抽出後 (人)
検査	3839	115
病名 *	16334	3446
注射	1427	401
手術	541	41
薬品処方	3445	1418
診断 *	1918	688
観察 *	1849	823
部位	772	567
小計	31087	7499
非医学用語	-	1000
計	-	8499

表 1 語彙の統計. 3ヶ月のカルテ情報に存在する項目は病院のリストより少ないため, 抽出後に項目が減少するのはこの理由を 1 つとするが, その他, 表記揺れによる理由, 個々の語彙のリストが完全なリストではない理由, カルテ情報という性質上の記述漏れという理由などが存在する. なお, カルテ情報という性質上の記述漏れというのは, 統計用に用いる場合, 調査対象の項目が記載されておらず, 再度医師が個別に漏れを埋めることも必要なケースも多い. この意味から, カルテ情報は精査した場合に記述漏れが存在する可能性が高いと考えたほうがよいという意味である. * のデータは medis のデータを用いた. 部位のデータは 1 年生の解剖学辞典のデータを用いた. それ以外のデータは九州大学病院のデータを用いた. 「非医学用語」は, 上記の 9 つの医学用語から漏れる用語で, 頻度の高い用語を抜き出した. ここにはたとえば食事に関する用語「食事, 食べ, 食欲, 夕食後, 朝食後, 食, 食事摂取, 生食, 食べる, 食事再開, 絶食, 食後, 毎食後, 食事摂取良好, 間食, 食事開始, 食欲不振, 食物アレルギー, 食事療法, 飲ん, 運動, 睡眠, 外泊, 飲水, 飲酒, 生食, 飲み」を含む.

集合演算子メモリネットワークの「集合」という名称が示唆するように, 個々のトレーニング集合に応じた適用化を目的とする. 1 つ目, これは日常生活などに関係する言葉の語彙を増やすことに繋る. トレーニング集合において事前に定義された語彙にヒットしない部分を解析すれば未定義の語彙を増やすことになる. まず, 語彙は有限で既知であると仮定する. 次に, カルテ情報すべてを語彙をストップワードを除いて頻度順にならべ, 用語として登録されていないものを一般の語彙として登録する. カルテ情報は有限のドキュメントであるため, 用語を頻度順に並べることができる. すでに設定された語彙から漏れる単語で, ストップワードを除いたものは, 日常生活などに関係する言葉も多い. 食事の様子, 患者の変化などの患者の環境因子や生活習慣因子に関係する言葉を用いた記述である可能性が出る. なお, 鬱病などにおいては, 患者の言葉などによる疾患の記述や患者の変化などの記述が特に多く, われわれの意図したものとは若干性質を異とする. 2 つ目, 集合演算子は有限の語彙を対象とするため, 上記の漏れの中を探索することで表記揺れを発見すること

*3 その他 2 つは遺伝のモード (mode of inheritance), 生命的な危機 (mortality/aging) である.
*4 http://medis.or.jp
*5 「1 年生の解剖学辞典」には体の部位/器官系ごとの一覧が存在する. URL は以下の通りである. http://www5.atpages.jp/motoneuron/mobile/.
*6 手術の用語リストがなく, DPC データの一部から抽出したため数が少ない.

に繋る可能性をもつ. 表記揺れは事前知識としてなければ非常に難しいが, 使用語彙が有限なので, それぞれの語彙の集合をうまく管理することでこのことを可能とする.

いずれの疾患に対しても再設定することなく有効であるため拡張性があり, メモリネットワークのように複雑な質問に回答することで文の内部を解釈した上で回答を行なえるという長所をもつ. 一方, すべての探索を事前に行なうのではなく質問がなされたときに探索をし始めるとオーバーヘッドをもつという短所をもつ.

なお, カルテ情報は一般的なカルテを想定する. 患者レコードは, 患者のみの記述がなされており他人の記述はない. 一人の患者の情報は, 患者レコードにおいて複数存在する可能性がある. 語彙は病院ごとに統一されている場合が多いが, 基本的には医師が個人的な責任で記述するためこれから漏れる場合, 新出用語, 誤表記, 表記揺れも存在する. 表記揺れの例は C ペプチドである. C- ペプチド, C ペプチド, CPR, C-Peptide などの表記が存在し, さらに検体に血液と尿があるため、それを区別するために先頭に、血中、尿中を付ける場合がある。つまり, 尿-C ペプチド, 尿-CPR などとなる. 病名除外や薬剤除外においては, 部分的にあるグループの病名や薬剤を一つの集合として扱い, そのグループとそれ以外を区別する. たとえば, 薬剤除外は「SU 剤」,「グリニド剤」,「DPP4 阻害剤」を一つのグループとして扱うやり方である. また, 病名に関する記述には「確定」と「疑い」の 2 通りが存在する.

3·3 集合演算子メモリネットワーク

T をエンティティーのラベルづけをされたタプルの集合とし, x をクエリ (質問) とする. T にはトリプル (述部, 主語, 目的語)(トリプルは日本語の場合格フレームである), ペア (物, 属性)(ペアは is-a 関係を示す), クアドラプル (述部, 主語, 目的語, 時間や場所を示す修飾子) などが存在する. x は患者一人一人のスコープをもつ自由文から検索したい内容であり, 用語となる. 表 2 においては「1A-2Ab?」,「膵移植？」,「血糖？」などがクエリの例である. $S(x)$ は次の空のメモリスロット N を返すとし, 与えられた x に対する関連する情報に対して, m_1 と m_i の組み合わせが依存関係などの関係をもつという意味で, 語彙を仲介としてメモリの連鎖のリストを作る. したがって, 結果的に与えられた x に対して, 依存関係などの関係をもつ確率が確からしい記述を選択することを再帰的に行なう. 以上を数式で記述すると

$$
\begin{aligned}
m_N &= S(x_{N-1}) \qquad (1)\\
&\ldots\\
m_3 &= S(x_2)\\
m_2 &= S(x_1)\\
m_1 &= S(x)\\
T &\in [S(x), S(x_i)]
\end{aligned}
$$

対象	値の抽出	自由文の例
IA-2Ab	39.0U/ml	2012/X/X-2013/X/X 総合診療科に入院。GADA 478U/ml, IA-2Ab 39.0U/ml, sCPR 0.8ng/ml, uCPR 18.75ug/day、
入院	あり	
膵移植	あり	膵移植後の腎機能評価 1 型糖尿病に対し 2012/X/X に父をドナーとする生体膵単独移植を行っております。
血糖	50mg/dl 程度	14:00-16:00 頃にかけて血糖上昇の傾向強い。 19:00-22:00 頃に血糖降下の傾向強い。 7/2 夜間低血糖 50mg/dl 程度。無自覚。 暁現象は抑制できている。 シルエットの挿入具はまだ使っていない。

表 2 表に自由文の例を 2 つ示す. 表に自由文の例を 4 つ示す. 1 番目の例はカルテ情報内に頻出する検査項目とその値のペアに対するもので,「1A-2Ab?」というクエリに対して 39.0Uml という値を自由文から読み取っている. 2 番目の例は文の記述から読み取らねばならないが,「膵移植？」というクエリに対して, ありという回答を抽出している. また, 膵移植が予定や可能性という話ではなく, すでに実施されたということを読み取っている. 3 番目の「血糖？」という例からは目的のターゲットが複数文に存在し, いずれの値を取るか選択する必要があり, この場合は (血糖, 50mg/dl 程度) を選択した..

となる. なお, メモリとは抽象的だが, 個々の文章の記述を指してもよいし, メモリ中に取り込んで別の表現形態となっていても構わない. 要約すると, 質問文に含まれる語彙「痛み」や「感じる」などをカルテ文書中に同定し, それらの間に事前に定義した依存関係などの関係を探索することのできるような, 枝分かれのないメモリチェインを得る. 有限集合の語彙を用いているためこれは常に可能となる.

このようにして得られたメモリチェインに対して,

$$r = \arg\max_{w \in W} s_R([x, m_{O1}, m_{O2}], w) \qquad (2)$$

を行なうと, 質問 x に対する応答を得る. したがって, 以上の操作を (データが小さい問題や欠損データを考えずに) 個々人の記述に対して行うと, 理論的には関連する特徴を抽出できることとなる. ただし, 本研究のようにもともと患者のカルテ情報として記載されていない内容に対する質問も多く存在する場合もある. この場合, そのような質問に対しては欠損情報という返答を返す. なお, 上記の式の中において $\arg\max$ の例を示したが, この部分は $count$ という操作に置き換えることによりクエリが数に関する場合に回答を下せる [Neelakantan 17].

オリジナルのメモリーネットワークは, 文を単位として参照の範囲を広げる. たとえば, 最初のキーワードが文 1 に含まれている場合, 文 1 に含まれている別の単語を二段目の探索と設定して 2 段目の探索を開始する. また, 探

索する語彙はクエリ文にしたがってアドホックに抽出される. 集合演算子メモリーネットワークにおいては語彙を含むパッセージに対するリンクは一段目で設立している. 1 つ目, 検査項目に対応する検査値を探索する場合には, 探索されている単語の発見されたパッサージのロケーション的にこのリンクの最近辺を探索する。特に直後に表れる可能性を埋め込む. その後, 若干離れた場所を探索する. 別のパッサージは探索しない. 検査値は「検査をしたかしない」という 2 値の場合と「検査結果を示す値」という連続値の場合があり, 前者の場合には検査項目の記述が患者レコードに記されていればよい. 「uCPR 12.6, GLT 126/139/147, CPR 0.74/1.4/1.4」の場合, CPR のロケーション的に直後に数字/単位つきが存在する可能性をまず探索することにより,「0.74/1.4/1.4」を抽出する. 2 つ目, 「インスリン は 現在 ランタス 12 単位 と スライディング 。」という形で単位つきで抽出すべき場合があると同様,「1 型糖尿病の診断 で 強化 インスリン療法 導入 。 主治医 異動に 伴い 、 2014 / 04 / 11 当科 紹介 。」などのようにインスリンを使用している事実はありそうだが, 量的な記述はない場合が多く, この場合は使用しているか否かを返し, 同じ患者の別のロケーションに対する「インスリン」に対する探索結果とを考え合わせる. つまり, 薬に対する 1mg や 50mL といった投薬割合を探索する場合は, 直後をまず探索し, その後, 若干離れた場所を探索する. 別のパッサージは探索しない. 3 つ目, 観察用語の「しびれ」に対する is-a 関係の抽出の場合には, しびれに対する「ない、残存、あり、消失」などの叙述を探索する. この場合, 叙述がないことも考えられる. 別のパッサージは探索しない. なお, is-a 関係の記述を抽出する際に程度を表現する語には事前にチェックをつけ, ペア表現を抽出した. チェックをした語は以下の通りであり, これは [Aramaki 14] を参考にした. 「記号と単位. 記号のアップとダウン. 程度, ほど, 傾向, 等. 陽性, 陰性, 予防, 悪化, 出現, 低下, 拡大, 上昇, 正常, 拡張, 縮小, 憎悪, 変化, 減少, 再発, 継続, 既往, 可能.」である.

4. 抽出した特徴を用いた分類

一旦, 集合演算子メモリネットワークにより特徴を抽出すると, 次にランダムフォレストを用いて 1 型糖尿病か否かの分類を行なう. この際に重要度を観察することにより, 因子の貢献度を測る. なお, 健康な人のデータを得ることは困難なため負例は 1 型糖尿病が含まれない他の疾患をもつ患者とした. また, 他の疾患をもつ患者の中には, 明示的に 1 型糖尿病を登録病名としてカルテに記述されていない患者が存在するであろうことは推測ができるが, これは無視する. ランダムフォレストは無作為に選択した特徴を用いて相関の低い複数の決定木／弱学習器を生成し, これらの弱学習器のアンサンブルにより出力を形成する.

5. 実 験 結 果

5·1 カルテ情報データ

九州大学病院のカルテ情報のうち 2014 年 7 月 1 日から 9 月 30 日までのデータを用いた. 件数は 824,453 レコードあり, レコードを一人の患者に対する形にまとめ、53,065 人分とした. カルテ情報は csv ファイルの形で存在し, レコード毎に (No, 患者 ID, Date, 診療科, 職種, 記録区分, 内容) の 7 つのフィールドをもっている. 1 レコードは 1 回の回診に相当するもので, 1 患者は通常複数のレコードをもつ. 3ヶ月分 (=約 90 日) ではあるが, 1 人のデータが最高 640 レコードに及ぶものもあり, この場合, 1 日に複数の項目がある場合もある.

内容のフィールドに検査結果, 投薬情報, その他さまざまな情報が (1) 自然言語の形, (2) ペア情報 (たとえば, 検査と結果のペアの情報「Asthma(-) DM(-), HT(+), HL(+), HUA(-), Stroke(-), IHD(-)」, もしくは投薬された薬とその量のペアの情報「プロイメンド（ｄ 1)、アロキシ (ｄ 1) DEX（ｄ 1-9.9mg、ｄ 2〜5、6.6mg」) の形, (3) 項目化された半構造データ *7として存在する. 受診科によりカルテ情報の書き方はまちまちで, 記録する情報も多岐に渡る. 824,453 レコードを患者ごとに結合し, 53,065 人の患者毎にフラットな形の表とした. 異なる日付の内容データを単純に結合し (本実験では患者の 3ヶ月以内での時系列解析はしないため), いくつのレコードを結合したかのフィールドを結合文書数とし, また, データを結合したためにレコード No., Date, 診療科, 記録区分などは複数個存在する. フィールドは (患者 ID, 結合文書数, No リスト, Date リスト, 診療科リスト, 職種, 記録区分リスト, 内容, ...) となった. また, 九大病院で用いられている検査項目一覧, 薬一覧を用いてこれらの項目については自由文の記述から抜き出した.

5·2 診 療 デ ー タ

診療データはカルテ情報を元に, カルテには記載されていないが, 統計には必要なデータを医師自身が患者に再確認して補完したデータである. したがって, 補充された項目に関しては診療データの方がカルテ情報より情報を豊富にもつ. 糖尿病患者 9231 人分を含み, この中には 1 型糖尿病患者 268 人, 2 型糖尿病患者 3544 人のデータが含まれる. このデータはカルテ情報の期間より長いため, カルテ情報に含まれる 94 人分のデータを含むが, それ以外の患者 174 人分のデータは含まれない. 患者それぞれに対し, (性別, 年齢, インスリン処方, ケトアシドーシス, 血中 CPR(陽性), GAD 抗体 (陰性), IA2 抗体 (陰性), 自己

*7 項目化された半構造データはさまざまな形で項目化されたデータを指す. たとえば, (＃1 陳旧性脳梗塞 ＃2 高血圧 ＃3 脂質異常症) である. なお, この半構造化の書き方は診療科, 担当医師によりさまざまなバリエーションが存在する. バリエーションとはさまざまなシンボルを用いるという意味で, たとえば以下のような書き方が存在した (【家族歴】, ■入院時身体所見■, ＜生活歴＞, ◆骨密度 (2014/7/4):, ●検査結果).

抗体 (陰性), SU 剤処方, グリニド剤処方, DPP4 阻害剤処方, 薬剤除外, 病名除外, 膵移植, BMI, HbA1c, 尿中ケトン体 (陰性), 基礎インスリン処方, 成長ホルモン剤処方, 1 型 DM 病名転帰済) という 20 のフィールドが記述され, さらに, 分類のラベルとなる 1 型糖尿病, インスリン依存性, 1 型糖尿病かつインスリン依存性が記述されている.

5·3 設　　定

集合演算子メモリネットワークは Theano[TheanoDevTeam 16] を用いて実装し, GTX1070 上で走らせることにより実験を行なった. テスト集合, ディベロップメント集合は 0.1 分割を用いた.

5·4 実 験 結 果

カルテデータ, 診療データを用いて 1 型糖尿病の分類学習を行なった. カルテデータ, 診療データとも 1 型糖尿病に対する分類学習を行ない, 診療データに関してはさらにインスリン依存性, 1 型糖尿病かつインスリン依存性に対する分類学習も行なった. 特徴の数はカルテデータに対しては 8499 (3·1 節参照), 診療データに対しては 20 を用いた. 特徴抽出は, たとえば, 身長, 体重に対するものを図 1 に示した. 左図の身長に対しては 160cm 近辺の分布の山に加えて, 子供や生れて間も無い乳児の身長もあるため, 分布が 140cm 以下に対してもなだらかになっている. なお, 200cm や 0cm は読取りミスであった. 右図の体重に対しては分布の山が 2 つ存在し, これは体重の増減値, 体重の絶対値の山を示す. このような複数の山に対する特徴をフェノタイピングの分類に対して用いることは弊害があると思われたが, 本論文においてはこれ以上の吟味は行なわなかった. 図 2 は, 患者一人あたりに項目が存在するか否かの統計を取り、項目あたり何人の患者が関連するかの頻度を示す. この表においても, 10,000 人以下の項目が大半であり, 50 項目以降は 5,000 人以下である.

表 3 は診療データに対する 1 型糖尿病の PPV は 73.56、oob(out-of-bag) スコアは 77.7 であった. 一方, カルテデータに対する 1 型糖尿病の PPV は、oob(out-of-bag) スコアは 77.7 であった. 設定 A における PPV は 65.0, oob スコアは 64.7 であった. 診療データに対する追加の設定に対しては, インスリン依存性に対する PPV は 80.5, oob スコアは 77.8, 1 型糖尿病かつインスリン依存性に対する PPV は 78.2, oob スコアは 80.0 であった.

特徴の数がカルテデータに対しては 8,499, 診療データに対しては 20 であるにも拘わらず, PPV に大きな差が出たことにはいくつも原因が考えられる.

- 図 2 において, 患者一人あたりに項目が存在するか否かの統計において一患者あたりの頻度が 30 人以下 (0.06%) である項目が大多数であり, これはある患者集団に対するフェノタイピングの分類に対して十分に機能しなかった可能性があること,
- 図 1 に示すように, 特徴ごとの統計量は複数の原因を混合したものである可能性があること,
- 特徴を抽出する集合演算子メモリネットワークの精度が現地点では十分ではないこと,
- カルテ情報に医師が記載した内容は本論文のフェノタイピングの分類に対するような場合に, 診療データで行なわれたような再評価が必要な項目が多い可能性があること

	目的変数		
	1 型糖尿病	インスリン依存	1 型糖尿病かつインスリン依存
診療データ			
PPV	0.736	0.805	0.782
OOB	0.777	0.778	0.800
カルテデータ			
PPV	0.650	–	–
OOV	0.647	–	–

表 3　表は PPV と OOB を示す.

6. 結　　論

本論文では集合演算子メモリネットワークを導入することにより, 説明変数 8,398 の特徴を抽出して 1 型糖尿病のフェノタイピングを行なった. 実験結果にも述べたが, 集合演算子メモリネットワークを用いて抽出した特徴をそのまま用いただけでは PPV が 65.0 とあまり良い結果を得ることはできなかった. この原因としては, 個々の特徴に複数の特徴が混在した状態である可能性があること, 個々の特徴に関係する患者の数が母集団全体から見るとそれぞれ 0.06%以下の微々たるものが大半であることなどいくつかの原因が考えられた. 診療データや中島のアルゴリズムにあるように, 複数の特徴の集合のマージや個々の特徴集合の洗練が必要であろうと思われたが, 今回これは行なわなかった. これらのマージや洗練化には医学的な深い知識を必要とすることによる.

さらなる研究にはいくつかのルートがある. 一つ目, 集合演算子メモリネットワークは表記揺れを検知できる可能性, 集合演算子のもつ特徴 (∪, ∩ の演算子について閉じていること) まで用いたものではなかった. 語彙として用いた用語から漏れたものを非医学用語として再利用する, 医学用語のナイーブな表記揺れの検知程度以上はしなかった. このための拡張は将来行ないたい. 二つ目, 今回の実験においては, 九州大学病院の 80 万件 (3ヶ月分, 53,065 人分相当) のカルテ情報を用いたが, 図 2 に示したように項目が不足していると考えられデータを増加させる必要がある. データを拡張することにより我々の真のターゲットであるフェノタイピングの同定へ近づけたい.

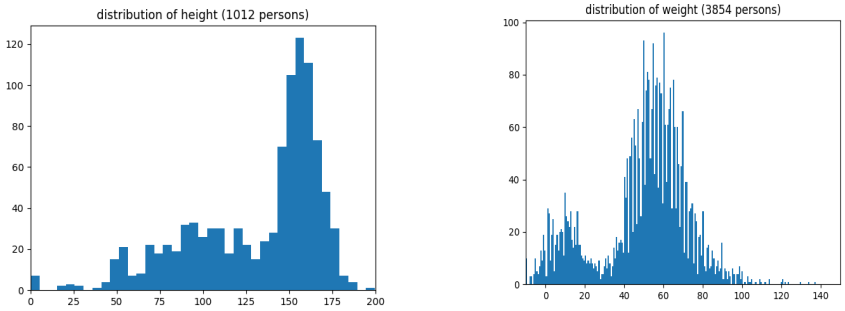

図 1 身長と体重の分布を示す. 左図の身長に対しては 160cm 近辺の分布の山がある. これに加えて, 140cm 以下に対してもなだらかな分布があり, これは子供や生れて間も無い乳児の身長を示す. なお, 200cm や 0cm は読取りミスであった. 右図の体重に対しては分布の山が 2 つ存在し, これは体重の増減値 (相対値), 体重の絶対値の山を示す.

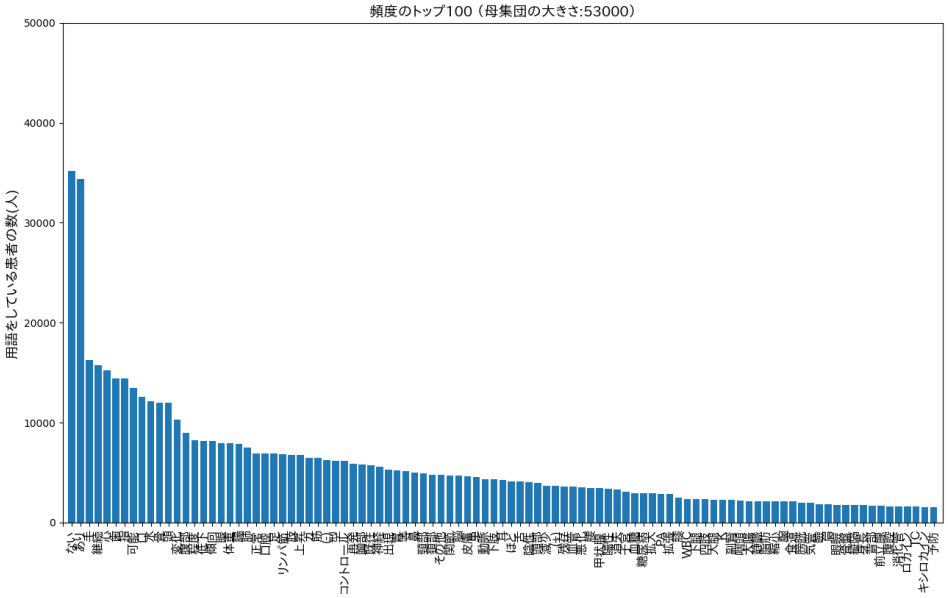

図 2 集合演算子メモリネットワークによる抽出は一項目につき通常複数エントリをもつ. これを患者一人あたりに換算して項目が存在するか否かの統計を取り、項目あたり何人の患者が関連するかの頻度を示したのがこの図である. 頻度の高い方から 100 項目を示す.

三つ目, 今回の実験においては, フェノタイピングの同定を行ない, その結果をフィードバックして, サブクラスの特定を行なうタスクは実行しなかった. こういうルートを辿ると 1 型糖尿病のサブクラスである SPIDDM(緩徐進行 1 型糖尿病) が存在などを検知できる可能性がある.

謝　　辞

田ジマ尚子先生には感謝の意を表します.

◇ 参　考　文　献 ◇

[Aramaki 14] Aramaki, E., Morita, M., Kano, Y., and Ohkuma, T.: Overview of the NTCIR-11 MedNLP-2 Task, *In Proceedings of NTCIR-11*, p. 8 pages (2014)

[Banko 07] Banko, M., Cararella, M., Soderland, S., Broadhead, M., and Etzioni, O.: Open information extraction from the web, *IJCAI* (2007)

[DaumeIII 09] DaumeIII, H.: Frustratingly Easy Domain Adaptation, *arXiv preprint*, p. arXiv:0907.1815 (2009)

[Delude 15] Delude, C. M.: Deep Phenotyping: The details of disease, *Nature (Macmillan Publishers)*, pp. S14, Vol 527 (2015)

[Fernald 11] Fernald, G., Capriotti, E., R, D., KJ, K., and RB, A.: Bioinformatics challenges for personalized medicine, *Bioinformatics*, pp. 27:1741–1748 (2011)

[Koehler 17] Koehler, S., Vasilevsky, N., Engelstad, M., and Foster, E.: The Human Phenotype Ontology in 2017, *Nucl. Acids Res.* (2017)

[Kumar 15] Kumar, A., Irsoy, O., Ondruska, P., Iyyer, M., Bradbury, J., Gulrajani, I., Zhong, V., Paulus, R., and Socher, R.: Ask Me Anything: Dynamic Memory Networks for Natural Language Processing, *arXiv*, p. arXiv:1506.07285 (2015)

[Lu 14] Lu, Y., D.B, G., M, A., and G., C.: Personalized Medicine and Human Genetic Diversity, *Cold Spring Harbor Perspectives in Medicine*, p. 4 (9): a008581a008581 (2014)

[Neelakantan 17] Neelakantan, A., V.Le, Q., Abadi, M., McCallum, A., and Amodei, D.: Learning a Natural Language Interface With Neural Programmer, *arXiv*, p. 1611.08945v4 (2017)

[Snow 05] Snow, R., Jurafsky, D., and Ng, A.: Learning syntactic patterns for automatic hypernym discovery, *NIPS* (2005)

[TheanoDevTeam 16] TheanoDevTeam, : Theano: A Python framework for fast computation of mathematical expressions, *arXiv e-prints*, p. abs/1605.02688 (2016)

[Wenzel 12]　Wenzel, S.: Asthma phenotypes: the evolution from clinical to molecular approaches, *Nature Medicine*, Vol. 18, No. 5, pp. 716–725 (2012)
[Weston 14]　Weston, J., Chopra, S., and Bordes, A.: Memory Networks, *arXiv*, p. 1410.3916v11 (2014)
[Zhou 14]　Zhou, J., Wang, F., Hu, J., and Ye, J.: From Micro to Macro: Data Driven Phenotyping by Densificationof Longitudinal Electronic Medical Records, *In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 135–144 (2014)
[河添 16]　河添 悦昌, 香川 璃奈, 山口 亮平, 桜井 亮太, 篠原 恵美子, 大江 和彦：電子的診療情報からの高次元特徴データを用いた EHR Phenotyping アルゴリズムの開発, 第 36 回医療情報学連合大会 (2016)
[田ジマ 15]　田ジマ 尚子：1 型糖尿病の疫学と生活実態に関する調査研究, 厚生労働科学研究循環器疾患糖尿病等生活習慣病対策総合研究研究成果発表会 (2015)

〔担当委員：××○○〕

19YY 年 MM 月 DD 日　受理