様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２７ 年　 ６ 月　 ５ 日現在

機関番号：３２６０７
研究種目：挑戦的萌芽研究
研究期間：2013～2014
課題番号：２５６７０１８０

研究課題名（和文）テキスト及び画像情報による客観的病理診断過程のモデル化と診断支援システムの開発

研究課題名（英文）Application of text mining techniques and image analysis in pathology diagnosis

研究代表者
原　敦子（Hara, Atsuko）

北里大学・医学部・講師

研究者番号：１０２７６１２３

交付決定額（研究期間全体）：（直接経費）　　2,600,000円

研究成果の概要（和文）：病理診断需要増大に伴い診断支援システムのニーズが増加しているが、病理診断は客観化＝数値化が難しく、ITによる病理診断支援システム開発は極めて遅れている。一方、病理診断は組織像から特徴量を解析する思考過程でありこれを数値化出来れば客観化は可能という発想の元、①診断報告書テキスト及び②標本画像を材料に「客観的病理診断モデル」構築を行った。具体的には①では診断報告書をテキストマイニング法等で解析・数値化、②ではバーチャルスライド装置によって得られた組織標本電子化データを機械学習方式で解析・数値化、さらに推定疾患の候補・診断確率・類似画像等を提示する病理診断支援システムの開発を行った。

研究成果の概要（英文）：We have developed a pathological information data base system and a diagnostic processing model with the use of pathology reports and images of digitized specimens. We first described an algorithm to enable the numeric transformation of pathology reports using both text mining and statistical analysis. Then, pathological diagnosis supporting system was provided, containing, (1) extraction and representation of the similar archival report, (2) calculation of probabilities for possible diseases, (3) consistency verification between diagnosis and details of the reports. Images of digitized specimen were divided into many small images and Wavelet transformation was performed. They were taken as training data and identified by pattern recognition by the K-nearest neighbour method. The result of this identification was used as a feature vector for a specimen image. Similar images were retrieved by comparing the feature vectors of the targeted images and the specimen images in the database.

研究分野：病理学

キーワード：病理診断　テキストマイニング解析　画像解析　診断支援システム

１．研究開始当初の背景

我国では癌患者数が増加傾向にある。それに伴い病理診断の重要性が高まり、大量の検体が病理部門に提出され最終診断が要求されている。一方、病理医数は慢性的に不足し、病理診断の現場は「量との格闘に追われ質の担保にまで手が届かない」のが現状である。この問題解決のためには病理診断を直接的にサポートする「診断支援システム」が不可欠であり、そのニーズはますます増加している。近年、IT 技術の進展で医療情報の電子化や創出されるビッグデータ利活用が活発化し、放射線画像分野ではコンピュータ診断支援システムの実用化が進んでいる。病理分野でも「病理診断報告書」および「組織ガラス標本」の電子化が可能となっているが、両者の電子化データを利活用した病理診断支援システムの開発は極めて遅れている。報告書からのテキスト情報を扱ったものは皆無、組織標本の画像解析は複数の施設で模索されるも悪性腫瘍の領域抽出のみにとどまり、実用診断レベルにまで達していない。その背景には、病理診断は全て人間の“目”で行われる作業であり時に主観的で再現性に乏しく、人工知能解析モデルによる数値化＝客観化が困難であることが挙げられる。

２．研究の目的

さて、病理医は組織標本に含まれる視覚的情報から多くの特徴量を抽出・解析するという複雑な思考過程を介して病理診断を行う。例えば観察された特徴量を **x**、特徴量の重要度係数を **a**、その総和を **f** とすると、

$$f = a_1x_1 + a_2x_2 + ... + a_px_p$$

という式が頭の中で形成され、**f** が一定以上なら悪性、それ以下なら良性等と判断し診断に至る。従ってこの思考過程を数値化出来れば、本来主観的な病理診断も原理的には客観化は可能であると考える。この発想の元、「病理診断報告書」電子化テキストデータ及びバーチャルスライド装置によって得られた「組織標本」電子化画像データを材料にテキストマイニング解析・画像解析を行い、病理医思考過程を数値化した全く新しい「客観的病理診断モデル」の構築を行った。さらにテキスト・画像情報の連携検索により診断確率、推定疾患、類似画像等の提示を行う「実用的病理診断支援システム」の開発も行った。図-1 に研究の具体的手法の概要を示す。

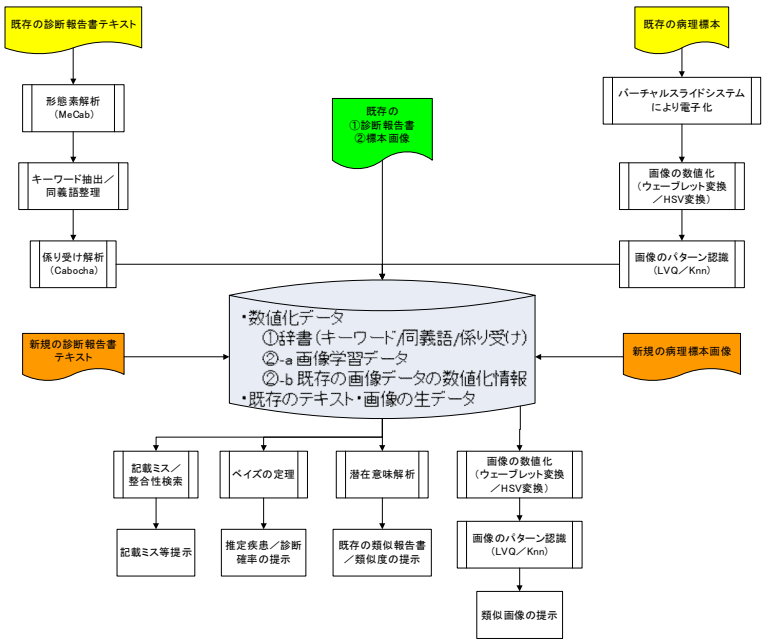

図-1

３．研究の方法

A. 病理診断報告書テキストマイニング解析

材料：北里大学病院で蓄積保管されている電子化された病理診断報告書(2007 年～2011 年、乳腺疾患 1500 症例)。疾患分類は乳癌取り扱い規約（第 17 版 2012 年)に準じた。報告書内容は、患者情報・臨床診断名・受付材料・採取法・病理診断名・病理所見など多岐にわたるが、ここでは病理診断名・病理所見をテキストデータとした。簡略化のため疾患名は記号化した（例： 硬癌→IB2a3 など）

方法：

(1)MeCab を用い、テキストデータを形態素解析（言語で意味を持つ最小単位への分割と品詞の判別）した。

(2)同義語を整理し、診断に関連する数百語の

キーワードを抽出し辞書を作成した。

(3)Cabocha およびオリジナルプログラムを用い、キーワード間の係り受け頻度解析を行った。

(4)以上から得られた数値化情報およびをテキストの生データを、病理情報データーベース(DB)に格納した。

(5)新規症例テキストが与えられた場合、DB 内の情報を基に①潜在意味解析(Latent Semantic Analysis)を用いた既存類似テキストや類似度の提示②ベイズの定理を用いた推定疾患や診断確率の提示③報告書内容の矛盾や記載ミスの提示、を各々行う「客観的病理診断モデル」を構築した。

(6)更にメニュー画面から簡便に上記①②③の各機能表示可能な「病理診断支援システム」を開発した。図-2 はシステム化した類似文章検索・診断確率の表示画面例である。

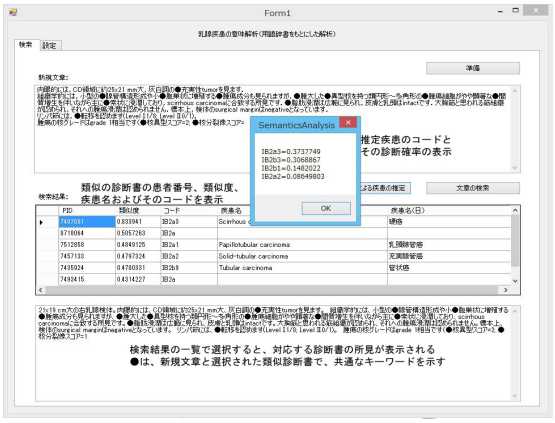

図-2

B. 組織標本電子データの画像解析

材料：北里大学病院で蓄積保管されている病理組織標本(2007 年～2014 年)で、報告書のテキスト解析を既に行った乳腺症例 1500 件。疾患分類は乳癌取り扱い規約(日本乳癌学会編第 17 版 2012 年発行)に準じた。

方法：(図-3 に画像解析概要を示す)

(1)選択された診断する部分の画像（100 倍で 1024×768 ピクセル）の数値化

画像ファイル（100倍画像）から1024×768ピクセルの領域を選択し、128×128ピクセルの大きさの165個の小画像（この小画像は疾患の

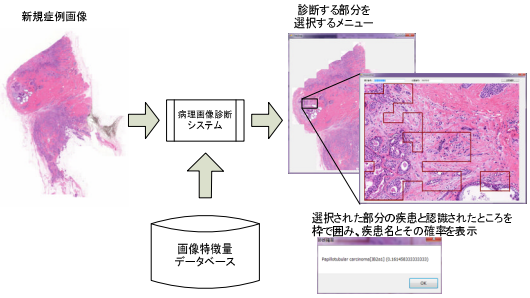

図-3

違いを認識できる大きさである）を抽出。その後、各小画像の特徴量をウェーブレット変換により数値化した（例えば、ウェーブレット変換で深さが 6 とすると、1 個の診断すべき画像は 18 変数×165 個の数値データに変換される)。

(2)k 近傍法を用いたパターン認識

各小画像が、画像特徴量データベースのどの特徴量に最も類似しているのかを検索し、各小画像がどの疾患に分類されるのかをパターン認識により調べた。画像特徴量データベースの構造は、既に診断されて疾患名が分かっている画像から小画像（128×128 ピクセル）を取り出し、学習データとしてウェーブレット変換しデータベースとして蓄積されたものである。ここでは、k 近傍法で k=1 の場合、かつ閾値 T を設けてパターン認識を行った。k 近傍法によるあるオブジェクトの分類は、その近傍の学習データ群の投票によって決定される（すなわち、k 個の最近傍の学習データ群で最も一般的なクラス（疾患）をそのオブジェクトに割り当てる)。k は正の整数で一般に小さい値であり、k = 1 なら最近傍の学習データと同じクラスに分類される。ただし、最近傍の学習データとの距離が T より大きな場合は、どのクラス（疾患）にも分類されない。

(3)高速処理のための並列化処理

学習データは疾患によっては約 21000 件あり、上記のウェーブレット変換と k 近傍法による処理は非常に時間がかかる。これを解決するため、複数のプロセッサによる並列処理

を行うことで処理時間を短縮した。(図-4 に並列処理の概要を示す)

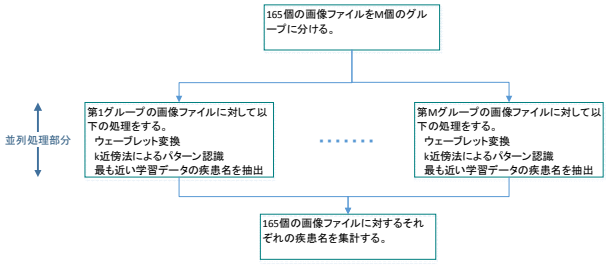

図-4

４．研究成果

(1)病理診断報告書テキストマイニング解析

①既存類似テキストや類似度の提示：提示された上位 5 位に新規症例と同一疾患名(診断名）の既存テキストが含まれていた場合を正解とすると、正解率（＝類似度、0.00～1.00 で表示）は全体で 0.91 であった。疾患ごとで見ると、特徴的キーワードを持つ疾患群（葉状腫瘍、乳腺症など）や件数が比較的多い疾患群（非浸潤性乳管癌など）で類似度が高かった。

②ベイズの定理を用いた推定疾患や診断確率の提示：本システムでは確率が 0.05～1.00 以上の場合に推定疾患・診断確率の高い順に表示する（IB2a3(硬癌): 0.38, IB2b3 (浸潤性小葉癌): 0.31 など)。疾患ごとで見ると症例数の過少とは無関係に特徴的キーワードを持つ疾患群で正解率が高かった（線維腺腫: 0.92, 異型乳管過形成: 0.91 など)。

③報告書内容の矛盾や記載ミスの提示：本システムでは、スペルミス・記載ミス(左右、臓器名、診断名、記号など)・論理的医学的矛盾(例:診断名が fibroadenoma(良性)→所見内に悪性所見ありと記載)の検出を可能とした。既存の全 1500 症例テキスト検索では 8 例で記載ミスを検出した。また 50 例のデモ例題では記載内容や論理的矛盾のミスを、組織学的波及度項目を除いてはほぼ 100%のミス検出が可能であった。

(2)組織標本電子データの画像解析

これまで採用していたニューラルネットワークや SVM などの方法では、分類する疾患の種類が増えたり学習データの個数が増えることにより、分類の精度が悪くなったり処理速度が多くかかった。またどこかに分類する必要があるため、疾患ではない部分（間質、血管、脂肪など）についても学習データとする必要があった。今回採用した k 近傍法（k=1）では閾値を設けているため、分類すべき疾患部分のみの学習データだけで十分であり、また、分類の種類が多い場合、学習データの件数が多い場合も問題なく処理が可能であった。分類の問題があるような症例があれば、その症例から新たに学習データを追加することにより精度を向上させることも可能となった。

高速処理のための並列処理についてはプロセッサの個数により、下表のように処理時間が短縮された。前処理と後処理にも時間がかかるために、プロセッサの個数は 4 個の場合が最適であった。

プロセッサの数	処理時間（秒）
1	11.388
2	6.926
3	5.7076
4	5.5536
5	5.880

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 1 件）

原　敦子、三枝　信、石橋雄一．病理診断におけるテキストマイニングの応用．計算機統計学．査読有．28(1), 2014, 1-12.

〔学会発表〕（計 5 件）

①原　敦子、石橋雄一：“病理診断における画像とテキストの解析　2014 年度統計関連学会連合大会(20140904). 会場名　東京大学(東京都文京区)

②石橋雄一、原　敦子：“大容量画像データの解析と検索　第 27 回大規模データ科学に関する研究第 1 回研究集会(20140718). 会場名　北海道大学(北海道札幌市)

③原　敦子、石橋雄一、三枝　信：“病理診断におけるテキストマイニングの応用－報告書の医学的・論理学的矛盾や記載ミスチェックシステムの開発　第103回日本病理学会総会(20140426). 会場名　広島国際会議場(広島県広島市)

④原　敦子、石橋雄一、三枝　信：“テキスト／画像情報による客観的病理診断のモデル化　　第 102 回日本病理学会総会(20130607). 会場名　さっぽろ芸文館(北海道札幌市)

⑤原　敦子、石橋雄一：“病理診断におけるテキストマイニングの応用　第 27 回日本計算機統計学会(20130517). 会場名　弘前大学(青森県弘前市)

〔産業財産権〕
○出願状況（計 1 件）

名称：病理診断報告書作成支援装置
発明者：石橋雄一、原　敦子
権利者：同上
種類：特許
番号：2013-244352
出願年月日：2013 年 11 月 11 日
国内外の別： 国内

６．研究組織
(1)研究代表者
原　敦子 (Hara Atsuko)
北里大学・医学部・講師
研究者番号（10276123）