

平成28年度北海道大学情報基盤センター共同研究成果報告書

1. 研究領域番号 A4 コンテンツ, データ, メディア教育
2. 研究課題名 大規模医療情報に対する高度統計技法の開発と可視化
3. 研究期間 平成28年 4月 1日 ~ 平成29年 3月31日

4. 研究代表者

氏名	所属機関・部局名	職名	備考
栗原 考次	岡山大学大学院・環境生命科学研究科	教授	

5. 研究分担者

氏名	所属機関・部局名	職名	備考
水田 正弘	北海道大学・情報基盤センター	教授	
南 弘征	北海道大学・情報基盤センター	教授	
富田 誠	東京医科歯科大学・医学部附属病院	准教授	
石橋 雄一	株式会社スタットラボ	代表取締役	
田中 豊	岡山大学	名誉教授	
石岡 文生	岡山大学大学院・環境生命科学研究科	准教授	

6. 共同研究の成果

下欄には、当該研究期間内に実施した共同研究の成果について、その具体的内容、意義、重要性等を、共同研究申請書に記載した「研究目的」と「研究計画・方法」に照らし、800字～1,000字で、できるだけ分かりやすく記載願います。文章の他に、研究成果を端的に表す図表を貼り付けても構いません。なお、研究成果の論文・学会発表等を行った実績（発表等の予定を含む。）があれば、あわせて記載して下さい。

本研究では、医療分野における画像、テキスト、ゲノムなどの大規模情報に対してデータベース化及び高度な臨床診断及び予測のための統計技法開発とその可視化を行うために、以下のような課題について検討した。

1. 画像及びテキスト情報に基づく医療診断と予測

病理組織標本画像及び病理診断書において、乳腺疾患患者に対する画像及びテキスト情報の効率的なデータベース化を行ない、新規の画像情報、テキスト情報に対して、類似の標本画像や病理診断書が検索可能なシステムになっている。画像情報では、ガラススライド標本の画像ファイルを数値化し、新規画像に対して可能性のある複数の疾患を確率で示すことと、データベースの類似の画像標本の検索を可能にした。1つの画像の解析には1時間以上要する場合もあるので、数台のマシンによる並列処理、あるいは1台のマシンの複数プロセッサによる並列処理により処理時間を短縮する試みも行った。さらに、ディープラーニングにより数値化せずに画像のままパターン認識を行うことを試みて、学習データの作成方法に工夫をすることが必要であることが認識された。テキスト情報では、病理診断書に関してテキストマイニングを用いた解析を行って、診断書テキストから特

(研究成果のつづき)

定の疾患である確率の計算や、類似の診断書をデータベースから検索することも可能にした。さらに、診断書の内容と診断名との医学的・論理的な整合性のチェックを可能にした。乳腺の診断書に加えて、胃や大腸などの臓器への拡張や、記号に関する矛盾のチェック、臓器の推定機能や学習機能など実用化に向けた機能強化を行った。

2. 空間構造を利用した遺伝・精神保健データの解析とモデル選択

SNP 間の組換え頻度により、組換えがほとんど起こっていない連鎖不平衡の領域を絞り込む方法をハプロタイプ・ブロック同定という。連鎖不平衡(LD)を含めた疾患関連候補座位の検出について、SKAT の理論に基づいた Echelon スキャンを適用するよりも線形モデルを応用したアプローチを試みた。また、連鎖不平衡(LD)についての連鎖不平衡係数の性質と利用について、テキストの形にまとめた。さらに、これら空間集積性の手法を、福島第一原発周辺の放射線量モニタリングポストデータに適用を試みた。具体的には、政府が定めた帰還困難区域を含む 40km 四方のエリアを 4 次メッシュ (500m×500m) に区分化し、その各々のメッシュにおける空間線量率を対象に Echelon スキャンを適用した。その結果、統計的に有意に高い放射線量を示す地域が同定され、さらにその範囲が天候の影響により大きく変化することが示された。

上記のテーマに関して研究会を 2 回開催した。各研究会では、全国から統計科学、情報科学、医学関係の研究者が参集し、有意義な討論と意見交換をすることができた。