様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２５年　５月　31 日現在

機関番号：13302
研究種目：基盤研究（Ｃ）
研究期間：2009～2011
課題番号：21500135
研究課題名（和文）：解釈指向マイニングによる診療情報からの医学的知見の発見

研究課題名（英文）：Finding Medical Knowledge from Clinical Data based on Interpretations

研究代表者
河崎　さおり（KAWASAKI SAORI）
北陸先端科学技術大学院大学・先端領域社会人教育院・特任准教授
研究者番号：40377437

研究成果の概要（和文）：
専門医の医学的関心に近づく医学データマイニングために、データマイニング結果と医学文献から獲得する背景知識と専門家の見解の連携を図り、血液検査結果と肝炎の治療法の著効性に関する傾向を調べた。また関連の問題に関する遺伝子型への医学的注目を踏まえ、肝炎ウイルスの配列パターンと治療法の著効性との関係に取り組み、主に公開配列データを対象とする準教師付学習手法を開発・改良し、計算手法による可能性を示した。

研究成果の概要（英文）：
The focus of this research is to obtain more interesting mining results for the medical experts, so that not only the data mining from the target clinical data itself but also the combining the background knowledge extraction from medical litarature and experts' interpretations have been pursued in finding the relationships among blood tests and the treatment effects. Additionally, in accordance with the high attention to sequence analysis in biomedicine, semi-supervised learning methods specially for the virus subgenotypes and the interferon/ribavirin treatments were developed to show the potentials of computational approaches.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2009 年度	1,200,000	360,000	1,560,000
2010 年度	1,100,000	330,000	1,430,000
2011 年度	1,100,000	330,000	1,430,000
年度			
年度			
総　計	3,400,000	1,020,000	4,420,000

研究分野：知識発見
科研費の分科・細目：情報学・知能情報学
キーワード：データマイニング，医療データ，専門知識，後処理，解釈モデル

１．研究開始当初の背景

データマイニングの戦略的な実践は多くの分野で当然のことになりつつあり、疫学をはじめ統計学的なデータ処理の伝統を持つ医学分野でも、病院情報の電子化の進展を受けた医学データを対象とする知識発見の試みが増えた。ライフサイエンス分野で、医学文献データベース MEDLINE、厚生労働省主催の統合 DB プロジェクト、各種遺伝子系 DB など様々なデータベースが公開され、ネット等を

経由して利用できる環境整備も進んできたことを受け、医学知識発見でも対象とする特定のデータベースだけでなく、医療文献などの他の情報源から得られる知識を組合せて結果の質の向上を図る研究も増えつつあった。
データベースからの知識発見(以下 KDD と略記)は、一般には①対象分野の理解とマイニング課題の設定、②データの前処理、③データマイニング（パターン/モデルの抽出)、④パターン/モデルの解釈と評価、⑤新知識の実利用、の 5 段階プロセスとされ、利用者の価値観を反映しつつ各段階を繰返すことで興味深く役に立つ新知識の発見を目指す。中でも④の「パターン/モデルの解釈と評価」は DM 結果の評価として利用者の果たす役割が大きく、このステップに対しては、様々な効果的な視覚化による支援(Xerox 社 Hyperbolic Tree 等)が一般的である。また、専門知識が特に重要な医学 DM では、結果の質をあげ評価者である多忙な医師の負荷を低減しるために重要度の高い結果を絞込むアプローチとして、客観的・主観的な種々の評価指標への医師の関心度の反映、学習結果に対する複合的な統計的有意性評価手法なども提案されてきた。
過去に、ユーザ中心データマイニングシステム D2MS、規則の統計的有意性フィルタ法の提示、知識管理的考え方を KDD 過程に導入した多種情報源を活用する統合的アプローチなどを専門医との医学データマイニングプロジェクトにおいて開発してきた経緯から、医師が直感的に理解しやすい表現を好むとともに統計的有意性を重視すること、医学データにデータマイニングの諸技法を適用し直に得られる関係やモデルと実用的な医学的知識との間には質的な隔たりがあること、規則同士の比較や医師自身の知識や経験を根拠として加味することで、評価対象の規則に対する確信を持って判断を施すことを確認した。こうした解釈のモデルを反映することで DM 結果をより豊かにし、医師の評価ステップが円滑化され、新知見の発見に近づき易くなると考える。

２．研究の目的
本研究は、診療情報からの新規性が高く有用な医学的知見の発見を目的とし、データマイニングの後処理(=解釈・評価)ステップに注目し、評価者である医師が診療マイニング結果を確信を持って解釈・評価・判断するための枠組みの開発を狙った。具体的には、図１に示すように、千葉大学医学部附属病院が蓄積する 28 年間の診療データから慢性疾患に関する計算処理を通じて得られる疾病分類や予後予測などのマイニング結果について、(1)医学的解釈のモデル化を試み、そのモデルを元にマイニング結果の規則に対し、例えば規則の条件をなす検査項目の文脈情報を組合せるなどの(2)情報連携・統合により DM 結果を再構成する手法を開発し、(3)実診療データからの新医学的知見発見を試みることを目指した。

３．研究の方法
本研究は、千葉大学病院の診療データ(初期的には慢性肝炎患者の検査結果)に関する長期時系列データを対象に、疾患の識別や予後予測などを分析するための時系列データに対する時系列抽象化手法を継続し、予測を目的とする学習手法から得られる結果を手がかりとして、「研究目的」に記載した(1)医学的解釈のモデル化、(2)情報連携・統合により結果を再構成する手法の開発、(3)実診療データからの新医学的知見発見について、まず既存の学習手法を適用し、結果の評価収集・分析による解釈モデルのプロトタイプの構築を中心に取り組み、その後、プロトタイプをデータ管理に統合する手法の開発、およびその統合情報を活用した診療データからの医学的知見の発見に取り組むこととした。

図１　本研究の３つの目的と研究体制

(1) 医学的解釈のモデル化
① データマイニング結果の評価収集・分析による解釈モデルのプロトタイプの構築：医師の評価時の指摘・疑問等の会議記録の収集のために、診療情報データについて課題の設定とそのデータマイニングを実施し、得られる識別規則の評価および手法への印象を確認する。長期時系列に関する特徴抽出と機械学習を組み合わてマイニングを行う。具体的には、千葉大病院からの診療データの提供を受け、医師からそのデータに対する医学的関心を確認し対応する学習手法を提案し課題を確認したうえで、データの前処理およびデータからの規則の学習を行い、獲得した学習結果の評価会を実施するとともに、評価の際の解釈の要件を整理する。
② テキスト処理による医学文献からの診療記録項目の文脈の抽出：診療データを構成す

る検査項目に関する文献上の記述を収集し本研究用に標準化し、医学記述に関する関連辞書を作成する。
主な実施項目：a)関心辞書の作成、b)医療文献から記述の収集、c)収集した記述の標準化、d) MeSH と共起を元にラフ集合の語彙近似などによる関係の取得し、パッセージ学習等を利用し、標準語彙に関する医学記述について辞書を作成する。
(2) 情報連携・統合により学習結果を再構成する手法： 解釈モデルのための規則学習結果の補完情報の表現:規則間の関係、診療データ項目間の関係などに関する構造を想定し、その表現について検討し、関係 DB 内にデータと学習手法と学習結果の動的な管理を試みる。
(3) 実診療データからの新医学的知見発見： (1)①の DM 結果の評価において、内容の蓋然性も含めて興味深いと判断されるものについては、専門医の追跡調査・実験に委ねる。その他、専門医との検討・評価過程で提起されるデータに関して医学的なデータ駆動型の関連課題に取り組み手法の提案を行う。

４．研究成果

(1) H21 年度は、データマイニング結果に関する医学的解釈モデルのプロトタイプ構築を中心課題と設定し、主に①診療データからのマイニング結果への専門医の評価に基づき、評価過程での議論や専門医の指摘事項をもとに結果の解釈に必要な要素等として収集・整理を実施したほか、②医学的背景知識の収集の一環として、医学文献、特に Medline からのテキストマイニングを行った。①に関しては、千葉大学病院と北陸先端科学技術大学院大学（JAIST)、また合同の国際学会発表の機会に、研究分担者および連携研究者との計 5 回の会合を実施し、(a)診療データのうち肝炎患者に着目してデータ提供を受けることの確認、(b)マイニング手法と疾病に関する相互説明と課題の設定、(c)診療データに対する前処理・時系列抽象化等のマイニング手法の適用、(d)マイニング結果の専門医の評価会を実施することで、評価過程での議論や専門医の指摘事項を収集し、専門家が結果を解釈するうえで必要な要素の洗い出しを実施した。また、この過程で、肝炎の専門医の観点から医学的新知識を考える場合、診療記録のみに関する規則やパターンの発見とは別に、HCV の RNA の変異など omics 情報も考慮したマイニングの可能性について提案があり、議論を進めている。②に関しては、診療記録上の主要検査項目に着目し、これらを関心辞書とし、各項目とそれに関する記述を抽出・収集・標準化し、文脈リストを作成した。また、項目間の関係・規則間の関係付けのための将来的な文脈情報を狙い、レランス・ラフ集合モデルに基づく代替表現を利用して強い関係を持つ項目の組合せパターンの獲得を試みた。
H22 年度は、本研究の目的である、A.医学的解釈のモデル化、B. 情報の連携と統合によるマイニング結果の再構成する手法、C.実診療データからの新医学的知見発見の試み、の中でも特に B および C について、前年度の肝炎専門家から提案をうけた HCV の RNA の変異など omics 情報も考慮したマイニングを試みることでアプローチした。特に、Omics データについては、標準的な治療法であるインタフェロン／リバビリン併用の効果に対し、HCV ウイルス内の NS5A 領域の変異の影響が医学的に注目されているという専門医からの示唆に基づき、ウイルスの配列データの取得・整備を行うとともに、ウイルス遺伝子亜型および著効性に特徴的なパターンを抽出するアルゴリズムの開発を行い、KICSS2010 にて準教師付学習によるアプローチと初期的な成果について発表した。これは、Los Alamos 研究所の HCV データに著効性情報を伴うデータが収集されているものの件数が少なく、一方、GenBank や名古屋市立大学では著効性情報がないものの数千件の HCV 配列が収集されているという状況を踏まえた 2 種類のデータを組み合わせる準教師付学習の枠組みで、ウイルス遺伝子亜型および著効性に関する識別パターンの獲得を試みたものである。また、診療データについて新たな取り組みのためにマイニング課題および新規データの提供を受けることとし、投薬量と期間のデータをこれまでの時系列パターンと組み合わせつつ、医学的関心にあわせたマイニング結果と医学文献からの背景知識との連携を再検討したものの、上位レベルの医学的な知識獲得には至らなかった。
医学的知識の発見に関しては、引き続き肝炎ウイルスの遺伝子亜型パターンと治療法の著効性との関係を明らかにするために、主に公開配列データを対象とする準教師付学習手法の改良に取組んだ。多量の遺伝子亜型配列が公開されながらも治療の効果が登録されている遺伝子亜型が極端に少ないという状況に対し、効果の有無を特徴付けるパターンのうち、配列中に 1 度しか出現しない特定配列パターン DOOPS を見つけるための exhaustive search 手法、および配列中に繰返し出現する特定パターン DMOPS を見つける separate-and-conquer 学習手法を提案し、C 型肝炎ウイルス内の NS5A ドメインへの適用について ECML/PKDD2011 併設ワークショップにて報告したほか、ACIIDS2012 でも予測精度の向上について報告した。
更に、医学的新知識の一例として遺伝子発現

抑制につながるsiRNA配列の識別問題に取り組み、Apriori をベースとする記述的学習手法および冗長な結果のフィルタリング手法を提案し ISKSS12 にて発表した。一方、HCV 遺伝子亜型の識別モチーフ発見のための準教師付学習については、クラス判断の際にクラスタ的性質を仮定するアンサンブル学習手法の開発による学習性能向上成果を雑誌報告予定である。

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７件）

1. Ho, B.H., Le, N.T., Ho, T.B.: Quantitatively assessing the effect of regulatory factors on nucleosome dynamics, *Journal of Ambient Intelligence and Humanized Computing*, Peer Reviewed, Vol. 3, Issue 4, 2013, pp.265-280
2. Nguyen, T.P., Ho, T.B.: Detecting Disease Genes Based on Semi-Supervised Learning and Protein-Protein Interaction Networks, *Artificial Intelligence in Medicine*, Peer Reviewed, Vol. 54, 2013, pp.63-71.
3. Le, N., Ho, T.B., Kanda, T., Kawasaki, S., Takabayashi, K., Wu, S., Yokosaka, O.: A Semi-Supervised Learning Method for Discriminative Motif Finding and Its Application, *Journal of Universal Computer Science*, peer reveiwed, Vol.19-No.4, 2012, pp 563-580.
4. Le, N.T., Ho, T.B., Ho, B.H.: Sequence-dependent histone variant positioning signatures, *BMC Genomics*, Peer reviewed, Vol. 11 (Suppl 4), 2011, pp. 1-9.
5. Luong, T.D., Ho, T.B.: Enhancing Privacy in Distributed Data Clustering, Journal of Computer Science and Cybernetics, peer Reviewed, Vol. 26, No. 2, 2011, pp. 1-15.
6. Kawasaki, S., Ho, T.B., Kanda, T.: Discovering Relationship between Hepatitis C Virus NS5A Protein and Interferon/Ribavirin Therapy, *Knowledge, Information and Creativity Support Systems KICSS2010 Revised Selected Papers*, Peer Reviewed, LNAI 6746, 2011, pp.79-90.
7. Le, N.T., Ho, T.B., Tran, D.H.: Characterizing nucleosome dynamics from genomic and epigenetic information using rule induction learning, *BMC Genomics*, Peer Reviewed, Vol.10(Suppl.3), 2009, pp.S27 (1-10).

〔学会発表〕（計 10 件）

1. Bui, N.T., Ho, T.B., Kawasaki, S. : An Effective Method for Generating siRNA Design Rules, *The 5th Asian Conference On Intelligent Information and Database Systems, ACIID 2013*, 18-20 March (2013), Kuala Lumpur/Malasia (LNAI 7803 pp. 196-205).
2. Ho, T.B., Takabayashi, T., Kanda, T., Kawasaki, S., Le, T.N., Bui, N.T., Than, Q.K.: From Clinical to Genomics Data in Hepatitis Study, *The First Asian Conference on Information Systems*, 6-8 December (2012), Siem Reap/Cambodia.
3. Bui, N.T., Ho, T.B., Kawasaki, S.: A Sequential Apriori Algorithm for Discriminative Design Rules of Effective siRNA Sequences, *13th International Symposium on Knowledge and Systems Science*, 19-20 November (2012), 石川県金沢市.
4. Than, K., Pham. N.K., Nguyen, D.K., Ho, T.B.: Supervised dimension reduction with topic model, *4th Asian Conference on Machine Learning 2012*, 4-6 November (2012), Singapore/ Singapore.
5. Le, N., Ho, T.B.: A Semi-Supervised Method for Discriminative Motif Finding and Its Application to Hepatitis C Virus Study, *4th Asian Conference on Intelligent Information and Database Systems ACIIDS 2012*, 19-21 March (2012), Kaohsiung/Taiwan.
6. Ho, T.B., Kawasaki, S., Le, N.T., Kanda, T., Le, T.N., Takabayashi, K., Yokosuka, O.: Finding HCV NS5A Discriminative Motifs for Assessment of IFN/Ribavarin Therapy Effect, *Workshop Data Mining in Genomics and Proteomics, International Conference ECML/PKDD*, September 5-9 (2011), Athens/Greece (pp.32-42)
7. Pham, N.K., Ho, T.B. Mining parallel documents across Web sites, The Sixth Asia Information Retrieval Societies Conference AIRS 2010, 1-3 December, (2010), Taipei.Taiwan.
8. Kawasaki, S., Ho, T.B., Kanda, T., Yokosuka, O., Takabayashi, K., Le, T.N.: Discovering Relationship between Hepatitis C Virus NS5A Protein and Interferon/Ribavirin Therapy, *Fifth*

International Conference on Knowledge, Information and Creativity Support Systems KICSS'10, 25-27 November (2010), ChiangMai/Thailand.

9. Ho, B.H., Le, N.T., Ho, T.B.: Quantitatively assessing the effect of regulatory factors on nucleosome dynamics, *IEEE-RIVF International Conference on Computing and Communication Technologies*, 1-4 November (2009), Hanoi/Vietnam.
10. Ho, T.B., Takabayashi, K., Pham, T.H., Nguyen, T.P., Kawasaki, S., Tran, D.H.: Towards service-oriented knowledge discovery in biomedicine research, *International Workshop on Third Generation Data Mining, ECML/PKDD 2009*, 7-11 September (2009), Bled/Slovenia, (pp.100-113).

〔図書〕(計2件)

1. Ho, T.B.: CRC Press and Taylor & Francis. Knowledge Discovery (Chapter 4 of, *Knowledge Technology and Science* edited by Y. Nakamori), 2011, pp.57-81
2. Nguyen, T.P., Ho, T.B.: Springer-Verlag. Mining multiple biological data for reconstructing signal transduction networks (*Data Mining: Foundations and Intelligent Paradigms* edited by D.E. Holmes & L.C. Jain), 2011, pp.163-185 (380).

〔産業財産権〕
○出願状況(計0件)

名称:
発明者:
権利者:
種類:
番号:
出願年月日:
国内外の別:

○取得状況(計0件)

名称:
発明者:
権利者:
種類:
番号:
取得年月日:
国内外の別:

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
河崎　さおり (KAWASAKI SAORI)
北陸先端科学技術大学院大学・先端領域社会人教育院科・特任准教授
研究者番号:40377437

(2)研究分担者
Tu・Bao Ho (ツー・バオ・ホー)
北陸先端科学技術大学院大学・知識科学研究科・教授
研究者番号:60301199

(3) 連携研究者
高林　克日己 (TAKABAYASHI KATSUHIKO)
千葉大学・医学部附属病院・教授
研究者番号:90188079

神田　達郎 (KANDA TATSUO)
千葉大学・医学研究院・特任講師
研究者番号:20345002