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　２８　年　　６　月　２２　日現在

機関番号：　１２６０８
研究種目：　若手研究(B)
研究期間：　2012～2015
課題番号：　２４７００１４０

研究課題名（和文）複雑かつ大規模なデータ処理のためのデータマイニング及び機械学習法

研究課題名（英文）Scalable data mining for processing complex and large-scale data

研究代表者
田部井　靖生（Tabei, Yasuo）

東京工業大学・情報理工学（系）研究科・東工大特別研究員

研究者番号：２０５８９８２４

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：本研究プロジェクトでは，大規模データを処理する上で重要な検索技術と圧縮技術を応用した大規模機械学習技術に焦点をあてて研究開発を行い成果を挙げることにできた．主な成果としては，大規模ネットワーク検索技術，大規模データに対する分類器の学習手法の開発である． 開発した手法の論文は，分野におけるトップの国際会議に採択され研究発表を行ってきた．さらに，手法を実装したソフトウェアーを公開し，世界中の研究者・技術者に利用され始めている．

研究成果の概要（英文）：We developed scalable similarity search and machine learning techniques for processing complex and large-scale data in this project. For example, we developed a large-scale similarity search for network data and scalable learning method for training linear classifiers. Papers on the developed methods was accepted to premier conferences. In addition, softwares implementing the developed methods are publicly available and have been used by researchers and engineers in all over the world.

研究分野：　計算機科学

キーワード：　データマイニング　機械学習　類似検索　ケモインフォマティクス

１．研究開始当初の背景

データマイニング及び機械学習とは，データの背後に潜む規則を自動的に抽出，獲得させる技術の総称である．これにより，人間がタスクに合わせてコンピュータを制御するプログラムを直接つくる必要がなくなり，効率よくデータを処理することが可能となる．近年の様々な応用場面では，データが高度に大規模化，複雑化している．例えば，コンピュータビジョン分野では，米国マサチューセッツ工科大学のグループがインターネット上から集めた画像データベースには約 8,000 万の画像が蓄積されている．画像データは GIST や SIFT などの画像特徴抽出技術を用いて数百次元の実数値ベクトルとして表現することができる．ウェブ情報処理分野では，twitter やフェイスブックなどのソーシャルネットワークから作られた大規模ネットワークデータが存在する．ケモインフォマティクスにおいては，米国国立生物工学情報センターの化合物データベース PubChem には約 2,800 万の化合物が蓄積されている．ゲノムデータにおいては、同研究所のゲノムデータベースには次世代高速シーケンサーから読み取られたゲノム配列が大量に格納されている．

このような状況から，データから統計的な情報をもとにデータの背後に潜む規則を自動的に抽出，推定しようというデータマイニング及び機械学習の研究が近年盛んに行われている．従来の研究に共通することは，入力データがベクトルや集合として表現されていることを仮定し，その上でデータ中の規則抽出をするということである．

しかし，近年の応用場面ではこの前提が成り立たないことが良くある．例えば，上記のソーシャルネットワークは，数百万から数億のノードからなる大規模グラフとして表現される[2]．化合物は，各原子をノードとするグラフとして表現する事ができる．１つ１つのグラフは数十のノードからなる比較的小さいグラフであるがその数は膨大である．ゲノムデータはアデニン(A)，グアニン(G)，シトシン(C)，チミン(T)の４種類からなる文字列として表現され，文字種類数が少ないゆえ特徴がつかみにくく解析が困難であることが知られている．また，これらのデータは大規模であるが，研究は比較的小規模なデータを対象として行われているために，必ずしも実際の大規模データに対して適応可能とは限らない．それゆえ，データ中に含まれる有意義な情報を取り出す効率的な手法の研究開発が，現代社会における緊急の課題となっている．

２．研究の目的

本研究の目的は，このような複雑かつ大規模なデータに対して適応可能なデータマイニング，機械学習法を最新の理論計算機科学の成果を用いて開発することである．具体的には，簡潔データ構造，Locality Sensitive Hashing (LSH)，二分決定グラフなどを応用する．本研究プロジェクトに先立ち研究代表者は，大規模なベクトルデータに適応可能な高速な全点間類似度検索法に関する研究，大規模グラフのための高速な類似度検索法に関する研究を行なってきた．それらの研究成果はトップレベルの国際会議にアクセプトされ，タンパク質からのモチーフ抽出や創薬に応用するプロジェクトも立ち上がっている．しかしこれまでの研究は，大規模機械学習法やネットワーク検索などの研究に関しては議論されてこなかった．そこで本研究プロジェクトでは，これまで開発してきた大規模データマイニング手法をさらに発展させ，より複雑かつ大規模な対象に適応できるように拡張する．

３．研究の方法

平成 24 年度

研究は，簡潔データ構造の一つであるウェーブレッド木をネットワーク検索に応用するところからはじめる．2011 年に SIAM international conference on data-mining にて，我々は整数配列上の self-index の一つであるウェーブレッド木を用いて化合物のような大量の小さいグラフからクエリーグラフと類似するグラフを高速に検索する手法を提案した．ネットワーク検索問題は大きなグラフからエラーを許してクエリーグラフの埋め込みを検索する問題であり，ネットワークデータの大規模化に伴い近年データベース分野のトップカンファレンス SIGMOD などで毎年のように手法の提案がされる重要な問題の一つになりつつある．提案手法を拡張することにより先行研究よりも効率よくネットワーク検索問題をとけることができるようにする．

同時に，コンパクトなデータ表現に関する機械学習の研究にもとりかかる．Locality Sensitive Hashing(LSH)などの実数値ベクトルをコンパクトに表す方法の研究が理論計算機科学の分野が行われている．LSH を用いることにより，ユーグリッド距離で定義された実数値ベクトルの問題がハミング距離で定義されたバイナリ文字上の問題へ変換することが出来る．これまでは，情報検索の場面などでよく使われてきた手法であるが，これを機械学習の問題へと応用させる．機械学習手法の最も代表的な手法であるカーネル法は学習時に二次計画問題を解かなければならないことや分類時にクエリーあたり学習データ個数分のオーダーの時間がかかるため大規模問題に応用することができないなどの問題があった．この問題に対して LSH を適応しデータをコンパクトに表現することによりこの問題を解決する．カー

ネル法はいろいろな応用があるため，ここで開発された手法は，密度比推定や超高次元相関解析などに適応することが可能であると考えられる．

平成２５年度以降

初年度に行ったアルゴリズムを実問題に応用する．開発したネットワーク検索問題はケモインフォマティクスにおいても重要な問題である．開発した手法を化合物データベース PubChem に登録されている約 2,800 万の化合物に対して適応する．これにより，創薬の場面などで役立つシステムを開発することが可能になることが可能となる．また、バイオインフォマティクスにおいても，提案手法は応用可能である．提案法は，数は少ないながらもサイズが大きいタンパク質のようなグラフの検索に対しても有効であると考えられる．この研究は，産業技術総合研究所の研究員と協力して進めていく予定である．

コンパクトな表現による機械学習法の応用として化合物とタンパク質の相関解析がある．従来は，カーネルをテンソル積を計算することで作っていたために超高次元になり大規模データに対して応用することは困難であった．提案法を応用することによりデータがコンパクトになりより効率的にこの問題を解くことが可能となる．この研究はフランスのキュリー研究所の研究員と協力して進める．開発したアルゴリズムはウェブで公開することを念頭においているため，C++やＭＡＴＬＡＢなどの汎用的な高級言語を用いてソフトウェアーパッケージを作成することにする．その際，協力研究員と協力し，効率よくソフトウェアー開発を行う．開発したソフトウェアーパッケージは Google code などにオープンソースプロジェクトとして公開し，研究者のみならずエンジニアにもアクセスしやすいようにする．

応用研究を効率よく遂行するために，国内外の大学・研究所・企業の研究パートナーとの連絡を密に取り続けることに特に気をつける．必要があれば，直接現地を訪問し，共同作業などに参加し，研究プロジェクトの円滑な遂行を促すよう努力する．また，国内外の最新の関連研究動向の情報収集を積極的に行い，プロジェクトの情報発信を適切な時期に有効に行えるよう注意を払う．

４．研究成果

研究開始当初計画していたことが概ね達成できた．提案した手法は，KDD をはじめとするトップ国際会議で発表することができた．提案手法を実装したソフトウェアーを公開レポジトリーにて公開し国内外の研究者・技術者に利用できるようにした．

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 4 件）

1.Y.Tabei and Y.Yamanishi: Scalable prediction of compound-protein interactions using minwise hashing, BMC Bioinformatics, 7, S3, 2013, 査読あり, DOI: 10.1186/1752-0509-7-S6-S3

2.Y.Tabei, A.Kishimoto, M.Kotera, Y.Yamanishi: Succinct Interval-Splitting Tree for Scalable Similarity Search of Compound-Protein Pairs with Property Constraints, In Proceedings of the 19th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2013, 査読あり, DOI:10.1145/2487575.2487637

3.Y.Tabei, E.Pauwels, V.Stoven, K.Takemoto, Y.Yamanishi: Identification of chemogenomic features from drug-target interaction networks using interpretable classifiers, Bioinformatics, 28(18), i487-i494, 2012, 査読あり, DOI: 10.1093/bioinformatics/bts412

4.Y.Tabei: Succinct Multibit Tree: Compact Representation of Multibit Trees by Using Succinct Data Structures in Chemical Fingerprint Searches, 7534, pp201-213, 2012, 査読あり, DOI:10.1007/978-3-642-33122-0_16

〔学会発表〕（計 4 件）

1.Y.Tabei and Y.Yamanishi: Scalable prediction of compound-protein interactions using minwise hashing, 24th International Conference on Genome Informatics (GIW), Singapore, 2013 年 12 月 16 日-18 日

2.Y.Tabei, A.Kishimoto, M.Kotera, Y.Yamanishi: Succinct Interval-Splitting Tree for Scalable Similarity Search of Compound-Protein Pairs with Property Constraints, 19th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), Chicago, USA, 2013 年 8 月 11 日-14 日

3.Y.Tabei, E.Pauwels, V.Stoven, K.Takemoto, Y.Yamanishi: Identification of chemogenomic features from drug-target interaction networks using interpretable classifiers, 11th European Conference on Computational Biology, Basel, Switzerland, 2012 年 9 月 9 日-12 日.

4.Y.Tabei: Succinct Multibit Tree: Compact Representation of Multibit Trees by Using

Succinct Data Structures in Chemical Fingerprint Searches, 12th Workshop on Algorithms in Bioinformatics (WABI), Ljubljana, Slovenia, 2012 年 9 月 9 日-14 日.

〔図書〕（計　　件）

〔産業財産権〕
○出願状況（計　　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　　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 田部井靖生（Tabei Yasuo）
（　　）
東京工業大学・大学院情報理工学研究科・
東工大特別研究員
研究者番号：20589824

(2)研究分担者
（　　）

研究者番号：

(3)連携研究者
（　　）

研究者番号：