様式 C-19

科学研究費補助金研究成果報告書

平成　23 年　5 月　27 日現在

機関番号：14301
研究種目：若手研究(B)
研究期間：2008～2010
課題番号：20700134
研究課題名（和文）　部分的類似構造の重ね合わせに基づく不均質データの多義的探索法の開発
研究課題名（英文）　Multifaceted exploration of nonhomogeneous and ambiguous data by combining partial similarities
研究代表者
瀧川　一学（TAKIGAWA　ICHIGAKU）
京都大学・化学研究所・助教
研究者番号：10374597

研究成果の概要（和文）：インターネットや各種電子化を受けて、大規模なデータが自動的に収集できるようになってきた。しかし、これらのデータは容易に得ることができる半面、前もって用途が固定されておらず多目的用途であるため、様々なレベルの情報が混在し質が不均一であり、伝統的な統計解析において困難や不具合を生じる。本研究では、近年著しく発展している部分構造の列挙技術を背景に、データ間の局所的または部分的類似性に基づいて、このようなデータを解析できる統計解析法の提案・解析を行った。

研究成果の概要（英文）：Due to the computerization of industrial and scientific data, we can automatically collect or systematically obtain a huge dataset. Whereas these data are more easily accessible than before, their poor quality causes problems when we try to statistically analyze and utilize them. Many levels of information are collapsed into a single dataset since the purpose of use is ambiguous and unfixed in advance, and as a result, their quality is not sufficiently homogeneous. To address this problem, we developed novel statistical methods based on recently-emerged techniques for substructure enumeration, which can analyze those types of data by combining partial or local similarities in the data.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００８年度	1,500,000	450,000	1,950,000
２００９年度	800,000	240,000	1,040,000
２０１０年度	800,000	240,000	1,040,000
年度			
年度			
総　計	3,100,000	930,000	4,030,000

研究分野：計算生物学、情報数理工学
科研費の分科・細目：総合領域 情報学 知能情報学
キーワード：データマイニング、部分構造探索、アルゴリズム、機械学習

１．研究開始当初の背景

インターネットを用いたマーケティングやテキスト情報収集において、データは多量に手に入るが、こうしたデータでは一つ一つのデータの信頼性がまちまちであったり、全体としては質が不均一であったりすることによって、通常の統計分析では有用な情報がなかなか得られにくい。また、データが容易に手に入る状況では、しばしば予め利用用途を特定せず、様々なレベルの情報を様々な状況で一緒くたに収集するようなケースも良く

見られるようになってきた。

こうした傾向は生命科学データでも顕著に見られるようになってきた。その背景には高度な機器によって実験が半機械化され、細胞から複数かつ多種類の計測を得ることが技術的に容易になったことがある。例えば、マイクロアレイなどを用いて、ある生物が持つ全遺伝子について mRNA 等の量を、特定の条件下でスナップショットのように得ることができるようになっている。また、創薬に用いる何百万という規模の化合物についても各種活性や構造式はデータベースから容易に得られ、スクリーニングやアッセイの結果データもハイスループット機器による計測で多量に得られるようになってきた。

しかし前者の例と同様に、こうしたデータは必ずしも、用途が明確ではなく、むしろ多用途を想定して網羅的に採取される(ゲノムプロジェクトのような)場合が多い。従って、様々なレベル、様々な質の情報が混合されてデータとなっており、こうした不均質なデータを扱う統計手法が求められていた。

２．研究の目的

本研究では前節で挙げた不均質で多義的なデータを解析する一つのアプローチとして、通常の統計解析のようにそれらをモデルや数値(回帰や分類、クラスタリングなど)に集約しようとするのではなく、まず、データ間に「局所的な類似性」や「部分的な類似性」を考え、そのような類似性を持つ対象を全て調べ上げるというアプローチによって、この問題に対応する一つの方法論を提示した。

この背景にはデータマイニングの分野で、ある条件を満たす部分構造を厳密に全て列挙する(調べ上げる)技術が成熟を見せ始めていたことがある。このような方法を用いることによって、データ間の局所的/部分的な類似性を明示的に提示し、解析することができると考えた。そこで以下の２点を基本的な指針として、次節で述べる具体的な問題に関してこのアイデアについて研究を行った。

- ある条件を満たす対象を調べ上げる(列挙する)方法に基づいた解析
- 局所的な類似度に基づいて全体を解析できる方法

３．研究の方法

前節で挙げた「部分的/局所的類似性を定義し、それに関して特定の条件を満たす対象を全て数え上げる」方法を指針として、実際のデータに関する以下の主に 5 つの項目について、統一的な関心のもとで研究を実施した。

(1) 凸包被覆に基づく判別分析

通常の統計解析で質が不均一な場合の方法で混合モデルを用いるアプローチがある。混合モデルではデータ全体は各局所的な領域ごとに別々にモデル化される。通常、各成分モデルには正規分布のような簡潔なパラメトリックモデルが用いられるが、一般的には各成分が楕円状の分布をなすとは限らない。ここでは各局所領域を凸包としてより柔軟に構成して判別分析を行うノンパラメトリックな統計手法を提案する。

(2) 部分構造パターンの要約

データとして、各データ点が木構造やグラフ構造を持つ対象を考える。このようなデータは XML 文書セット、化合物データセット、糖鎖データセット、タンパク質構造データセット、など、近年幅広く見られるようになっている。そのような対象において、頻出する部分構造パターンを全て調べ上げる(全列挙する)アルゴリズム研究が「頻出パターンマイニング」として発展してきた。しかし、単純に頻出パターンを全列挙すると、出力が膨大になり、出力をどのように理解すれば良いのか、という二次的困難を生じる。そこで、木構造データおよびグラフ構造データを対象に頻出部分構造パターン集合を意味のある形で集約する方法を考察する。

(3) 部分構造ペアの列挙と分子間相互作用解析への応用

項目(2)と同様に、データとして、各データ点が木構造やグラフ構造を持つ対象を考える。特に、生命科学におけるデータでは、遺伝子、化合物、タンパク質、糖鎖といった生命現象の要素分子ごとに、例えば構造データがグラフとして得られる。

一方で、近年の生命科学の興味は、こうした分子要素の間のネットワークの理解である。生命現象は通常、複数の分子がお互いに関係し合い形成される。関与する分子間をエッジで結んで分子間ネットワークとして様々な対象を理解するというアプローチは様々な対象に応用されている。各分子の構造に興味がある場合の分子ネットワークはある構造とある構造の間の関係を抽象的に示

していると言える。例えば、薬剤とその標的タンパク質のネットワークを考えると、どのような化学構造を持つ薬剤がどのような配列特徴を持つタンパク質を標的にしているのか、といった事がこのネットワークとノードに関連づけされた構造(グラフ構造、木構造、配列など)を分析することで特徴づけできると考えられる。

そこで本研究の目的に沿って、この構造・構造ペアのデータの解析手法の提案を行う。具体的には、薬剤の化学構造式(グラフ構造)と標的タンパク質のアミノ酸配列(文字配列)のペアとして、既知の薬剤-標的ネットワークを考える。ここでこのペアデータに出現している有意な部分構造-部分構造ペア(部分グラフ-部分配列ペア)を列挙する方法を考察する。この場合、通常の頻出パターン列挙の方法の拡張に加えて、それぞれの部分構造単独の効果ではなくペアとしての効果(因子が組み合わされて生じる効果)であることを判定するための統計的方法が必要であり、これらの問題点を含めて、統一的な方法の提案を行い、薬剤-標的タンパク質ネットワークを構造特徴から実際に分析する。

(4) 交互作用を持つペアの列挙

マイクロアレイなどの技術を用いて、生物種が持つ全遺伝子についてそれぞれの遺伝子の発現量をまとめて測定することができるようになっている。各遺伝子の発現プロファイルはまとめて通常の多変量解析によって統計解析される。しかし、遺伝子の発現量は各々独立ではなく、一般的には交互作用(組み合わされた時に生じる効果)を解析し、どの遺伝子とどの遺伝子に有意な交互作用があるか調べる必要があると考えた。そこで、遺伝子間の局所的関係の分析として、遺伝子セットと各発現プロファイルが与えられた場合に、与えられた因子に関して遺伝子ペアの交互作用の強さを調べ、最も高いものから順に列挙する方法を考察する。実際上の問題としては大規模な計算が必要になるため計算効率化を行う必要がある。

(5) アクティブパス列挙に基づく生物ネットワーク解析

代表者らは代謝ネットワークにおけるアクティブパスを酵素遺伝子の発現量に基づいてスコア順に列挙する方法を提案しており、これらも部分的特徴(ネットワーク上で隣接する遺伝子間の発現プロファイルの相関など)に基づいて、パターンを調べ上げる方法である。そこでこの方法を拡張することで、ネットワーク全体での挙動を分析する方法を考案し、実際のデータを用いて代謝系の挙動に対する大規模解析を行う。

４．研究成果

従来の集約的統計解析ではなく、部分的/局所的類似性を定義し、それに関して特定の条件を満たす対象を全て数え上げるという方法は、前節に挙げた複数の問題において、新しい方法の提案に結びつき、実際の問題への応用も含めて、十分な成果を得ることができた。具体的な点について各項目問題ごとに以下にまとめる。

(1) 凸包被覆に基づく判別分析

与えられた点群を正例の部分集合を含み負例を含まない凸包の族で被覆する枠組によるノンパラメトリックなパターン分類法・探索的データ解析法およびその構成のための厳密アルゴリズムおよび効率的確率化アルゴリズムを示し、データ点群の局所分解(混合モデル)を与えながら、パラメタ設定に過敏な従来法と同等の精度が得られることが確認できた。(論文⑨)

(2) 部分構造パターンの要約

生物学で解析が待たれている DNA,タンパク質,脂質とならぶ重要高分子である糖鎖が木構造であることを鑑みて、現在利用できる糖鎖構造データの部分的類似構造として、頻出部分木パターンを分析し、定義から出力が肥大しがちなこのマイニング問題に対して、delta-tolerance に基づく出力のパラメトリック要約法を提案し、実際の糖鎖データの分析を行った。その結果、要約された頻出部分構造パターンは既知の様々な機能モチーフと多く合致が見られ、幅広いレベルの生物学的情報を内包する糖鎖構造データの部分的類似構造分析の多面的有効性が示唆されたと言える。(論文⑪、⑧)

また、こうした木構造データに関するアルゴリズム開発に基づいて、delta-tolerance による要約法をさらにグラフデータに適用するための効率的アルゴリズムについて研究を行った。

(論文②)

(3) 部分構造ペアの列挙と分子間相互作用解析への応用

化合物-タンパク質間相互作用データは従来思われていたように1化合物-1標的の関係ではなく多対多の複雑な相互作用を示すことが近年示唆されている。特に現行の有効承認薬剤の多標的性が明らかになるにつれて polypharmacology の重要性が再認識されているが未だこの相互作用データについては化合物-標的ネットワークのトポロジー分析が主な分析技術であった。この相互作用データの解析に取り組み、本計画で掲げている部分構造類似性からのアプローチとして、作用する化合物とタンパク質のペアについて、化合物は分子グラフ、タンパク質はアミノ酸配列とみなし、どのような部分構造ペア(部分グラフ-部分配列ペア)が有意に頻出しているかを厳密かつ効率的に調べる方法論を確立し現行のデータの大規模な分析を行った。結果として現行の化合物-標的ネットワークは特徴的な部分構造ペアによっていくつかの polypharmacology パターンとして捉えられることが示され、GPCR などにおける立体構造との比較の観点からの機能モチーフ候補も発見された。
(論文③)

(4) 交互作用を持つペアの列挙

特定の SNP の有無に応じて与えられた2遺伝子の発現量変化が有意に変化するか、言い換えれば、特定の SNP の有無を2遺伝子の遺伝子発現量で説明する際に交互作用が存在するかどうか、をヒューリスティクスを導入して効率的に行い、実際の超大規模データについて分析を行った。(論文⑩)

また、より一般的に特定の因子(ガン細胞か正常細胞かなど)に関して与えられた2遺伝子の発現パターンが因子の有無で逆相関する典型的交互作用パターン(クロスパターン)を網羅的に同定する方法について、より頑健な方法を提案した。
(論文①、⑦)

(5) 生物データへの応用と解析

代表者は、ある化合物を別の化合物に変換する代謝の連鎖反応について、与えられた条件で代謝ネットワークで可能な経路のうちどの経路が実際にアクティブであったかを、酵素遺伝子の発現パターンと代謝ネットワークの構造から、スコアの高い順番に列挙する方法を提案していた。この手法は2点間のパス探索に限定していたが、さらに全2点ペアで探索を行い、その大規模出力を集約するためのクラスタリングを統合することで、代謝ネットワーク全体を酵素遺伝子発現量の観点から分析する方法を構築し、実際のデータの分析を行った。
(論文⑤)

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 11 件）

① M. Kayano, I. Takigawa, M. Shiga, K. Tsuda, H. Mamitsuka, ROS-DET: robust detector of switching mechanisms in gene expression, Nucleic Acids Research, 査読有, 2011, 印刷中. doi: 10.1093/nar/gkr130.

② I. Takigawa, H. Mamitsuka, Efficiently mining delta-tolerance closed frequent subgraphs, Machine Learning, 査読有, Vol. 82, No. 2, 2011, pp. 95-121.

③ I. Takigawa, H. Mamitsuka, Mining significant substructure pairs for interpreting polypharmacology in drug-target network, PLoS One, 査読有, Vol. 6, No. 2, 2011, e16999.

④ M. Shiga, I. Takigawa, H. Mamitsuka, A spectral approach to clustering numerical vectors as nodes in a network, Pattern Recognition, 査読有, Vol. 44, No. 2, 2011, p. 236-251.

⑤ T. Hancock, I. Takigawa, H. Mamitsuka, Mining metabolic pathways through gene expression, Bioinformatics, 査読有, Vol. 26, No. 17, 2010, pp. 2128-2135.

⑥ A. Nakamura, T. Saito, I. Takigawa, H. Mamitsuka, M. Kudo, Algorithms for finding a minimum repetition representation of a string, Lecture Notes in Computer Science, 査読有, Vol. 6393, 2010, pp. 185-190.

⑦ M. Kayano, I. Takigawa, M. Shiga, K. Tsuda, H. Mamitsuka, On the performance of methods for finding a switching mechanism in gene expression, Genome Informatics, 査読有, Vol. 24, 2010, pp. 69-83.

⑧ I. Takigawa, K. Hashimoto, M. Shiga, M. Kanehisa, H. Mamitsuka, Mining

patterns from glycan structures, Proceedings of the International Beilstein Symposium on Glyco-Bioinformatics, 査読無, 2010, pp.13-24.

⑨ I. Takigawa, M. Kudo, A. Nakamura, Convex sets as prototypes for classifying patterns, Engineering Applications of Artificial Intelligence, 査読有, Vol. 22, No. 1, 2009, pp.103-108.

⑩ M. Kayano, I. Takigawa, M. Shiga, K. Tsuda, H. Mamitsuka, Efficiently finding genome-wide three-way gene interactions from transcript- and genotype-data, Bioinformatics, 査読有, Vol. 25, No. 21, 2009, pp. 2735-2743.

⑪ H. Hashimoto, I. Takigawa, M. Shiga, M. Kanehisa, H. Mamitsuka, Mining significant tree patterns in carbohydrate sugar chains, Bioinformatics, 査読有, Vol. 24, No. 16, 2008, i167-i173.

〔学会発表〕(計 6 件)

① I.Takigawa, K. Tsuda, H. Mamitsuka, Mining significant substructure-substructure pairs in structural associations, The 20th International Conference on Genome Informatics (GIW2009), 14-16 December, 2009, Yokohama, Japan.

② M.Kayano, I. Takigawa, M. Shiga, K. Tsuda, H. Mamitsuka, Genome-wide three-way gene interactions from transcript and genotype data, The 20th International Conference on Genome Informatics (GIW2009), 14-16 December, 2009, Yokohama, Japan.

③ I. Takigawa, K. Hashimoto, M. Shiga, M. Kanehisa, H. Mamitsuka, Efficiently finding significant substructural patterns conserved in glycans, 2008 Annual conference of the Japanese Society for Bioinformatics, 15-16, December, 2008, Osaka, Japan.

④ M. Kudo, I. Takigawa, A.Nakamura, Classification by reflective convex hulls, 19th International conference on pattern recognition (ICPR2008), 8-11, December, 2008, Tampa, Florida, USA.

⑤ K. Hashimoto, I. Takigawa, M. Shiga, M. Kanehisa, H. Mamitsuka, Mining significant tree patterns in carbohydrate sugar chains, ECCB'08 European Conference on Computational Biology, 22-26, September, 2008, Cagliari, Italy.

⑥ 瀧川一学, 酵素遺伝子の発現情報に基づく効率的な代謝経路ランキング, 2008年度統計関連学会連合大会, 2008 年 9 月 7 日～10 日, 慶應義塾大学.

〔その他〕

プロジェクトホームページ

① http://www.bic.kyoto-u.ac.jp/pathway/grasp/

② http://www.bic.kyoto-u.ac.jp/pathway/kayano/bioinfo_three-way.html

６．研究組織

(1)研究代表者

瀧川 一学 (TAKIGAWA ICHIGAKU)

京都大学・化学研究所・助教

研究者番号：10374597