様式 C-19

科学研究費補助金研究成果報告書

平成２２年　５月２８日現在

研究種目：若手研究（Ｂ）
研究期間：２００８　～　２００９
課題番号：２０７００８１
研究課題名（和文）　グラフ表現による構造化データの可視化

研究課題名（英文）　Visualization of structured data based on graph representation

研究代表者
森　康久仁（MORI YASUKUNI）
千葉大学・大学院融合科学研究科・助教
研究者番号：４０３６１４１４

研究成果の概要（和文）：本研究では，複雑な構造や分布を持つデータを適切に解釈する一つの方法を提案した．提案した手法は，データの局所的な部分集合をグラフの一つの頂点として表し，データの部分集合間の関係を辺で表現する．実際のデータに対して本提案手法を適用した結果，データの性質を非常によく表現することができた．本手法により，いわゆる学習やパターン認識の問題に対する認識率や学習性能を高めるための方法論の解明や，データ間の関係が複雑に依存しあっているようなネットワーク構造の解明につながると考えられる．

研究成果の概要（英文）：A methodology which extract an effective information from data with complicated structure or distribution was proposed in this study. The proposed method is based on graph-representation which express a subset of data as a node and a relationship between subsets as a edge. The proposed method was applied to some real data sets. As a result, the method can represent the characteristics of data effectively. It is expected that the learning performance or recognition rate of classification problem is improved and that we find a clue to elucidate the network structure of data with dependencies among them.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００８年度	１，１００，０００	３３０，０００	１，４３０，０００
２００９年度	１，０００，０００	３００，０００	１，３００，０００
年度			
年度			
年度			
総　計	２，１００，０００	６３０，０００	２，７３０，０００

研究分野：パターン認識
科研費の分科・細目：情報学・メディア情報学・データベース
キーワード：データ可視化，パターン認識，データマイニング，特徴選択

１．研究開始当初の背景

コンピュータ技術の急速な発展に伴い，高速の中央演算処理装置や十分な容量の記憶装置を安価に利用できるようになり，さらにはインターネットの発展などにより，世界中の情報が簡単に利用できるようになってき

た．そこで，それら世界中に蓄積されているデータから情報を検索する技術や，ハードディスクなどに保存された膨大な量のデータから有効な情報を取り出すデータマイニングなどの研究が盛んに行われるようになってきている．それに伴い，かつてなかった規模のデータの蓄積と処理が必要になってきている．

そのようなデータを有効に活用するためには，人間の優れたデータ解析能力や学習能力を利用し，それを事前知識として利用することが重要である．従って，通常では直感的に捉えることができない膨大な量のデータの情報や構造を人間の知覚できる形で視覚化することは，データ解析を必要とするあらゆる分野において非常に有効な手段となる．このような研究の背景が本研究の出発点である．

２．研究の目的

本研究では様々な種類のデータに対して，それらが持つであろう特徴的な構造を可視化する手法の提案を行う．もし，それらのデータが十分精度良く，また，物理的な法則に乗っ取ってデータ計測がなされたのならば，これら実際に扱われるデータの多くは，局所的または大域的に，そのデータ特有の特徴を持つであろうことは容易に想像できる．そのような特徴を捉えることは，様々に分野におけるデータを実際に処理する際，最も重要な情報になることは間違いない．特にシステム設計の初期の段階では，対象とするデータの特質をわかりやすい形で確認し，データの性質を理解することは，システム全体の性能を向上させるためには非常に重要なポイントとなる．このような問題意識から，本研究では次の２点を研究の目的とした．

（１）データの持つ特徴的な構造を可視化する新たな手法の提案．

（２）可視化結果の実問題への応用．

３．研究の方法

本研究では，多数の数値属性を持ち各データに割り当てるべきクラス属性が付いた，いわゆる教師付きデータを研究対象とした．しかしながら，全てのデータが数値属性のみで表現されているわけではない．そのため，将来的に数値以外のデータ属性を持つ場合に拡張できるような枠組みを考える．

多数の属性値からなる，多変量のデータの分布状況を把握するためには，我々が直接知覚することができる，２，３ 次元空間に射影し可視化することがよく行われる．本研究では従来の一般的なアプローチのような，高次元空間に点在する各々のデータについて可視化することを考えずに，局所的なデータ集合を一つの単位として考え，その集合の関係をグラフ表現を用いて視覚情報とすることを考える．このように表現することで，データの持つ特徴的な構造や，カテゴリ間の関係等を的確に表現することできる．さらに，あるデータの集合を一つのまとまりとして見ることで，大量のデータの可視化を行う際に起こる，結果の可読性の悪化を防ぐことができる．これは従来のデータ可視化手法では困難なことであり，本研究課題が極めて独創性が高く，有効な点であると言える． また，今後の展開を考慮し，様々なデータ形式に対応できるようにデータ間の類似度をカーネル関数により導出する．これにより，一般的な数値データのみならず，文字列データなどの特殊なデータ型にも対応できるような手法を目指す．図１にカーネル関数を用いたデータの分布領域の推定結果を示す．カーネル関数による分布領域の推定により，データの非線形な構造を捉えることができる．

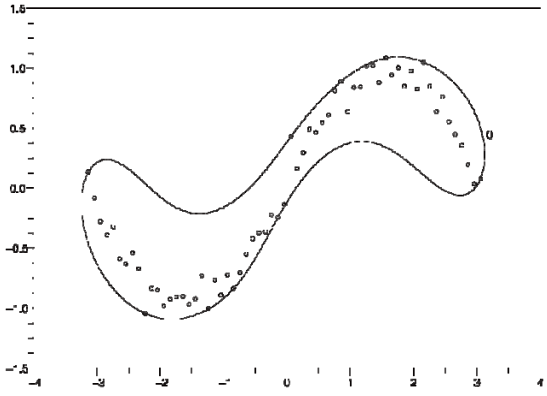

図１：データ領域の推定

本研究では，カーネル関数を用いて推定したデータの分布領域の情報を可視化することで，データの分布している各々の領域間の関係を適切に表現する手法を提案する．提案手法は，従来の手法と異なり，データ点そのものを低次元空間に射影するのではなく，各クラス毎にカーネル関数を用いた領域推定によりデータの分布領域を求め，その領域をグラフの頂点として表現し，集合間の近接の度合を辺で表す．つまり，各クラスにつき一つの頂点が割り当たることになる．図２にグラフ表現の例を示す．この例では説明のために2次元空間に散らばる点をグラフ表現しているが，一般には高次元に分布する点をグラフ表現する．このように表現することによって，

データ点自体の詳細な情報は表現できないものの，他の手法に比べてクラスの大域的な情報，特にクラス間の分離状況を的確に表現することができる．

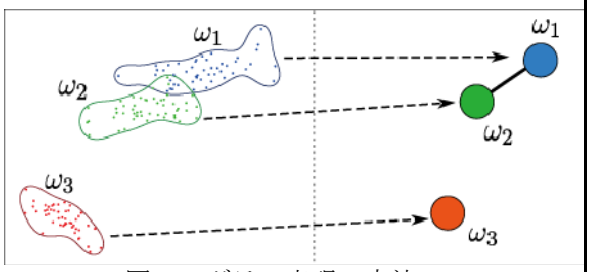

図２：グラフ表現の方法

４．研究成果

本研究の成果として複雑に構造化したデータの可視化方法の提案，および実際の分類問題への適用し，分類問題における特徴の評価に用いることを示した．以下に概要を示す．

（１）グラフ表現による有効なデータ可視化手法を提案した．提案手法は，多数の属性を持つデータの情報を過不足なく表現できるようにするためグラフによる表現を用いた．さらに，グラフの構成方法(頂点の配置，隣接関係の表現方法)について複数の方法を提案した．また，これらの可視化を段階的に行う手法についての検討を行い，実際の遺伝子発現データのデータ構造を可視化することで定性的な評価を行った．この成果は第７回情報科学技術フォーラムにて報告した．

図３に実際に作成したグラフ表現アプリケーションの例を示す．このアプリケーションはマウス等の操作によって，表示を制御する閾値を選択することができ，リアルタイムに結果を表現することができる．

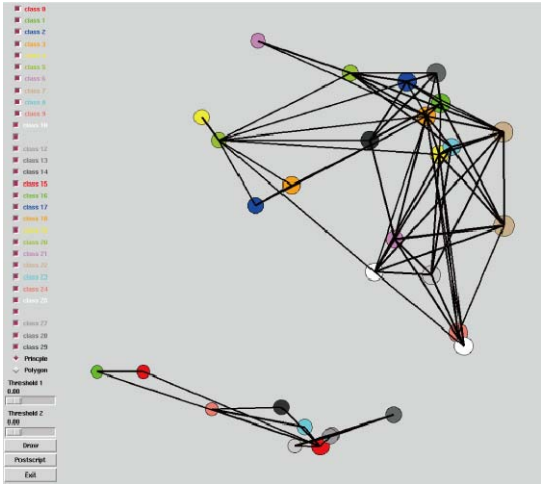

図３：可視化アプリケーションの例

（２）分類問題に対する特徴の評価に関する検討を行った．提案した可視化手法により，実際の分類問題に対して，現在用いているデータの特徴が十分な分類性能を有しているかどうかの判断をする材料を与えることができることを示唆した．また，ある特定のクラスが，そのクラスと間違えやすいかなどの情報も視覚的にみることができた．図４に類似漢字文字１５対の手書き漢字データを可視化した結果を示す．このように表現することにより，データ間の関係が直感的に把握できるようになった．これにより，学習や識別システム全体の性能を向上させるための情報を提供できると考えられる．

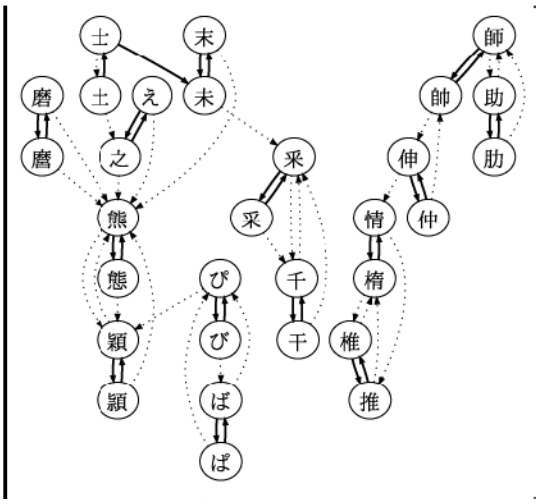

図４：実データによる可視化例

現在のデータ解析において望まれることは，従来の方法論がそのまま適用できないような大規模な問題(カテゴリ数，変量数，データ数が多い場合）に対して，十分効率良く，かつ高精度な解析が行えることである．この問題に対して本研究の成果では，複雑な構造を持つデータを適切に解釈する一つの方法を与えええることができたといえる．また，従来の可視化手法に比べ，可視化結果の可読性が非常に高いことを明らかにした．これは現在のデータ量の大幅に増大している状況において非常に有効なデータ解析手法の一つとなる可能性がある．

本提案法を用いてデータの構造を適切に解析することにより，いわゆる学習やパターン認識の問題に対する認識率や学習性能を高めるための方法論の解明や，データ間

の関係が複雑に依存しあっているようなネットワーク構造の解明の一端を担える可能性があることを示した.

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）

〔学会発表〕（計２件）
①小田原　平，森　康久仁，松葉　育雄，相互情報量フィルタリングを用いた遺伝的アルゴリズムによる特徴選択，第７回情報科学技術フォーラム，2008 年 9 月 3 日，慶応義塾大学
②桑原　俊，森　康久仁，松葉　育雄，分割最適化クラスタリングの階層的可視化，第７回情報科学技術フォーラム,2008 年 9 月 3 日，慶応義塾大学

６．研究組織
(1)研究代表者
森　康久仁（MORI YASUKUNI）
千葉大学・大学院融合科学研究科・助教
研究者番号：４０３６１４１４

(2)研究分担者
（　　）

研究者番号：

(3)連携研究者
（　　）

研究者番号：