HOKKAIDO UNIVERSITY

Title	多様性を考慮したビッグデータの分類手法の研究
Author(s)	米森, 力
Degree Grantor	北海道大学
Degree Name	博士(情報科学)
Dissertation Number	甲第11290号
Issue Date	2014-03-25
DOI	https://doi.org/10.14943/doctoral.k11290
Doc URL	https://hdl.handle.net/2115/55441
Type	doctoral thesis
File Information	Chikara_Yonemori.pdf

Hokkaido University Collection of Scholarly and Academic Papers : HUSCAP

平成 25 年度 博士学位論文

多様性を考慮したビッグデータの分類手法の研究

米 森　　力

北海道大学大学院　情報科学研究科

複合情報学専攻

目次

第 1 章　序論 **9**

1.1　本論文の背景 9

1.2　本論文の構成 10

第 2 章　ビッグデータ **13**

2.1　ビッグデータとは 13

2.1.1　ビッグデータの定義 13

2.1.2　ビッグデータに対する期待 14

2.1.3　ソフトウェア基盤 15

2.1.3.1　Apache Hadoop 15

2.1.3.2　NoSQL 16

2.1.3.3　CEP 17

2.1.4　分析手法 18

2.1.4.1　データマイニングにおける分類 18

2.1.4.2　シンボリックデータ解析 26

2.2　ビッグデータ分析の事例 27

2.3　ビッグデータ分析における課題 28

2.4　ビッグデータ分析に使用されるアプローチとその限界 29

2.5 データの多様性における分析の困難性 31
2.6 提案アプローチ 31

第 3 章 企業ランキングにおける分類 33

3.1 はじめに 33
3.2 企業評価におけるランキング 34
3.3 部分空間に基づくランキングモデルと関連属性の自動抽出 36
3.3.1 ランキングモデルの構築 36
3.3.2 反復学習 39
3.3.3 関連属性の自動抽出方式 40
3.4 実験 41
3.4.1 公開データを用いた実験 41
3.4.2 企業評価への応用 43
3.4.2.1 実験に用いるデータ 43
3.4.2.2 ランキングモデルの検証 45
3.4.3 関連属性の自動抽出方式の評価 46
3.5 本章のまとめ 53

第 4 章 クリーク表現を基にした企業分類と取引関係の抽出 55

4.1 はじめに 55
4.2 企業間関係におけるネットワーク分析 56
4.3 クリークを用いた企業コミュニティ理解 56
4.3.1 極大クリーク全列挙アルゴリズムの適用 56
4.3.2 擬似クリークによる企業コミュニティ抽出 57
4.3.3 結束度に基づく企業コミュニティ理解 58
4.4 実験 58

4.4.1 実験に用いたデータ 58
4.4.2 クリークを用いたコミュニティ抽出 60
4.4.2.1 極大クリークの抽出 60
4.4.2.2 擬似クリークによるコミュニティ抽出 62
4.4.3 結束度による取引関係の構造分析 64
4.5 本章のまとめ 66

第 5 章 分布値データの非階層的シンボリッククラスタリングと道路特性の抽出 69

5.1 はじめに 69
5.2 シンボリックデータ解析について 71
5.2.1 データ表現 71
5.2.2 従来の解析手法とその課題 71
5.3 分布値データの非階層的クラスタリング手法 72
5.3.1 分布値データの定義 72
5.3.2 分布値データによるクラスタの定義 73
5.3.3 分布値データに対する k-means アルゴリズム 74
5.4 数値実験 75
5.4.1 人工データによる動作確認 75
5.4.2 交通データによる検証結果 79
5.5 本章のまとめ 81

第 6 章 結論 87

6.1 考察 87
6.1.1 提案分類手法についての有効性 87
6.1.2 データ表現の決定に関する対処について 89

6.1.3 分析自動化について 91

6.2 まとめと今後の課題 91

業績一覧 **95**

参考文献 **97**

図目次

2.1 NoSQL の分類 18

2.2 CEP によるリアルタイム処理のイメージ 19

2.3 数値データの分類手法 19

2.4 決定木による分類 21

2.5 k-NN による分類 22

2.6 SVM による分類 23

2.7 適合率と再現率 25

2.8 分類手法選択の観点 26

2.9 提案アプローチのフロー 32

3.1 提案モデルによるランキング処理 37

3.2 条件付自己相関行列の算出対象となる事例 40

3.3 学習前（破線）と学習後（実線）の再現率のリフト図（open） 47

3.4 提案モデル（実線），ロジットモデル（破線），判別分析（点線）の再現率のリフト図 48

3.5 属性に対する関連度の推移 50

4.1 クリークのサイズに対する擬似クリークの頻度 62

4.2 擬似クリークに含まれる頻度上位 20 社におけるサイズの構成と結束度 65
4.3 結束度の値が大きい企業の取引構造 (密型) の例 67
4.4 結束度の値が小さい企業の取引構造 (星型) の例 68

5.1 人工データ発生元の混合正規分布 (上から p_1, p_2, p_3) 76
5.2 クラスタリングの結果得られたクラスタ中心 (上からクラスタ 1, 2, 3) 78
5.3 交通ビッグデータのクラスタリングの結果得られたクラスタ中心 (上から順にクラスタ 1, 2) 83
5.4 交通ビッグデータのクラスタリングの結果得られたクラスタ中心 (上から順にクラスタ 3, 4) 84
5.5 交通ビッグデータのクラスタリングの結果得られたクラスタ中心 (クラスタ 5) 85

表目次

2.1 ビッグデータの特徴の各観点における解決のアプローチ . . . 29

3.1 公開データ (housing data) に対する提案モデル (proposed), ロジットモデル (logit), 判別分析 (LDA) の 11 点平均適合率 43
3.2 実験で用いた財務指標の例 . . . 44
3.3 知的潜在力指数の構成要素 . . . 45
3.4 提案モデル (proposed), ロジットモデル (logit), 判別分析 (LDA) の 11 点平均適合率 . . . 49
3.5 高い関連度を持つ属性の上位 15 指標 . . . 52

4.1 高い次数を持つ企業 (上位 20 社) の業種 . . . 59
4.2 サイズの降順に並べた極大クリーク (上位 20 個) と属する企業の業種 . . . 61
4.3 サイズの降順に並べた擬似クリーク (上位 20) と属する企業の業種 括弧内は企業件数を表す. . . . 63

5.1 分布ごとの混合正規分布のパラメータ . . . 77
5.2 人工データに対するクラスタリングによる分類: 正解率=0.985 77
5.3 使用した交通情報の特性 . . . 79

5.4 交通ビッグデータのクラスタリングの結果得られたクラスタ中心 80
5.5 交通ビッグデータのクラスタリングによる分類結果 80

第 1 章

序論

1.1　本論文の背景

インターネットの普及・活用により，世界に蓄積されるデータ量は急速に増加している．このような膨大なデータを活用する試みはビッグデータと呼ばれ，量的な変化が，様々な分野の研究やビジネスの質的変化をもたらしつつある．この背景には，スマートフォンの普及により，様々なシーンにおいて人々の行動を詳細に把握できるログデータや，生活のあらゆる場面にセンサが溢れ，実世界の状態を計測・制御することができるセンサデータが蓄積されるようになったことがある．またこれらの膨大なデータを処理可能にするため，分析用ツールも改良が進んだ．従来の統計解析やデータマイニングツールに加え，Google が開発した MapReduce やそのオープンソース技術の Hadoop のような安価に大規模分散処理を行える環境が整ったことにより，膨大な量のデータを処理できるようになってきている．

一方，ログデータやセンサデータは従来の表形式で数値により整理された構造的なデータと違い，テキストベースであったり関係性を持つなど非構造的で

あるという特徴も併せて持つ．そのため，今後は非構造データのような多様なデータ表現に対して，分析を行う必要性が増すと考えられる．

本研究ではこの多様なデータ表現に対して分析する方法を提案し，ビッグデータから有用な情報を抽出することを大きな目的として以降の議論を進める．

1.2　本論文の構成

本研究では，ビッグデータにおける多様性に焦点を当て，表形式のデータでは分析が困難であるデータ表現として部分空間，グラフ構造，分布値データの 3 つを研究対象として取り上げ，分類手法を提案する．提案に当たり，第 2 章でビッグデータ分析を概観した後，各データ表現に関する研究の成果について述べていく．以下，本論文の構成を示す．

第 2 章では，本研究の背景で述べたビッグデータについて概観し，ビッグデータの定義および特徴について整理する．その上でビッグデータの特徴として大規模，多様性，速度の 3 つを対象にビッグデータを処理可能にするため現在取り組まれている手法を述べ，その課題を整理する．併せて本研究でも対象としている多様性について，より詳細に言及し本研究における研究対象を明確にする．最後にビッグデータの活用事例について述べる．

第 3 章では，企業ランキングを対象として部分空間表現を活用することで高精度に分類を行う手法を提案する．企業ランキングを行う上で分析対象となる財務データには，貸借対照表や損益計算書等に含まれる財務指標間の相関が高いこと，およびまたランキング上位と下位とで企業の特性が違うことから線型モデルを用いた分析が適用しにくい課題がある．この課題を解決するために部分空間を用いた分類モデルを提案する．

第 4 章では，企業間の取引関係を想定し，この取引関係の特性を適切に分類する手法を提案する．企業間の取引関係はグラフ構造としてデータ表現が可能である．企業間の取引において与信に問題が生じた際は影響範囲が意思決定に重要となるため，このグラフ構造から取引関係の型を星型と密型に分類できる指標を提案する．

第 5 章では，GPS データから取得されるプローブデータを想定し，このデータを分布値データとして表現した際に，複雑な道路特性を適切に分類する方法として非階層的シンボリッククラスタリングを提案する．本手法では分布値データが持つ縮約性とパラメータを通した特性理解が実現される．

第 6 章では，第 3 章から第 5 章で得られた本研究の成果をまとめる．

第 2 章

ビッグデータ

本章では，前節で述べたビッグデータを概観し，その定義および特徴について整理する．その上でビッグデータの特徴として大規模，多様性，速度の 3 つを対象にビッグデータを処理可能にするため現在取り組まれている手法を述べ，その課題を整理する．併せて本研究で焦点を当てている多様性について，より詳細に言及し本研究における研究対象を明確にし，研究のアプローチを述べる．

2.1　ビッグデータとは

2.1.1　ビッグデータの定義

インターネットの普及や計算機の処理速度の向上などに伴い，これまで十分に活用されていなかった大容量のデジタルデータを分析を行おうとするトレンドが生まれ，ビッグデータはこのトレンドを表す用語として使われてきている．2014 年現在，ビッグデータに関するコンセンサスを得た明確な定義は存在しないとされている．しかしながらこのビッグデータを提唱した Adrian は

「ビッグデータは一般的に使われているハードウェア環境とハードウェア環境とソフトウェアツールでは，ユーザー層が許容できる時間内にキャプチャ，管理，処理できないデータ」と定義した [43]．またマッキンゼーでは，「ビッグデータとは典型的なデータベースソフトウェアのキャプチャ，格納，管理，分析能力を超えるサイズを持つデータセットである」と定義した [42]．これらの定義は分析者の所属する組織や，そこで扱い得る計算機の性能によりビッグデータとされるサイズが変化するものとなっており，データ量の増大は 2014 年現在，単一データとしては最大で数ペタバイト程度とされる．

一方，データの特徴からビッグデータを定義するアプローチとして 3V と呼ばれる，ボリューム（volume，データ量），速度（velocity，入出力データの速度），多様性（variety，データ種類とデータ源の範囲）がガートナー社より提案されている．これらの概念は分析で用いられるデータの特徴の記述にも使われ，ビッグデータの定義としてコンセンサスを得つつある．ビッグデータの特徴の記述において，データ量の定義はマッキンゼーの定義と同様であるが，速度はデータ処理に求められるレイテンシーの短さ，および多様性は組み合わせられるデータの種類の多さとされる gartner．本論文においても，この観点でデータの特徴を分類し，考察を行うこととする．

2.1.2　ビッグデータに対する期待

ビッグデータに対する期待が高まった理由には，高度かつ複雑な分析手法を用いるよりもデータを大量に集めることで精度が高い分析結果が得られる事実が発見されたことによる．2006 年にアメリカ国立標準技術研究所（NIST）で開催された自動翻訳のアルゴリズムを競うコンテスト [41] では，英語で書かれた文章を他言語に翻訳し，さらに英語に再翻訳し，最初の英語文章と最後の翻訳文章の一致度を測るタスクにおいて，自然言語処理による翻訳を使用せず，

Web 上の大量のデータから単語間の関連性を計算し，関連性の近い単語を単純に割り当てるという単純なアプローチが最も高い精度が得られた．この結果が，分析手法は単純であっても大量のデータが利用できれば，高精度が得られるという知見となり，これまで行われてきた分析手法を高度化するアプローチよりも大量のデータ活用がより付加価値を生むだろうという期待に繋がったものと考えられる．

このように大量のデータを扱うことへの期待が高まってきたことにより，政府が保有しているデータ等を公開し，より有効に活用するというコンセプトの下，オープンデータという潮流も生まれてきた．これらのデータを組み合わせた分析が進展することで，今後ますますビッグデータ活用への期待は高まっていくものと考えられる．

2.1.3　ソフトウェア基盤

ビッグデータを支えるソフトウェア基盤として特徴的な 3 つの基盤を取り上げる．いずれも単独のハードウェアでは処理限界に達するため分散処理を行いビッグデータを取り扱える基盤となっている．

2.1.3.1　Apache Hadoop

Apache Hadoop は大規模データの分散処理を支える Java ソフトウェアフレームワークである．Hadoop はアプリケーションが数千ノードおよびペタバイト級のデータを処理することを可能としている．分散処理環境としては MPI や CORBA 等も知られているが，Hadoop はプロセス同士の通信，監視，障害時の対応も基盤の中で考慮されるため，広く使われるようになった．

Hadoop は，分散ファイルシステム「HDFS　(Hadoop Distributed File System)」と分散処理フレームワーク「MapReduce」から成り，プロセス同士

の通信・監視，障害対応といった分散に必要な機能を提供し，スケールアウトできることが特徴である．Hadoop 上でのアプリケーション開発は，Java 言語等で実装する．

Hadoop の有用性が認められるとともに，Hadoop の使用を容易にし，活用範囲を拡張するツール群が開発され改良が進められている．例えば Hadoop のインターフェースとして Hive や Pig などのツールにより SQL 言語に近い言語で Hadoop の機能が使用できる．また分析においても Mahout や Dremel などの機械学習や SQL による分析が可能な環境が整備されつつある．

ビッグデータの処理や分析において Hadoop は中心的な役割を担うことが多いが，Hadoop の強みは，Map-Reduce を使って，数時間以上かかるようなバッチ処理を複数マシンに分散して高速化できることにある．すなわち，単体のマシンで数時間以上かかるような，ボリュームの大きいデータを読み込んで解析する処理，大量の計算が必要な処理に向いている．一方で Hadoop は速度が高いデータには適しておらず，そのようなデータに対しては CEP など別の処理基盤が適している．

2.1.3.2　NoSQL

NoSQL(Not only SQL）とは，リレーショナルデータベース管理システム (RDBMS) 以外のデータベース管理システムを指すコンセプトである．関係モデルではないデータストアの特徴として，固定されたスキーマに縛られないこと，関係モデルの結合操作を利用しないこと，スケーラビリティが確保しやすい事が多いこと，トランザクションを利用できないものが多いことなどが挙げられる．

NoSQL 系データベース管理システムには，データの格納および取得が高度に最適化されているものが多い．その最適化のために機能性を最小限に

しているものもある．値およびそれを取得するためのキーだけを格納できるKey-Value 型データベースは，その例である．NoSQL 系データベース管理システムの分類を図 nosql に示す．NoSQL 系データベース管理システムが有用な場面は，関係モデルを必要としないデータを扱う時や，大量のデータを扱う時である．用途は多様であり，数百万の key-value ペアを格納したり，数個程度の連想配列を格納したり，数百万の構造的データを格納したりと様々に使われる．この構造は，大規模なデータを統計的に解析したり，増えつづける情報をリアルタイムに解析するのにも利用可能である．産業界での有名な実装として，Google の BigTable，アマゾンの Amazon DynamoDB などがある．オープンソースの実装も数多く存在し，例えば MongoDB， Redis，Apache HBase，Apache Cassandra などがある．ビッグデータの処理に当たってはデータの読み書きを高速に行うことが求められることから，ビッグデータの種類に合わせて高いパフォーマンスで処理可能なデータ基盤として利用が急速に進んできている一方，データの特性に合わせた製品選択が必要である．

2.1.3.3　CEP

CEP(Complex Event Processing) は「複合イベント処理」と訳される．大量に発生し続けるデータをメモリー上でリアルタイムに（ミリ秒レベル）処理する技術で製品の多くが，ロジック実装のための開発環境や入出力アダプターを備えており，開発効率性を高めている．CEP の内部ではデータに対する処理条件や分析シナリオを，あらかじめ CEP エンジンに設定しておき，入力されたデータがその条件に合致すると即座に決められたアクションを実行する仕組みとなっている．従来型の処理がデータをデータベースに蓄積しておき，そこから必要なデータを取り出して，集計や分析をするのに対し，CEP は事前に定義した分析シナリオに，リアルタイムにデータを流し込むことができるた

	キーバリュー型	列指向型	ドキュメント指向型	グラフ指向型
特徴	一つのキーに対して、いくつかの値を対応させる、単純な構造で管理される。高速な読み書きを得意とする。	列別に集計をするような場合に、高速性を発揮する。	任意の長さの文書を保持できるようにしたもの。Webページでテキストや画像などをまとめて扱える柔軟性を重視している。	人同士、モノ同士の関係のように多対多の関係にあるようなデータの取り扱いを得意としている。
代表的な製品	・SimpleDB （米アマゾン）	・Cassandra （Apache SW Foundation、米DataStax） ・HBase （Apache Foundation）	・CouchDB （Apache Foundation） ・MongoDB （米10gen）	・Neo4j （米Neo Technology） ・AllegroGraph （米Franz） ・InfiniteGraph （米Objectivity）
適用例	Amazon： ECサイトのショッピングカート	Twitter、Facebook： ツィートの分類やランキング	FourSquare： SMSチェックイン	Semantic Web、 ゲノム解析

図 2.1　NoSQL の分類

め，速度の高いビッグデータには特に適している基盤といえる．

2.1.4　分析手法

2.1.4.1　データマイニングにおける分類

データマイニングは知識発見とも称され，特に CRM 分野を中心としたビジネスへの応用も近年盛んに行われている．データマイニングの定義としては様々なものが提唱されているが，Bigus による「大規模なデータの集まりから，価値があり自明でない情報を効率的に発見すること」の定義はよく知られている [48]．データマイニングにおいては既に蓄積されている大規模なデータを対象とすることからもデータ収集のコストは深刻な問題ではなく，また「自明でない情報を発見」するという仮説発見型のアプローチは仮説検証型のアプローチとは対照的である．データマイニングは統計学の枠組みにおける探索的デー

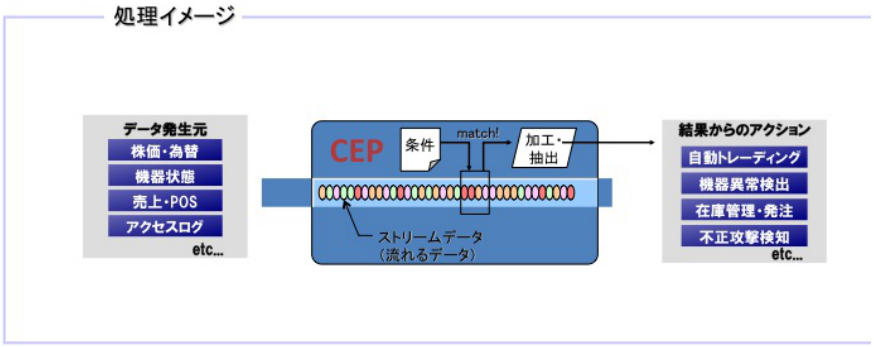


図 2.2　CEP によるリアルタイム処理のイメージ

タ解析に近い目的を持っていると位置づけることができる．したがってデータマイニングで用いられる手法には伝統的な統計解析の他に，相関ルール分析や時系列パターン分析など様々なものが挙げられるが，特に分類手法としては学習モデルを用いるものとクラスタリングのように記述的に分類を行うものがある (図 2.3).

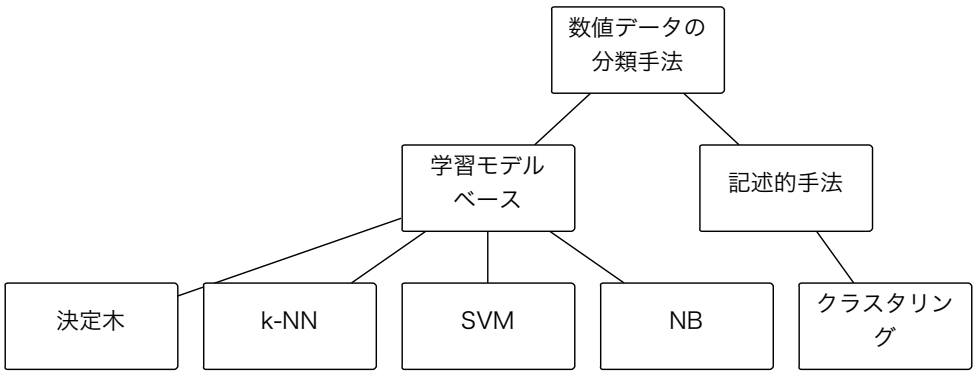


図 2.3　数値データの分類手法

ビッグデータの分析においてもこれらデータマイニングの分析手法は使われ，とくに最近では機械学習のような高度な分析手法の利用への期待が高まっている．ここではデータマイニングで使われる分類手法を紹介する．

データマイニングにおける 分類手法は統計ベースと規則ベースとの 2 つに分けられる．例外的な学習事例の影響を減らし，安定した精度で分類したい場合は，全体の特徴を要約する統計ベースの手法を選択するとよい．また，システムが付与した分類の付与根拠を明確にしたい場合は，分類結果の根拠をある規則として得られる規則ベースの手法がよい．実システムに適用する場合には，それぞれの特徴を踏まえて，利用場面に応じた選択をすることが求められる．以下にそれぞれについての代表的な手法を説明する．

- 規則ベースの手法 [決定木]

  決定木は機械学習の分野でよく知られた手法で，カテゴリ空間を分割する木を作る 2.4．木の節にはカテゴリを弁別するために用いられる規則が記述されている．テキストに規則を順に適用して弁別を繰り返すことで，テキストを分類できる．この方法は計算時間が高速であり，分類付与の根拠が規則として理解できる利点がある．

- 統計ベースの手法 [k-NN(k-Nearest Neighbor)]

  k-NN はパターン認識の分野で研究が進んできた統計的手法である 2.5．k-NN ではまず，テスト事例に類似した k 個の学習事例を検索する．次に k 個の学習事例のカテゴリ毎に類似度の和をとり，最大の値を持つカテゴリをテスト事例のカテゴリとする．k-NN は分類精度が高いことで知られており，実装も簡単である．しかし，高い分類精度を得るには，学習事例を大量に必要とする．k-NN は学習事例をコンピュータのメモリに記憶するため，大量の学習事例を扱う際には，メモリの残り容量に

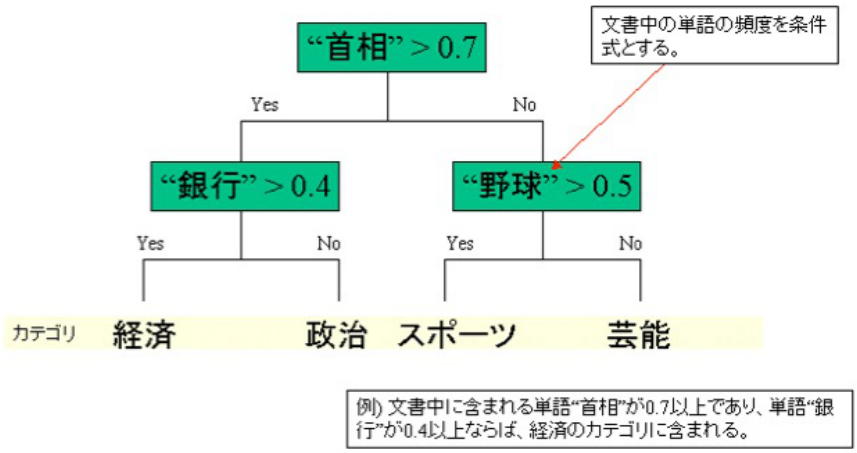


図 2.4　決定木による分類

注意を要する．k-NN では直接分類の根拠になる規則は得られないが，分類に使用した学習例を分類結果の根拠として使用できる．

- 統計ベースの手法 [SVM(Support Vector Machine)]

  SVM は近年盛んに研究されている手法である．もともと二つのカテゴリのどちらに分類すべきか判別するための分類手法であるが，現在は多カテゴリ化の研究も進められている．SVM では異なるカテゴリの学習事例同士の間隔を最大化するような境界線を引き，テスト事例が境界線のどちら側にあるかで分類する 2.6．SVM はカテゴリの増加に伴って学習時間が指数的に増加するが，一般化性能と認識誤りとの両面から最適化が行われるため，過学習 しにくい特徴を持ち，かつ分類精度が高い．SVM の過学習しにくい特徴は，比較的少数の学習事例でも十分な精度が得られるという長所にも繋がる．しかし，SVM は学習に時間が

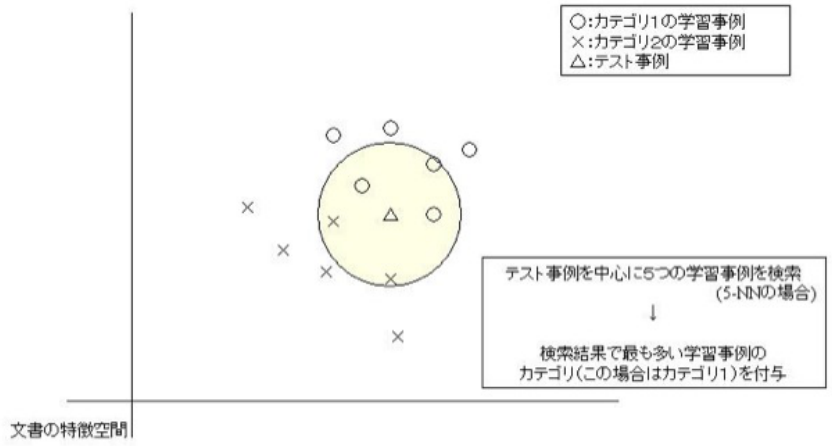


図 2.5　k-NN による分類

かかり，また分類の根拠が得られにくい短所も持つ．

- 統計ベースの手法 [NB(Naive Bayse)]

  NB は，ベイズ統計学の理論に基づき，各カテゴリに分類される確率を学習事例のカテゴリから推定する手法である．この方法では，テスト事例を与えたときベイズの法則により最も分類される確率の高いカテゴリをテキストに付与する．NB の理論では学習事例が母集団から独立に抽出されていること，学習事例が無限にあることが仮定されているが，現実には無限個の事例を用意することは不可能であるため，実際の分類精度はベイズ理論で示される理論的な精度より低くなる．一般に NB の仮定を満たす学習事例の収集は難しく，過去の実験では，NB の分類精度は k-NN や SVM と比べて，低くなることが報告されている．

- 記述的手法 [クラスタリング]

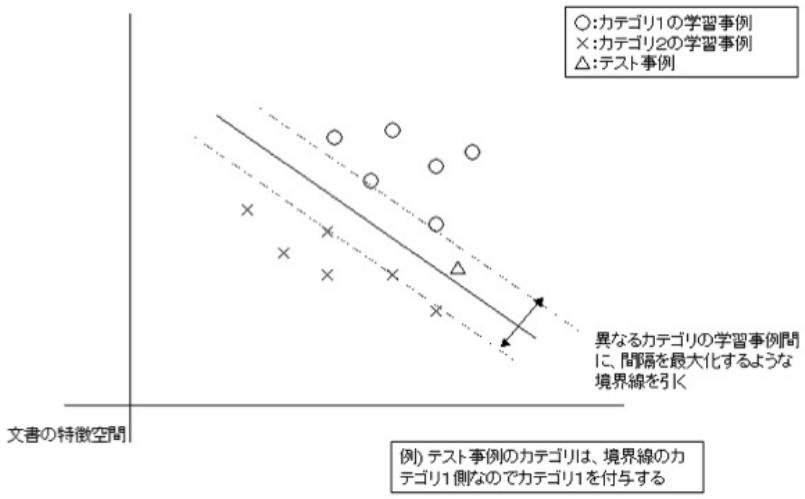


図 2.6　SVM による分類

クラスタリングは，与えられたデータを外的基準なしに自動的に分類する手法である．クラスタリングのアルゴリズムは，大きく分けるとデータの分類が階層的になされる階層型手法と，特定のクラスタ数に分類する非階層的手法とがある．それぞれの代表的な手法としてウォード法，K 平均法などがあり，探索的データ解析手法としても良く使われる手法である．

以上のようにデータマイニングの分類手法は多数あるため，その手法選択の方針が必要になる．分類手法の選択では，分類精度のみに注目してしまい易い．しかし，分類を実システムとして運用する場合はユーザの目的や環境などにも注意を払う必要がある．以下に類手法を選択する際に考慮すべき事項を挙げる．

- 分類精度

  分類手法の精度の指標には，一般に適合率と再現率が用いられる図 3.5. 適合率は分類付与したデータ中に，どの程度正確に分類付与できたかを表す指標で，図中では (A ∩ B) / A にあたる．一方，再現率は分類付与すべき文書にどの程度漏れなく分類付与できたかを表す指標で，図中では (A ∩ B) / B にあたる．一般に適合率と再現率はトレードオフの関係にあるため，ユーザの要求条件に応じて，付与漏れの少なさと正確さとでどちらが重要か考慮し，適合率と再現率のどちらを重視するかを決定する必要がある．

- 学習事例の量

  分類対象となる文書の種類によっては，十分な量の学習事例が得られないことがある．この場合は，少ない学習量でも高い精度が保証される手法がよい．

- 学習時間

  学習事例の特徴表現を頻繁に更新する必要があるシステムでは，更新時間の短縮が要求される．分類エンジンの更新時間短縮には，特徴表現の更新にかかる時間が短い手法を選択する必要がある．

- 分類速度

  短期間に大量の文書を分類することとを要求されるシステムでは，単位時間当たりの処理件数が多い手法がよい．

- メンテナンス性

  分類手法の選択では，ユーザからのフィードバックをシステムに反映する必要が有るかどうかや，分類体系の変更に伴う見直しへの対応が容易か否かなどに注意することが必要である．メンテナンスにかかる手間や時間が少ない手法が有利となる．

- 分類根拠の可読性

  アプリケーションによっては，テスト事例を分類した根拠を要求されることがある．この場合には，分類に際して使用した基準や規則などがわかりやすい手法がよい．

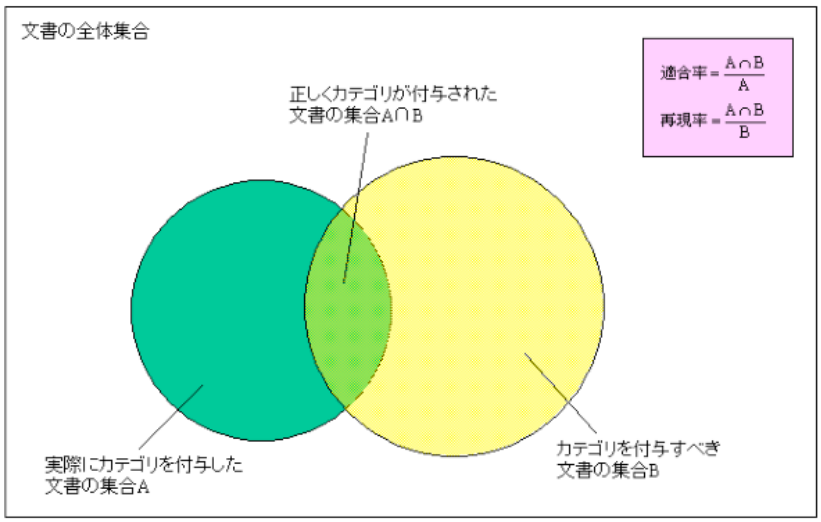


図 2.7　適合率と再現率

ここで述べた手法について分類手法の選択の観点で整理した結果を図 2.8 に示す．分類システムの設計で重要な観点は分類精度である．分類精度は SVM と k-NN が優れており，他の観点における制約が小さければ，この 2 手法のどちらかを選択するのが妥当である．ユーザの知識をフィードバックしたいという要望には，決定木が適切である．決定木はユーザの知識を規則化し，分類エンジンが持つ規則を更新することができる．専門家の知識をフィードバックできれば分類精度の向上が期待できる．学習事例数や処理時間に制約がある場合

	分類精度	学習事例の量	学習時間	分類速度	メンテナンス性	根拠の可読性
決定木	多くの条件で弁別を繰り返すことで、、NBよりも高い精度が得られる。	正しく分類できる木を構成するには、十分な量の事例が必要である。	木の節にある規則の作成方法によって異なるが、学習時間は短くない。	カテゴリにいたる節ごとの条件の比較のみでよいため、計算速度は速い。(※)	規則を一度作成した木に追加でき、ユーザの知識を反映できる。(※)	弁別に使用した木の節を規則で表現された分類根拠として利用可能である。(※)
k-NN	事例の距離をスコアとしたランキング方式を用いており、高い精度が得られる。(※)	学習事例がそのまま距離計算に使われるので、十分な量の事例が必要である。	学習事例の数だけ記憶域にデータを記憶する必要がある。事前に特徴表現を作る必要がなければ、学習時間は必要ない。(※)	距離の近い事例の検索の速度に依存する。	フィードバックを活かすには、パラメータkの調整で柔軟に対応可能である。	分類に使用した事例を根拠として提示できる。
SVM	最大マージン(間隔)の境界線を用いており、高い精度が得られる。(※)	一般化性能が高く、精度を得るための学習事例の量は、他手法より少なくて良い。(※)	Support Vectorの選別、カーネル空間の計算コストが高いので、学習時間がかかる。	2値分類では、テスト事例が境界線のどちら側かの判断ですみ、計算速度は速い。	—	境界線が分類の根拠になるが、解釈が難しい。
NB	学習事例を大量に集めるのが難しく、k-NNやSVMよりも高い精度を得るのは難しい。	NBが本来持つ性能を出すには、学習事例が大量に必要である。	事例の発生毎に分類エンジンの更新が可能なため、学習時間が短い。	計算コストは単語頻度の積分計算であり、速度は早くない。	—	各カテゴリに付与される確からしさが分類の根拠となるが、解釈が難しい。

図 2.8　分類手法選択の観点

は，学習事例が少ないときは SVM が，高速に分類したいときは決定木がこの目的に適している．以上のように，ユーザの要求を満たすには，分類の目的や環境を念頭に置き，適切な分類手法の選択が必要である．

2.1.4.2　シンボリックデータ解析

従来までの表形式を中心としたデータ分析を一般化した新しいデータの捉え方として Diday らによるシンボリックデータ解析 [34] がある．これは，従来のデータ構造の枠組みを拡張し，多様なタイプのデータを許容するシンボリックデータを定義し，それを解析する方法である．例えば，通常の質的データ，量的データ以外に，区間データ，分布値・関数も解析対象とすることが可能である．

シンボリックデータ解析は扱うデータを拡張したものであることから，そ

の解析方法も従来の多変量解析の拡張が提案され，関数データに対する主成分分析，分布値データに対するクラスタリングなどが提案されている．なお，シンボリックデータ解析用ソフトウェアに SODAS (Symbolic Official Data Analysis System) がある．

ビッグデータ分析において，シンボリックデータ解析の持つ柔軟な記述力は，多様性を持つデータを分析するために有用と考えられる．

2.2　ビッグデータ分析の事例

ビッグデータの活用は，その言葉が生まれる前より，バイオテクノロジーにおけるゲノム解析や気象予測など政府系・アカデミック分野に限らず航空機の運航管理やシステムログの分析など民間分野でも行われてきたが，スーパーコンピュータなど高価な資源を用いた限定された利用方法であった．これが前述のハードウェアおよびソフトウェア的な基盤が整備され安価な資源が利用可能になることで，ビッグデータ分析の事例が蓄積された状況にある．

蓄積されたビッグデータの事例は様々な分野にわたることもあり，各省庁や民間企業が整理分類を進めている．総務省が中心としてとりまとめたストック情報とフロー情報および全体と個別という観点で事例を分類している [?].

このうち，政府系のデータ活用やスマートシティなど社会システムに関わる分野は，インパクトが広範囲にわたるためビッグデータの活用への期待が非常に高い分野である．例えば，ビルや住宅のエネルギーマネジメントシステム（BEMS/HEMS）は，電力の使用状況を蓄積，分析することで利用状況に応じた最適な空調，電気製品の制御等を可能とし，電力消費量の削減等に寄与を目指している．また自動車やスマートフォンの位置情報を収集して分析し，交通状況をリアルタイムで提供できるプローブ交通情報サービスが実施され，従来

の交通情報システムに比べ地上設備（車両検知用センサー等）への投資が軽減を目指している．本研究における分析においてもこれらの分野から財務・企業データと交通データを取り上げ，分析を進める．

2.3　ビッグデータ分析における課題

ビッグデータは，従来のデータと比べて以下の 3 点で大きく異なり，分析を困難にしている．

- マシンが作っている
- 新しいタイプのデータソース
- 構造化されていない

1 点目について，ビッグデータの生成のされ方として，人間が介在せずハードウェア基盤で述べたようにセンサから自動的にデータが取得される場合が多い．例えば，GPS データは人が意識せずとも自動的にデータが通信機器を通じて収集され，位置情報に関する状況を知ることができる．マシンから生成されるデータは人の意図とは関係なく収集されるため目的を持ったデータ収集が行われていないことを意味している．そのため収集されたデータのうち意味を持つ部分が少ないことにつながり，分析時には大量のデータから必要な部分を抽出する必要に迫られる．

2 点目については，大量のデータが蓄積できることによって，従来のデータソースで主に扱っていた結果系のデータだけではなく，その結果を生み出した過程に関するデータが利用可能になってきたことが挙げられる．例えば，オンラインショッピングにおける取引においては従来は購買履歴データとして，購買した商品と支払い情報が取引の結果として収集されていたが，それに加えて

表 2.1　ビッグデータの特徴の各観点における解決のアプローチ

No	特徴	アプローチ
1	ボリューム	スケールアウト可能なソフトウェア基盤の利用
2	速度	即時処理可能なソフトウェア基盤の利用
3	多様性	分析シナリオの検討によるデータソースの組み合わせ検討

ショッピング Web サイトのページ遷移の情報も利用されるようになってきた．このようなトラッキングが可能となるような過程に関するデータは非常に膨大な量となるため，データを効率的に蓄積する基盤が必要である．

3 点目については，ビッグデータにおいては，データを組み合わせる際にそのデータを利用する際に想定していた主目的とは異なる目的で蓄積されているため，扱いづらい形式となっていることがある．例えば Twitter などの SNS データは発言内容は自由であり構造化はされていない．2 点目で述べた蓄積基盤の他に，新たに設定した分析目的に合わせ，分析で扱い得る形式にするための適切な変換処理をする必要がある．

2.4　ビッグデータ分析に使用されるアプローチとその限界

前節で述べた課題に対する解決のアプローチは冒頭で述べたビッグデータの特徴であるボリューム，速度，多様性の観点で整理できる．

ボリュームについては，これまでのスケールアップのアプローチからスケールアウトのアプローチに変えることによって問題を解決している．このための基盤ツールとして複数の計算機にデータの処理や蓄積を分散させることが

容易に行える基盤ツールが数多く提供されている．データ蓄積基盤としてはHadoop や NoSQL データベースが利用される．またデータ処理基盤としてはMahout や Dremel などの分析を並列で行える環境も提供されている．このようにボリュームに対しては，このような分散処理ソフトウェア基盤を適切に設計，運用することで対処が行われている．これによりデータを効率的に蓄積・処理することが可能になる．

速度については，データを蓄積して処理するのではなくデータが入力されると即時に処理するアプローチにより解決される．このための基盤ツールとしてCEP エンジンやによる基盤ソフトウェアが使用される．これらの基盤ソフトウェアでは即時処理を可能とするためにルールベースの分析を適用したり，オンライン学習を使用している．これにより分析時に大量のデータから必要な部分を抽出できる．

多様性について，データ変換ツールにより複数のデータソースの統合を行う基盤が提供されている．しかしながら新しい分析目的に合致したデータソースの組み合わせは分析者による設計が必要である．そのため，現在は利用可能なデータと分析目的を結びつける分析シナリオを分析者とデータ保有者とで議論した上で構築するアプローチがとられている．

以上のように，ビッグデータの特徴であるボリュームと速度については，ソフトウェア基盤により分析が実行可能となるが，多様性については，人間による組み合わせの検討を必要とするため，ソフトウェア基盤による解決には限界がある．したがって，あらかじめ利用可能なデータの特徴を詳細に捉えておくことが必要になる．

2.5　データの多様性における分析の困難性

ビッグデータにおける多様性をとらえるため，メタデータマネジメントの考え方を援用する．Leney では複数のデータソースにおける多様性として「incompatible data format」，「non-aligned data structures」，「inconsistent data semantics」を考慮し，管理する必要があると指摘している [?]．分析の文脈においては，データの意味はそのデータソースが生成された業務・目的などを表しており，用途を特定するための情報となる．データの構造はそのデータが持つ統計的特性や構成を表しており，例えばデータに含まれる変数間の相関やデータの偏り，分布，変量数などが該当する．フォーマットはデータの表現方法を表しており，数値データ・カテゴリカルデータ表現，グラフ表現，分布表現などである．実際に分析を行う際にはデータの構造・意味・フォーマットの違いを考慮し，データと分析手法を選択した上で，分析シナリオを構築する．しかしながらデータの構造やデータの表現によっては，既存の分析手法が適用できない場合もある．例として，変数間の相関やデータの偏りが大きい場合は，数値演算上の問題で統計分析でよく使われる多変量解析がうまく動作しない場合がある．このようにビッグデータの多様性のため，従来の分析手法では扱いにくいデータも増えてきている状況にある．

2.6　提案アプローチ

本研究ではビッグデータの多様性に焦点を当て，従来の手法では対応が難しい性質をもつ解析対象に適した分類手法を提案，実データでの検証を通じて，ビッグデータ分析の適用範囲を拡大する．特に，入力データの特徴を表すデータ表現として，部分空間表現，グラフ表現，分布値データ表現を取り上げ，従

来の分類手法より高精度の分類を実現する．提案するアプローチのフローを図 2.9 に示す．

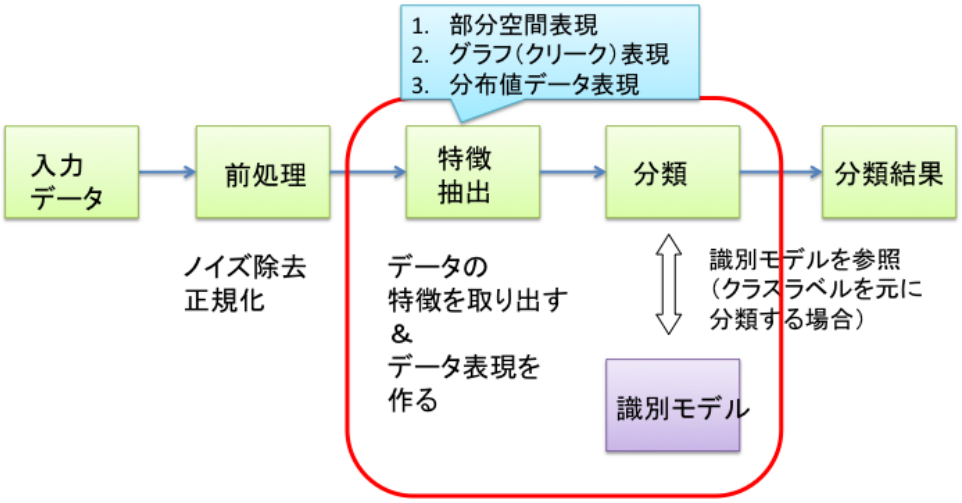


図 2.9　提案アプローチのフロー

第 3 章

企業ランキングにおける分類

3.1 はじめに

ランキングは膨大な情報からの効果的な提示の手段として，レコメンデーションやユーザモデリングなど，幅広い分野で活用されている [1]．こうしたランキングはまた，企業評価の分野においても期待が大きい [2]．例えば，信用リスクの管理においては，顧客のリスクの度合を把握する必要からランキングが活用される．しかしながら，これまでの企業評価においては，専ら業務上の経験をよりどころに手作業で行われ，評価基準で明らかに上位または下位企業などの特徴的な一部の企業しか分析対象にできない課題がある．また特に，その評価基準も分析者に内在しており，知識共有ができない問題がある．

本論文では，統計的機械学習 [7] により得られる部分空間の基底を用いたランキングモデルと，モデルへの関連度を用いることで関連属性の自動抽出を可能とする方法を提案する．この方法では，上位と下位企業それぞれに対する財務データの特徴を表す部分空間の基底を，元の属性の基底変換により得ることで, 統計的な機械学習により，ランキングモデルを構築できる．このランキン

グモデルでは，上位に対応する部分空間と下位に対応する部分空間が表現されることで，評価対象企業に両者の特性が考慮された評点を与える．また，関連属性の自動抽出法を適用することで，ランキング基準に関連する指標を自動的に抽出でき，ランキング基準の解釈を可能としている．

3.2　企業評価におけるランキング

ランキングは膨大な数の企業に対する情報提示の手段として活用され，Forbes や日本経済新聞などの経済誌では，企業の成長性や資産価値など，様々な観点からのランキング結果が報告されている．ランキングの企業評価への応用として，与信管理や投資リスク評価への活用例が挙げられる．

従来，こうしたランキングに対してはロジットモデル [6] がしばしば適用される．ある評価対象企業 i に対するロジットモデルによる評点 $S_{logit}^{(i)}$ は次式のように定められる．

$$S_{logit}^{(i)} = \frac{1}{1 + \exp(-(a_0 + a_1 x_1^{(i)} + ... + a_m x_m^{(i)}))}, \tag{3.1}$$

ここで，$x^{(i)}$ は企業 i を表す特徴ベクトルの属性値，m は属性の数，a は上位，下位のカテゴリーを学習することで得られる係数である．ロジットモデルによる評点は，exp の括弧内における特徴ベクトルと属性との関連を表す係数との積和として表現される．

また，上位と下位とのカテゴリーに基いて企業を分類する古典的な方法として，判別分析の適用が知られている [7]．判別分析の評点 $S_{LDA}^{(i)}$ は次式のように定められる．

$$S_{LDA}^{(i)} = b_0 + b_1 x_1^{(i)} + \ldots + b_m x_m^{(i)}, \tag{3.2}$$

ここで，b は上位，下位企業群において群内分散と群間分散の比の最小化から得られる判別軸による係数である．判別分析における評点もまた，ロジットモ

デルと同様に特徴ベクトルと属性との関連を表す係数との積和として表現される。

これらのモデルでは、与えられたデータを学習することで得られる係数に基づく一つの基準により，評点を付与する。しかしながら，一般に上位企業群と下位企業群とを特性付ける特徴は，カテゴリー毎に異なることが指摘され [?]，上位と下位のカテゴリーで共通した一つの基準を用いるアプローチには、限界があるといえる。

本論文では，上位企業群と下位企業群とで異なる特徴をそれぞれ部分空間の基底で記述することにより，両者の視点を考慮したランキングモデルを提案する。これにより，評価対象企業に対する評点は，上位企業群からなる空間と下位企業群からなる部分空間とに対する相対的な帰属度として定義され，カテゴリー間で異なる基準を反映したランキングモデルが実現できる。

また，ランキングの基準を解釈するための関連属性の抽出に関して，従来モデルでは，上位企業群と下位企業群とで共通した一つの基準を介して解釈するアプローチをとっている。しかしながら，前述のように，一つの基準を用いるアプローチでは，上位と下位との特性の違いが表現しきれないため，適切な属性が挙げられないことが考えられる．部分空間を用いるランキングモデルでは，ランキング基準との関連が，特徴ベクトルを構成する属性の基底ベクトルと上位企業群及び下位企業群の部分空間の基底とから定量化され，上位と下位との特性の違いを踏まえた属性抽出が実現でき，企業評価における効果的なアプローチといえる．

3.3　部分空間に基づくランキングモデルと関連属性の自動抽出

3.3.1　ランキングモデルの構築

本章では，説明のため，財務データを用いた企業評価を例に提案モデルの導出について述べる．

提案するランキングモデルでは，企業 i は m 個の属性を持つ特徴ベクトル $x^{(i)}$，与えられた学習事例は X（$x^{(1)}, x^{(2)}, \ldots, x^{(n)}$，$n$ は企業数）として記述される．また，上位と下位に含まれる企業群のカテゴリーをそれぞれ，h と l とし，それぞれのカテゴリーに含まれる企業群の集合を C_h，C_l と記述するとき，企業群の特徴を部分空間の基底として記述でき [5, 8]，事例として与えられる上位企業群の集合 C_h と下位企業群の集合 C_l それぞれに属する特徴ベクトルから上位企業群と下位企業群のモデルが構築される．ある企業の特徴ベクトルに対するカテゴリーへの帰属度は特徴ベクトルと部分空間を構成する基底ベクトルとの間の余弦値から計算される．一般に，同じカテゴリーに属する企業群は共通の特性を持つと考えられ，こうした共通の特性は類似した基底として，異なるカテゴリー同士の特徴の違いは互いに異なる基底として表現され、上位と下位とで特性の異なる特徴を多変量で表現することができる．

提案するモデルによるランキング処理を 3.1 に示す．3.1 のように提案モデルによる処理は，上位企業群と下位企業群とから部分空間の基底を構成するモデル構築部と構築したモデルに対して評点を付与するためのランキング部とから構成されている。

ある評価対象企業 i の評点 $S^{(i)}$ は，上位と下位の部分空間 $V^{(h)}$，$V^{(l)}$ と評価対象企業 i の特徴ベクトル $x^{(i)}$ を用いて，次式により定義される．

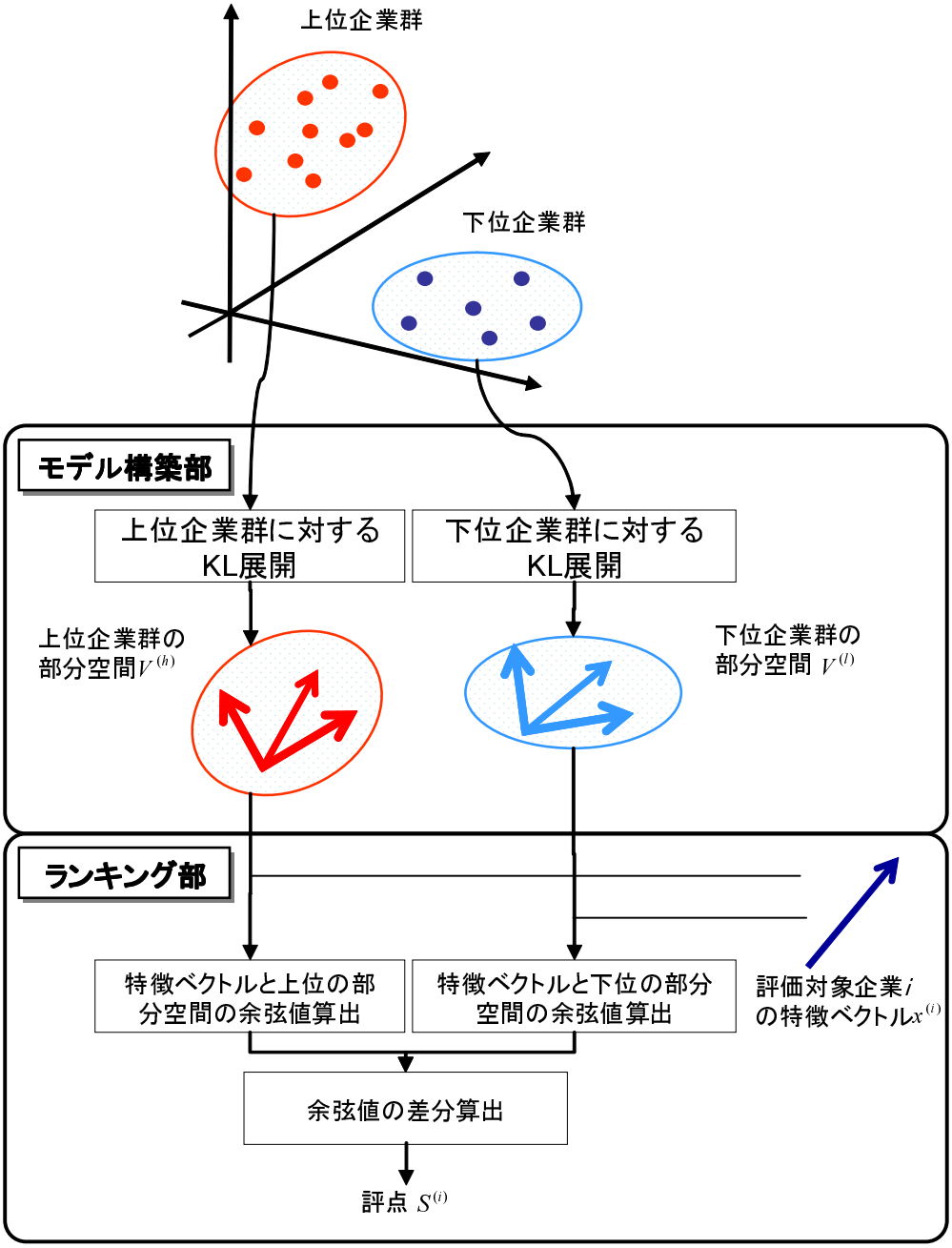


図 3.1　提案モデルによるランキング処理

$$S^{(i)} = S(V^{(h)}, x^{(i)}) - S(V^{(l)}, x^{(i)}), \tag{3.3}$$

提案モデルでは，評点 $S^{(i)}$ を全ての評価対象企業に対して算出し，算出された評点を降順に並べることで，ランキング結果が得られる．ここで，$S(V^{(h)}, x^{(i)})$ は，上位企業群の部分空間に対する企業 i の評点であり，$S\left(V^{(h)}, x^{(i)}\right)$ は次式で定義される．

$$S\left(V^{(h)}, x^{(i)}\right) = \sum_{p=1}^{d_k^{(h)}} \frac{(x^{(i)T}\psi_p^{(h)})^2}{x^{(i)T}x^{(i)}}, \tag{3.4}$$

ここに，$S(V^{(h)}, x^{(i)})$ は直交基底の選び方に依らず決定され，$d_k^{(h)}$ は $V^{(h)}$ の次元数を表す．下位企業群の部分空間に対する評点 $S\left(V^{(l)}, x^{(i)}\right)$ も同様に算出される．

上位企業群の部分空間 $V^{(h)}$ は次式の自己相関行列

$$R^{(h)} = \frac{1}{n^{(h)}} \sum_{x^{(i)} \in C_h} x^{(i)} x^{(i)T}, \tag{3.5}$$

($n^{(h)}$ はカテゴリー h の企業数) の KL 展開から得られる固有値 $\mu_p^{(h)}(\mu_1 \geq \mu_2 \geq, \ldots, \geq 0)$ に対応する固有ベクトル $\psi_p^{(h)}(p = 1, 2, \ldots, m)$ を基底として次式より計算される。

$$V^{(h)} = \left(\psi_1^{(h)}, \ldots, \psi_{d_k^{(h)}}^{(h)}\right), \tag{3.6}$$

ここで，モデルの基底数を表す $d_k^{(h)}(\leq m)$ は次式の固有値の累積寄与率を基準に決定される．

$$\frac{\sum_{p=1}^{d_k^{(h)}} \mu_p^{(h)}}{\sum_{p=1}^{m} \mu_p^{(h)}} \leq \kappa < \frac{\sum_{p=1}^{d_k^{(h)}+1} \mu_p^{(h)}}{\sum_{p=1}^{m} \mu_p^{(h)}}, \tag{3.7}$$

この式のパラメータ κ は，分析対象とするデータから実験的に定められる．下位企業群の部分空間 $V^{(l)}$ も同様に計算される．

3.3.2　反復学習

部分空間を用いたモデルを用いたランキングの精度向上を目的に，各カテゴリーごとに与えられる自己相関行列を次の式（3.8），式（3.9），式（3.10）により更新することで部分空間の基底を更新する反復学習 [5] を導入する．カテゴリー k に対する反復学習は，与えられた学習事例を用いて次の手順で実行される．

$$R_{t=0}^{(k)} = \sum_{x \in C_k} xx^T, \tag{3.8}$$

$$R_t^{(k,j)} = \sum_{x \in C_k, x \to j} xx^T (k \neq j), \quad R_t^{(j,k)} = \sum_{x \in C_j, x \to k} xx^T (j \neq k), , \tag{3.9}$$

$$R_{t+1}^{(k)} = R_t^{(k)} + \alpha \sum_{j(\neq k)} R_t^{(k,j)} - \beta \sum_{j(\neq k)} R_t^{(j,k)}.$$

ここで，$x \to j$ は x がカテゴリー j に分類されることを示す．反復学習では、まず式 (3.8) により自己相関行列の初期値が計算され，t 時点の自己相関行列から導かれる部分空間を用いて，学習事例に対して k を上位とするランキングを行い，各学習事例に対して t 時点のカテゴリーを次の分類基準

$$\begin{cases} \text{カテゴリー } k : \text{学習事例がランキングの上位 } n^{(k)} \text{個に含まれる場合} \\ \text{カテゴリー } j : \text{学習事例がランキングの下位 } n^{(j)} \text{個に含まれる場合} \end{cases} \tag{3.10}$$

（$n^{(k)}$ はカテゴリー k の企業数，$n^{(j)}$ はカテゴリー j の企業数）によって付与する。3.2 のように付与された t 時点のカテゴリーが学習事例のカテゴリーと異なる事例を対象として，条件付自己相関行列 $R_t^{(k,j)}$，$R_t^{(j,k)}$ が算出され、自己相関行列 $R_t^{(k)}$ は，カテゴリー k に属する事例は k の空間に近づけるように，カテゴリー j に属する事例はカテゴリー k の空間から遠ざけるように更新

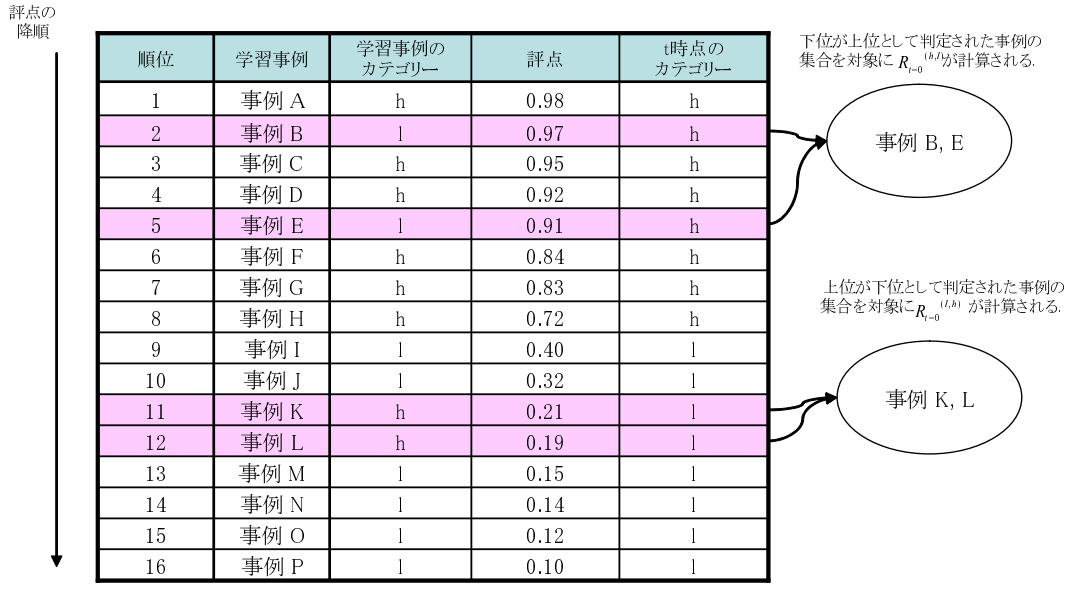


図 3.2　条件付自己相関行列の算出対象となる事例

される．反復学習の収束条件は L 回続けて式 (3.10) による正しいカテゴリーに分類された数が更新されなかったときとする[*2]．

提案モデルにおける反復学習では、上位企業群の自己相関行列 $R^{(h)}$ は k を h、j を l として計算され、下位企業群の自己相関行列 $R^{(l)}$ は k を l，j を h と入れ替えて計算される．

3.3.3　関連属性の自動抽出方式

提案モデルにおけるランキング基準を解釈するには、部分空間の基底と特徴ベクトルを構成する属性の関連の解釈が必要である。これらの関連を定量化する方法として、部分空間の基底と属性を表す基底ベクトル間の角度の余弦値を用いる方法が知られ [9, ?]，提案モデルではに，これを適用してランキン

[*2] 一般に，κ は記憶領域や処理時間の許す範囲で許容される大きな値として 0.9 のまわりで指定される．また，α, β は更新幅を表し，文献 [5] では 0.5 が指定されている．また，L は収束するのに十分な回数を指定すればよい．

グの関連属性を抽出する．ここで，属性に対応する基底ベクトルを, それぞれ $w_q(q = 1, 2, ..., m)$ とすると, 基底ベクトル w_q と上位企業群の部分空間 $V^{(h)}$ との間の余弦値 $\theta(V^{(h)}, w_q)$ は次式で定義される．

$$\theta(V^{(h)}, w_q) = S\left(V^{(h)}, w_q\right), \tag{3.11}$$

下位企業群に対する $\theta(V^{(l)}, w_q)$ も同様に算出される．属性 q のランキングに対する関連度は次式で得られる．

$$Rel(w_q) = |\theta(V^{(h)}, w_q) - \theta(V^{(l)}, w_q)|, \tag{3.12}$$

($|\cdot|$ は絶対値)．$\theta(V^{(h)}, w_q)$ と $\theta(V^{(l)}, w_q)$ は，属性 q がなすベクトルを，上位企業群の部分空間の基底と下位企業の部分空間の基底とに射影したときの余弦値にそれぞれ対応し，これらの違いが大きい属性は，評価の対象となる企業を上位・下位の部分空間のどちらかに帰属させる寄与の度合が大きいことを意味しており，この $Rel(w_q)$ が大きい属性 q を，ランキング基準を特徴付ける属性とすることでランキング基準を解釈することができる．

3.4 実験

本章では，4.1 にて公開データを用いた実験により，従来手法に対するランキング精度の優位性を確認した上で，4.2 にて実際の財務データを用いた実験を通して，知の潜在力企業 [10] を例にランキングモデルと関連属性の自動抽出方式の有効性を検証する．

3.4.1 公開データを用いた実験

本節では，従来モデルとしてロジットモデルおよび判別分析を取り上げ，精度比較により提案モデルの有効性を検証する．

実験に用いる公開データとして UCI machine learning repository[*3]にて公開されている housing データを取り上げる．本データは 506 の事例と 14 の変量からなる住宅価値を予測するためのデータである．本データの所有済み住居の評価額の中央値 (MEDV) によるランキングを評価対象とする．本実験では，その上位群 200 個，下位群 200 個のうち，各群より 100 個をランダムに抽出した 200 個を学習用，残りの 200 個と残りの事例 106 個の計 306 個を評価用として使用した．本実験において，提案モデルの学習に用いるパラメータは，予備実験より，$\kappa = 0.9$，$\alpha, \beta = 0.4$，$L = 300$ を用いた．

ランキング結果の評価には，情報検索分野で用いられる 11 点平均適合率 [11] を用いる．11 点平均適合率は各再現率レベル（0, 0.1, 0.2, . . . , 0.9, 1）における適合率の平均であり，適合率 Pr と再現率 Re は次式で定義される．

$$Pr(l) = \frac{\text{上位 } l \text{ 内に含まれる正解事例数}}{l},$$

$$Re(l) = \frac{\text{上位 } l \text{ 内に含まれる正解事例数}}{\text{評価対象事例数}},$$

ここで，l はランキング上位からの事例数である．

評価用データの 306 事例に対する各々のモデルによる 11 点平均適合率の結果を 3.1 に示す．ロジットモデルおよび判別分析では，AIC 基準に基づくステップワイズ法 [**?**] によって変数選択[*4]を行っている．実験の結果，3.1 から，提案モデルの精度がロジットモデルおよび判別分析を上回っていることが示され，提案モデルの有効性が確認できる．

[*3] http://www.ics.uci.edu/ mlearn/MLRepository.html

[*4] 事前の実験で，全変量では多重共線性のため正しく動作しないことが確認されている．

表 3.1 公開データ (housing data) に対する提案モデル (proposed)，ロジットモデル (logit)，判別分析 (LDA) の 11 点平均適合率

		proposed	logit	LDA
pt.	再現率	適合率	適合率	適合率
1	0.0	1.000	1.000	1.000
2	0.1	1.000	0.962	1.000
3	0.2	1.000	0.962	0.968
4	0.3	1.000	0.878	0.939
5	0.4	1.000	0.870	0.909
6	0.5	1.000	0.855	0.867
7	0.6	0.921	0.847	0.849
8	0.7	0.910	0.809	0.724
9	0.8	0.833	0.734	0.661
10	0.9	0.638	0.592	0.581
11	1.0	0.391	0.386	0.392
11 点平均適合率	-	**0.881**	0.809	0.808

3.4.2 企業評価への応用

3.4.2.1 実験に用いるデータ

本実験では，企業データとして，日本国内の企業の 2002 および 2003 年度の財務情報を用いる．実験では，貸借対照表および損益計算書の項目より算出した一般的な企業評価で使われる財務指標 89 指標を用いた．財務指標の例を 3.2 に示す．また，分析対象企業として，日本標準産業分類で製造業に分類さ

表 3.2　実験で用いた財務指標の例

属性	単位
固定資産回転期間増加率	%
総資本企業収益率	%
総資本経常利益率	%
固定資産回転期間	日
総資産経常利益率	%
総資本回転期間	日
有形固定資産回転期間	日
経常収支比率	%
売上高経常利益率	%
税引前自己資本利益率	%

れる企業から欠測値を除いた 774 社（2002 年度），757 社（2003 年度）を対象とした．ここで，分析データは平均 1，分散 1 に規格化されている．実験では，モデルの学習と評価のため，2002 年度データを学習用（closed）に，2003 年度データを評価用（open）として使用する．

企業評価の例として知の潜在力指数を取り上げる．知の潜在力指数は 6 つの要素から構成され，企業が技術・ノウハウなどの知的資産を使ってどれだけ価値を創出する力を備えているかを表すとされる指数 [10] である．知の潜在力指数の構成要素を 3.3 に示す．

本実験では，この知の潜在力指数を基に専門家によって定められた上位 100 社を知の潜在力を持つ企業としてカテゴリーを付与した．ただし，上位 100 社のうち，2002 年度は 3 社，2003 年度は 5 社を欠測値のため分析の対象外とし，それぞれ，97 社と 95 社を上位企業として使用した．また，下位 100 社のカテ

表 3.3 知的潜在力指数の構成要素

1. 技術革新力
= 過去 2 年分の研究開発費合計
2. 取引の効率性
= 棚卸資産回転日数
3. 従業員の生産性
= 営業利益 / 従業員数
4. 設備の活用力
= 営業利益 / 有形固定資産
5. 将来収益期待
= 当期利益 + 研究開発費
6. 市場の期待
= 株式時価総額

ゴリーを便宜的に，上述した 6 つの要素を用いた第 1 主成分の評点に基づいて付与した[*5]．

3.4.2.2 ランキングモデルの検証

学習用のデータを用いて，3.1 で述べた反復学習の効果を評価する．ここで，学習に用いるパラメータおよび評価基準は，前節の実験と同様である．なお，上位企業群の部分空間の次元数は 23，下位企業群は 31 であった．実験結果で

[*5] 下位 100 社については，専門家によるラベルは与えられていないが，その計算法として表 3.3 で与えられる構成要素 6 個を入力として主成分分析を行い，その第 1 主成分得点を評点として，その値の降順にランキングする方法が紹介されており [10]，本実験ではこの計算法に基づき，下位 100 社を決定した．

は，11 点平均適合率は学習前が 0.983，学習後が 1.000 が得られており，学習が適切に動作していることが確認できる．

また，学習の効果を評価用のデータで確認する．ランキング結果の再現率のリフト図を，3.3 に示す．グラフの横軸にはランキングの評点の降順に事例が並んでおり，縦軸はランキング上位 j 番目に対する再現率を表す．リフト図は，図中の左上の位置に寄るほど高い精度を持つことを示す．ここで，図中の実線は学習後，破線は学習前のモデルである．3.3 に示されるように，学習により，評価用データに対しても精度が向上することがわかる．また，11 点平均適合率は学習前が 0.447, 学習後が 0.571 となり，実際の適用が想定される次年度のデータに対しても，妥当なランキング基準が構成されたことが確認できる．

次に，従来との比較として，ロジットモデルおよび判別分析を取り上げ，精度比較により提案モデルの企業評価への有効性を検証した．3.4 では，評価用データの 757 企業に対する，各々のモデルによる再現率を示した．3.4 と 3.4 からは，提案モデルの精度がロジットモデルおよび判別分析を上回っており，提案モデルが企業評価におけるランキングにおいても有効に機能することが確認できる．

3.4.3　関連属性の自動抽出方式の評価

本節では，提案する関連属性の自動抽出方式の有効性を，知の潜在力を用いたランキングを例に検証する．

検証では，提案方式で上位に選択された指標と表 2 に示した知の潜在力指数の構成要素とを対照させる．ここで，知の潜在力指数の構成要素のうち，全 89 指標（4.2 節参照）に含まれる属性は「従業員の生産性」に対応する 1 人あたり営業利益と「取引の効率性」に対応する棚卸資産回転日数の 2 指標であ

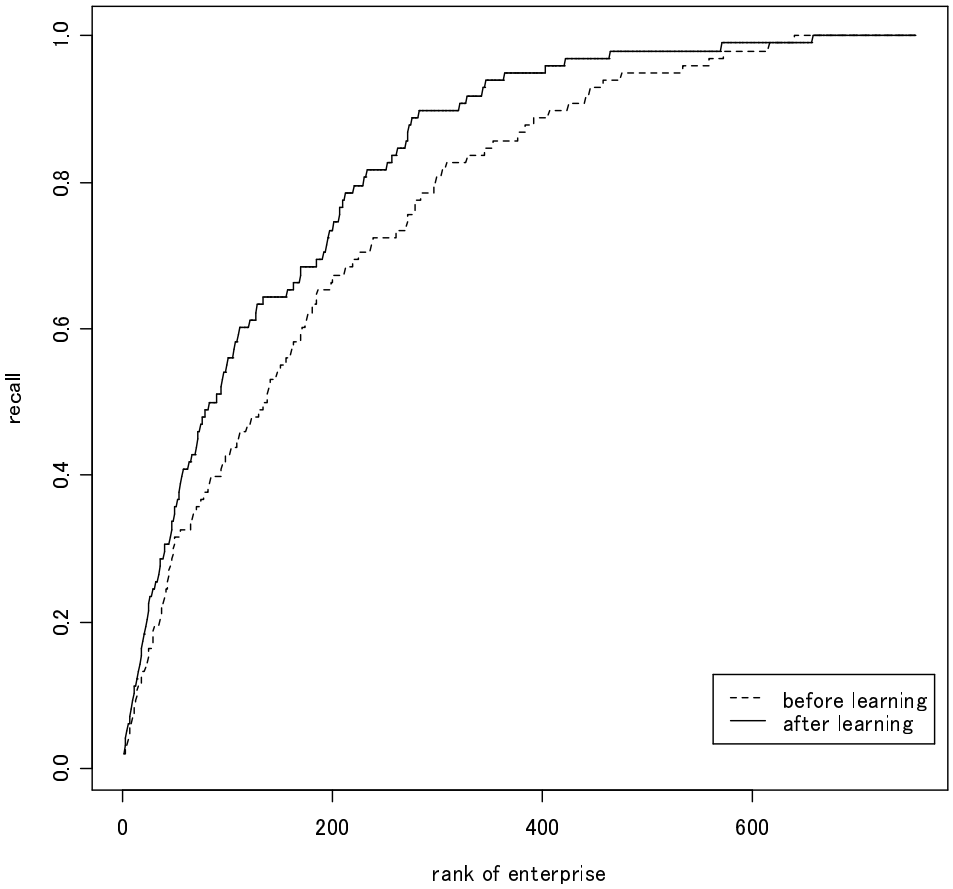


図 3.3 学習前（破線）と学習後（実線）の再現率のリフト図（open）

る．これらが関連属性として上位に順位付けされるかどうかが重要な検証対象である．

3.5 は全 89 指標に対する関連度 (3.2 節参照) を値の降順で示している．また，3.5 は提案方式によって上位に挙げられた属性の 15 指標である．3.5 には，提案方式では関連属性として，知の潜在力を構成する要素である棚卸資産回転日数が 2 位に，1 人あたり営業利益が 6 位に抽出されている．ランキング基準に用いられている要素と一致する指標が上位に挙げられていることは，特

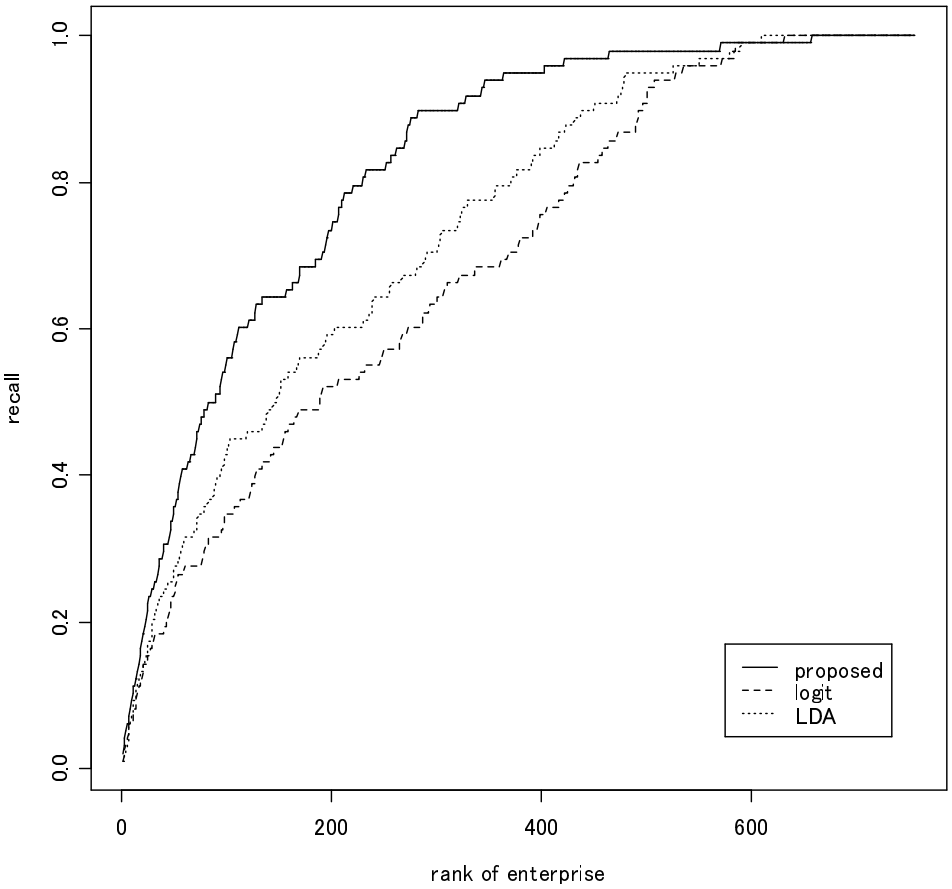


図 3.4　提案モデル（実線），ロジットモデル（破線），判別分析（点線）の再現率のリフト図

筆すべき結果といえ，提案方式が妥当に機能していることが確認できる[*6]．

次に，提案方式で選択された残りの指標について，知の潜在力の残り 4 つの構成要素との関連をみると，「技術革新力」は，提案方式により 1 位となった使用総資本利払前利益率（Return On Assets: ROA）で代替されたものと考えられる．ROA は研究開発費との相関が高いことが経済産業省の通商白書 [13] で示されている．また，「将来収益期待」および「市場の期待」について

[*6] 関連属性を抽出するための手法に，F 値を用いる方法 [?] と AIC による方法が知られるが，いずれも知の潜在力の構成要素となる 2 つの財務指標は選択されない結果が得られている．

表 3.4 提案モデル (proposed)，ロジットモデル (logit)，判別分析 (LDA) の 11 点平均適合率

		proposed	logit	LDA
pt.	再現率	適合率	適合率	適合率
1	0.0	1.000	0.714	0.727
2	0.1	0.909	0.625	0.688
3	0.2	0.846	0.472	0.625
4	0.3	0.684	0.361	0.500
5	0.4	0.684	0.305	0.417
6	0.5	0.544	0.262	0.336
7	0.6	0.488	0.208	0.258
8	0.7	0.360	0.185	0.234
9	0.8	0.339	0.183	0.210
10	0.9	0.274	0.179	0.195
11	1.0	0.147	0.153	0.159
11 点平均適合率	-	**0.571**	0.332	0.395

は，ROA が企業の将来性や企業価値を判断するための重要な指標として知られていることから，技術革新力と同様に ROA により代替されていると判断できる．また，3 位となった総資本営業利益率，4 位となった総資本経常利益率および 9 位の総資本企業収益率は，ROA の簡易指標として知られており，ROA と同様に技術革新力と将来収益期待および市場の期待を代替しているものと考えられる．残る「設備の活用力」については，株主資本を元に投資した資本と有形固定資産との相関が知られており [14]，5 位の株主資本経常利益率

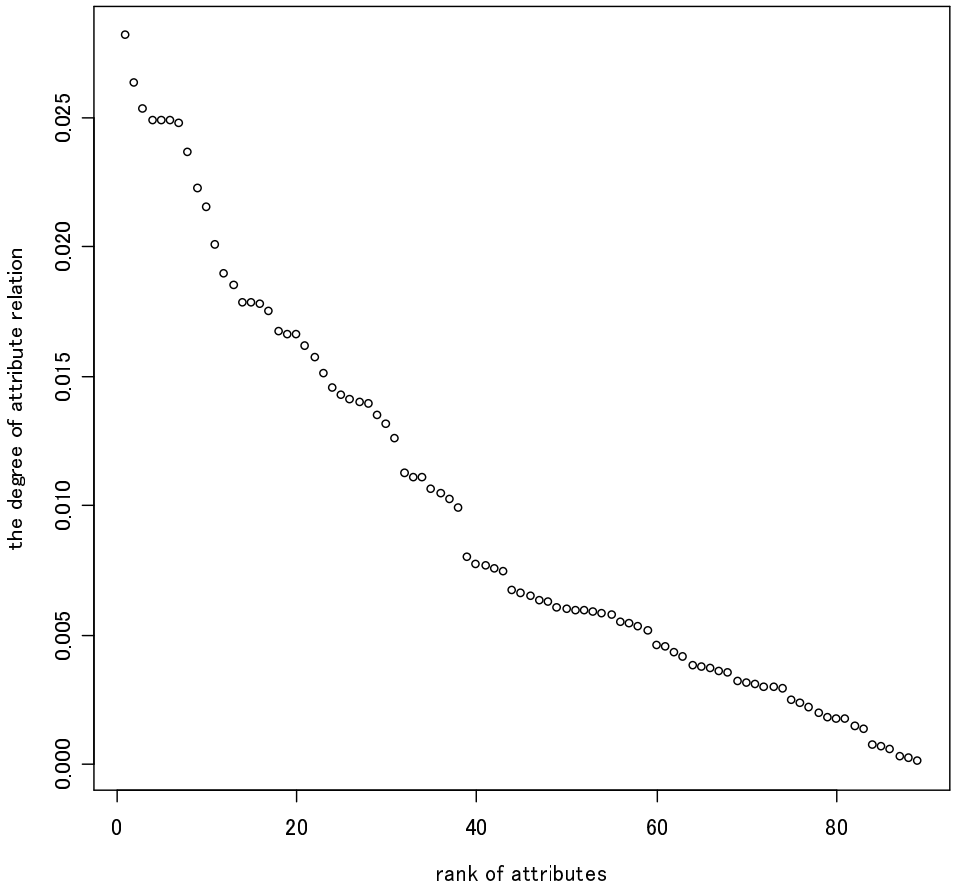


図 3.5　属性に対する関連度の推移

（ROE）が代替指標として対応付けられると考えられ，知の潜在力を構成する 6 つの構成要素と 3.5 の上位の指標との関連が確認できる．

さらに，10 位の売上高営業費用率は，特許権取得によるライセンス支出や広告宣伝費など無形資産に対する支出が営業費用として会計上扱われることから，営業費用の知的財産に対する関連が，通商白書 [13] で指摘されている．また，12 位の売上高営業利益率と 13 位の資本回転期間についても，資本を利益に結びつける能力との関わりから，知的資本との関連が通商白書で示されており，これらの専門家による知見と対応する指標が自動処理で見出されることが

示されている．

　以上のように，抽出された指標は知の潜在力指数の構成要素との対応に加えて専門家の知見と一致する結果が得られており，知識発見に繋がるといえる．

表 3.5　高い関連度を持つ属性の上位 15 指標

順位	財務指標	単位	計算式	分析の観点	関連度
1	使用総資本利払前事業利益率 (ROA)	%	（営業利益 + 受取利息・配当金）/ （負債・株主資本合計）	収益性	0.0282
2	棚卸資産回転日数	日	棚卸資産 / (売上原価 / 365)	効率性	0.0263
3	総資本営業利益率	%	営業利益 / 総資本	総合収益力	0.0253
4	総資本経常利益率	%	経常利益 / 総資本	総合収益力	0.0249
5	株主資本経常利益率 (ROE)	%	経常利益 / 株主資本	総合収益力	0.0249
6	1 人当たり営業利益	円	営業利益 / 従業員数	生産性	0.0249
7	総資本棚卸資産比率	%	棚卸資産 / 総資本	安全性	0.0248
8	1 人当たり売上総利益	円	売上総利益 / 従業員数	生産性	0.0237
9	総資本企業収益率	%	(経常利益 + 支払利息割引料) / 総資本	総合収益力	0.0223
10	売上高営業費用率	%	(売上原価 + 販管費) / 売上高	収益性	0.0216
11	経常損益比率	%	(売上高 + 営業外収益) / (売上原価 + 販管費 + 営業外費用)	流動性	0.0201
12	売上高営業利益率	%	営業利益 / 売上高	効率性	0.0190
13	資本回転期間	日	資本金合計 / (売上高 / 365)	収益性	0.0185
14	経常収支比率	%	経常収入 / 経常支出	流動性	0.0178
15	受取手形回転期間	日	(受取手形 + 割引手形 + 裏書譲渡手形) /(売上高/365)	効率性	0.0178

3.5　本章のまとめ

本論文では，部分空間を用いたランキングモデルと，そのモデルから関連属性を自動抽出する方式を提案した．提案モデルを適用することで，計算機処理によりランキングが実現できること，および関連属性が自動的に抽出でき，妥当な指標が得られることを，実データを用いた実験を通して示した．

また，今回の実験では，企業評価の分析対象データを製造業に絞って実施したが，財務分析では年度間の比較や業種間の比較により，新たな知見が得られるとされており，使用するデータを拡大することによる年度間・業種間でのランキングや関連属性の比較について，今後取り組む予定である．

提案したランキング方式は，財務データへの適用に制約されるものではなく，他のデータマイニングのアプリケーション，例えば，CRM 分野におけるランキングを用いた優良顧客のセグメンテーションなどの適用可能性については，今後の検討課題である．

第 4 章

クリーク表現を基にした企業分類と取引関係の抽出

4.1 はじめに

企業評価は, 取引時の与信をはじめ, 昨今の企業の吸収合併を代表例に一層重要度合を増し, 情報分析の対象として期待されている. ここに, 企業間の関係は, 経営上の意志決定に関する重要な情報源として活用される. 一般に, 投資先の決定や融資先の限度額の決定では, リスク把握の観点から対象となる企業の資本関係や提携関係が考慮される [15] がこれまでは取引先のみの把握に留まり, 取引先の取引先など隣接した企業間関係を超えるような全体的な視点からは, 取引関係を理解することができない状況にあった [16]. 取引先のみではなく取引構造が与信等の企業評価における意思決定に今後ますます重要となると考えられる.

本章では, 極大クリークの全列挙によるグラフ構造の特徴の定量化を通して, 企業の取引構造の把握を図る. 企業をノード, 企業間の取引関係をエッジとす

るグラフに対して, 極大クリーク全列挙アルゴリズムを適用する. 個々の企業におけるグラフ構造の特徴をクリークのサイズに着目した指標により算出し, 企業の取引構造が企業コミュニティの視点から定量化される.

4.2　企業間関係におけるネットワーク分析

従来のネットワークに対する分析方法としてコミュニティ抽出が挙げられ [20], また, グラフ構造に応じた部分グラフの抽出やクリークの列挙のアプローチ [21, 22] が知られている.

企業の活動は一般的に企業間の関係から記述される. 例えば企業間の取引活動においては, 何らかの材料を仕入れて, 製品やサービスを生みだし, 販売先に提供するという商取引が成立している. こうした活動は, 企業間取引のネットワークとして, グラフで記述することができる. 極大クリークはノード間が全てエッジで結ばれる完全部分グラフであり, 企業間取引関係においては密な取引構造を持つ企業コミュニティに対応する. 従って, 個々の企業に着目したとき, 当該企業がノード数の大きい極大クリークに属しているほど, 密な構造を持つコミュニティに属していると見ることができ, 属するクリークの大小に関する構成比を見ることで, 当該企業の持つ取引関係構造のタイプ分類に繋がる定量化が期待される.

4.3　クリークを用いた企業コミュニティ理解

4.3.1　極大クリーク全列挙アルゴリズムの適用

グラフ G の部分グラフのうち, 完全グラフとなっているものはクリークと呼ばれ, そのうち，真に小さいクリークをその中に包含しないクリークは極大クリークと呼ばれる. 一般に, 極大クリーク全列挙問題は NP 困難な問題として

知られる [21][22]. ここでは, 極大クリークを列挙する方法として極大クリーク全列挙アルゴリズム CLIQUES[21] を用いる. 大規模なグラフで最も高速に処理しうることが確認されており, 本検討においてもふさわしいと考えた.

本節では, ノードの有限集合 V, 異なるノードの対 $(v,w)\{v \in V, w \in V, v \neq w\}$ を接続するエッジの有限集合 E より構成される無向グラフを $G = V \times E$ と表記する. G に関して極大クリークを全列挙した結果集合を $C = \{C_1, \ldots, C_m\}$ とする. ここで m は抽出された極大クリーク数である. また極大クリーク C_l のサイズ $|C_l|$ は, 極大クリークに含まれるノード数とする.

4.3.2 擬似クリークによる企業コミュニティ抽出

前節で述べた極大クリークにおいて, 構成ノードがわずかに異なる極大クリークが多数出力されることが指摘されており, こうしたわずかな違いを吸収するため, 擬似クリークの列挙というアプローチ [24] が知られている. 擬似クリークは, 極大クリーク同士の重複を許すように極大クリークの定義を拡大することで, 極大クリーク同士を統合し, より解釈のし易いクリークを構成するものである.

ある分析対象のグラフ G より得られた極大クリークの集合 C' の重複度を次式

$$overlap(C') = \min_{C'_l \in C'} \left(\frac{|\bigcap_{C'_h \in C'} C'_h|}{|C'_l|} \right) \tag{4.1}$$

で定義する. ここで $\cap$ はクリークの積集合を表し, 重複度は定義より, $0 \leq overlap(C') \leq 1$ の値をとる. 与えられた極大クリーク C' に関する擬似クリーク集合 $PC(C')$ は, 極大クリークの組み合わせとして, 次式

$$PC(C') = \bigcup_{C'_h \in C'} C'_h, overlap \left(\bigcup_{C'_h \in C'} C'_h \right) > \alpha \tag{4.2}$$

より定義される. ここで $\cup$ はクリークの和集合を表し, α は重複度を制御する

閾値であり, この閾値を超えるように擬似クリークが構成される. ここでは, 得られた擬似クリークを企業コミュニティとみることとする.

4.3.3　結束度に基づく企業コミュニティ理解

本節では, 企業コミュニティの抽出結果を用いて, 企業の持つ取引構造をタイプ分類するための指標として, 結束度を定義する.

結束度の算出ではまず, グラフ G に含まれるノード v に関して, 極大クリーク C'_l のサイズ $|C'_l|$ に含まれる頻度 $f_{(v,|C'_l|)}$ を要素に持つ行列を次式

$$F = f_{(v,|C'_l|)}, v \in V, |C'_l| = 1, \ldots, max(|C'_l|) \tag{4.3}$$

とし, あるノード v の結束度 u_v は, 次式

$$u_v = \frac{\sum_k \{f_{(v,k)}\} - f_{(v,2)}}{\sum_k \{f_{(v,k)}\}}, k \leq 2 \tag{4.4}$$

で算出され, この値により当該ノードに対して, 接続関係が相互に密接な密型か単純な星型かのタイプ分類が可能となる.

4.4　実験

4.4.1　実験に用いたデータ

本実験では，実際の企業データとして，国内の企業の 2003 年度の仕入並びに販売の取引情報 (日経 NEEDS) を用いる．実験は, 上場企業を対象とし, 仕入先または販売先の取引関係が連結する 2,903 社を用いた.

次数の高い企業例を 4.1 に挙げる. 表には各企業の属する業種を示した. 4.1 からは取引先の多い企業として, 物流関係を持つ商社や系列を持つ自動車業が上位に挙げられることが示されている.

表 4.1 高い次数を持つ企業 (上位 20 社) の業種

順位	次数	業種
1	288	総合商社
2	234	総合商社
3	154	総合商社
4	124	総合商社
5	118	自動車
6	107	通信機（含通信機部品）
7	102	総合電機
8	91	自動車
9	76	家庭電器（含音響機器）
10	74	総合電機
11	72	印刷
12	62	自動車関連
13	62	機械金属商社
14	59	大手建設
15	57	自動車
16	54	印刷
17	54	総合商社
18	51	他電気機器
19	48	家庭電器（含音響機器）
20	48	化学商社

4.4.2　クリークを用いたコミュニティ抽出

4.4.2.1　極大クリークの抽出

表 4.2　サイズの降順に並べた極大クリーク (上位 20 個) と属する企業の業種

順位	クリークのサイズ	業種名 1	業種名 2	業種名 3	業種名 4	業種名 5
1	5	内燃機	自動車	自動車関連	建材商社	自動車部品
2	5	内燃機	自動車	自動車関連	建材商社	自動車部品
3	4	総合商社	大手化学	酸素	化工機械	
4	4	総合商社	塩素・ソーダ	塗料	他化学	
5	4	自動車部品	自動車	車体・他	自動車部品	
6	4	自動車部品	自動車	車体・他	建材商社	
7	4	自動車部品	自動車	自動車関連	建材商社	
8	4	大手化学	総合商社	化学商社	合成樹脂	
9	4	大手化学	総合商社	化学商社	石油精製及び販売	
10	4	他化学	自動車	自動車関連	自動車	
11	4	他ゴム製品	自動車	自動車	車体・他	
12	4	内燃機	自動車	自動車関連	自動車部品	
13	4	内燃機	自動車	自動車関連	電機関連商社	
14	4	内燃機	自動車	化工機械	建材商社	
15	4	内燃機	自動車	自動車部品	建材商社	
16	4	内燃機	自動車	自動車部品	自動車部品	
17	4	軸受	自動車	自動車	自動車部品	
18	4	軸受	自動車	建材商社	自動車部品	
19	4	通信機（含通信機部品）	電子部品	電子部品	他商社	
20	4	家庭電器（含音響機器）	自動車	車体・他	自動車部品	

対象データから得られた極大クリークのうち, サイズ上位の極大クリークに含まれる企業の業種を 4.2 に示す. 得られた極大クリークの数は 6,302 個であり, サイズの最大値は 5 が得られた. 4.2 より, サイズが大きいクリークに含まれる企業の特徴として自動車業および総合商社が含まれていることがわかる. 4.1 との対比により, 高い次数を持つ総合商社よりも自動車業がサイズの大きい極大クリークとして挙げられる傾向が示されている.

4.4.2.2　擬似クリークによるコミュニティ抽出

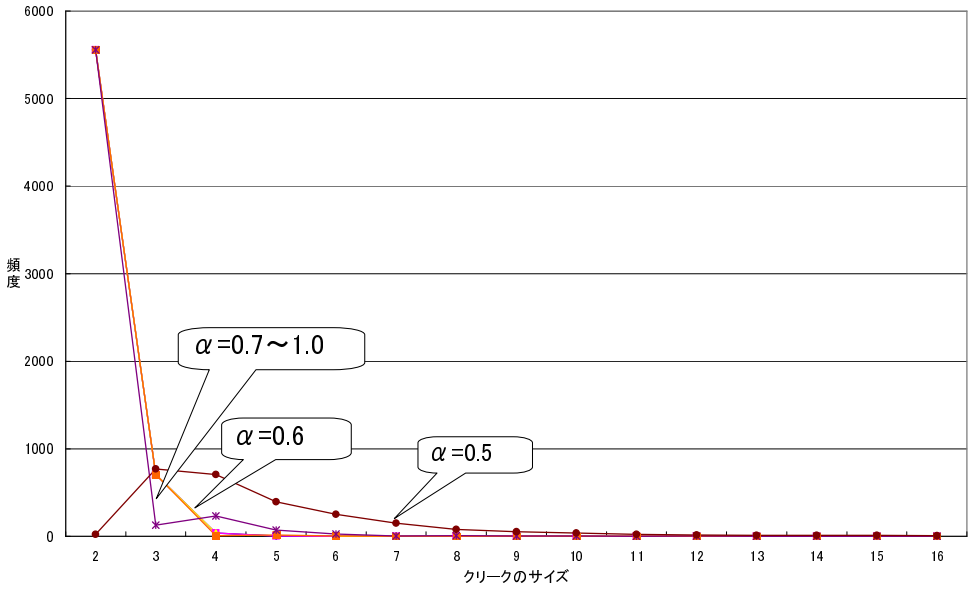


図 4.1　クリークのサイズに対する擬似クリークの頻度

表 4.3　サイズの降順に並べた擬似クリーク (上位 20) と属する企業の業種 括弧内は企業件数を表す.

順位	サイズ	業種								
1	13	他金属製品	繊維機械	化工機械:(1)	軸受:(3)	自動車:(2)	自動車部品:(3)	自動車関連	自動車:(2)	車体・他
2	12	倉庫	繊維二次加工	他化学	化合繊:(2)	他繊維:(3)	繊維二次加工:(3)	繊維機械	総合商社	
3	12	他ゴム製品	繊維二次加工	軸受	他機械	他電気機器	自動車関連:2	車体・他	自動車部品:(2)	自動車:(2)
4	11	化合繊	肥料	大手化学	塩素・ソーダ	他化学:(4)	銑鋼一貫	他製造業	総合商社	
5	10	運輸関連	石炭石油製品	建材商社	特殊鋼	工作機械	自動車関連	自動車	自動車部品:(2)	建材商社
6	10	バイオ他	自動車部品:(4)	他金属製品:(2)	他電気機器	自動車	車体・他			
7	10	倉庫	他化学:(3)	石油化学	塗料:(2)	他製造業	総合商社	食品商社		
8	9	他鉱業	陸運	外航・他	石油精製及び販売:(2)	他商社:(2)	石炭石油製品	化工機械		
9	9	大手海運	他商社	陸運	大手製錬	他パルプ・紙	石油精製及び販売	石炭石油製品	化工機械	総合商社
10	9	調味料	油脂・洗剤	塗料:(2)	化粧品・歯磨	農薬・殺虫剤	他金属製品:(2)	総合商社		
11	9	他化学	他建設	他サービス業	プレス機械	工作機械	総合電機	造船:(2)	総合商社	
12	8	他サービス業	機械金属商社	小売業・紙	他製造業スタッフ	他機械	電機関連商社	自動車関連	自動車	
13	8	電機関連商社	内燃機	自動車	自動車部品:(3)	自動車部品	建材商社			
14	8	自動車部品	内燃機	化工機械	自動車関連	自動車	自動車部品	自動車部品	建材商社	
15	8	自動車部品:(3)	機械金属商社	内燃機	自動車関連	自動車	建材商社			
16	8	他輸送用機器	自動車部品:(2)	鋳鍛鋼	軸受	自動車:(2)	車体・他			
17	8	内燃機	自動車関連	自動車:(3)	自動車部品	時計	建材商社			
18	8	他化学:(2)	肥料	塩素・ソーダ	塗料	他サービス業	化工機械	総合商社		
19	7	他化学	自動車関連	自動車:(2)	自動車部品:(3)					
20	7	自動車	車体・他	自動車部品:(4)	機械金属商社					

対象データに対して, 閾値を 0.8 より 0.1 刻みで 0.5 まで変化させて抽出された擬似クリークのサイズの頻度分布を 4.1 に示す. 本節では, クリークのサイズをある程度大きく保ってコミュニティ理解を図る意図から閾値 0.6 を採用し, そのときの擬似クリークを分析対象とした. 得られた擬似クリークの数 6,038 個において, 擬似クリークのサイズの最大値は 13 が得られた.

4.3 は擬似クリークサイズの上位 20 個に含まれる企業の業種を示す. 最もサイズの大きい擬似クリークに自動車業と化学業に属す企業が抽出された. またこれまで同様, 商社が上位によく現れる結果となっている. ここで, 擬似クリークによる企業コミュニティでは, 極大クリークでは上位に現れなかった繊維業 (4.3 の第 2 位), 鉱業 (4.3 の第 8 位), 海運業 (4.3 の第 9 位) 等, 多様な業種が得られている. いずれの企業も仕入・仲介・販売や物流, 生産・加工・販売等企業の活動に即した企業コミュニティが得られており, 企業の取引構造の理解に繋がる結果が得られていると判断できる.

4.4.3　結束度による取引関係の構造分析

本節では, 擬似クリークに基づき得られた企業コミュニティを用い, 結束度に基づき企業の取引構造の定量化について評価する. 4.2 は前節で抽出した擬似クリークにより, 各企業が擬似クリークのサイズ毎に含まれる頻度と各企業の結束度を重ね合わせた結果であり, 各企業が擬似クリークに含まれる頻度合計の上位 20 社を左から属する業種により示す. 4.2 からは, 結束度の高い業種として自動車業界が, 低い業種として商社や卸・小売業等が挙げられる. 結束度において高い値および低い値を示した企業のグラフ構造の可視化結果を, それぞれ 4.3 および 4.4 に示す. これらの図より, 結束度は企業の取引が相互に依存する度合を表しており, 互いに取引し合う密型となるか, 取引関係が特定企業が中心に行われる星型となるかの構造に関してタイプ分類に繋がる量と

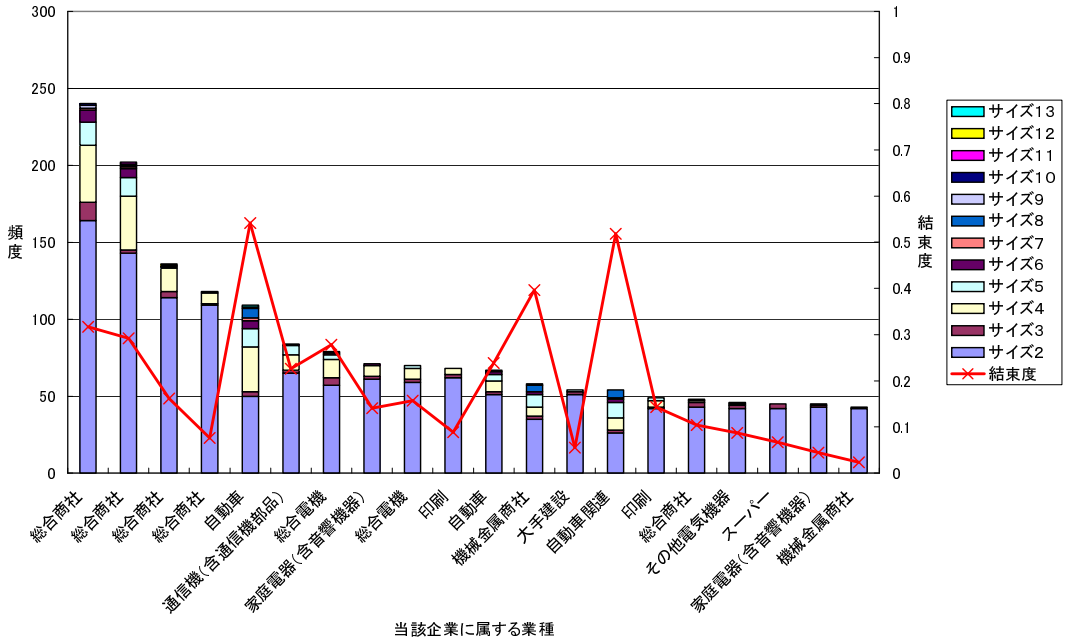


図 4.2　擬似クリークに含まれる頻度上位 20 社におけるサイズの構成と結束度

なっていることがわかる.

4.3 で示されているような密型は自動車業に顕著であり, 部品の仕入先や販売先など業務プロセス上で密接に協調して生産を行っていることが取引構造に反映されていることが示唆される. こうした取引構造では, 一旦どこかの企業がサービス提供できない状況になると悪影響が全体へ波及するというリスクを持つため, 高い結束度を持つ企業に対する与信では, 取引先の取引先のようにコミュニティを捉えた評価が有効と考えられる.

4.4 で示されているような星型は商社や卸売業など仲介を行う企業に顕著であり, 中心には仲介する企業が, 取り巻きには仕入先・取引先となる企業が位置づけられる取引構造となっている. こうした取引構造では, 円滑に商品の仕入・販売ができるが, 売上不振になると取引が停止されるというリスクも持つことが知られる. また取引に当たっては, 与信評価を行う必要性があるが[*1], バック

[*1] 物流系企業では, 売掛金の回収等リスク審査が経営上の課題となっている [32]

ボーンとなる企業の存在がないため, 与信評価は従来通りの取引先にのみに着眼することで構わないと判断し得る.

以上のように, 結束度として定義した指標の大小に応じて, 企業の取引構造を密型や星型にタイプ分類することができ, 企業のコミュニティにおける活動形態の理解と共に企業評価への活用可能性がいえる.

4.5　本章のまとめ

本章では, 極大クリーク全列挙アルゴリズムを用いた企業コミュニティ抽出について述べ, クリークを用いて関係構造の特徴をタイプ分類する結束度を考案した. 考案した指標を用いた実データに対する実験により, 互いに取引し合う密型と特定企業を中心に取引が行われる星型とのタイプの定量化を可能にし, クリーク抽出が企業コミュニティ理解に有用であることを示した.

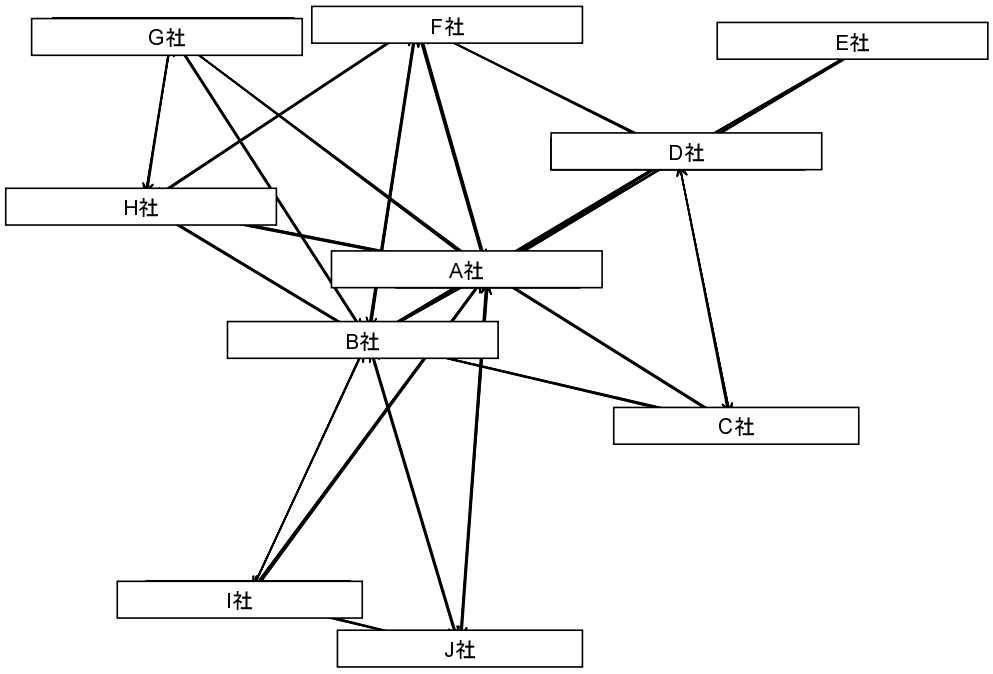


図 4.3　結束度の値が大きい企業の取引構造 (密型) の例

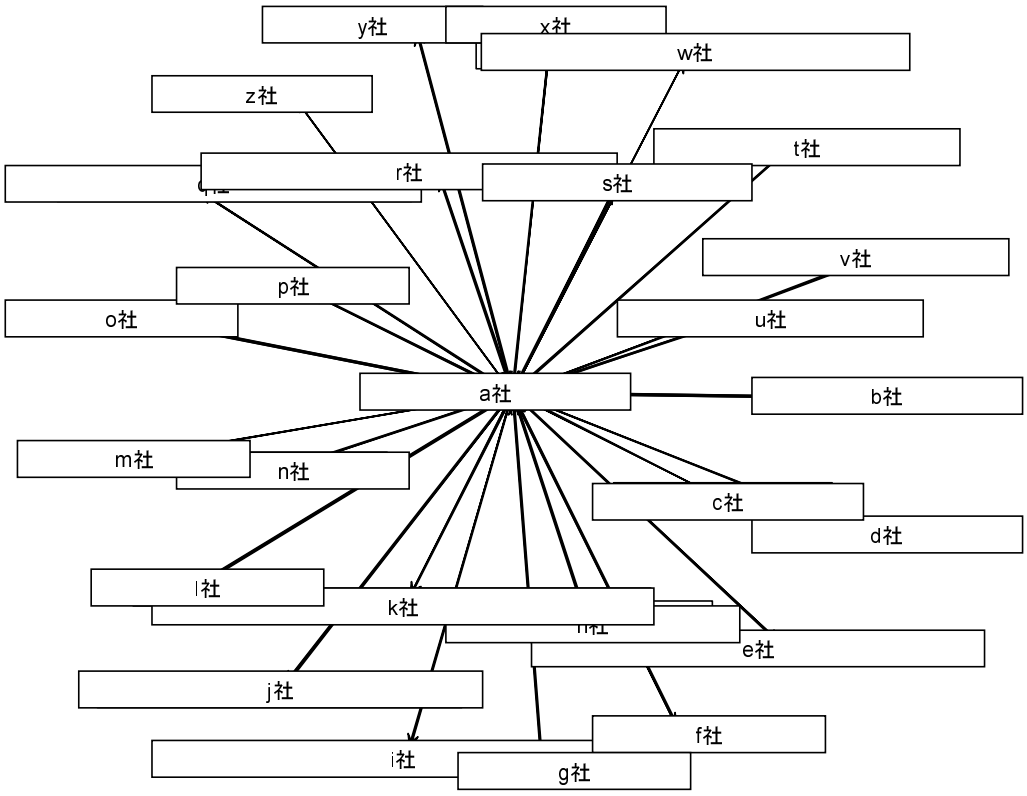


図 4.4　結束度の値が小さい企業の取引構造 (星型) の例

第 5 章

分布値データの非階層的シンボリッククラスタリングと道路特性の抽出

5.1 はじめに

近年, インターネットや各種センサを備えたデバイスの普及に伴い, 蓄積されるデータ量が爆発的に増大している. 特にスマートフォンや携帯電話などのモバイル端末の普及に伴い, その端末上で利用できる SNS や GPS・地図情報を利用したサービスが提供され, これらサービスから発生するデータも激増している状況である.

これら大規模 (Volume)・多様性 (Variety)・高発生頻度 (Velocity) の特徴を持つデータは「ビッグデータ」と呼ばれている [35]. ビッグデータを解析するには, 従来の関係データベースやビジネスインテリジェンスツール上での分析では規模と速度が不十分なため, Hadoop などの分散処理環境がよく用いら

れる．現在, 分散環境上での解析技術として, バッチ処理を中心とした単純集計, 可視化がよく用いられているが, データから価値を引き出すため, 予測やパターン認識など高度な統計分析や機械学習への要望が高まっている．

一方, ITS(Intelligent Transportation System) に代表される交通インフラもビッグデータの活用が期待されている分野の 1 つである [37]．我が国においては高速道路や幹線道路に設置されているトラフィックカウンタ等のセンサから取得した情報を基に, 渋滞や交通規制などの道路交通情報をリアルタイムに送信する VICS(Vehicle Information and Communication System) が普及している．しかしこのシステムではセンサが設置されていない道路は分析対象外となる課題がある．そのため, 車や人に関する移動履歴を, カーナビゲーションシステムやスマートフォンに搭載された GPS 機能により記録したプローブデータを束ねて「ビッグデータ」とすることで, 分析可能な道路を拡大することが試みられている [39]．

プローブデータの分析活用例として交通センサスや渋滞情報生成が知られている．これらの例においては道路特性を表す重要な指標として走行速度が用いられる．走行速度は曜日・時間帯等の交通状況やドライバーの運転特性等で変動するため, 代表値として, 集計処理による平均速度等が利用される．しかしながら, 代表値を用いた場合, 変動情報を直接扱うことができないため複雑な道路特性を見逃す可能性がある．

本章では, 典型的なビッグデータであるプローブデータに関して, プローブデータの複雑な道路特性を表現し, 特性を抽出する手法として分布値データに対する非階層的シンボリッククラスタリングを提案する．本手法ではデータ表現として代表値ではなく確率分布を用い, 確率分布の類似性に基づいて特性を分類するアプローチをとる．これにより道路の複雑な特性が確率分布のパラメータから直感的に理解することができ, また元のプローブデータを少数のパ

ラメータで縮約表現できるため, 計算コストや記憶領域の面でも効率的と考えられる.

5.2　シンボリックデータ解析について

本節では Diday によって提案されたシンボリックデータ解析 (Symbolic Data Analysis: SDA[34]) の基本概念と用語を説明し, SDA の解析手法のうちクラスタリングの先行研究を紹介する.

5.2.1　データ表現

本章では解析対象のデータ表現に, SDA におけるデータ表現であるシンボリックデータを用いる. シンボリックデータの特徴に解析対象データ表現の柔軟性がある [38]. 一般のデータ解析において, 分析対象は個体であり, 各個体はスカラーやベクトル形式で表記されている. それに対し, SDA では分析対象の個体をインディヴィジュアルとして最小解析単位とし, インディヴィジュアルを構成要素とする複合体を解析する. その代表的な形態として, 多次元ベクトル, 区間, ヒストグラム等が知られているが, 本章ではデータ表現に確率分布を扱い, 以降確率分布で表現されたコンセプトを分布値データと呼ぶ.

5.2.2　従来の解析手法とその課題

SDA におけるクラスタリングは Chavent や Gowda and Diday で提案されている [33][36]. これらの手法ではデータ表現として区間を用いている. この場合, 少数の代表値を用いてクラスタを構成するため, 元データが保持していた密度やばらつきなどの情報が反映されない可能性がある.

元データの持つ密度やばらつきを活かしたクラスタリングを行う方法とし

て片山らによる確率分布に対する階層的クラスタリング [38] や Terada and Yadohisa よる確率分布に対する非階層的クラスタリング [40] がある．これらの手法では確率分布をデータ表現として用い, 分布値データ間の類似度や重心分布を新たに定義することで変量間の関係やばらつきを考慮した解析を実現している．

一方, 片山らの階層的クラスタリングにおいては分布値データ間の類似度行列が必要となるため, 分布値データサイズが n^2 のオーダーで拡大し, 大規模データの分析には限界があると考えられる．また Terada and Yadohisa の非階層的クラスタリングにおいては重心分布が分布関数形式で定義されているため, 対象とする区間の確率をすべて保持する必要があるという課題がある．

5.3　分布値データの非階層的クラスタリング手法

本章では, 解析対象コンセプトの散らばりを考慮し, 大規模データでの解析を実現するため, コンセプトの表現として確率分布を想定し, 平均・分散などのパラメータをデータの縮約表現として保持したまま, 分析を行うクラスタリング手法を提案する．

5.3.1　分布値データの定義

コンセプトが確率分布で与えられた場合, 階層的クラスタリングにおけるコンセプトの併合を行うにあたり確率分布の混合を考慮する必要がある．そこで本章ではコンセプトを記述するための確率分布として一次元の混合正規分布を用い, n 個のコンセプトの記述を

$$p_i = \sum_{j=1}^{m} \pi_{ji} N(\mu_{ji}, \sigma_{ji}^2) \qquad i = 1, 2, ..., n \tag{5.1}$$

とおく．ここで $N(\mu, \sigma^2)$ を正規分布, m を混合数, π_{ji} は混合率とし, $\mu_{1i} \leq \mu_{2i} \leq \ldots \leq \mu_{mi}$ とする．

5.3.2　分布値データによるクラスタの定義

コンセプトを分布値データとしたクラスタを定義するために, まず分布値データ間の距離を定義する．計測対象の分布値データがそれぞれ q, r で与えられている．このとき分布値データ間の距離を L^2 空間のノルム

$$L^2(q, r) = \int (q(x) - r(x))^2 dx \tag{5.2}$$

で定義する．

また, 混合正規分布として表された分布値データに対するクラスタ C_k の中心 P_k を

$$P_k = \sum_{j=1}^{m} \pi_{jk} N(\mu_{jk}, \sigma_{jk}^2) \tag{5.3}$$

と定義する．ただし $\pi_{jk} = \frac{1}{\#C_k} \sum_{i \in C_k} \pi_{ji}$, $\mu_{jk} = \frac{1}{\#C_k} \sum_{i \in C_k} \mu_{ji}$, $\sigma_{jk}^2 = \frac{1}{\#C_k} \sum_{i \in C_k} \sigma_{ji}^2$ とする．

式 (5.2) と式 (5.3) に基づき, k 番目のクラスタ C_k はクラスタ中心の集合 C のうち最も距離が小さいクラスタ中心 P_k に属する分布値データ p_i の集合

$$C_k = \{p_i | P_k = \min_{P_k \in C} L^2(p_i, P_k)\} \tag{5.4}$$

として定義される．

上記のようにクラスタ中心の表現を混合正規分布とすることで, 平均・分散などのパラメータをデータの縮約表現として保持することができ, クラスタリング結果の解釈をパラメータを通じて行うことができる．

5.3.3　分布値データに対する k-means アルゴリズム

解析対象とする分布値データが大量になる場合, 階層的クラスタリングは計算量が多いことや, 解析結果表示に用いるデンドログラムが煩雑になり, 視認による解釈が難しいことから実用に適さない. そのため, 本章では非階層的クラスタリングの k-means アルゴリズムに対する拡張を提案する. アルゴリズムでは, 以下の目的関数 Q を最小化するように各対象の属するクラスタを決定する. ここでクラスター数は k とする.

$$Q = \sum_{j=1}^{k} \sum_{i=1}^{n} L^2(p_i, P_j) \tag{5.5}$$

具体的なアルゴリズムは以下の通りである.

＜ステップ１＞　クラスタ中心の初期値を決める. 分析対象として与えられた分布値データ $p_i(i = 1, \cdots, n)$ からランダムに k 個の分布値データを選択し, クラスタ中心の初期値とする.

＜ステップ２＞　各クラスタ中心から各分布値データの距離を式 (5.2) に基づいて算出し, 式 (5.4) に基づき, 各分布値データを割り当てる.

＜ステップ３＞　割り当てたクラスタごとに式 (5.3) に基づいてクラスタ中心を算出する.

＜ステップ４＞　上記の処理で各分布値データのクラスタの割り当てが変化しなかった場合は処理を終了する. それ以外の場合はステップ 2, 3 の処理を繰り返す.

5.4 数値実験

本章では, 人工データにより提案法の動作を確認した後, 実際の交通ビッグデータを用いた数値実験を通して, 提案法の有効性を確認する.

5.4.1 人工データによる動作確認

提案した手法を人工データに対して適用する. ここでは以下の p_1 から p_3 の混合正規分布それぞれに対し, 分布値データを 1000 個ずつ合計 3000 個生成し, これを実験に用いる. 平均や分散を考慮したクラスタリングが実現できるかどうかを確認するため, 混合正規分布を構成する正規分布の平均と分散は同じ値を用い, 混合率のみを変えている. 分布値データは, 分布値データが従う混合正規分布からランダムに 100 個標本を生成し, 混合正規分布のパラメータを推定した結果を用いている.

$$p_1 = 1/2 * N(3,1) + 1/4 * N(6,1) + 1/4 * N(9,1) \tag{5.6}$$
$$p_2 = 1/4 * N(3,1) + 1/2 * N(6,1) + 1/4 * N(9,1) \tag{5.7}$$
$$p_3 = 1/4 * N(3,1) + 1/4 * N(6,1) + 1/2 * N(9,1) \tag{5.8}$$

本実験における評価としては, 上記 3 つの混合正規分布から発生させた分布値データに対して提案法を適用した結果が, 発生元の混合正規分布に分類される割合から, 提案法の妥当性を確認する. 評価指標としては, 以下の式で定義する正解率を用いる.

$$precision = \frac{\text{発生元の混合正規分布に分類された分布値データ数}}{\text{全分布値データ数}} \tag{5.9}$$

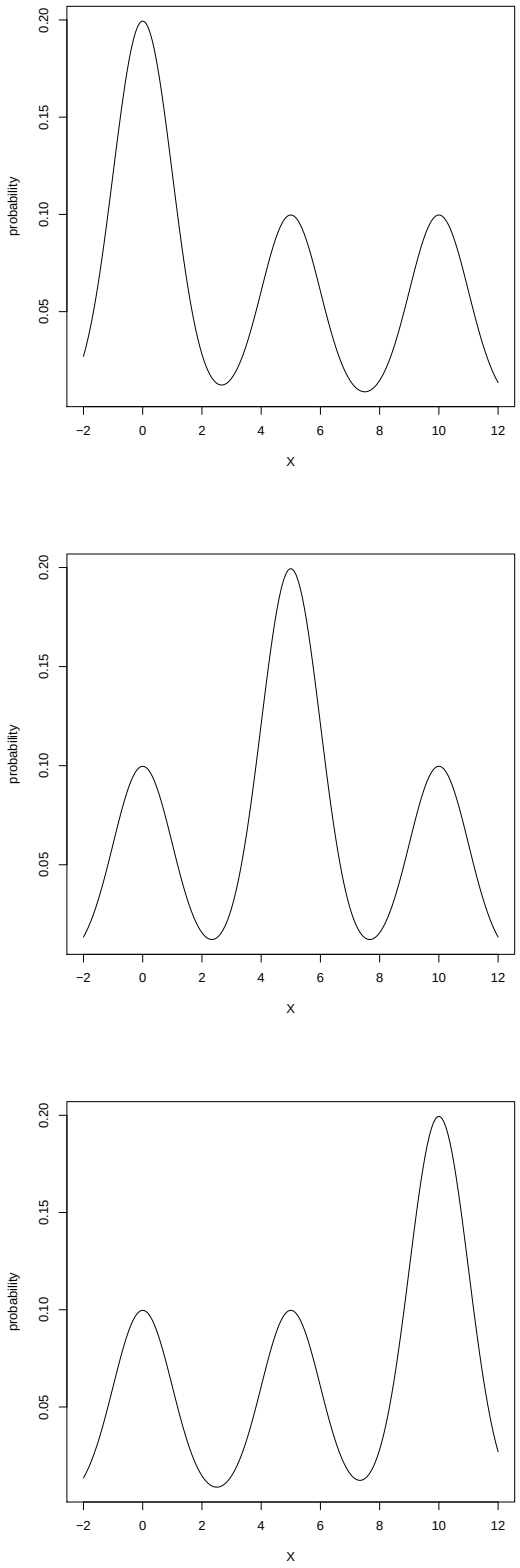


図 5.1　人工データ発生元の混合正規分布 (上から p_1, p_2, p_3)

表 5.1 分布ごとの混合正規分布のパラメータ

クラスタ	平均			分散			混合率		
	分布 1	分布 2	分布 3	分布 1	分布 2	分布 3	分布 1	分布 2	分布 3
1	-0.756	4.617	9.460	0.939	1.465	1.564	0.459	0.256	0.284
2	0.561	5.417	10.07	1.568	1.425	0.941	0.285	0.256	0.459
3	0.391	5.003	9.562	1.431	1.331	1.400	0.264	0.465	0.271

上記人工データに対し $k = 3$ として実験を行った結果得られたクラスタ中心を表 1 に示す. 表 1 にはクラスタ中心の平均, 分散, 混合率を示している. これらのパラメータを発生元の混合正規分布のパラメータを比較すると, 発生元と対応する混合正規分布が構成されていることがわかる. また表 1 のパラメータを基にした混合正規分布の可視化結果 (図 5.2) と発生元の分布の可視化結果 (図 5.1) を比較すると妥当なクラスタ中心が算出されていることが確認できる.

クラスタ中心からの距離により割り当てられた分類結果を表 5.2 に示す. 付与されたクラスタの割り当て結果は, 発生元のクラスタと 0.985 の割合で符合しており, 本クラスタリング手法が適切に動作していることが確認できる.

表 5.2 人工データに対するクラスタリングによる分類: 正解率=0.985

付与クラスタ＼正解クラスタ	1	2	3	計
1	992	8	10	1010
2	4	985	11	1000
3	4	7	979	990
計	1000	1000	1000	

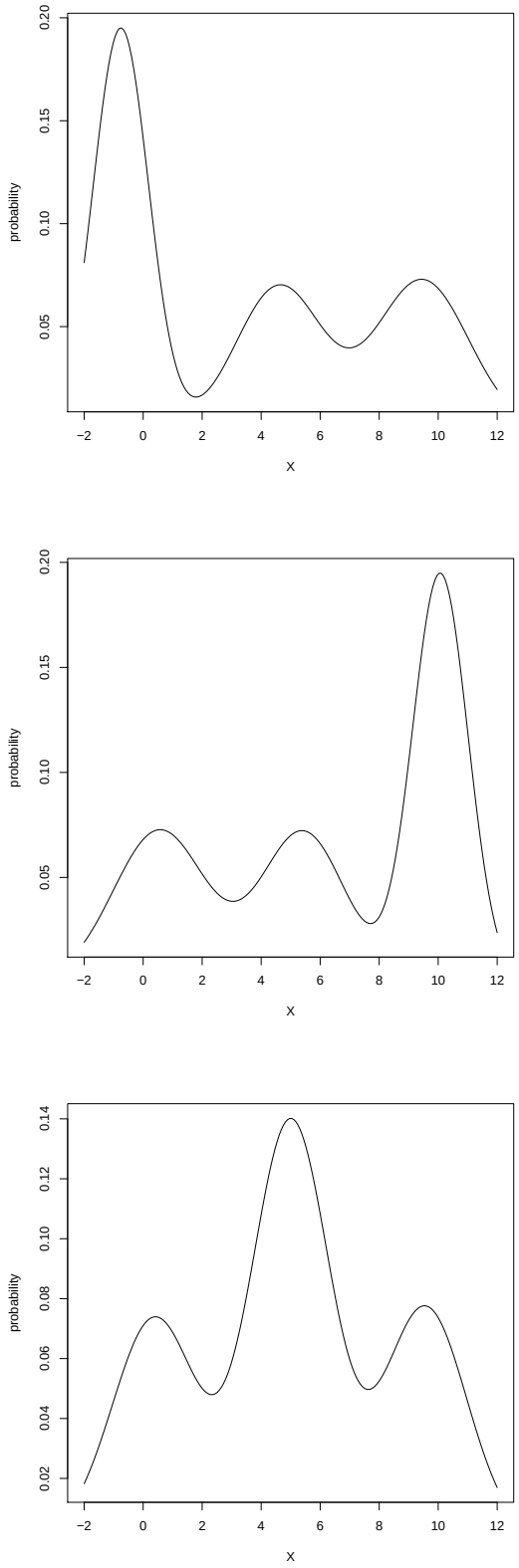


図 5.2　クラスタリングの結果得られたクラスタ中心 (上からクラスタ 1, 2, 3)

5.4.2　交通データによる検証結果

本実験では, 中華人民共和国のある都市で取得された道路情報および 1 ヶ月, 約 3000 台分のプローブデータ (表 5.3) を用いる. 実験では, 道路ごとの車両走行特性の把握を目的とする. そこで, 分析対象とするコンセプトを道路単位とし, 当該道路における走行速度ごとの車両発生確率を分布値データとする. 分布値データに用いる混合正規分布の混合数は事前の検討によりピークが 2 つある分布形状であることから 2 とした. 例としてある道路における走行速度の分布を示す. 上記の設計に基づき分析対象データは以下の手順で作成する.

＜ステップ 1 ＞　停止しているプローブをデータから除外する

＜ステップ 2 ＞　プローブデータを走行している道路に割り当てる

＜ステップ 3 ＞　各道路における走行速度の混合正規分布を EM 法にて推定する

上記の処理により分析対象のサイズは 2.52G バイトから 12M バイトに縮約される. $k = 5$ として, 本データに提案手法を適用し, 得られたクラスタ中心を表

表 5.3　使用した交通情報の特性

特性	数値
道路数	169,161
データ取得期間	2010/5/28-6/28
プローブ数	3,220
プローブレコード数	30,408,640
プローブデータサイズ	2.52G バイト

5.4 に示す．表 5.4 はクラスタ中心の平均, 分散, 混合率を示している．またクラスタ中心からの距離により割り当てられた分類結果を表 5.5 に示す．

表 5.4　交通ビッグデータのクラスタリングの結果得られたクラスタ中心

	平均		分散		混合率	
クラスタ	分布 1	分布 2	分布 1	分布 2	分布 1	分布 2
1	9.404	30.847	2.149	2.740	0.475	0.525
2	19.183	41.305	2.905	2.650	0.459	0.541
3	35.454	66.542	2.925	3.027	0.366	0.634
4	7.058	20.115	1.858	2.385	0.569	0.431
5	16.970	36.726	2.825	3.037	0.491	0.509

表 5.5　交通ビッグデータのクラスタリングによる分類結果

クラスタ	道路数
1	19,611
2	57,878
3	10,869
4	56,768
5	24,035
計	169,161

クラスタリングの結果は分布の特徴を表すパラメータ値を介して評価できる．クラスタ 3 に該当する道路は他のクラスタに比べ 66km/h 付近にピークがあるため, 高速道路や幹線道路など快適に走行できる道路と推察できる．それとは逆にクラスタ 4 は 2 つのピークがともに低速であり, 幅が狭い道路など速

度が出せない道路に対応すると考えられる.

クラスタ 2, 5 はピークが 40km/h と 20km/h 付近にあり低速になる傾向がなく, 渋滞の少ない道路の可能性がある. 一方クラスタ 1 はピークが 10km/h と 30km/h と速度差があり, 渋滞などの理由により速度の変動が大きい道路と考えられる.

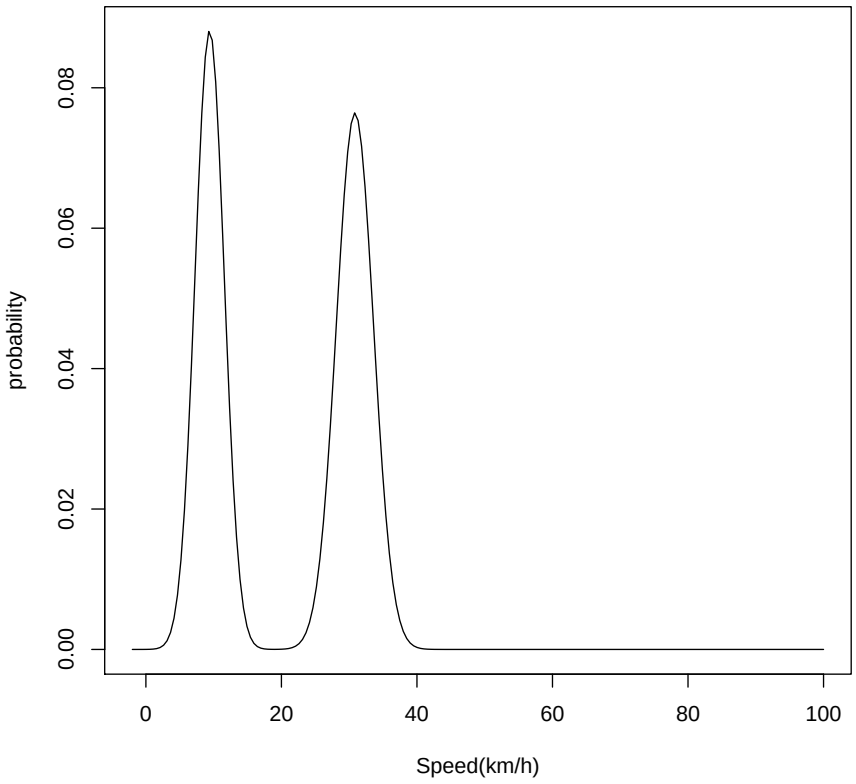
以上のように, 提案手法により得られたクラスタ中心に対応するパラメータを解釈することで, 道路特性を把握し, 道路の特性に基づく分類が実現できることを示した.

5.5　本章のまとめ

本章では, SDA において, コンセプトが分布値データで表される場合の非階層的シンボリッククラスタリングについて提案した. また交通ビッグデータに対する実験により, ビッグデータの大規模・多様性を効率よく扱う縮約表現として確率分布を用い, クラスタリングが実現できることを示した. またクラスタ中心の確率分布のパラメータを検討することでクラスタの解釈につながることも示した.

今回提案したクラスタリング手法は, クラスタ中心を決定する際の混合正規分布の和を計算する必要から, コンセプトの表現は一次元の混合正規分布を対象としていたが, 先行研究に見られるようにパラメータ間の相関を捉えるには, 多次元の確率分布をコンセプトと扱うことができるとより有用になると考えられる. また, 混合数の決定にも AIC 等の情報量規準による決定が考えられ, 提案法の拡張に今後取り組む予定である. また, クラスタ中心からの距離を用いることでコンセプトの外れ値の検出など, クラスタリング以外の手法への利用も適用可能と考えられ, 今後検討していきたい.

提案したクラスタリング手法は, 交通データへの適用に制約されるものではなく, 他の分野に対しても適用可能であると考えられる．他のビッグデータとして活用が期待されるセンサデータやライフログデータにおいても適用し, 応用範囲の拡大を考えたい.



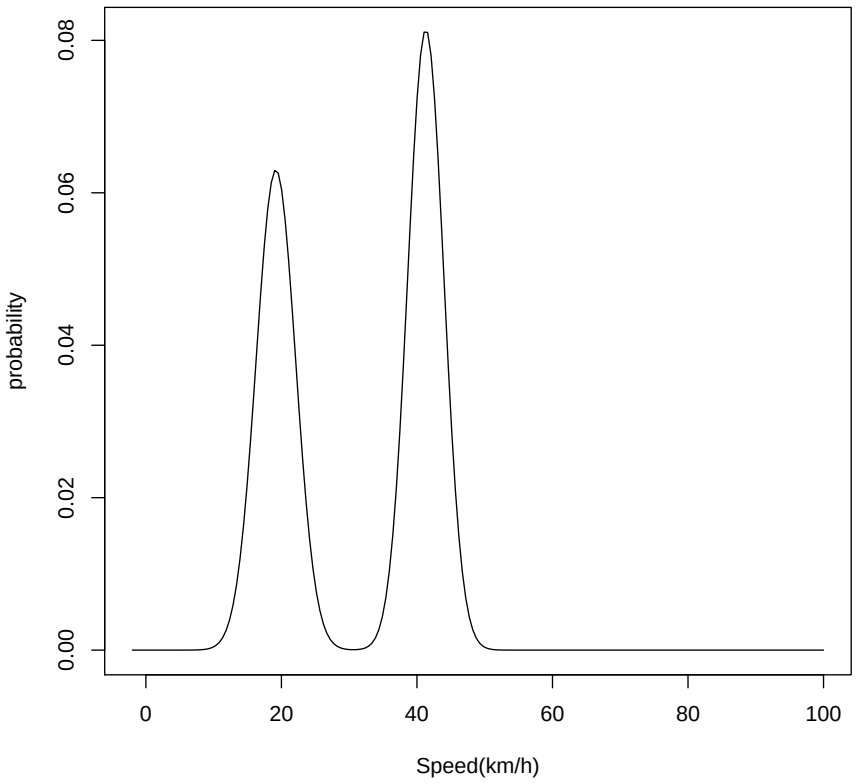
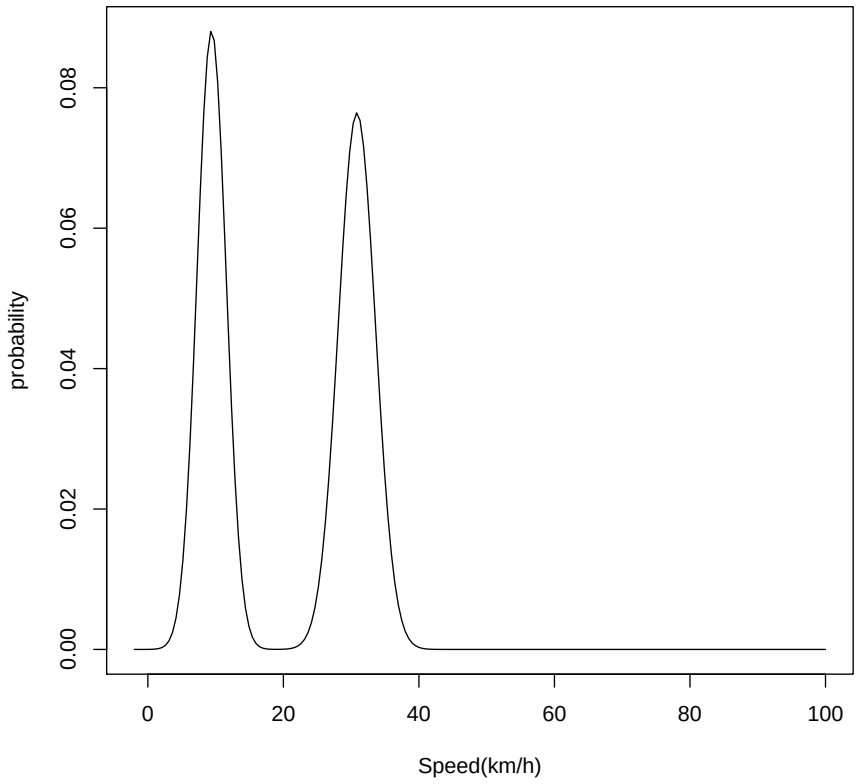


図 5.3　交通ビッグデータのクラスタリングの結果得られたクラスタ中心



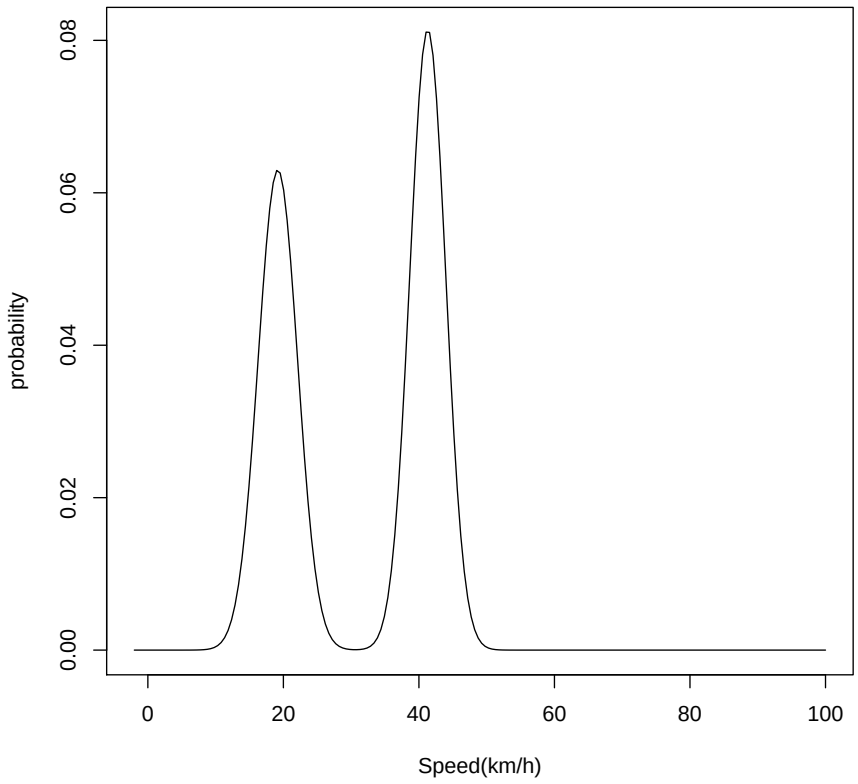


図 5.4　交通ビッグデータのクラスタリングの結果得られたクラスタ中心

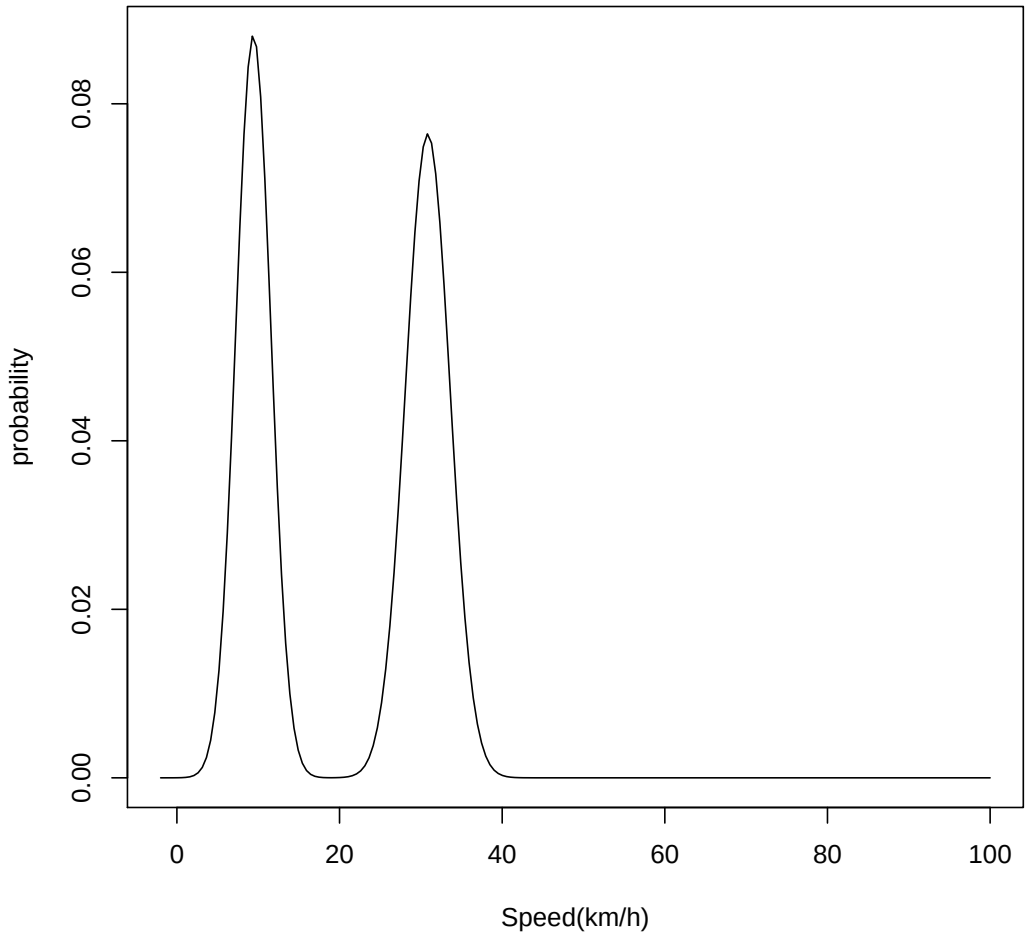


図 5.5　交通ビッグデータのクラスタリングの結果得られたクラスタ中心 (クラスタ 5)

第 6 章

結論

本章では第 3 章から第 5 章までの実験結果についてビッグデータの 3 つの特性ごとに考察を行い，本論文における到達点を述べ，本論文に関連するデータ分析の取り組みを述べ，今後の動向を考察する．その上で残された課題を整理し，今後の研究の方向性について述べる．

6.1　考察

6.1.1　提案分類手法についての有効性

まず，第 3 章から第 5 章までに述べた 3 種類の分類手法の結果についてビッグデータの 3 つの特性ごとに有効性の考察を行う．多様性に関して第 3 章では，部分空間表現において多変量解析手法と比較して多重共線性を持つデータに対して頑健であることが確認できた．第 4 章では，グラフ表現においては極大クリークを活用した分類により取引構造の違いを確認できた．第 5 章では，分布値データにより道路特性の分類を実現できた．以上より，今回扱った財務データや交通データについて，適切なデータ表現を選択することにより，分類

精度向上および構造や特性の把握に繋がる分類結果を得ることができたといえる．

一方，様々なデータ表現から何を選択するかについては，分析目的とデータの特性により変わるため，明確でな指針は存在しないと考えられる．従って，データ表現選択に当たっては，データの特性を把握するための集計分析やデータの分布や相関構造を算出するなどの事前調査を行った上で，いくつかのデータ表現を試行することが必要となる．このためにもビッグデータの処理基盤を活用し，試行を高速に行うことがデータ表現の決定に有益であると考えられる．

ビッグデータの多様性を扱うには，あらゆる分類目的とデータ種別・特性の組み合わせで適切なデータ表現を選ぶことができることが理想といえる．しかし，今回取り扱った分類手法は第 2 章で述べた手法のうちの一部にとどまるため，さらに分析例を追加することが求められる．

次にボリュームに関しては，本論文で提案した手法ではデータの持つ情報を集約したアプローチをとっている．それぞれ，第 3 章のランキングの上位と下位を部分空間，第 4 章のグラフ分類では極大クリークのサイズの集計結果，第 5 章の分布値データのクラスタリングでは，分布のパラメータとなり元のデータサイズより集約できることから，ボリュームの問題が解消されているといえる．

最後に速度に関しては，本論文で検証に用いたデータがいずれも速度が遅いことが前提となっており，バッチ処理においては有効であったが，速度が速い場合についても提案法が有効とは必ずしもいえない．特に速度が速いデータの場合は，処理にかけられる時間が少ないため，適切なデータ表現に変換する処理コストもかけることができないと考えられる．従って速度の速いデータの多様性については，集約を行う今回のアプローチではない，別のアプローチの検

討が求められると考えられる．

6.1.2 データ表現の決定に関する対処について

前節の考察にてデータ表現はその決定に当たり試行錯誤を必要とすることを論じたが，実務上は試行錯誤に必要な業務知識と分析を熟知する専門家の不足や，分析の設計や前処理にかけられる時間的制約が課題となる．特にビッグデータの分析においては，通常のデータ分析よりも試行錯誤を行うための分析環境の準備や処理に時間を要する．

また，データ分析の目的は，その分析を行う業種や業務を特定することで絞り込むことも可能である．例えば，小売業におけるマーケティング業務では，POS データを活用した併売分析により併売率向上が目的になるであろう．併売分析を行うためのアソシエーションルール抽出のために POS データは，トランザクション形式のデータ表現に変換される．このように業種・業務を特定することで，分析目的が絞ることで扱えるデータ種別，必要な分析，データ表現が決定される．

そこで，ある業種，業務における分析目的ごとに分析事例をノウハウとして蓄積，再利用することを考える．分析を行う際にこれらノウハウを利用することで，データ種別の特定，分析手法の選択，データ表現の設計に係る試行錯誤を減らし，分析を効率化することができる．同時に蓄積したノウハウのうち有用なものはベストプラクティスは同じ業務で横展開できるため，今後はこれらをテンプレートとして整備することが進むと考えられる．

2014 年現在はこのようなテンプレート化の取り組みが行われているところである．例えば，NTT データによる分析目的ごとに有用な分析フローをテンプレートとして定義した BICLAVIS の取り組み [45] や，SAS によるの取り組み [46] 等がある．これらは集計分析やデータマイニングできる範疇のデータ

を対象としており，ビッグデータの分析に必要な環境やデータ表現を対象とはしていない．テンプレートのさらなる充実させるにあたり，本論文で扱ったデータ表現をテンプレートへ組み込むことは有用であろう．

一方，テンプレートは有用ではあるもののその整備・活用は容易ではない．問題の 1 つとして，テンプレートを作るための分析事例の収集が挙げられる．一般に分析事例はそれ自体が競争力の源泉であり，テンプレートを作るための汎用化においてデータ保有者の許可を必要とするため，これらのノウハウを蓄積できるのは分析を長年行っている分析ソフトウェア販売会社や分析代行会社などに一部に限られる．また別の問題として，テンプレートに組み込むビッグデータやそれを処理するコード等のリソースの共有がある．データを蓄積するストレージの費用や公開に関する費用が必要になることからリソースは十分共有されていない．

しかしながら，東日本大震災におけるビッグデータ分析など公共性が高く，分析の信頼性が重要な対象のノウハウ蓄積・テンプレート構築に役立つ環境整備は重要な課題と思われる．この環境整備推進に役立つと思われる，興味深い産官学の取り組みがいくつか現れてきている．

データ分析事例の蓄積とリソースの利用を推進する取り組みとしては，第 2 章でも紹介したオープンデータの取り組みがある．政府が保有しているデータを一般ユーザにも利用できるように公開する他，オープンデータを用いた分析のコンテストなどを行うことで分析事例の蓄積が促進されている．また Kaggle[49] と呼ばれる企業や研究者がデータを投稿し、世界中の統計家やデータ分析家がその最適モデルを競い合うサービスでは，これらの分析に取り組んだものから模範解答が提示される．この結果を見ることで分析事例の蓄積も促進される．これらの取り組みは従来十分共有されてこなかったノウハウの公開を促進することができ，新たなテンプレート構築の一助となることが期待さ

れる．

6.1.3　分析自動化について

前節で述べたテンプレートを蓄積することにより，将来的には分析モデルの構築と分析作業の自動化の実現に繋がるものと考えられる．この分析自動化は分析における一つの理想型をなすものであり，いくつか自動化を目指した技術が現れてきている．その中で代表的なものとして米 SAP 社の KXEN が挙げられる [47]．このシステムではマーケティング業務に用途を絞り，データマイニング実施前のデータ加工を Vapnik の SRM 原理を用いてサンプリング・変数選択などを自動的に行うことができる．今後は，この例のように分野に特化し，テンプレートをベースにした分析自動化の取り組みが進んでいくものと考えられる．

6.2　まとめと今後の課題

本論文ではビッグデータの多様性に焦点を当て，従来の手法では対応が難しい 3 種類のデータ表現として部分空間表現・グラフ表現・分布値データ表現を扱える分類手法を提案した．また本提案手法について実データでの検証を通じて有効性を確認した．これにより，多様性を考慮した分析の適用範囲拡大が示唆される．

適用範囲の拡大を狙うには，実データでの検証例を増やすことが必要になる．今回取り扱わなかったセンサやライフログを対象とした検証を追加して行いたい．また本論文で検討したデータ表現の拡大を発展させ，シンボリックデータとして統一的に取り扱う一般化に取り組みたい．そして将来的には分析自動化につなげるための一助としたいと考えている．

謝辞

本論文は，著者が北海道大学大学院　情報科学研究科　複合情報学専攻　博士後期課程在学中および株式会社 NTT データ技術開発本部での研究成果をまとめたものである．本研究を遂行するにあたり，学内，学外，社内，社外問わず数多くの方々に援助を頂いた．ここに感謝を意を表します．はじめに，本研究を進め，論文としてまとめるにあたり，北海道大学大学院　情報基盤センターの水田正弘教授には，研究遂行に当たり多くの有益な御助言を頂いた．水田先生には，著者の博士前期課程においては学術的な助言を与えていただき，博士後期課程においては指導教官として御指導頂きました．また，社会人学生として，本学で学位を得る機会を与えていただき，入学後も社会人として研究活動を進める上で，温かいお言葉，ご配慮をいただきました．本研究での研究活動を成し遂げることができたのは，水田先生のご指導によるところが大きく．心より感謝の意を表します．そして，副指導教官として本論文の執筆にあたり，有益なご助言と熱心な御指導を賜わりました，同情報基盤センター南 弘征准教授，小宮由里子助教におかれましても厚く御礼申し上げます．特に社会人学生として遠隔地からの指導を頂く必要があるところ，大変感謝いたします．

著者は株式会社 NTT データに在籍しながら，博士後期課程への入学の機会を頂き，本研究をまとめることができました．業務の忙しい時期にもかかわらず入学を薦めてくださり，業務が筆者の入学をご快諾してくださいました株

式会社 NTT データ技術開発本部木谷 強本部長に心からの感謝の意を表します．本部長にはたびたび博士後期課程の状況をについてお気遣いいただき，筆者の最大の理解者であり，支援者でいてくださいました．重ねて深く感謝致します．

最後に，社会人学生として二足の草鞋を履いた 2 年間，家庭での務めを十分に果たせない私を許し，常に私を支えてくれた最愛の妻と娘，そして，私の家族を支えてくれた私と妻の両親に心から感謝致します．

業績一覧

[1] 部分空間を導入したランキング方式の企業評価への応用、米森力，西村正寿, 松永務, 関根 純、日本データベース学会論文誌　Vol.7, No.4, pp.49-54, 2009 年 3 月採録

[2] クリークを用いた企業間関係の構造分析、米森力，松永務，関根純，富田悦次、日本データベース学会論文誌　Vol.7, No.4, pp.55-60, 2009 年 3 月採録

[3] 分布値データに対する非階層的シンボリッククラスタリングと交通ビッグデータへの適用について、米森 力, 南 弘征, 水田 正弘、日本統計学会和文誌 Vol.43, No1, pp59–67, 2013 年 9 月採録

参考文献

[1] Adomavicius, G. and Tuzhilin, A., Toward the Next Generation of Recommender Systems, A Survey of the State-of-the-Art and Possible Extensions, IEEE trans. Knowledge and Data Engineering, Vol.17, No.6, pp.734-750 2005.

[2] Thomas, L. C., A survey of credit and behavioral scoring, forecasting .financial risk of lending to consumers, International Journal of Forecasting, Vol.16, pp.149-172 2000.

[3] 高橋久尚, 山下智志, 中小企業リスク情報データベースを用いて, 統計数理, Vol.50, No.2, pp.241-258 2002.

[4] 安川武彦, 平行性の仮定と格付けデータ, 順序ロジットモデルと逐次ロジットモデルによる分析, 統計数理, Vol.50, No.2, pp.201-215 (2002).

[5] エルッキ・オヤ, パターン認識と部分空間法, 産業図書 1986.

[6] Altman, E.I. and Saunders, A., Credit risk measurement, Developments over the last 20 years, Journal of Banking, Vol.21, pp.1721-1742 1998.

[7] Duda, R. O., Hart, P. E. and Stork, D. G., Pattern Classification, John Wiley & Sons 2003.

[8] 松永 務, 大規模知識構造化のための知識表現モデルと遺伝子機能解析へ

の応用, 信学論 D-II, Vol.86, No.8, pp.1196.1204 2003.

[9] 米森力, 末永高志, 原正巳, 松永務, 部分空間比較による変量選択法, 信学技報 PRMU, Vol.103, No.296, pp.61-66 2003.

[10] 岡田依里, 知財戦略経営, 日本経済新聞 2003.

[11] Buckley, C. and Voorhees, E.M., Evaluating Evaluation Measure Stability, Proc. 23rd. ACM SIGIR conference on Research and development in information retrieval, pp.33-40 2000.

[12] Yang, Y. and Barron, A.R., An Asymptotic Property of Model Selection Criteria, IEEE trans. information theory, Vol.44, No.1, pp.95-116 1998.

[13] 経済産業省, 2004 年度版通商白書, 行政 2004.

[14] 経済産業省経済産業政策局調査統計部, 産業活動分析平成 15 年 1-3 月期, 行政 2003.

[15] Barabási, A.-L., *LINKED: The New Science of Networks*, Perseus Book Group 2002.

[16] 安田 雪, 自動車部品サプライヤーの取引構造にみる「べき法則」, *University of Tokyo MMRC Discussion Paper*, No. 33, 2005.

[17] 松永 務: 情報間参照構造に基づく関係性解析による疾患関連遺伝子の探索, 信学論 D-I, Vol. J87-D-I, No. 111, pp. 991–1000, 2004.

[18] J. Kleinberg: Authoritative sources in a hyperlinked environment, *J.of the ACM* Vol. 46, pp. 604–632, 1999.

[19] M. Girvan and M. E. J. Newman: Community structure in social and bilogical newtworks, *PNAS*, Vol. 99, No. 12, pp. 7821–7826, 2002.

[20] 米森力，松永務: グラフスペクトル法を用いた企業コミュニティ抽出, 日本ソフトウェア科学会, JWEIN’06, 2006.

[21] E. Tomita, A. Tanaka and H. Takahashi: The worst-case time complexity for generating all maximal cliques and computational experiments, *Teoret. Comput. Sci.* 363, pp. 28–42, 2006.

[22] K. Makino and T. Uno, New algorithms for enumerating all maximal cliques, *Proc. of 9th Scandinavian Workshop on Algorithm Theory*, Denmark, 2004.

[23] 堀田一弘, 富田悦次, 高橋 治久: 最大クリーク抽出に基づく向きの変化に依存しない人物の顔検出法, 情報処理学会論文誌, 数理モデル化と応用, Vol. 44 No. SIG 14(TOM 9) pp.57–70, 2003.

[24] M. Haraguchi and Y. Okubo: A method for pinpoint clustering of web pages with pseudo-clique search, *Federation over the Web, International Workshop, Dagstuhl Castle, Germany, May 1 - 6, 2005, Revised Selected Papers*, Springer-LNAI 3847, pp. 59–78, 2006

[25] S. Sarkar and K. L. Boyer, Quantitative Measures of Change Based on Feature Organization: Eigenvalues and Eigenvectors, *Computer Vision and Image Understanding*, Vol. 71, No. 1, pp. 110–136, 1998.

[26] Inoue, K. and Urahama, K.: Sequential fuzzy cluster extraction by a graph spectral method, *Pattern Recognition Letters*, No. 20, pp. 699–705, 1999.

[27] 増田直紀, 今野 紀雄, 複雑ネットワークの科学, 産業図書, 2005.

[28] Barabási, A.-L., Albert, R., and Jeong, H.: Mean-field theory for scale-free networks, *Physica A*, Vol. 272, pp. 173–187, 1999.

[29] Kamada, T. and Kawai, S.: An algorithm for drawing general indirect graphs, *Information Processing Letters*, No. 31, pp. 7–15, 1989.

[30] みずほ銀行産業調査部: みずほ産業調査 特集：2003 年度の日本産業動向,

Vol. 10, No.7, pp. 173–187, 2003.

[31] 帝国データバンク: TDB Watching 主要建設会社 200 社 法人申告所得推移調査, 帝国データバンク, 2003.

[32] G. エメリー, 有賀健, 河口品彦: 企業間信用と企業間関係の分析, 大蔵省財政金融研究所 フィナンシャル・レビュー, 1993.

[33] Chavent, M. and Lechevallier, Y. . Dynamical clustering algotrithm of interval data: oprimization of an adequacy criterion based on Hausdorff distance. *Classification, Clustering and Data Analysis, Springer*, 53–59 2002.

[34] Diday, E. The symbolic approach in clustering and related methods of Data Analysis, *Proceedings of IFCS, Classification and Related Methods of Data Analysis*, 673–684 1998.

[35] Eaton, C. and Zikopoulos, P. Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data *McGraw-Hill Companies*, 2012.

[36] Gowda, K.C. androgen Diday, E. Symbolic clustering using a new dissimilarity measure. *Pattern Recognition*, **24**, 6, 567–578 1991.

[37] Hunter, T., Moldovan, T,, Zaharia, M., Merzgui, S., Ma, J., Franklin, M.J., Abbeel, P and Bayen, A.M. Scaling the Mobile Millennium System in the Cloud. *SOCC 2011* 2011.

[38] 片山琴絵, 南　弘征, 水田正弘.「分布値データに対する階層的シンボリッククラスタリングについて」『計算機統計学』, **22**, 2, 83–89 2009.

[39] 増田有孝「ビッグデータを活用した高精度の道路交通情報サービス」『ITソリューションフロンティア, 16–19 2012.

[40] Terada, Y. and Yadohisa, H. Non-hierarchical clustering for

distribution-valued data. *COMPSTAT 2010: Proceedings in Computational Statistics, Psysica-Verlag, Heidelberg*, 1653–1660 2010.

[41] NIST 2006 Machine Translation Evaluation Results *http* : *//nist.gov/speech/tests/mt/*2006*/doc/mt*06*eval*$_o$*fficial*$_r$*esults.html*

[42] McKinsey Global Institute, Big data: The next frontier for innovation, competition, and productivity, *http* : *//www.mckinsey.com/insights/mgi/research/technology*$_a$*nd*$_i$*nnovation/big*$_d$*ata*$_t$*he*$_n$*ext*$_f$*rontie* 2011.

[43] Big Data *http* : *//www.gartner.com/it − glossary/big − data*

[44] Doug, L, Application Delivery Strategies, *http* : *//blogs.gartner.com/doug − laney/files/*2012*/*01*/ad*949 − 3*D* − *Data* − *Management* − *Controlling* − *Data* − *Volume* − *Velocity* − *and* − *Variety.pdf* 2001.

[45]「戦略的な情報活用を組み込んで顧客の変革を加速させる NTT データの BI サービスを体系化して提供開始」,*http* : *//www.nttdata.com/jp/ja/news/release/*2010*/*082000.*html* 2010.

[46] 松永佳丈「SAS による“ビッグデータ”活用の新しい挑戦と、先進的な取り組み」*http* : *//www.it − yugo.jp/program/PDF/*2 − 1$_m$*atsunaga.pdf* 2012.

[47] KXEN *http* : *//www.kxen.com/*

[48] Bigus, Hoseph P.,Datamining with nural networks, McGraw − Hill, 1996.

[49] kaggle *http* : *//www.kaggle.com/*