Intrinsic Dimensional Design and Analysis of Similarity Search

審 査 の 結 果 の 要 旨

論 文 提 出 者 氏 名　ネット ミヒャエル

データに潜む規則性を発見し、これを新たな知見の創出や意思決定に役立てるためのデータ解析技術は、情報の溢れる現代社会における情報処理技術の要であり、その研究は機械学習やデータマイニング、パターン認識、統計科学、あるいは各々の応用領域などの様々な分野において精力的に進められてきた。情報通信技術や計測技術の進歩に支えられた電子的な情報基盤が急速に整いつつある現在、集められたデータを有効に活用するためのデータ解析技術は益々重要になってきており、近年のビッグデータの潮流はまさにその象徴といえる。

与えられたデータに対して類似あるいは関連したデータをデータベースの中から見つける類似検索は、分類、クラスタリング、異常検知などの機械学習・データマイニングの主要課題においてしばしば現れる基本問題のひとつであり、理論的・実用的に効率的なアルゴリズムの設計やその性能保証、データの特徴づけなど様々な観点からの研究が行われてきた。特に高次元空間における類似検索は、実用上の重要性と技術的な困難さから現在も盛んに研究がおこなわれている中心的な研究課題のひとつである。

本論文は「Intrinsic Dimensional Design and Analysis of Similarity Search」（内在的次元に基づく探索手法の設計と解析）と題し、９章からなる。

第１章「Introduction」(序論)では、類似検索問題とこれにまつわる様々な課題を概観した後、本論文の主要な成果を概説している。

第２章「Related Work」(関連研究)では、所謂「次元の呪い」と呼ばれる高次元データを扱う際に起きる深刻な性能低下の問題と、データの複雑さを定量化するための様々な指標やその考え方を纏めている。

第３章「Generalized Expansion Dimension」(一般化拡張次元)では、データの複雑さの指標のひとつである拡張次元を、類似検索の文脈に拡張した一般化拡張次元を提案し、いくつかの具体的な距離空間において解析を行っている。 また一般化拡張次元を実際のデータから推定するための頑強な方法を与えている。

第４章「Continuous Intrinsic Dimension」(連続的内在的次元)では、一般化拡張次元の連続化として位置づけられる連続的内在的次元に対し、そのデータ解析の文脈における役割を論じるとともに、極値理論に基づく具体的な推定法を構成している。

第５章「Discussion of Intrinsic Dimensional Models」(内在的次元についての議論)では、これまでの章で登場したいくつかの内在的次元の指標の違いを実験的に考察するとともに、雑音がその推定精度に及ぼす影響についても論じている。

第６章「Interpretation of Shared-Neighbor Distances」(共通近傍距離の解釈)では、近傍探索において重要な役割をもつ共通近傍距離の性能保証に対して、連続的内在化次元を用いた解釈を与えている。

第７章「Rank Cover Trees for Nearest-Neighbor Search」(近傍探索のためのランクカバー木)では、近傍探索のためのデータ構造であるランクカバー木の性能に対して一般化拡張次元を用いた保証を与えるとともに、実験的な考察も行っている。

第８章「Dimensional Testing for Reverse Neighbor Search」(逆近傍探索のための次元テスト)では、拡張一般化次元を用いた枝刈りに基づく、逆近傍探索問題のための効率的な手法を提案し、その性能解析を行っている。

最後に第９章「Implications and Future Work」(議論と今後の課題)では、本論文の成果を簡潔に纏めると共に、今後の研究課題を提示している。

以上を要するに、本論文は、データのもつ複雑さを内在的次元という軸でとらえ、これに基づく様々な探索アルゴリズムの設計と性能解析、また具体的なデータからの推定方法を与えることで、類似検索等のデータ処理の基礎技術の性能向上に寄与したものであり、数理情報学の発展に大きく貢献するものである。

よって本論文は博士（情報理工学）の学位請求論文として合格と認められる。