カテゴリカルデータと混合データを対象とする並列クラスタリングアルゴリズム

NGUYEN Thi Minh Hai
School of Information Science,
Japan Advanced Institute of Science and Technology
February 13, 2003

キーワード:クラスタリング、カテゴリカルデータ、混合データ、大規模データベース、並列アルゴリズム

1. はじめに

クラスタリングはイメージプロセシング、パターン認識、機械学習などのような多くの研究分野での基本で重要な技術である。クラスタリングは、分類されていないデータを、互いに似ているデータのかたまり（クラスタ）として分類することであり、ユーザのデータセットの取り扱いの単純化とデータ構造の発見を支援するために行われる処理である。クラスタとは、ある分類条件を満たすデータオブジェクトの集合である。データオブジェクトは、同一クラスタ内の他のオブジェクトとの類似性および他のクラスタに属すオブジェクトとの非類似性を条件として、適するクラスタに分類される。

実際的見地から、様々なタイプのデータベースにクラスタリング方法を適用することが可能である。一般に、データはカテゴリカルデータと混合データの2つのタイプに分類される。混合データは、数値とカテゴリカルな属性の両方を含んでいる。そのため、カテゴリカルデータのクラスタリングと混合データのクラスタリングを対象とする研究が多くなされてきた。しかし、カテゴリカルデータに対しては演算操作や順序の比較を行うことは出来ない。したがって、カテゴリカルデータと混合データのための良い評価規準を見つけることは困難でる。多くの研究はカテゴリカルデータを数値データに変換し、通常の類似性で評価しょうとしている。しかしこの変換によって細部の情報が失われ、誤ったクラスタを生成する可能性がある。例えば、カテゴリカルな属性をバイナリに変換して得られる二つの非常に異なったカテゴリカル値が、それら以外のある値と近い類似性を持つ可能性がある。

良く知られたクラスタリングアルゴリズムであるk-meansは、数値データベースへの適用を目的に作られているため、データマイニングで用いられるカテゴリカルデータに直接適用することが難しい。カテゴリカルデータと混合データを取り扱うために、k-meansを拡張したアルゴリズムがいくつか提案されている。しかしこれらのアルゴリズムは、局所的最適に陥ったり、初期設定として与えるグループに強く依存するという問題点がある。また、階層的なクラスタリングを行なった場合、これらのアルゴリズムのほとんどが長い処理時間

を必要とする。このため、大規模データへ適用することが難しい。近年になって、カテゴリカルデータと混合データのためのアルゴリズムとして、初期選択に影響されずにクラスタリングを行なうことができるk-setsが提案された。さらにk-setsは、k個のクラスタにk個のコアを作ることで、局所的最適に陥ることを免れている。しかしk-setsには、常に高い精度クラスタリングを行うわけではなく、大規模データベースに適用できないという問題がある。

近年、データベースのサイズは数ギガバイトの大きさとなっており、将来はさらに膨大になることが予想される。したがって、クラスタリングアルゴリズムには、複雑な構造を持つデータを扱うことに加えて、データサイズの増加に対応することが求められている。さらに、データベースサイズが非常に大きくなるとデータは単一のメモリに格納することが不可能になる。複雑で巨大なデータをクラスタリングするために、最近の高機能技術を用いることが有望である。近年提案されている並列アルゴリズムのほとんどは数値データの処理用であり、複雑なデータを取り扱う並列アルゴリズムが求められている。

まとめると、カテゴリカルデータと混合データに対するクラスタリングは、実際のデータベースの取り扱いと分析において非常に重要である。加えて、クラスタリングアルゴリズムに巨大なデータベースを扱う能力があることも同様に重要である。カテゴリカルな属性を持つ大規模なデータベースを対象とした、効果的で効率的なクラスタリングアルゴリズムが強く期待されている。

2. カテゴリカルデータと混合データを対象にした並列HACアルゴリズム

本論文では、カテゴリカルデータと混合データに対して良好なクラスタリングを行なえる高精度アルゴリズム(HAC)を提案した。また本論文では、非常に大規模なデータベースを扱うために、HACアルゴリズムのための2通りの並列アプローチを示した。

2.1. 新しいグループ化アプローチに基づくカテゴリカルデータと混合データのための高精度クラスタリングアルゴリズム

k-setsアルゴリズムは、他のアルゴリズムより優れているが、いかなる場合も高い精度でクラスタリングを行なえるわけではない。よって、本論文ではクラスタリング精度の向上のために、k-setsアルゴリズムの再パーティションステップを改良する方法を示す。オリジナルのk-setsアルゴリズムは、クラスタリング中に未分類のデータオブジェクトをクラスタに割り当てる（再パーティション）際に、N-ESCセットのオブジェクトに最も近いクラスタを選択する。ここでN-ESCセットは、k-setsの類似性の尺度でより近いオブジェクトから構成されている集合である。結果として再パーティション後に、互いに最も近いオブジェクトがそれぞれ異なったクラスタに分類されてしまう可能性がある。したがってクラスタリングの精度は低下する。この問題を克服するために同一のN-ESCの全てのオブジェクトを同じクラ

スタに割り当てる方法を提案する。これによってクラスタリング精度の向上が期待される。

2.2. 並列HACアルゴリズム

クラスタリングアルゴリズムは、複雑で巨大なデータを扱える必要がある。クラスタリングアルゴリズムの処理時間短縮に有効な手段の一つは並列化であり、大規模データベースを許容できる時間で取り扱うための方法としてHACアルゴリズムに並列化手法を適用することを考える。

HACアルゴリズムはデータベースをk個のグループにクラスタリングするアルゴリズムである。クラスタリングにはいくつかのパラメータを使用するが、高精度の結果を得るために、HACアルゴリズムは入力パラメータ全てから使用するパラメータを自動的に決定する。クラスタリングの実行には膨大な処理時間がかかるため、本研究ではHACアルゴリズムに並列化し、シングルレベル並列HAC、マルチレベル並列HACとした。前者は一つの入力パラメータについて並列実行を行なうのに対し、後者は2つ以上のパラメータについて並列実行を行なうため、混合データの処理でより効果的である

3. クラスタリング精度と処理時間の評価

HACアルゴリズムの有効性を示すために、HACアルゴリズムの精度と時間を評価した。比較のため、関連研究で用いられているものと同じデータベースを用いた。また、並列アルゴリズムの性能を評価する実験は、Cray T3E上で数種類のデータセットに対して評価を行なった。本研究では平均実行時間、速度向上、効率および並列実行での損失時間について詳細に議論している。

性能評価により、HACアルゴリズムは従来のアルゴリズムよりも常に良い精度でクラスタリングすることがわかった。また、提案した並列HACアルゴリズムは実行時間を大幅に削減することが可能で、120プロセッサを使った場合、90倍以上の速度向上が得られることが分かった。

4. 結論

この論文では、カテゴリカルデータと混合データを対象として、k-setsアルゴリズムの再パーティションステップの改良による新しいクラスタリングアルゴリズムを提案した。提案したHigh Accuracy Clustering (HAC) アルゴリズムは、新しいグループ化アプローチに基づいている。このアルゴリズムは、k個のクラスタをk個の最も大きいグループを選ぶことで決定する。このため、局所的最適解に陥ることを免れることができ、データの出現順序に影響を受けない。HACアルゴリズムはカテゴリカルデータと混合データに対して、既存のアルゴリズムとの同じデータセットを用いた比較で最も精度の高いクラスタリングを行なうしめした。さらに、非常に大規模なデータセット向けに2つの並列HACアルゴリズムを提案した。提案した並列アルゴリズムは、逐次アルゴリズムと比較して実行時間が大幅に短縮可能なこ

とから、スケーラブルな並列アルゴリズムであると考えられる。また実験では高い速度向上が得られており、効果的な並列アルゴリズムであると言える。提案した並列アルゴリズムは、複雑で大規模なデータを扱うことができると結論できる。今後の課題としては、HACアルゴリズムの有効性と効率性を向上させるのに適した評価基準の考案と、複雑なデータを対象としたk-meansアルゴリズムの論理に基づく、よりスケーラブルな並列アルゴリズムの開発が挙げられる。