平成29年度 情報工学コース卒業研究報告要旨

石川 研究室	氏　　名	志 村　薫
卒業研究題目	RDBの構造を考慮したデータベースからの学習手法に関する研究	

ビッグデータ時代の今日，機械学習によるデータからの新たな知見の発見や獲得が研究分野のみならず実社会においても盛んに取り沙汰されており，さまざまな機械学習用のツールが開発，公開されている．しかし，その多くはRDBMS（relational database management system）において一般的なデータ形式であるマルチテーブルでのデータ入力には対応しておらず，シングルテーブルでのデータ入力のみの対応となる．したがって，ユーザが機械学習において全特徴を得るためには，主キーと外部キーを用いてマルチテーブルを結合し，シングルテーブルへと変換する必要がある．この処理によって生じる遅延はデータサイズとともに増大し，ビッグデータを取り扱う上で無視できないものとなっている．

Kumarらはスタースキーマのデータ構造を想定し，この問題の解決策として特定の外部キーを，その外部キーが参照するディメンションテーブル内の全特徴の代表とみなすことでファクトテーブルとディメンションテーブルの結合を一部省略する手法を提案した．各ディメンションテーブルについて，結合省略基準の決定にはtuple ratioを利用する．ファクトテーブルの行数をn_S，i番目のディメンションテーブルの行数をn_{R_i}としたとき，i番目のtuple ratioは$TR_i \equiv \frac{n_S}{n_{R_i}}$と定義される．tuple ratioとは，テーブルの結合を省略することで生じるリスクを評価した指標であり，tuple ratioが小さいほどリスクが大きいことを表す．また，tuple ratioは$TR_i \equiv \frac{n_S}{n_{R_i}}$という定義からわかるように，データの中身というインスタンスには依存せず，テーブルの行数というメタデータに依存する．したがって，tuple ratioにより決定された結合省略基準はデータが更新され中身が書き換わった場合でも，同様なスキーマ構造のデータセットに対して，更新以前の結果を再利用することが可能である．

本論文ではKumarらのtuple ratioに基づく教師あり学習の高速化手法の有用性を検証する．まず，実験としてファクトテーブルと結合するディメンションテーブルの各組み合わせごとに機械学習を行い，予測精度および実行時間を計測し，それらとtuple ratioとの関係を調査する．この手法を活用することにより，精度を落とすことなくテーブルの結合を省略することが可能となり，学習時間の削減が望める．最後に，実験結果を踏まえデータベースと機械学習の連携について今後の展望を述べる．