4 UCI KDDアーカイブ：データマイニング研究と実験のための大規模データ集合のアーカイブ

翻訳：鈴木英之進（横浜国立大学工学部電子情報工学科）
suzuki@dnj.ynu.ac.jp

The UCI KDD Archive of Large Data Sets for Data Mining Research and Experimentation

Stephen D. Bay (Univ. of California, Irvine)
sbay@ics.uci.edu
Dennis Kibler (Univ. of California, Irvine)
kibler@ics.uci.edu
Michael J. Pazzani (Univ. of California, Irvine)
pazzani@ics.uci.edu
Padhraic Smyth (Univ. of California, Irvine)
smyth@ics.uci.edu

はじめに

データベース技術の商業的な成功と比較的安価な計測，保存，および計算ハードウェアが入手可能となったことにより，オンラインのデータ記録が過去20年間にわたって爆発的に増えている．これらの巨大データベースは，巨大データからの構造の探索，すなわちデータマイニングと知識発見が急速に発展する動機となっている．

科学と産業の諸分野でデータ収集活動の規模が拡大したのに対し，統計学と機械学習における伝統的なデータ解析研究はこの挑戦的な状況に立ち向かうのが比較的遅く，多くの研究と公表される論文はまだ比較的小さなデータ集合を対象としている．

巨大データ集合は，データ解析研究において長期的には当たり前の存在となることは明らかだが，この傾向は現在比較的ゆっくりと進んでいる．たとえば，近年の知識発見とデータマイニング（KDD）国際会議における大学研究者たちによる多くの論文は，数百例程度と大変小さなデータ集合を使って実験を行っている[7]．

巨大かつ高次元で複雑なデータ集合がデータマイニング研究において用いられる状況が早く実現するように，我々はNSF（National Science Foundation）が公募する情報とデータ管理プログラムの援助を受けて最大1,000MBにものぼる巨大データ集合のオンラインデータアーカイブを開発した．このアーカイブには高次元データ集合だけではなく時系列データ，空間データ，およびトランザクションデータなど，種々のデータ型を持つデータ集合が含まれる．

我々はこのデータアーカイブに3つの役割を期待している．第1にこのアーカイブは，計算機科学者，統計学者，エンジニア，および数学者を含むデータマイニング研究者たちが自分のデータ解析アルゴリズムを大規模データ集合に対応させるためのテストベッドの役割を果たすだろう．巨大データ集合の標準的なコレクションが利用できることにより，データマイニング研究の系統的な進歩が刺激され育成されると期待される．

第2に，アーカイブによりデータマイニングと知識発見アルゴリズムの評価方法が進歩するだろう．アーカイブは研究者たちが研究結果を再現することや各種手法を定量的に比較することを可能にする共通問題を提供するだけではなく，発見された知識を評価する共通環境を提供する．知識発見手法を開発するにあたって，発見された知識の評価が難しいことが問題となっている．発見されたパターンの質を明示的に計測することは不可能でないにしても困難であるため，我々は研究者たちがある手法で新しく興味深いパターンを発見したときに，新しい発見をデータについてすでに知られていることと比較できることを期待する．これは解析者が領域専門家ではなくデータをあまりよく知らない場合には特に重要である．

第3に，アーカイブによりデータマイニングと知識発見における探索的研究が可能となる．幅広い種々の応用

分野から挑戦しがいがある問題を集め，種々の関心を持つ研究者たちを集めて引き合わせることにより，データの中に隠された情報を発見する新しい手法の開発が促進されると考えられる．

KDDアーカイブは1999年6月にhttp://kdd.ics.uci.eduにおいてオンラインで公開され，現在幅広い種類のデータ型とタスクにわたる26個のデータ集合を含む．以下においては，まずアーカイブ構築の背景と現在の構成を説明する．次に研究分野に与えた最初の影響を議論し，単一の中心的なアーカイブを持つことにより起こり得る欠点に関して読者の注意を喚起する．最後に，アーカイブに関する将来の計画を議論する．

背景

KDDアーカイブは，1987年に構築され機械学習における実験的研究を育ててきた先輩格のUCI機械学習データレポジトリ[1]（http://www.ics.uci.edu/~mlearn）と同様に発展している．UCI機械学習データレポジトリは，工学，分子生物学，医学，金融および政治を含む幅広い分野における120個以上のデータベースを含む．

機械学習（ML）レポジトリは産業界と大学の両方の研究者たちに一般的に使われている．このレポジトリは人工知能研究において広く引用されており，いわゆる「UCIデータ集合」は新規既存を問わず学習アルゴリズムの経験的な評価において最も広く用いられているベンチマークである．Web上で入手できる800本以上の論文がこのレポジトリを引用している．なおこの数字は，自動引用インデックスシステムであるCiteSeer（http://citeseer.nj.nec.com）[5]から得た．

このレポジトリが作られる前は，機械学習の典型的な論文は新しいアルゴリズムを説明しそのアルゴリズムを1つの問題に適用するだけであった．複数のアルゴリズムを比較する論文や，提案するアルゴリズムを種々の問題に適用する論文は，ほとんどなかった．このような状況では，新規アルゴリズムやアルゴリズムの改良が本当に進歩に相当するかを評価することは難しかった．レポジトリにより，研究分野全体において各種アルゴリズムが最も適する問題のクラスが理解されるようになった．このレポジトリの重要な科学的貢献は，理論研究を補完し理論研究に貢献する経験的方法[4]を可能にしたことである．

レポジトリはデータ解析研究の発展に重要な役割を果たしたが，現実的なデータマイニング研究の有用なリソースとなるためには多くの限界がある．最初に，レポジトリにある多くのデータ集合は規模が小さすぎる．データベースに含まれるレコード数の中央値は1,000より小さく，属性数の中央値は15より小さい．第2に，データ集合は主に分類問題に焦点が当てられており目的が限られているうえ，レポジトリは画像，時系列，あるいは他の複雑なデータ型を含んでいない．

KDDアーカイブの設計

このアーカイブの目的は幅広い範囲のデータ型とタスク問題にわたる大規模データ集合を保存することである．本章では興味深いデータ型とタスク問題を簡単に議論し，各データ集合の説明文書についても述べる．

データ

一般的なレベルでは，データ集合はその規模と型によって分類される．規模はデータ集合に含まれる個別のオブジェクト数（N）（あるいはサンプル，レコード，個体，例数）と，各個別オブジェクトの次元数（d）（すなわち，各オブジェクトについて記録される測定項目，変数，フィーチャ，あるいは属性数）によって定まる．

我々の目的は，Nとdの関数として表される規模についての振る舞いに関する課題を既存のアルゴリズムに対して課すほど大規模であるが，インターネット経由でのダウンロードが現実的な時間内で不可能であるほどは大規模でないデータ集合を保存することである．したがって，各データ集合の規模はたかだか1,000MBとした．これは大雑把にいえば，各計測値を8バイトとしてN=500,000レコード×d=100次元のデータ集合を圧縮なしで記録できる規模に相当する．

データ「型」は，データ表現の基本構造を表す．機械学習と統計学においては伝統的に，「フラットなファイル」あるいは属性－値表現（Nとdが定まっておりベクトル空間と見なせる）が圧倒的に多い．この表現では各オブジェクトは同一の計測値集合で表される．たとえば人口・社会統計データにおいては，各個人を年齢，職業，および年収で記録できる．この形式のデータはしばしば，複数個の計測変数を持つので多変量，あるいは行を例，列を変数とすればデータがテーブルとして表せるのでテーブル形式と呼ばれる．もっとも実際の応用では次に示すものなど他のデータ型が数多く存在する．

画像データは，顔や指紋のコレクション，あるいは興味深い地域に注釈をつけた大規模画像などを表す．たとえばNASAのJPLは，アーカイブに金星のレーダー画像を提供したが，画像には惑星地理学者たちが火山の場所に注釈をつけていた．

関係データは，通常は複数のテーブルで表される互い

に関連し合うデータを含む．たとえば，アーカイブの映画データベースには，多くの映画についてタイトル，監督，および種類などの情報を記述したメインテーブルがある．メインテーブルは，テーブル中の各要素をより詳しく記述する他のテーブル（たとえば，映画のキャストや製作者たち）にリンクされている．

空間データは，2次元か3次元の格子点上に位置する観測物の集合を表す．たとえばアーカイブのエルニーニョデータベースには，太平洋の赤道付近一帯に設置された一連のブイで計測された海洋学と水面気象学で扱われる値が保存されている．

テキストデータは，Webページや新聞記事などを表す．たとえば，SyskillとWebertデータ集合には，4個の異なる分野（音楽バンド，生体臨床医学，ヤギ，およびヒツジ）に関するWebページが保存されている．

時系列データと配列データは，アーカイブの脳波データ集合などのように，連続した順序つきの観測に関するものである．時系列データは，株価や経済指標などのように連続値属性の値に関する変化を計測したものである．一方配列データは，DNAあるいはタンパク質配列，Webログにおけるファイル要求など名目属性に関する順序つきの値の集合を記録したものである．

トランザクションデータは，スーパーマーケットの記録や小売業における購買記録などである．

ヘテロジニアスデータは，複数個のデータ型を含む．たとえば毎日の海上温度を記録したエルニーニョデータ集合などのような空間データを時系列的に記録したものや，1970, 1980, 1990年の人口・社会統計値を含むIPUMS国勢調査データなどのように多変量データを一定間隔で記録したものである．

タスク

前節で列挙したデータ型は，幅広い解析タスクに用いることができる．ここでは現在アーカイブに含まれるデータ集合に基づき，我々が関心を持つタスクを具体例とともに紹介する．

分類学習では，名目属性を目的変数としその値を予測する．たとえば保険ベンチマークデータ集合は製品使用データと人口・社会統計学情報に基づいて，ある保険をかけることに興味を持つ顧客を予測するタスクに使用された．

回帰学習では，連続値属性を目的変数としその値を予測する．たとえばKDDカップ1998で用いられたデータは，ダイレクトマーケティングキャンペーンに対して個人が寄付する金額の予想に用いることができる，人口・社会統計学的データを含んでいる．

時系列と配列予測では，時系列データ（実数値）あるいは配列データ（名目値）において次に起こる値を予想する．

クラスタリングは，レコードについて意味があるグループを作る．たとえば新聞記事を，それぞれが共通の話題を持つグループにクラスタリングすることができる．

探索的データ解析／パターン発見は，グラフ学的手法や探索に基づく手法を用いてデータから属性間の関係など未知の構造を発見する．

異常値／異常検出は，データにおいて他とは異なるレコードやイベントを検出する．たとえばUNIXユーザデータは多くの人のコマンド履歴を保存しており，侵入の検出に使用された．この応用では，あるアカウントについて許可がないユーザが，そのアカウントの所有者とは異なるコマンドを（たぶん）使用するだろうと仮定している．

説明文書

データが最大限に有用となるように，各データ集合に説明文書を体系的につけた．アーカイブの各データ集合には，(a) データ自体を説明する説明ファイルと，(b) そのデータに関して行われた解析を説明するタスクファイルがついている．

説明ファイルでは，データの収集方法とデータの一般的な性質が説明されている．一般的な性質とは，用いられている計測属性の説明と，欠落値や検閲値などの存在や用いられた前処理方法などの関連情報である．説明データは，そのデータ集合を使った研究発表のリストと関連Webサイトへのリンクなどの引用情報を含んでいる．

タスクファイルには，データのクラスタリングなど特定の解析タスクに関する結果が載っている．このファイルには，用いられた手法，実験の設定条件，および結果に関する議論が説明されている．ここでも関連する文献とWebサイトへのリンクが列挙されている．

アーカイブのメンテナンスとして，我々は新しい情報を入手し次第，各データやタスクファイルを更新している．

結果

アーカイブは1999年6月に初めて公表されまだ新しいが，研究分野に重要な影響を及ぼし始めている．2000年10月現在，15,000人（固有のIPアドレスで測定）以上の人たちがアーカイブを訪問した．Webログファイルの解析による訪問者数の推定は不正確な場合があるが，この数字はアーカイブの影響を大雑把にだが示すことに注意していただきたい．サーチエンジンGoogle

(http://www.google.com) によれば，約180個のWebサイトがアーカイブにリンクを張っている．CiteSeer (http://citeseer.nj.nec.com) [5] によれば，1999年には10件の引用があり，これは出版までに時間がかかることを考えればかなり多いと思われる．

アーカイブのデータ集合を使用した論文には，興味深いものが数本ある．たとえば，Fan, Lee, Stolfo, およびMiller (2000) は，実時間のネットワーク侵入検出システムにおいて，操作コスト（システムを実行するコスト）を軽減する問題に取り組んだ．彼らは，処理を加えたtcdumpファイルに基づきプロービング，サービス不能(DoS)，不正なローカルアクセス，および不正なルートアクセスなどの侵入を含むKDDカップ1999のデータ集合を使用した．彼らは侵入を検出するために，それぞれ異なるコスト範囲の属性を用いる複数のルール集合を構築し，単一モデルを用いた方法に比較して操作コストを97%削減した．

KeoghとPazzani (1999) は，オーストラリア手話の発話に関する手の動き（X, Y, Z座標の位置；ロール，ピッチ，ヨーの角度；および手の傾き）についての時系列データであるAustralian Sign Languageデータ集合を使用し，ダイナミックタイムワーピング (DTW) をセグメントに分割する手法を構築した．ダイナミックタイムワーピングは，時系列データのマッチングにおいて局所的なゆがみを考慮するために時系列データを局所的に伸ばしたり縮めたりするために用いられ，データ間の距離をより正しく頑健に評価することを可能とする．彼らが提案するDTWのセグメント分割手法を用いることにより，ユークリッド距離を用いる場合とほとんど同じくらい早く，しかもダイナミックタイムワーピングを用いる場合に匹敵するほど正確にデータをクラスタリングできた．この結果は約20倍の速度向上に相当する．

Pavlov, Mannila, およびSmyth (2000) はマイクロソフトWebデータを使い，2値トランザクションデータに関する近似問合せへの応答に関する種々の確率モデルと確率アルゴリズムを開発・テストした．彼らの結果は，モデルの複雑さ，問合せ近似の正確さ，および問合せに応答する時間に関する一般的なトレードオフを明らかにした．

表-1に，アーカイブのデータ集合を簡単な説明と公表以来のWebページアクセス数とともに示す．アーカイブに長く存在するデータ集合はアクセス数が多いことに注意されたい．表におけるヒット数は，該当するデータ集合のメインのWebページへのリクエスト数から，UCIドメインからのリクエストと同一IPアドレスからの複数回のリクエストを除いて推定した．

考察と教訓

KDDアーカイブはまだ新しく研究分野に影響を与え始めたばかりである．現在の状況は，機械学習レポジトリが最初に構築されたときの機械学習分野における状態と似ている．前例があるため，機械学習レポジトリで得た経験を熟考してKDDアーカイブに関する教訓をまとめることが可能である．

最初に，機械学習レポジトリは研究分野をより実験重視にし分析をより綿密にする手助けとなることで，機械学習コミュニティに影響を及ぼした．我々は，KDDアーカイブがこの影響を同様に及ぼすと予想している．

しかし機械学習レポジトリは，データベースの標準集合が存在することが有害となる場合があることを示した．Salzbergは，このような問題の多くを詳しく論じている[8]．標準のベンチマークがあることにより，あるアルゴリズムの性能が良いか悪いかの理由を理解しようとするのではなく，定量的な比較を過度に強調し「どれ」が良いアルゴリズムかを決める順位付けを奨励してしまう恐れがある．さらに過去の最良結果に勝つことを目的として，研究分野が既存のアルゴリズムに対する漸増的で些細な改良に重点を置いてしまう場合がある．最後に，アルゴリズムの開発と評価のサイクルを繰り返すことにより，アルゴリズムが有名なデータ集合に適合し他の領域ではうまくいかなくなってしまうという過学習の危険性がある．

これらのすべての問題点はKDDアーカイブが陥る可能性がある落し穴である．研究者とユーザはデータを使用するにあたって注意し，これらの問題点を心に留めておくべきである．

結論と将来の計画

UCI機械学習レポジトリは実験における評価の質と綿密さを向上することにより，機械学習における研究の進め方に革命をもたらした．我々はKDDアーカイブがデータマイニングと知識発見に同じようなインパクトを与えることを望んでおり，KDDアーカイブがこの研究分野にとって重要な財産であると信じている．

今後，主に3つの計画を予定している．最初に，なるべく早くアーカイブを拡大する予定である．KDDにおける大規模データベースへの強い興味は他の研究分野にも広がり，多くの国際会議がそれぞれ独自のデータマイニングトラックを開くようになった．我々はこの関心を利用し，多くの研究者たちが見たことがないと思われる新しいデータ集合と新しい問題を集めたいと考え

名　称	内　容	規模	ヒット数
画像			
CMU Faces	顔画像	33 MB	965
Volcanoes	火山に注釈をつけた金星の画像	187 MB	892
多変量			
Census-Income	年収と人口・社会統計学のデータ	156 MB	783
COIL 1999	河川での化学物質の濃度と藻類の密度	< 1 MB	802
Corel Features	画像データベースから抽出したフィーチャ	57 MB	1,074
Forest Covertype	森林での30m×30m単位のセルごとの植生	75 MB	2,004
IPUMS	1970，1980，1990におけるロスアンゼルスとロングビーチでの人口・社会統計学データ	45 MB	598
Insurance Benchmark	顧客情報	2 MB	75
Internet Usage	インターネットユーザの人口・社会統計学データ	2 MB	1,203
KDD CUP 1998	寄付金額と人口・社会統計学データ	136 MB	2,649
KDD CUP 1999	ネットワークへの侵入データ	743 MB	1,620
配列			
Entree Chicago	ユーザとレストラン推薦システムのやりとり	3 MB	571
Microsoft Web Data	ユーザが訪問するWebサイトの分野	2 MB	2,684
UNIX User Data	UNIXのコマンド履歴	1 MB	1,335
時空間			
El Nino	海洋学と海上の気象学におけるデータ	23 MB	1,183
関係			
Movies	映画情報	7 MB	1,575
テキスト			
20 Newsgroups	Usenetのニュースへの投稿	61 MB	1,238
Reuters 21578	ロイターのニュース記事	28 MB	1,144
Syskill & Webert	4つの話題領域に関するWebページ	2 MB	1,596
時系列			
Australian Sign Language	オーストラリア手話についての手の動き	59 MB	1,278
EEG Data	脳波計の計測結果（64電極）	~3 GB	1,617
Japanese Vowels	男性話者の母音発音	1 MB	315
Pioneer	ロボットの環境についてのセンサ計測値	< 1 MB	826
Robot Failure	フォースとトルクの計測値	< 1 MB	654
Synthetic Control	プロセス制御のチャート	< 1 MB	1,367
Synthetic TS	時系列的なインデクシング性能評価	16 MB	635

表-1　アーカイブのデータベース

ている．我々は，この論文の読者が自分の大規模で複雑なデータ集合をアーカイブに提供することを考慮し，他の研究者たちに同様に行うように勧めることを強く望む．

次に，我々はデータマイニングにおける新しい規格を利用したいと考えている．たとえばデータ文書化イニシアティブ（http://www.icpsr.umich.edu/DDI/）は，メタデータすなわちデータ自身に関するデータや情報に関する基準と規格を確立しようとしている．標準形式が定まることにより，他分野に存在すると思われる多くのユーザたちが大いに得すると考えられる．データマイニンググループ（http://www.dmg.org）は予測モデルを共有するための，予測モデルマークアップ言語と呼ばれるXMLの規格を開発している．アーカイブの対象は単なる予測モデルよりも幅広いが，この言語は予測モデルを用いる研究者たちにとってデータだけではなく他の研究者たちのモデルも利用できるようになるので大変有用となると予想される．

最後に，我々は最終的にはKDDアーカイブとより有名な機械学習レポジトリを統合する予定である．現在は，新しいデータ集合とそれらに関連するタスクを強調するために，これらを区別している．

謝辞　KDDアーカイブと機械学習レポジトリにデータを提供していただいた方々に感謝したい．新しいアーカイブは機械学習レポジトリの成功なしには構築できなかった．したがって機械学習レポジトリの創設者と管理者たち，すなわちPatrick Murphy, David Aha, Chris Merz, Catherine Blake, そしてEamonn Keoghに特に感謝したい．本研究は一部NSF補助金IIS-9813584の援助を受けている．

参考文献

1) Blake, C. and Merz, C. J.: UCI Repository of Machine Learning Databases, http://www.ics.uci.edu/~mlearn/MLRepository.html, University of California, Department of Information and Computer Science, Irvine, Calif. (1998).
2) Fan, W., Lee, W., Stolfo, S. and Miller, M.: A Multiple Model Cost-Sensitive Approach for Intrusion Detection, Proc. Eleventh European Conf. on Machine Learning (ECML), pp.142-153 (2000).
3) Keogh, E. and Pazzani, M. J.: Scaling up Dynamic Time Warping to Massive Datasets, Proc. Third European Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD), pp.1-11 (1999).
4) Langley, P. (ed.): Machine Learning as an Experimental Science, Machine Learning, 3 (1), pp. 5-8 (1988).
5) Lawrence, S., Giles, C. L. and Bollacker, K.: Digital Libraries and Autonomous Citation Indexing, IEEE Computer, 32 (6), pp.67-71 (1999).
6) Pavlov, D., Mannila, H. and Smyth, P.: Probabilistic Models for Query Approximation with Large Sparse Binary Data Sets, Proc. Sixteenth Conf. on Uncertainty in Artificial Intelligence (UAI) (2000).
7) Ramakrishan, R. and Stolfo, S. (eds.): Proc. Sixth ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (KDD) (2000).
8) Salzberg, S.: On Comparing Classifiers: Pitfalls to Avoid and a Recommended Approach, Data Mining and Knowledge Discovery, 1 (3), pp. 317-328 (1997).

（平成13年3月31日受付）