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

Stephen D. Bay, Dennis Kibler, Michael J. Pazzani, and Padhraic Smyth
Department of Information and Computer Science
University of California, Irvine
Irvine, CA 92697
{sbay, kibler, pazzani, smyth}@ics.uci.edu

ABSTRACT

Advances in data collection and storage have allowed organizations to create massive, complex and heterogeneous databases, which have stymied traditional methods of data analysis. This has led to the development of new analytical tools that often combine techniques from a variety of fields such as statistics, computer science, and mathematics to extract meaningful knowledge from the data. To support research in this area, UC Irvine has created the UCI Knowledge Discovery in Databases (KDD) Archive (http://kdd.ics.uci.edu) which is a new online archive of large and complex data sets that encompasses a wide variety of data types, analysis tasks, and application areas. This article describes the objectives and philosophy of the UCI KDD Archive. We draw parallels with the development of the UCI Machine Learning Repository and its affect on the Machine Learning community.

Keywords

data mining, data archive

1. INTRODUCTION

The commercial success of database technology and the availability of relatively inexpensive sensing, storage, and processing hardware has led to explosive growth in online data storage over the last two decades. In turn, these large databases have motivated the rapid development of data mining and knowledge discovery, namely, the search for structure in large volumes of data.

While science and industry have scaled up their data gathering activities, traditional data analysis research in statistics and machine learning has been relatively slow to take up the challenge and much research and published work is still focused on relatively small data sets.

While it is clear that in the long run large data sets will eventually become commonplace in data-analytic research settings, this is currently occurring rather slowly. For example, many papers by academic researchers in the recent conferences on Knowledge Discovery and Data Mining experiment on data sets that are as small as a few hundred examples [7].

To accelerate the infusion of large, high-dimensional, and complex data sets into the data mining research environment we have developed (under the sponsorship of the NSF Information and Data Management program) an online data archive of large data sets (up to 1000 megabytes). The archive includes high-dimensional data sets as well as data sets of varying data types such as time series, spatial data, transaction data, and so forth.

We envision three roles for this archive. First, the archive will serve as a testbed to enable researchers in data mining, including computer scientists, statisticians, engineers, and mathematicians, to scale data analysis algorithms to very large data sets. We believe that the availability of a standard set of large data sets will directly stimulate and foster systematic progress in data mining research.

Second, the archive can lead to improvements in the evaluation of data mining and knowledge discovery algorithms. In addition to providing a common set of problems that allows researchers to replicate study results and to make quantitative comparisons between methods, it also provides a common environment for evaluating discovered knowledge. A difficult aspect of developing knowledge discovery techniques is evaluating the discovered knowledge. Because explicit measures of the quality of a discovered pattern are difficult if not impossible to define, we hope that when researchers find new and interesting patterns with a given technique, they can compare the discovered findings with what is already known about the data. This is especially important when the analyst is not a domain specialist and is not intimately familiar with the data.

Finally, the archive will enable exploratory research in data mining and knowledge discovery. By bringing together challenging problems from widely different application areas and researchers with varying interests, it will promote the development of new techniques to discover hidden information in the data.

The KDD archive opened publicly in June 1999 and is online at http://kdd.ics.uci.edu. We currently have 26 data sets, spanning a wide variety of data types and tasks. In the remainder of this article, we will describe the background under which the archive was developed and its current or-

*Reprinted with permission from Information Processing Society of Japan Magazine, 2001, volume 42, number 5.

ganization. We will then discuss its initial impact on the field and caution readers about the potential drawbacks of a central archive. Finally, we discuss future plans for the archive.

2. BACKGROUND

The KDD Archive follows in the footsteps of the existing UCI Machine Learning Data Repository [1] which was created in 1987 to foster experimental research in machine learning (http://www.ics.uci.edu/~mlearn/MLRepository.html). It contains over 120 databases from a broad range of problem areas including engineering, molecular biology, medicine, finance and politics.

The Machine Learning (ML) Repository is commonly used by industrial and academic researchers. It is widely cited in the artificial intelligence literature and the "UCI data sets" are the most widely used benchmark for empirical evaluation of new and existing learning algorithms. Over 800 papers available on the web cite the repository. [1]

Prior to the creation of this repository, a typical machine learning paper described a new algorithm and the application of that algorithm to a single problem. Few papers included comparisons of multiple algorithms or the use of an algorithm on a variety of problems. Under such circumstances, it was difficult to evaluate whether a new algorithm or an enhancement to an algorithm was an improvement. The repository has permitted the research community as whole to gain understanding into the classes of problems for which each type of algorithm is most appropriate. An important scientific contribution of the repository has been the enabling of experimental methods [4] which complement, and contribute to, theoretical research.

Although the repository has played an important role in advancing research into data analysis, it has a number of limitations in serving as a useful resource for realistic data mining research. First, the data sets it contains are generally too small. The median number of examples in a database is less than 1000 and the median number of fields is less than 15. Second, the data sets are limited in scope, focusing mainly on classification, and the repository does not contain image, time series or other complex types of data.

3. KDD ARCHIVE DESIGN

The goal for this archive is to store large data sets that span a wide variety of data types and problem tasks. In this section, we briefly discuss the types of data and the problem tasks of interest. We also describe how each data set is documented.

3.1 Data

On a general level, we can characterize data sets by their size and type. Size can typically be measured by *the number* (N) *of individual objects* (or samples, individuals, or examples) contained in the data set, and *the dimensionality* (d) *of each individual object* (i.e., the number of measurements, variables, features, or attributes recorded for each object).

Our goal is to store data sets that are large enough to provide challenges to existing algorithms in terms of scaling behavior as a function of N and d, yet that are not so large as to make downloading via the Internet in reasonable time impossible. Thus, we target individual data sets up to 1000 Megabytes in size, which roughly allows for the storage of an $N = 500,000$ measurements $\times$ $d = 100$ dimensional data set with 8 bytes per measurement, and no compression.

Data "type" refers to the underlying structure of the data representation. Traditionally, in machine learning and statistics, the "flat file" or attribute-value representation has dominated (fixed N, fixed d, and can be viewed as a vector-space). This representation consists of an identical set of measurements for each object. For example, in a demographic data set we may record for each person their age, occupation, and salary. This type of data is often referred to as *multivariate* since we have multiple measurement variables, or *tabular* since the data can be represented as a table with rows representing examples and columns representing variables. In practical applications, however, there may exist many other commonly occurring data types including the following.

- **Image data** such as face or fingerprint collections, or large images containing annotations marking regions of interest. For example, the Jet Propulsion Laboratory at NASA donated radar images of Venus to the archive, where the images are annotated with the locations of volcanoes by planetary geologists.
- **Relational data** which consists of interrelated data usually represented by multiple tables. For example, the Movies Database in the archive contains a main table with information on the title, director, category, etc of many movies. It is linked to other tables which describe elements of the main table in more detail (i.e., the casts and people involved with making the movie).
- **Spatial data** which represents a set of observations located on a 2 or 3 dimensional grid. For example, the El Nino Dataset in the archive contains oceanographic and surface meteorological readings taken from a series of buoys positioned throughout the equatorial Pacific.
- **Text Data** such as webpages or newspaper articles. For example, the Syskill and Webert data set contains webpages dealing with four different topic areas (music bands, bio-medical, goats, and sheep).
- **Time series and Sequence** data which consists of a consecutively ordered set of observations, such as the EEG data set in the archive. Time series measure changes in the value of a continuous variable such as stock prices or economic indicators whereas sequence data records an ordered set of categorical variables such as DNA or protein sequences, or file requests in a weblog.
- **Transaction data** such as records of supermarket or retail purchases.
- **Heterogeneous data** refers to a mixture of several data types. Examples include, spatial data that is collected

[1]This figure is derived from CiteSeer [5] an autonomous citation index (http://citeseer.nj.nec.com).

over time, such as daily sea-surface temperature measurements as in the El Nino data set, or multivariate data collected at regular intervals such as the IPUMS Census data which contains demographic samples from 1970, 1980, and 1990.

3.2 Tasks

The data types listed above can each be used for a wide range of analysis tasks. We present here some of the tasks that we are interested in along with examples based on data sets currently in the archive.

Classification: predict the value of a categorical target variable. For example, the Insurance Benchmark data set was used to predict which customers were interested in buying an insurance policy based on product usage data and demographic information.

Regression: predict the value of a continuous target variable. For example, the data from the 1998 KDD CUP contains demographic data which can be used to predict the amount of money a person will donate in response to a direct marketing campaign.

Time Series and Sequence Prediction: predict the next value that will occur in the time series (real values) or sequence (categorical).

Clustering: develop a grouping for the examples that is meaningful. For example, we can cluster newspaper articles into groups, each with a common topic.

Exploratory Data Analysis/Pattern Discovery: find unknown structure in the data such as relationships between variables using graphical or search based techniques.

Deviation or anomaly detection: detect unusual examples or events in the data. For example, the UNIX user data contains command histories of different people and was used to investigate intrusion detection: an unauthorized user of an account would (hopefully) use commands differently from the account owner.

3.3 Documentation

To maximize the usefulness of the data, we try to ensure that each data set is well documented in a structured manner. For each data set the archive contains both (a) a description file that describes the data itself and (b) task files that describe analyses performed on the data.

The description file explains how the data was collected and describes its general characteristics including an explanation of the measurement variables used and other relevant information such as the presence of missing or censored values, and preprocessing steps taken. It also contains references to further information such as a list of publications which use the data set and links to related web sites.

The task files show the results for a particular type of analysis, such as clustering the data. It describes the approach used, the experimental setup and discusses the results. Again, there is a listing of related publications and web sites.

As part of ongoing maintenance of the archive, we update each data or task file with new information as it becomes available.

4. RESULTS

Although the archive is quite young (it was first publicly announced in June 1999), it has begun having a measurable impact on the field. As of October 2000, over 15000 people (as measured by unique IP addresses) have visited the archive. Note that analyzing web log files to estimate the number of visitors can be tricky, but should give a rough estimate of its impact. The search engine Google (http://www.google.com) reports that approximately 180 websites link to the archive. CiteSeer (http://citeseer.nj.nec.com) [5] reports about ten citations for 1999 which we feel is significant given publishing delays.

There have been several interesting papers that have used data sets in the archive. For example, Fan, Lee, Stolfo, and Miller [2] worked on the problem of reducing operational costs (the cost of running the system) for a real-time network intrusion detection system. They used the KDD CUP 1999 data set which is based on processed tcpdump files and includes intrusions such as probing, denial of service, illegal local access, and illegal root access. They developed multiple rule sets to identify intrusions, with each set using features from different cost ranges to reduce the operational cost by 97% compared with a single model approach.

Keogh and Pazzani [3] used the Auslan data set, which consists of time series that represent hand motions (X,Y,Z positions, roll, pitch, yaw, and finger bends) for utterances in Australian Sign Language, to develop a segmented approach to dynamic time warping (DTW). Dynamic time warping is used to locally stretch or shrink time series to account for distortions when matching time series and it generally allows for improved and more robust distance measures. Using their segmented approach to DTW allowed them to cluster the data almost as fast as using a Euclidean distance metric but achieved performance on par with dynamic time warping (a speed up factor of approximately 20).

Pavlov, Mannila, and Smyth [6] used the Microsoft Web data to develop and test a variety of probabilistic models and algorithms for approximate query answering on binary transaction data. Their results illustrated general trade-offs between model complexity, accuracy of the query approximation, and time taken to answer a query.

Table 4 shows the data sets in the archive along with a brief description and the number of webpage accesses since data was donated. Note that data sets that have been in the archive longer will have higher access counts. The number of hits was estimated by using the number of requests for the main web page for the given data set with requests from the UCI domain as well as multiple requests from identical IP addresses removed.

5. REFLECTIONS AND LEARNED LESSONS

The KDD Archive is still new and is just beginning to have an impact on the field. Its position is similar to the state

Table 1: Archive Databases

Name	Description	Size	Hits
Image			
CMU Faces	face images	33 MB	965
Volcanoes	images of Venus annotated with volcanoes	187 MB	892
Multivariate			
Census-Income	demographic and salary data	156 MB	783
COIL 1999	river chemical concentrations and algae densities	<1 MB	802
Corel Features	features extracted from image database	57 MB	1074
Forest Covertype	type of forest cover for 30m x 30m cells	75 MB	2004
IPUMS	demographic data for Los Angeles/Long Beach in 1970, 1980, and 1990	45 MB	598
Insurance Benchmark	customer information	2 MB	75
Internet Usage	demographic data on internet users	2 MB	1203
KDD CUP 1998	demographic data and donation amount	136 MB	2649
KDD CUP 1999	network intrusion data	743 MB	1620
Sequence			
Entree Chicago	user interactions with restaurant recommendation system	3 MB	571
Microsoft Web Data	areas of web site that user visits	2 MB	2684
UNIX User Data	Unix command histories	1 MB	1335
Spatio-Temporal			
El Nino	oceanographic and surface meteorological readings	23 MB	1183
Relational			
Movies	information on movies	7 MB	1575
Text			
20 Newsgroups	Usenet news postings	61 MB	1238
Reuters 21578	Reuters news articles	28 MB	1144
Syskill & Webert	webpages on 4 topic areas	2 MB	1596
Time Series			
Auslan Data	hand motions for Australian sign language	59 MB	1278
EEG Data	electroencephalogram measurements (64 electrodes)	$\sim$3 GB	1617
Japanese Vowels	vowels spoken by male speakers	1 MB	315
Pioneer	robot sensor readings during environmental interaction	<1 MB	826
Robot Failure	force and torque measurements	<1 MB	654
Synthetic Control	process control charts	<1 MB	1367
Synthetic TS	time series for indexing performance testing	16 MB	635

of the Machine Learning Repository in its community when it was first developed. Because we have the advantage of hindsight, we can reflect on our experience with the ML Repository for lessons relevant to the KDD Archive.

First, the ML Repository affected machine learning significantly by helping to make the field more experimental and the analyses more thorough. We anticipate that the KDD archive will have this affect as well.

However, the ML Repository also showed that having a standard set of databases can be harmful. Salzberg [8] discusses in depth many of these problems. Having a standard benchmark can overly emphasize quantitative comparisons and promote the ranking of algorithms to determine which are "better" as opposed to trying to understand why a particular algorithm worked well or poorly. It can also focus the field on incremental and minor improvements to existing algorithms in an effort to beat the past best result. Finally, with a repeated cycle of algorithm development followed by testing, there is a danger of overfitting as algorithms become tailored to the popular data sets but no longer perform well on other domains.

All of these issues are potential pitfalls for the KDD Archive. Researchers and users should be careful in their use of the data and remain aware of these issues.

6. CONCLUSIONS AND FUTURE PLANS

The UCI ML repository revolutionized the way research is conducted in machine learning by improving the quality and thoroughness of experimental evaluation. Our hope is that the KDD archive will have a similar impact on data mining and knowledge discovery, and we believe that it represents an important resource for the field.

For the future, we envision three major thrusts. First, we will continue to expand the archive as rapidly as possible. The enthusiasm for large databases in KDD has spread over to other fields and many conferences have started their own data mining track. We want to capitalize on this interest and recruit new data sets and new problems that many researchers may not have been exposed to. We strongly encourage readers of this article to consider donating their large complex data sets to the archive and to encourage other researchers to do likewise.

Second, we want to take advantage of emerging standards in data mining. For example, the Data Documentation Initiative [2] is attempting to establish criteria and standards for meta data, which is data or information about the data itself. A standard format would greatly benefit users who may be in many different fields. The Data Mining Group [3] is developing an XML standard for sharing predictive models called the Predictive Model Markup Language. Although the scope of the archive is broader than just predictive models, this would be very beneficial to researchers working on predictive algorithms as they could have access not only to the data, but also to other researcher's models.

Finally, we eventually plan to merge the KDD archive with the more well known UCI Machine Learning repository. For the present, we are keeping them separate to emphasize the new data sets and associated tasks.

[2] http://www.icpsr.umich.edu/DDI/

[3] http://www.dmg.org

Acknowledgments

We would like to thank all of the donors who contributed data both to the KDD Archive and the Machine Learning Repository. The new archive would not have been possible without the success of the ML Repository and thus its creators and librarians deserve special thanks: Patrick Murphy, David Aha, Chris Merz, Catherine Blake, and Eamonn Keogh. This work was funded in part by the National Science Foundation Grant IIS-9813584.

7. REFERENCES

[1] C. Blake and C. J. Merz. UCI repository of machine learning databases. Irvine, CA: University of California, Department of Information and Computer Science. [http://www.ics.uci.edu/~mlearn/MLRepository.html]., 1998.

[2] W. Fan, W. Lee, S. Stolfo, and M. Miller. A multiple model cost-sensitive approach for intrusion detection. In *Proceedings of the Eleventh European Conference on Machine Learning*, 2000.

[3] E. Koegh and M. J. Pazzani. Scaling up dynamic time warping to massive datasets. In *Proceedings of the 3rd European Conference on Principles and Practice of Knowledge Discovery in Databases*, 1999.

[4] P. Langley. Editorial: Machine learning as an experimental science. *Machine Learning*, 3(1):5–8, 1988.

[5] S. Lawrence, C. L. Giles, and K. Bollacker. Digital libraries and autonomous citation indexing. *IEEE Computer*, 32(6):67–71, 1999.

[6] D. Pavlov, H. Mannila, and P. Smyth. Probabilistic models for query approximation with large sparse binary data sets. In *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*, 2000.

[7] R. Ramakrishan and S. Stolfo, editors. *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. The Association for Computing Machinery, 2000.

[8] S. Salzberg. On comparing classifiers: A critique of current research and methods. *Data Mining and Knowledge Discovery*, 1:1–12, 1999.