Dataset cataloging metadata for machine learning applications and research

Sally Jo Cunningham
Department of Computer Science
University of Waikato
Hamilton, New Zealand
email: sallyjo@cs.waikato.ac.nz

Abstract: As the field of machine learning (ML) matures, two types of data archives are developing: collections of benchmark data sets used to test the performance of new algorithms, and data stores to which machine learning/data mining algorithms are applied to create scientific or commercial applications. At present, the catalogs of these archives are ad hoc and not tailored to machine learning analysis. This paper considers the cataloging metadata required to support these two types of repositories, and discusses the organizational support necessary for archive catalog maintenance.

1. Introduction

Most successful machine learning applications reported in the literature are based on considerable input from both a domain expert and a machine learning practitioner: the domain expert providing an understanding of the semantic structure and meaning of the data, and the machine learning expert using this information to guide the transformation of the raw data to a form suitable for ML algorithms to process. Unfortunately, the domain expert's dataset description is rarely formally recorded, and even more rarely stored with the data if it is archived. This loss of descriptive metadata can negatively impact future research in a number of ways: replication of results is more difficult; results often cannot be thoroughly reviewed; the data cannot be effectively re-analyzed by other researchers using different techniques; and it is more difficult to re-use all or part of the data in combination with new datasets to explore different probems.

Similar issues in acquiring and recording dataset description exist for datasets archived as benchmarks for fundamental ML research, in which the emphasis is often on comparing the performance of algorithms. For this kind of dataset usage, a different type of metadata is useful: rather than detailed descriptions of the ways in which the data attributes model the domain, users generally require extensive documentation of the information that has previously been mined from the data and the level of accuracy achieved by prior ML experiments. Again, this metadata is often difficult to find, or indeed is not formally maintained.

This paper provides an overview of dataset metadata to support both fundamental and applied ML research. Section 2 describes static metadata for ML applications, based on the content of the dataset as well as its bibliographic description; Section 3 discusses dynamic metadata, based on dataset usage; Section 4 discusses ways that metadata can be used to support ML research; and Section 4 provides an overview of organizational supports for metadata maintenance that have been suggested in the literature.

2. Types of metadata for machine learning

Sheth and Vashyap present a classification of metadata to describe structured databases for text or image retrieval [Sheth and Kashyap, 1996]. The following set of metadata types applies their

taxonomy to the requirements of machine learning, and extends this scheme with additional items useful for ML:

content-independent:

Content-independent metadata corresponds to "physical" cataloging descriptors in bibliographic catalogs. This type of information does not capture the content of the data, but instead characterizes its form, location, and any unique identifiers. In a conventional library catalog, for example, a bibliographic entry would include a tag identifying a document's type (that is, journal article, book, dissertation, etc.), its location on the library shelves (usually denoted by a call number derived in part from a standardized classification scheme such as the Library of Congress scheme), and unique identifiers such as an ISBN and document title.

Examples of content-independent attributes of a machine learning data set include file name, location, file size, date of creation, data set "author", etc. These attributes directly correspond to the traditional bibliographic descriptors; indeed, the AACR2 (*Anglo-American Cataloger's Rules*, the cataloging "Bible") has included a chapter on software and data files since 1984 [ALA 1984, chapter 9]. Data sets held in libraries or by government institutions tend to provide the most complete and standardized metadata of this type, while less formally managed repositories such as the UCI repository for machine learning datasets [Merz and Murphy, 1996] offer varying levels of description for datasets. At UCI, for example, the documentation records range from the highly specific (for example, the abalone dataset record in Figure 1a) to the nearly non-existent (the "undocumented databases" directory at UCI, as typified by the economic sanctions data description in Figure 1b). This variability is not surprising, given that the UCI administrators are volunteers and that the majority of the cataloging is provided by dataset donors. Formal cataloging is no trivial task, and requires training and a significant amount of effort to achieve consistency.

Why is this type of cataloging important for ML research and applications? As will be discussed below, standard ways for referencing a data set would provide significant support for efforts to provide authorship credit to data set providers—and this credit is argued to be an important tool in supporting greater standardization and availability of data.

Of more immediate concern is the greater possibility for data re-use that more substantive cataloging could enable, by allowing potential users to merge useful portions of existing data with their own (for example, to incorporate historic temperature and rainfall data into a dataset of crop growth measurements). Computational chemistry provides an extreme example of this process, since workers in this field rarely create raw data from physical experiments. Instead, research is based on existing physical data or generated data and models; experiments often require the location of small bits of data from many files, all the pieces of which must be merged into new files or processed for use as parameter settings for calculations [Keller and Jones, 1996]. Given the amount of effort required to construct useful datasets, parsimony demands that the data be used to the greatest extent possible—and this re-use requires effective cataloging. Content-descriptive metadata, described below, is also essential in fully exploiting data.

1. Title of Database: Abalone data
2. Sources:
(a) Original owners of database:
Marine Resources Division
Marine Research Laboratories - Taroona
Department of Primary Industry and Fisheries, Tasmania
GPO Box 619F, Hobart, Tasmania 7001, Australia
(contact: Warwick Nash +61 02 277277, wnash@dpi.tas.gov.au)
(b) Donor of database:
Sam Waugh (Sam.Waugh@cs.utas.edu.au)
Department of Computer Science, University of Tasmania
GPO Box 252C, Hobart, Tasmania 7001, Australia
(c) Date received: December 1995

(a) portion of abalone data description

I think you'll find some limited documentation on Mike's database in his papers. His dissertation would be a good reference (UCLA). Perhaps

pages 152-153 in the EWSL-1988 proceedings should help with understanding
the data format. Pages 713-718 of IJCAI-1989 should help even more.

Date: Wed, 27 Sep 89 16:03:50 -0700
From: Michael Pazzani <pazzani@ICS.UCI.EDU>

;;; -*- Mode: LISP; Syntax: common-lisp; Base: 10 -*-
;;;totally undocumented sanctions database.
;;;Pazzani AAAI-86 or EWSL-88

(b) economic sanctions dataset header

Figure 1. Content independent catalog samples from the UCI ML dataset repository

content-dependent:

In comparison to content-independent metadata, content-dependent descriptors are directly related to or derived from the contents of a file. The simplest type of content-dependent description is the data file itself, in the form of attribute-value pairs representing each line of data. More usefully, derived descriptors provide information about each attribute over the entire dataset: distributions of values, average value, data type (real, integer, symbolic, etc.), percentage of missing values, range of values, etc. Some types of context-dependent metadata are used directly by the ML schemes (such as data type). Most, however, are currently used by ML experts to select attributes for inclusion in a ML experimental run. Data repositories (such as UCI) tend to provide good support for this type of cataloging, although the catalog records may not be in a rigorously enforced or standardized format.

content-descriptive

The least-supported type of metadata is the content-descriptive: characterizations of columns in a file that cannot be directly derived from the data set contents. In a previous paper [Cleary et al, 1996], we define two content-descriptive categories:

- *basic*: is the attribute discrete or continuous? are values ordered, or unordered? does the value range have a zero point?

- *relational*: what relationships exist between this attribute and other attributes? For example, is this attribute functionally dependent on another attribute? Can this attribute only be interpreted in the presence of a second attribute? Was this attribute value derived from one or more other attributes?

Basic and relational descriptors are essential in both applying ML algorithms to domain data and interpreting ML results. For example, if two functionally dependent attributes are included in a data set, many ML schemes will attempt to rediscover the (already known) dependency. The effect of this is twofold: meaningless rules are generated, and other patterns which are not part of functional dependencies will be ignored in favour of the functionally related items. Either a subset of the original data file (with the functional dependency removed) should be presented to the ML algorithm, or the redundant or common-sense rules produced by the functional dependency must be manually identified.

Additional content-descriptive metadata items include:

- *measurement-specific*: what is the unit of measurement? how was the item measured? at what date, with what instrument? is any measurement bias or error suspected? was the attribute taken from another source (a previous experiment, for example)?

- *semantic*: what is the meaning of the attribute? what part of the real world does it describe? Some semantic descriptors may be general to a domain and cut across a large class of data sets: for example, a definition of "relief" for geographic data [Sheth and Kashyap, 1996]. Domain-applicable metadata may be tied into existing domain ontologies (see, for example, [Mena et al, 1996]).

Measurement error or bias metadata is of primary importance in deciding whether an attribute is trustworthy enough to be included in domain modelling, and unit or information is necessary in constructing derived attributes and in interpreting results. Semantic descriptions are, of course, the basis for model construction: the researcher must ensure that all factors that contribute to an effect are included in the dataset processed by the ML algorithm. Domain semantics are also a prerequisite for evaluating the plausibility of the induced model, and in applying the model's derived information back in the domain.

Note that the study of relational metadata for ML has been restricted primarily to the consideration of connections and associations between values in columns, rather than between rows. Most ML algorithms require that each instance (row) be independent of all other instances, and so it is tacitly assumed in documentation that this requirement holds true.

3. Dynamic or usage metadata

The above types of cataloging metadata are primarily static, documenting the values and attributes contained in a file. However, a data set should also be viewed in terms of its usage, as well as its contents. Descriptors of dynamic attributes include citation information, process documentation, and analytical descriptions.

citation information:

Data sets are referenced by articles that utilize the data. Bibliographic descriptions of these reports should be part of the data set documentation, so that future researchers can compare their work with the results of earlier experiments. At present, no formal mechanisms exist in the ML community for recording dataset use. While many datasets at the UCI repository contain descriptions of usage prior to the donation of the dataset, post-donation usage is not maintained. Even more unfortunately, articles that include experimental work over these benchmark datasets often do not include references to the original papers describing the data sets, much less the body of work that also utilized the sets.

process documentation:

Using ML algorithms with real data is primarily an exploratory process; an initial data set is tested against an initial algorithm, the data set is modified, another algorithm may be chosen, and the new set is processed. This experimentation cycle should be recorded, particularly given that the final "results" may actually be a merged set of rules drawn from several runs over several data sets, or selected decision tree portions mined from a number of experiments. Tools such as Omega-Stat exist to organize and document statistical experiments [Harner and Galfalvy, 1995], and similar software has been developed to support general types of scientific investigation [Keller and Jones, 1996]. The WEKA machine learning toolbox currently incorporates simple experimentation management tools [Cleary et al, 1996], and more sophisticated documentation methods are under development.

analytical description

For a given data set, what results (statistical or machine learning) have been extracted from it in the past? This type of metadata provides a more direct link to previous research than citation information. Analytical descriptions may be as important for basic research into machine learning as for developing machine learning applications. It appears that the additional inductive power offered by ML algorithms is not required to process many benchmark data sets – as illustrated by the recent, startling discovery that a classifier that bases its decisions on a single attribute often performs as well as more sophisticated algorithms! At least part of the basis for this result is that many of the standard test databases used in the machine learning community embody very simple underlying structures (Holte, 1993). This problem could have been foreseen if earlier statistical analyses of some of these databases had been recorded in the data archives.

4. Using metadata in ML research

Creating and maintaining a metadata description of a dataset is a significant task; how, then, can this amount of effort be justified? At present, the primary use of metadata is in guiding the construction of the data subset that is presented to the ML algorithm (for a case study of this type of data cleansing and file construction, see [Garner, et al, 1995]). For example, many raw data files contain attributes that are essentially symbolic, but which are encoded as numerals—such as cow breed information stored so that 1 means the animal is a Friesian, 2 that it is a Jersey, 3 that it is a Murray Grey, etc. If these numbers are fed into many ML algorithms, the algorithm willl happily produce rules based on nonsensical calculations such as that a Friesian is less than a Jersey. Other, more complex, uses for a high-level and semantic understanding of the data include the construction of derived attributes, selection of clustering granularities for real valued attributes, and elimination of attributes from consideration for ML.

Similarly, metadata is useful in the interpretation of the results of a ML run. As noted above, any functional dependencies present in the dataset are likely to be discovered, but will not be of interest (and, as we have found to our embarrassment, should not be presented to a domain expert as an example of the analytical power of ML!). On the other hand, certain common sense relationships will generally be known in advance, and a ML run that fails to note them should be regarded with caution.

Finally, it is hoped that ML algorithms can be developed that will themselves autonomously use available metadata. One promising technique for instance based learning is vary the similarity measures used to calculate the closeness of attributes according to the attribute types; for example, to use a Manhattan rather than a Euclidean distance where appropriate, or to recognize that “day of the month” is modulo rather than strictly ordered from 1 to 31. The K* instance-based algorithm is currently being used to test a number of different similarity measures [Cleary and Trigg, 1995].

Of course, not all of this information will be useful for all data sets and all intended users. At present, two types of machine learning data archives are emerging: data stores intended for research into a particular domain, to which machine learning algorithms will be applied (most likely, along with other analysis techniques); and archives of standard or “interesting” data sets intended for use in conducting basic research in machine learning itself, to serve as benchmarks in testing new algorithms or implementations. The former, domain-oriented archive requires a greater depth of cataloging, since the semantics of the domain and data will guide the analysis. All of the above types of metadata can be of use in domain applications of ML. While the ML algorithms directly utilize relatively little even of the basic content-dependent descriptors, the more complex metadata is important in hand-crafting the tables used by the ML schemes as input and in interpreting the ML output.

In contrast, when using data sets as benchmarks the researcher is rarely interested in the informational content of a data set or even in the generalizations produced by a machine learning

run; instead, the experimental method concentrates on comparing summary measures (such as classification accuracy) across algorithms and data sets. For basic research in ML, then, the archives should support analytical, citation, and simple content-dependent metadata, but do not necessarily require process documentation or content-descriptive metadata. At present, benchmark archives provide some of this information, but in an ad hoc format with little standardization across data sets.

5. Organizational and structural support for metadata acquisition and maintenance

In setting up a useable catalog for a data repository, the first requirement is a standard cataloging format (however informal). Selection of an existing format or development of a new one will be contingent on the intended users and use of the data sets: will the users be domain experts? domain novices? will the repository be of interest to researchers in a variety of domains, or a single field? Finally, who will perform the cataloging — data set providers? professional data archivists? Clarification of these issues will dictate the degree and level of data description required. It is important that a good match exists between users, providers, and a standard, or the standard will be ignored. For example, the US Federal metadata standard for geographic data has been criticized as too extensive (requiring a great deal of effort to complete a catalog entry) and at too high of a semantic level (requiring an intimate knowledge of the domain, which effectively prevents the use of support staff to catalog a data set) [Foresman et al, 1996]. At present few geographic data sets completely conform to this standard, largely because of these difficulties.

Given a standard, a cataloging aid should be developed to guide a user through the metadata documentation process. This aid may be as simple as a form, to remind the user what descriptors to record. More complex tools may be required if the metadata standard is extensive or requires consultation of an outside source (such as a thesaurus of controlled vocabulary terms). One example of a simple machine learning data set description tool is provided by WEKA: as a data file is created, attribute descriptors are stored detailing the origins of derived attributes, data type information, and annotations providing content descriptions [Cleary et al, 1996]. The integration of data analysis and data file documentation/management software is particularly attractive, since the documentation process can be partially automated and tuned to the analysis method (here, machine learning).

Powerful tools and well-designed standards are not sufficient, however, to ensure that a data repository is continuously maintained or the standard adhered to. The providers and catalogers of the data must also have a strong personal or organizational investment in the development of the data archive. Unfortunately, this type of support has been extremely difficult to build — particularly at the organizational level. US Federal requirements are viewed as an unfunded mandate, and are a low priority with data providers struggling under high workloads and low funding [Foresman et al, 1996]. This problem is particularly pernicious when the data providers will not themselves be users of the data repository, as they have no incentive other than altruism to provide high quality cataloging.

In the research world, the scientific/academic culture provides rewards for publication of papers, but not usually for the creation or maintenance of data sets (though data set providers may be indirectly compensated with a greater visibility for their work, and data archive maintainers with informal professional recognition). Given this situation, it is hardly surprising that the machine learning and statistics archives have a low level of cataloging. One suggested solution to this problem is to formalize dataset archiving as a part of the scientific publishing process, as is currently being attempted by the Australian Geological Survey Organisation [Callahan et al, 1996]. Obviously, modifying the entire scientific culture is a daunting task. However, this approach may work with commercially-oriented organizations such as the AGSO (in which the reward structure can be tied to performance appraisals), or in small, highly developed “invisible colleges” within a discipline (see, for example, the revolution in publishing occurring within the high energy physics

community as WWW-based technical report archives are overtaking print journals as the preferred publication mechanism [Ginsparg, 1994].

Current recognition in the ML field for data set archiving is spotty, at best. Maintainers of the UCI archives, the most widely used set of benchmark datasets, suggest a pseudo-APA citation style for pseudo-APA style of reference for repository (see citation for [Merz and Murphy, 1996]). However, this reference style is not uniformly applied; for example, in the 1995 proceedings of the XII ICML conference [Prieditis and Russell, 1995], one of the primary publication venues for ML research, only 8 of the 22 papers that used datasets from the UCI repository *formally* referenced the archive (although several of the 22 mentioned the archive in the body of the paper). The original data set donors or creators also received varying degrees of referencing credit from papers that did not cite the UCI repository. It remains to be seen whether nor not the ML community will standardize around formal referencing of benchmark datasets and the archives themselves.

References

ALA (1995) Anglo-American Cataloging Rules, 2nd Edition. Published jointly by the American Library Association, the Canadian Library Association, and the Library Association

Callahan, S., Johnson, D., and Shelley, P. (1996) "Dataset publishing — a means to motivate metadata entry," *Proceedings of the First IEEE Metadata Conference* (Silver Spring, MD, USA). <URL: http://www.nml.org/resources/misc/metadata/proceedings/callahan/callahan.html>

Cleary, J., Holmes, G., Cunningham, S.J., and Witten, I.H. (1996) "MetaData for database mining", Proceedings of the First IEEE Metadata Conference (Silver Spring, MD, USA). <URL:http://www.nml.org/resources/misc/metadata/proceedings/holmes/DataBaseMining.html>.

Cleary J.G. and Trigg L.E. (1995) "K*: An Instance-Based Learner Using an Entropic Distance Measure," Proc Machine Learning Conference, Tahoe City, CA, USA, pp. 108-114.

Foresman, T., Porter, D., and Wiggins, H. (1996) "Metadata myth: misunderstanding the implications of federal metadata standards," *Proceedings of the First IEEE Metadata Conference* (Silver Spring, MD, USA. <URL: http://www.nml.org/resources/misc/metadata/proceedings/wiggins/foresman_final.html>.

Garner S.R., Cunningham S.J., Holmes G., Nevill-Manning C.G. and Witten I.H. (1995) "Applying a Machine Learning Workbench: Experience with Agricultural Databases," Proc Machine Learning in Practice Workshop, Machine Learning Conference, Tahoe City, CA, USA, pp. 14-21.

Ginsparg, P. (1994): "First steps towards electronic research communication," *Computers in Physics 8(4)*, p. 390-401.

Harner, E.J., and Galfalvy, H.C. (1995) "Omega-Stat: An environment for implementing intelligent modeling strategies," *Proceedings of the Fifth International Workshop on Artificial Intelligence and Statistics* (Ft. Lauderdale, FL, USA), 252-258.

Holte, R.C. (1993) "Very simple classification rules perform well on most commonly used datasets. " *Machine Learning 11*, 63-91.

Keller, T., and Jones, D. (1996) "Metadata: the foundation of effective experiment management," *Proceedings of the First IEEE Metadata Conference* (Silver Spring, MD, USA). <URL: http://www.nml.org/resources/misc/metadata/proceedings/keller/metadata.html>

Mena, E., Kashyap, V., Sheth, A., and Illarramendi, A. (1996) "OBSERVER: an approach for query processing in global information systems based on interoperation across pre-existing ontologies," *Proceedings of the First IFCIS International Conference on Cooperative Information Systems* (Brussels, Belgium). <URL: http://lsdis.cs.uga.edu/~kashyap/coopis.ps>.

Merz, C.J., & Murphy, P.M. (1996). UCI Repository of machine learning databases

[http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science.

Sheth, A., and Kashyap, V. (1996) "Media-independent correlation of information: what? how?" Proceedings of the First IEEE Metadata Conference (Silver Spring, MD, USA). Available at <URL: http://www.nml.org/resources/misc/metadata/proceedings/sheth/index.html>.