In Proceedings of the 6th Australian Joint Conference on Artificial Intelligence, World Scientific, 281-286, 1993.

A BENCHMARK FOR CLASSIFIER LEARNING

ZIJIAN ZHENG
Basser Department of Computer Science, University of Sydney, NSW 2006
Email: zijian@cs.su.oz.au

ABSTRACT: Although many algorithms for learning from examples have been developed and many comparisons have been reported, there is no generally accepted benchmark for classifier learning. The existence of a standard benchmark would greatly assist such comparisons. Sixteen dimensions are proposed to describe classification tasks. Based on these, thirteen real-world and synthetic datasets are chosen by a set covering method from the UCI Repository of machine learning databases to form such a benchmark.

1. Introduction

Many algorithms for classifier learning from examples have been developed and applied to problems from a variety of fields such as medical science, biology, and linguistics.

All of these algorithms perform differently on different domains. A great amount of literature concerning evaluating or comparing classifier learning algorithms has been published. One common problem is that only a few domains are used in each evaluation and different papers use different domains. It is very hard to judge an algorithm by seeing its performance on only a few arbitrarily selected domains.

When evaluating or comparing classifier learning algorithms the following questions are raised: how many domains are reasonable and, furthermore, what kind of domains should be used? The objective of this paper is to address this problem by proposing a benchmark for classifier learning. Section 2 gives a set of dimensions for describing classification tasks. Then, Section 3 proposes a set of datasets to form a benchmark. Finally, Section 4 briefly describes related work. Due to lack of space, we can only give a very brief description of the benchmark and its generation; further information can be found in a technical report[6].

2. A Taxonomy of Datasets

Generally speaking, the dataset of a classification problem has three aspects: the form of the attributes, the form of the classes, and the form of the instances.

The following 16 dimensions will be used to describe datasets. Four concern the attributes, five regard the instances (attribute/class values), and seven relate to the classes. One reason for adopting the 16 dimensions is that different classifier learning algorithms behave differently along each dimension. For example, some systems can deal with, but some others have difficulty with: discrete values, continuous values, missing values, noise, multiple classes, or irrelevant attributes. Some systems slow down very rapidly as the number of training examples and/or the number of attributes increases, but others do not.

To make it easier to choose the datasets for the benchmark, we first define all the dimensions using nominal values. In this paper, we use terms such as *small*, *low*, *medium*, *large*, and *high* as values of the dimensions. *Small* and *low* mean *in the first quartile* of all the candidate values. *Large* and *high* mean *in the last quartile* of all the candidate values. *Medium* means *between the small (or low) and the large (or high)*.

The dimensions and their nominal values are described briefly below, with two examples for each dimension value.

1. *Type of attributes* (4 values):
   - *Binary attributes only*, e.g. LED-24 and LED-7
   - *Nominal attributes only*, e.g. Promoter and NetTalk (Phoneme)
   - *Continuous attributes only*, e.g. Breast Cancer (W) and Diabetes
   - *Mixed attributes*, e.g. Soybean (16 binary attributes, 19 nominal attributes) and Hepatitis (13 binary attributes, 6 continuous attributes)
2. *Number of attributes* (3 values):
   - *Small*, e.g. NetTalk (Phoneme) (7) and Breast Cancer (W) (9)
   - *Medium*, e.g. Mushroom (22) and Hepatitis (19)
   - *Large*, e.g. Promoter (57) and Waveform-40 (40)
3. *Number of different nominal attribute values* (3 values):
   - *Small*, e.g. Promoter (4) and Monks-2 (4)
   - *Medium*, e.g. Soybean (7) and Lymphography (8)
   - *Large*, e.g. Mushroom (12) and NetTalk (Phoneme) (27)
4. *Number of irrelevant attributes* (2 values):
   - *With irrelevant attributes*, e.g. LED-24 and Waveform-40
   - *Without irrelevant attributes*, e.g. NetTalk (Phoneme) and LED-7
5. *Level of noise or indeterminacy*:
   This dimension is further divided into two:
   - *In attribute values* (2 values):
     - *With noise*, e.g. LED-24 and Waveform-40
     - *Without noise*, e.g. NetTalk (Phoneme) and Monks-2
   - *In class memberships* (2 values):
     - *With noise or indeterminacy*, e.g. NetTalk (Phoneme) and Breast Cancer (W)
     - *Without noise or indeterminacy*, e.g. LED-24 and Waveform-40
6. *Frequency of missing attribute values* (3 values):
   - *None*, e.g. Lymphography and NetTalk (Phoneme)

- *Few*, e.g. Mushroom (1.39%) and Breast Cancer (W) (0.25%)
- *Many*, e.g. Soybean (9.78%) and Thyroid (6.74%)

7. *Data-set size* (3 values):
   - *Small*, e.g. Promoter (106) and Lymphography (148)
   - *Medium*, e.g. Diabetes (768) and Thyroid (3163)
   - *Large*, e.g. NetTalk (Phoneme) (5438) and Mushroom (8124)
8. *Data-set density* (3 values):

$$\text{Density} = \frac{\text{Number-of-examples}}{\text{Size-of-description-space}}$$

$$\text{Size-of-description-space} = \prod_{i=1}^{n} N_i$$

   where n is the number of attributes and N_i is, for the i-th attribute, the number of different values (if it is a binary or nominal attribute), or the number of different values in the dataset (if it is a continuous attribute).
   - *Low*, e.g. Promoter (5.10×10^{-33}) and Waveform-40 (8.70×10^{-92})
   - *Medium*, e.g. Soybean (5.47×10^{-13}) and Breast Cancer (W) (7.77×10^{-7})
   - *High*, e.g. LED-24 (1.19×10^{-5}) and LED-7 (1.56)
9. *Number of classes* (3 values):
   - *Binary*, e.g. Hepatitis and Promoter
   - *Small or medium*, e.g. Lymphography (4) and LED-24 (10)
   - *Large*, e.g. Soybean (19) and NetTalk (Phoneme) (52)
10. *Default accuracy* (3 values):
    - *Low*, e.g. Soybean (13.7%) and NetTalk (Phoneme) (14.7%)
    - *Medium*, e.g. Lymphography (54.7%) and Breast Cancer (W) (65.5%)
    - *High*, e.g. Hepatitis (79.4%) and Thyroid (95.0%)
11. *Entropy* (3 values):

$$\mathit{Entropy} = -\sum_{i}^{N} P(C_i) \times log_2 P(C_i) \text{ bits}$$

    where N is the number of classes and $P(C_i)$ is the prior probability of class C_i. It is the expected amount of information for classifying an example.
    - *Low*, e.g. Hepatitis (0.73) and Thyroid (0.29)
    - *Medium*, e.g. Lymphography (1.23) and Diabetes (0.93)
    - *High*, e.g. Soybean (3.84) and NetTalk (Phoneme) (4.74)
12. *Difficulty of the domain*:
    We use the following two dimensions to discuss the *difficulty of the domain*. The highest accuracy achieved by some existing algorithms, recorded in "Past Usage" in UCI ML Databases, is used as *predictive accuracy*. The corresponding *relative accuracy* is computed as

$$\text{Relative accuracy} = \frac{\text{Predictive accuracy} - \text{Default accuracy}}{100\% - \text{Default accuracy}} \times 100\%$$

- *Predictive accuracy* (3 values):
  - *Low*, e.g. Lymphography (78.4%) and LED-24 (70.0%)
  - *Medium*, e.g. Hepatitis (83.0%) and Soybean (97.1%)
  - *High*, e.g. Mushroom (99.4%) and Monks-2 (100%)
- *Relative accuracy* (3 values):
  - *Low*, e.g. Hepatitis (17.5%) and Diabetes (39.3%)
  - *Medium*, e.g. Lymphography (52.3%) and Breast Cancer (W) (84.9%)
  - *High*, e.g. Soybean (96.6%) and Mushroom (98.7%)

13. *Information score*:
    To take the class prior probabilities and class distribution of a dataset into account, Kononenko and Bratko introduce the average information score and relative information score criteria[2,6]. Here, we use the well-known decision tree learning algorithm C4.5[4] as the classifier to calculate the information score, and have the following two dimensions:
    - *Average information score* (3 values):
      - *Low*, e.g. Hepatitis (0.14) and Thyroid (0.21)
      - *Medium*, e.g. Lymphography (0.61) and Mushroom (0.99)
      - *High*, e.g. Soybean (3.35) and LED-24 (2.07)
    - *Relative information score* (3 values):
      - *Low*, e.g. Hepatitis (19.1%) and Promoter (42.1%)
      - *Medium*, e.g. Lymphography (54.3%) and NetTalk (Phoneme) (78.6%)
      - *High*, e.g. Breast Cancer (W) (87.5%) and Soybean (91.4%)

3. Data Sets

It is impossible to evaluate or compare learning algorithms by using them on all possible domains. Here we use the dimensions discussed above to propose the least number of representative[I] domains as a benchmark that covers the space described by the dimensions as completely and evenly as possible. In order to cover the dimensions adequately, a requirement was imposed that every value of each dimension be represented by at least two datasets.

To minimize subjectivity when selecting domains, a two step process was followed:

- The candidates were all real-world domains and well-known synthetic domains from the UCI Repository of machine learning databases[3]. To ensure that the benchmark does not contain relatively unknown datasets, only datasets that have been referenced in the last three Proceedings of the International Workshop/Conference on Machine Learning were considered. This dual constraint guarantees that all datasets are commonly known and easily accessible. Thirty-one such datasets were identified.

[I] *representative* here means that the domain has some common characters with several other domains, that is, it can be a prototype of a group of domains.

Table 1. Datasets in the benchmark

Name	Description
Breast Cancer (W)	Medical diagnosis applied to breast cytology (Wisconsin)
Hepatitis	Predicting whether a patient will die from hepatitis
LED-24	LED display with 24 segments (17 irrelevant)
LED-7	LED display with 7 segments
Lymphography	Lymphography database
Monks-2	The second Monks' problem
Mushroom	Mushrooms classified as poisonous or edible
NetTalk (Phoneme)	NetTalk Corpus for the phonetic transcription of the 1000 most common English words (prediction of phoneme)
Diabetes	Pima Indians diabetes database for diagnosing diabetes
Promoter	Promoter gene sequences (DNA)
Soybean	Large soybean database
Thyroid	Hypothyroid database (thyroid disease records)
Waveform-40	Waveform database with 40 attributes (19 irrelevant)

- A set covering algorithm was used to select the smallest subset of them that satisfies the coverage requirement. The algorithm was run 50 times with different random orderings of the 31 datasets and the frequency of occurrence of each dataset in the selected subsets was noted. With the datasets sorted by this frequency, the set covering algorithm was run once more to identify the final subset.

The result was a collection of thirteen datasets described briefly in Table 1.

4. Related Work

Almost all the developers of learning algorithms use some domains to demonstrate the performance of their algorithms and to compare theirs with other methods. For example, Weiss and Kapouleas[5] use four domains to compare statistical, connectionist, and machine learning classification algorithms. To the best of our knowledge, however, there is no generally accepted benchmark for classifier learning.

Brazdil[1] uses 23 domains from the Esprit Project StatLog to compare and characterize 22 different learning algorithms using meta level learning. The test results of the algorithms on the domains together with domain descriptions are used as input data for the meta level learning and, for each algorithm, rules predicting the applicability of the algorithm are synthesized. The domains are characterized by using 3 simple measures such as the number of examples, 6 statistical measures such as SD-ratio (geometric mean ratio of standard deviations of the individual populations to the pooled standard deviation), and 11 information based measures such as Cx_class (complexity of classification scheme).

5. Conclusion

A benchmark for classifier learning is very important for evaluating or comparing algorithms for learning from examples. In this paper, sixteen dimensions are developed to describe classification tasks. Based on these, thirteen real-world and synthetic datasets are chosen from UCI Repository of machine learning databases to form a benchmark.

The benchmark provides a basis for evaluating and comparing algorithms for learning from examples. It also makes it possible to characterize any given learning algorithm. With the benchmark, we can pinpoint particular strengths and/or weaknesses of different classifier learning algorithms with respect to measures such as accuracy, theory size, and learning time.

6. Acknowledgements

This research was supported by an ARC grant (to Ross Quinlan) and by a research agreement with Digital Equipment Corporation. The author is supported by EMSS scholarship. Many thanks to Ross Quinlan for his advice and suggestions, particularly about doing this work. I'd also like to thank Mike Cameron-Jones, Alen Varšek, and Kai Ming Ting for their helpful discussions and suggestions. Finally, P.M. Murphy and D. Aha are gratefully acknowledged for creating and managing UCI Repository of machine learning databases.

7. References

1. P. Brazdil, J. Gama, and B. Henery, Comparison of ML and statistical approaches using meta level learning, *Workshop Notes on Real-World Applications of Machine Learning, ECML,* (1993).
2. I. Kononenko and I. Bratko, Information-based evaluation criterion for classifier's performance, *Machine Learning*, **6** (1991), p. 67-80.
3. P.M. Murphy and D.W. Aha, *UCI Repository of machine learning databases* [Machine-readable data repository]. Irvine, CA: University of California, Department of Information and Computer Science, (1991).
4. J.R. Quinlan, *C4.5: Programs for Machine Learning*, (Morgan Kaufman, 1993).
5. S.M. Weiss and I. Kapouleas, An empirical comparison of pattern recognition, neuralnets, and machine learning classification methods, *Proceedings of the 11th IJCAI*, (Morgan Kaufman, 1989), p. 781-787.
6. Z. Zheng, A benchmark for classifier learning, *Technical Report, Basser Department of Computer Science, The University of Sydney*, (1993).