Working Paper Series
ISSN 1170-487X

Applying machine learning to subject classification and subject description for information retrieval.

by Sally Jo Cunningham, Brent Summers

Working Paper 95/20
June 1995

Department of Computer Science
The University of Waikato
Private Bag 3105
Hamilton, New Zealand

Applying machine learning to subject classification and subject description for information retrieval

Sally Jo Cunningham, Brent Summers
Department of Computer Science
University of Waikato
Hamilton, New Zealand
email: {sallyjo,bcs}@waikato.ac.nz

Abstract: This paper describes an experiment in applying standard supervised machine learning algorithms (C4.5 and Induct) to the problem of developing subject classification rules for documents. These algorithms are found to produce surprisingly concise models of document classifications. While the models are highly accurate on the training sets, evaluation over test sets or through cross-validation shows a significant decrease in classification accuracy. Given the difficult nature of the experimental task, however, the results of this investigation are promising and merit further study. An additional algorithm, 1R, is shown to be highly effective in generating lists of candidate terms for subject descriptions.

1. Introduction

Subject classification of documents has long been recognised as an important element of document retrieval systems. Keyword searches, while useful, can't do it all; often a document surrogate (such as title and abstract) doesn't contain one or more of the terms that the document is "about", and a subject search is used to retrieve these documents without degrading search precision by including additional, potentially ambiguous, terms in the search query. Currently, however, these subject headings must be assigned to documents by highly trained human cataloguers. Manual document subject classification is necessarily time-consuming and expensive, and constitutes a formidable bottleneck in the cataloguing of a collection (Salton and McGill, 1983).

Similarly, services which monitor news wires for documents of interest to a given user generally require that user to define a profile of terms pertaining to the user's topic. Constructing an appropriate term list can be difficult, particularly for novice users. Automated techniques may provide more principled (and hopefully more effective) means of building a concept description.

This paper explores the potential of supervised machine learning techniques for automating document classification and for building concept descriptions. We extend an experiment performed by Crawford et al (1991) that uses supervised machine learning algorithms to construct topic descriptions. Crawford et al constructed a test set of 50 Reuters articles on terrorism and 50 other randomly chosen Reuters documents. The CART algorithm was then used to construct a decision tree which could be used to classify new documents as "terrorist" or "non-terrorist". In Section 2, this paper examines the performance of two other machine learning algorithms – C4.5 and Induct – in discriminating between documents about machine learning and neural networks. This task is obviously more demanding, as the two techniques are based in the same discipline and can be used to solve similar types of problems. Section 3 examines the reliability and accuracy of the subject classification models that are developed.

The efficacy of Holte's 1R algorithm (Holte, 1993) as a tool to generate sets of candidate subject descriptors is explored in Section 4. The goal in this experiment was not subject classification per se, but rather an attempt to semi-automate the task of producing rich concept descriptions.

The WEKA machine learning workbench (Holmes et al, 1994) provided a testbed for experimenting with the applicability of these algorithms to information retrieval problems. WEKA is an integrated system comprising tools for developing and manipulating data sets, running and cross-validating several common machine learning algorithms, and viewing and analysing results.

2. Similarity-based algorithms: C4.5 and Induct

To test the effectiveness of machine learning algorithms for subject classification and description, two sets of document titles and abstracts were constructed: the first consisting of 37 neural network and 44 machine learning documents, and the second set of 58 neural network and 21 machine learning articles. There is no overlap between the documents in the two data sets, and no article could be classified with both subject descriptors.

Machine learning algorithms require data to be in the form of a table of examples, where each example is described by a group of single-valued attributes. For these experiments, each document is treated as an example. To construct the attribute set, a program extracts a list of all the unique words from a document collection. This list is culled of high frequency functional "stop words" (such as "a", "and", "the", etc.) and words occurring only once in the entire document set. We also remove

the words "machine", "learning", "neural", and "network", since they were the search terms used to construct the data sets (and therefore at least two of the four are present in each document). The remaining words are used as attribute names. Training sets are then generated in which each example contains a Boolean value denoting whether each attribute appears in the original document. For the first data set, then, we have 1283 attributes and 81 examples, and the second data set contains 1204 attributes describing 79 examples.

2.1 C4.5

C4.5 is a popular machine learning algorithm that induces a decision tree from examples, and produces an equivalent rule set from the decision tree (Quinlan, 1992). Applying C4.5 to the two data sets produces the following rule sets to determine whether a document is about machine learning (ml) or neural networks (nn):

Data set 1:

Rule 1: If "arm" = yes
Then class = nn

Rule 2 If "associated" = "no" and "input = "yes"
Then class = nn

Rule 3 If "trained" = yes
Then class = nn

Rule 4 If "feed-forward" = yes
Then class = nn

Rule 5 If "local" = yes
Then class = nn

Rule 6 If "initial" = yes
Then class = nn

Rule 7 If "discrete-time" = yes
Then class = no

Rule 8 If "arm" = no and "discrete-time = no and "feed-forward" = no and "initial" = no and "local" = no and "trained" = no
Then class = ml

rule	# of examples rule classifies	# of examples classified incorrectly
1	12	0
2	7	0
3	3	0
4	4	0
5	3	0
6	3	0
7	2	2
8	47	3

Data set 2

Rule 1 If "inductive" = yes
Then class = ml

Rule 2 If "decision" = yes
Then class = ml

Rule 3 If "ai" = yes"
Then class = ml

Rule 4 If "under" = yes
Then class = ml

Rule 5 If "effectively" = yes
Then class = ml

Rule 6 If "constraints" = yes
Then class = ml

Rule 7 If "ai" = no and "constraints" = no and "decision" = no and "effectively" = no and "inductive" = n and "under" = no
Then class = nn

rule	# of examples rule classifies	# of examples classified incorrectly
1	7	0
2	4	0
3	3	0
4	2	0
5	2	0
6	2	0
7	59	1

Discussion

Both data sets can be described to a high degree of accuracy by surprisingly small rule sets. For data set 1, the rules 1-7 pick out characteristic terms used to describe neural networks, and rule 8 declares that the absence of these terms defines a machine learning document. The six terms used to characterise neural networks documents all convey semantic information: "arm" refers to a common neural network application, driving a robotic arm; nets execute in "discrete-time" increments and are "feed-forward"; various network parameters must be given an "initial" setting; networks frequently settle into a "local" minima; and networks are "trained" rather than programmed.

For data set 2, the default rule 7 defines a neural network document, and rules 1-6 characterise machine learning papers in that training set. Here, only 3 of the 6 terms used by the rule set are semantically meaningful when describing machine learning research: "constraints", "decision", and "inductive". The other terms, "effectively", "under", and "AI" are more general words (at least

in this context), and their inclusion is most likely an artefact of the composition of the training set. This hypothesis is supported by the fact that only 7 of the documents are classified by these rules, and is confirmed by an examination of the documents themselves.

2.2 Induct

Unlike C4.5, Induct infers a rule set directly without first creating a decision tree. The two algorithms also differ in their approach to selecting "significant" attributes; C4.5 primarily uses information theoretic measures, while Induct relies more heavily on statistical tests (Gaines, 1991). Using Induct on the two data sets, we generate the rule sets below:

Data set 1

Rule 1 If "reasoning" = no and "framework" = no and "here" = no and "concept" = no and "discuss" = no and "evaluating" = no and "observations" = no and "agent" = no and "ai" = no and "attempt" = no and "characterizations" = no
Then class = nn

Rule 2 If "reasoning" = yes or "framework" = yes or "here" = yes or "concept" = yes or "discuss" = yes or "evaluating" = yes or "observations" = yes or "agent" = yes or "ai" = yes or "attempt" = yes or "characterizations" = yes
Then class = ml

rule	# of examples rule classifies	# of examples classified incorrectly
1	37	0
2	44	0

Data set 2

Rule 1 If "domain" = no and "decision" = no and "inductive" = no and "approximately" = no and "intelligence" = no
Then class = nn

Rule 2 If "function" = yes or "agents" = yes
Then class = nn

Rule 3 If "domain" = yes and "activation" = no
Then class = ml

Rule 4 If "decision" = yes or "inductive" = yes
Then class = ml

Rule 5 If "approximately" = yes and "algorithm" = yes
Then class = ml

Rule 6 If "intelligence" = yes and "able" = no
Then class = nn

rule	# of examples rule classifies	# of examples classified incorrectly
1	55	0
2	15	0
3	9	0
4	12	0
5	2	0
6	3	0

Discussion

Like C4.5, Induct achieves a high accuracy in classification over the training set – in this case, 100% accuracy for both data sets. Again, the rule sets formed are small and concise. The Induct rules, however, contain a larger number of terms that are unlikely to contain semantic subject information ("here", "discuss", "evaluating", "ai", "framework", "approximately", "able", and "algorithm"). Intuitively, the rules containing the highest degree of subject-specific words are likely to be more robust in classifying new documents. Further study is needed to explore the relative suitability of these algorithms to this specific task.

Note also the small overlap between the rule sets inferred by the two algorithms: only two terms appear in common ("decision" and "inductive", for data set 2). As will be discussed in Section 4, this disjunction between the term sets is likely due to the fact that several terms will possess approximately the same discriminatory power. The different attribute selection methods employed by the two algorithms (information theoretic and statistical) appears to lead them to choose different terms from the equivalency sets.

3 Evaluating the rule sets

While the rule sets above achieve exceptional classification accuracy over their respective training sets, this performance may not be predictive of their accuracy on new documents. We applied two standard validation techniques to these models: cross-validation and evaluation over an independent test set.

3.1 Cross-validation

Once a rule set is constructed, how do we know how accurate that model is likely to be on new data? One method for estimating rule set accuracy is a resampling technique called *cross-validation*. A random sampling of the original data set (the training set) is used to construct a model, and the model is tested on the remainder of the original data (the test set). This sampling process is carried out

a number of times, and the error rates for the test sets are averaged. This test-and-train technique provides a simulation of the system's performance on new cases.

For the C4.5 rules, we performed 25 resamplings over both of the two data sets for three levels of training set/test set ratios: 66/34, 50/50, and 33/67 The results of this cross-validation are presented below. Note that while the accuracy over the test sets degrades significantly as compared to the accuracy over the training set, the models developed are relatively stable at the three training/test set ratios. These results are promising, in that they indicate that a relatively small number of sample documents can be used to construct a useful subject description. The relatively low classification accuracy over the test set indicates that, not surprisingly, machine learning techniques are not likely to be sufficient to perform subject classification autonomously. However, this method for building subject descriptions appears useful as an adjunct to human subject description.

% in test set	training set % correct	std. dev.	test set % correct	std. dev.
66	97.43	1.88	69.14	7.46
50	97.17	2.08	66.70	7.52
33	97.34	2.27	66.89	6.64

C4.5 rules, 25-fold cross-validation, data set 1

% in test set	training set % correct	std. dev.	test set % correct	std. dev.
66	97.46	1.33	80.30	7.46
50	98.20	1.35	77.95	8.76
33	96.61	2.56	74.49	7.13

C4.5 rules, 25-fold cross-validation, data set 2

For the Induct rule sets, 25-fold cross-validation yields the following classification accuracies:

	rule set 1	rule set 2
Number of examples	81	79
correctly classified	50 (61.7%)	68 (86.1%)
incorrectly classified	28 (34.6%)	10 (12.7%)
no classification	0 (0.0%)	1 (1.3%)
multiple classifications	3 (3.7%)	0 (0.0%)

The C4.5 rules achieved generally higher accuracies. This differential may be an artefact of the data sets used in these experiments, or may reflect a greater suitability of the C4.5 algorithm for this type of classification task. Further investigation is needed on this point.

3.2 Evaluation over an independent test set

A second method for examining the accuracy of a rule set is to test the model developed from a training set over an independent test set; for example, to test the rules constructed for data set 1 on the documents in data set 2, and vice-versa. This method can provide a more realistic view of the rule set's accuracy when a high degree of variability can be expected in new data, as in the case of document descriptions.

Due to software problems, this validation technique could only be used with C4.5 rule sets. Applying the rules derived from data set 1 to data set two achieves a classification accuracy of 48.2% (38 errors). Only one of the errors entailed misclassifying a machine learning document as a neural network document. The rules formed over data set 2 obtained a 72.8% classification accuracy when applied to the documents in data set 1. Of the 22 classification errors, 21 occurred in rules 1-4 (mis-classifying a neural networks document as a machine learning article).

While these classification accuracies would not be acceptable for an automatic classification scheme, they are promising as an adjunct to human classification efforts. Further, this type of rule set accuracy could prove useful as a mechanism for extending a user's query so as to improve its recall. We envision the user providing a sample set of documents meeting the user's information needs, and a machine learning scheme inferring a descriptive rule set that can retrieve additional potentially relevant documents from the collection. Again, it appears unlikely that these schemes can produce subject models robust enough for automated classification, but these results are indicative that they can be useful in semi-automated classification and retrieval.

4. Building descriptive term sets: Holte's 1R

Typically, query construction or subject description involves manually constructing a set of candidate terms to describe the information need or topic. This process is extremely labour-intensive, however, and the subject terms selected may not be useful for a given document collection (if those terms do not appear in the document set, for example). Ideally, a pre-processor would infer a set of predictive words, and this set would be edited manually to construct a subject description – a far less daunting task than developing a term list from scratch.

As the experiments above illustrate, standard machine learning algorithms produce rule sets that succinctly describe subjects. We could use the terms from the rule set to construct or augment a subject description, but the very conciseness of the rules limits their usefulness. However, the terms appearing in the rule set generally possess only a slightly higher degree of discriminatory power than other terms in the document collection – the rule set terms effectively serve as proxies for many other words. To extract a list of these alternative terms, we turn to another machine learning algorithm: Holte's 1R technique (Holte, 1993).

1R was originally developed to illustrate the lack of complexity of the standard datasets used to test new machine learning algorithms. 1R builds rules based on a single attribute of a dataset, for each attribute in the dataset. The one-feature rules are then ranked on their classification ability. Holte showed that for the standard datasets, selecting the best one-feature rule achieves similar performance as the state-of-the-art machine learning techniques!

The WEKA workbench includes a version of 1R developed by Holmes and Neville-Manning (1995). This implementation provides a ranked listing of the attributes (here, the terms in the document collection), an estimate of each term's discriminatory power, and the rules associated with each term. Examination of this output indicates that the ranked terms are indeed a rich source of subject descriptors. In addition, the estimate of classificatory power is useful in establishing a cut-off point in the list. Effectively, the user can choose the best N terms, where N can vary according to the user's needs.

As with the rule sets developed by C4.5 and Induct, terms are included that have little semantic relation to the topic (and appear as artefacts of the contents of the training set). These terms must be identified manually. Interestingly, 1R provides an intuitively reasonable distinction between machine learning and neural networks descriptors. Given that it is particularly difficult for humans to produce term sets that differentiate between two relatively similar subjects, the 1R algorithm appears useful as a pre-processor to generate a set of candidate terms for subject descriptions or user query augmentation.

5. Summary

In this paper we explore the applicability of three supervised machine learning algorithms to the problem of subject classification and developing subject descriptions for an information retrieval system. The results of our experiments are promising, in that the rule sets derived are sufficiently accurate to provide an adjunct to human classification or subject description. Note, too, that these experiments are worst-case scenarios: the two topics that the machine learning algorithms must distinguish between are semantically close, and the most useful distinguishing terms have been stripped from the data sets ("machine", "learning", "neural", and "networks"). These latter terms were not used to form the subject description models since the data sets were constructed by using these words as search terms. It would be expected that adding these terms to the rule sets would increase their classification accuracy on new documents.

Further research in this area will include: larger-scale testing of the ability of machine learning algorithms to classify over more than two subjects; an examination of the affects of word stemming on the document terms, as it appears that some concepts are being buried by being represented by several attributes (for example, "cluster, clustering, clusters"); and a consideration of the effects of using word pairs as a single attribute (for example, a single attribute for "machine learning" rather than two attributes, "machine" and "learning").

References

Crawford, S.L., Fung, R.M., Appelbaum, L.A., and Tong, R.M. "Classification trees for information retrieval." *Proceedings of the eighth international workshop on machine learning*, 1991, pp. 245-249.

Gaines, B.R. "The tradeoff between knowledge and data in knowledge acquisition in knowledge discovery in databases." In *Knowledge discovery in databases*, G. Piatetsky-Shapiro and W.J. Frawley (eds.), AAAI Press, 1991, pp. 491-505.

Holmes, G., Donkin, A., and Witten, I.H. "WEKA: a machine learning workbench." *Proceedings of the Second Australia and New Zealand Conference on Intelligent Information Systems*, Brisbane (Australia), 1994, pp. 357-361.

Holmes, G. and Nevill-Manning, C.G. "Feature selection via the discovery of simple classification rules." To appear in the *Proceedings of the International Conference on Intelligent Data Analysis*, Baden-Baden (Germany), July 1995.

Holte, R.C. "Very simple classification rules perform well on most commonly used datasets." *Machine Learning*, vol. 11, 1993, pp. 63-90.

McQueen, R.J., Neal, D.L., DeWar, R.E., Garner, S.R., and Nevill-Manning, C.G. "The WEKA machine learning workbench: its application to a real world agricultural database." *Proceedings of the Canadian machine Learning Workshop*, Banff (Canada), 1994..

Salton, G., and McGill, M.J. *Introduction to modern information retrieval*, McGraw-Hill Book Company, 1983.

Quinlan, J.R. *C4.5: programs for machine learning*, Morgan Kaufmann, 1992.