MDL and Categorical Theories (Continued)

J.R. Quinlan
Basser Department of Computer Science
University of Sydney
Sydney Australia 2006
quinlan@cs.su.oz.au

Abstract

This paper continues work reported at ML'94 on the use of the Minimum Description Length Principle with non-probabilistic theories. A new encoding scheme is developed that has similar benefits to the ad-hoc penalty function used previously. The scheme has been implemented in C4.5RULES and empirical trials on 25 real-world datasets reveal a small but useful improvement in classification accuracy.

1 INTRODUCTION

When classifiers are induced from data, the resulting theories are commonly interpreted as functions from attribute values to classes rather than to class distributions. So, for example, we talk of the accuracy of the learned classifier on unseen cases, measured as the percentage of such cases for which the classifier predicts the actual class. Such theories and their interpretation will be described as *categorical*, although the synonym *deterministic* is also in common use.

A concern when learning in real-world domains is that the theory should not overfit the data because overly complex theories often have lower accuracy on new cases.[1] Among the many techniques for overfitting avoidance such as cost-complexity pruning [Breiman, Friedman, Olshen, and Stone, 1994] and reduced-error pruning [Quinlan, 1987], those based on the Minimum Description Length Principle [Rissanen, 1983] or the analogous Minimum Message Length Principle [Georgeff and Wallace, 1985] are particularly attractive because they have both an intuitive interpretation and a strong theoretical base. In the MDL approach, possible theories $\{T_i\}$ derived from data D are characterized by their *description length*, the number of bits needed to encode both the theory and the data from which it was learned. Choosing the theory T_i with minimum description length is equivalent to maximizing the probability $Pr(T_i|D)$ of T_i given the data.

This raises an immediate problem because the best theory learned from noisy data would not be expected to fit that data exactly. If T_i is interpreted categorically and does not fit D, then $Pr(T_i|D)$ is zero. As Pednault [1991] puts it,

> In the deterministic case, any theory that does not absolutely agree with the observations can be ruled out.

In such situations, MDL makes sense only if the theories are interpreted probabilistically:

> The objective sought by MML ... is not the correct classification of the maximum number of [unseen] cases, but the minimization of the amount of information needed to determine the class once the category is known. [Wallace and Patrick, 1993, p18]

Despite this, MDL is often used in situations where the learned theory is assessed on the categorical accuracy of its predictions, e.g. [Quinlan and Rivest, 1989].

Examples of tasks in which MDL leads to poor choices among competing categorical theories are given in [Quinlan, 1994]. For those tasks, theories with larger categorical error rates tend to assign an unexpectedly high or low prior probability to the described class. That paper recommended an additional bias in favor of theories whose predicted class distribution matches that observed in the data. I can offer no theoretical justification for this preference, but it could be argued from a more philosophical perspective that a theory learned from data should accurately summarize that data. If a theory is intended to be interpreted categorically, it should not misrepresent the prior probabilities of the classes.

Although limited empirical trials showed that this bias is effective in selecting theories with a lower categorical

[1] However, Schaffer [1993] points out that all overfitting avoidance is a form of bias that must lead to worse performance in some situations.

error rate on unseen cases, its implementation using a penalty function was rather ad-hoc and the paper concluded:

> ... a new encoding scheme that reflected categorical performance and reasonable prior assumptions would be more satisfying.

An encoding scheme along the lines envisaged has now been developed. The following section defines the kind of theories considered here and their use with MDL. After summarizing the problem and the previous approach, the paper introduces the new encoding scheme that has been incorporated into a learning program C4.5RULES [Quinlan, 1993]. Experiments on 25 real-world domains demonstrate the benefit of the scheme.

2 CLASS DESCRIPTION THEORIES AND MDL

Symbolic classifiers come in many forms including decision trees [Hunt, Marin, and Stone, 1966], decision lists [Rivest, 1987], CNF and DNF expressions [Pagallo and Haussler, 1989], and concepts described in special-purpose logics [Michalski, 1980]. Like [Quinlan, 1994], this paper concerns two-class tasks in which the learned theory is a description of one of the classes, called the *target* class, although the formalism in which this description is expressed is not important. A theory *covers* a case if the case matches the description; cases so covered are predicted to belong to the target class while all other cases are assigned to the non-target class.

The MDL Principle can best be explained in terms of a communication model in which a sender transmits to a receiver a description consisting of a theory T and the data D from which it was derived [Quinlan and Rivest, 1989]. The description length associated with T consists of the cost of a message encoding T itself (the *theory cost*) and then the data given T. Intuitively, the length of the former component measures theory complexity and that of the latter the degree to which the theory fails to account for the data, so that description length represents a balancing of model fit against complexity. If there is a choice among several theories, the MDL Principle states that the theory associated with the shortest description length should be preferred.

We assume some agreed language in which all theories are expressed, so that the theory cost is the number of bits needed to transmit the particular sentence representing T. The cost of encoding D given the theory can be broken down into the bits needed to transmit the attribute values for each case plus the bits required for the cases' classes. The former is the same for all theories and can be ignored, since description lengths are used only to compare possible theories. For the latter, identifying each case's class given a theory comes down to identifying the cases misclassified by the theory, since their classes can be inverted under the two-class assumption. The number of bits needed to identify the errors made by a theory is referred to as its *exceptions cost*.

Several methods for encoding exceptions are discussed in [Quinlan, 1994]. Instead of specifying such schemes in detail, this paper follows Wallace and Patrick [1993] in adopting a more abstract perspective. If messages $\{m_1, m_2, ...\}$ occur with probabilities $\{p_1, p_2, ...\}$, we postulate an encoding scheme in which message m_j requires $-\log(p_j)$ bits (all logarithms being taken to base 2). Of course, this assumes that the probability of a message occurring is independent of the previous messages and that the receiver also knows the relevant probabilities $\{p_j\}$.

For instance, suppose that T misclassifies e cases in D. The errors can be identified by sending one of the messages $\{$*correct, incorrect*$\}$ for each case in D with probabilities $e/|D|$ and $1 - e/|D|$ respectively. Since the receiver must know these probabilities in order to decode the messages, we first transmit e (which ranges from 0 to $|D|$). The total number of bits to be transmitted is then

$$\begin{aligned}&\log(|D|+1)\\&+ e \times (-\log(\frac{e}{|D|}))\\&+ (|D|-e) \times (-\log(1-\frac{e}{|D|})).\end{aligned} \tag{1}$$

This will be called the *uniform* coding strategy since errors across D are identified as a single group.

An alternative *divided* strategy identifies separately the errors in the cases covered by the theory (the *false positives*) and those in the remaining cases (*false negatives*). If there are fp and fn of these respectively, and C and U are the numbers of cases covered and not covered by the theory respectively, the exceptions cost is

$$\begin{aligned}&\log(C+1)\\&+ fp \times (-\log(\frac{fp}{C}))\\&+ (C-fp) \times (-\log(1-\frac{fp}{C}))\\&+ \log(U+1)\\&+ fn \times (-\log(\frac{fn}{U}))\\&+ (U-fn) \times (-\log(1-\frac{fn}{U})).\end{aligned} \tag{2}$$

Although the divided strategy often requires more bits than the uniform strategy, the approach of identifying errors in subsets of the data is used in both [Quinlan and Rivest, 1989] and [Wallace and Patrick, 1993].

Table 1: Exceptions costs for five competing theories

Theory	False Pos	False Neg	Cases Covered	Uniform Encoding	Divided Encoding	Biased Encoding
T_1	19	28	291	283.5	289.1	282.0
T_2	24	24	300	287.8	289.2	281.4
T_3	47	10	337	325.4	289.0	293.2
T_4	74	0	374	390.6	286.2	333.4
T_5	681	272	709	283.5	289.1	546.7

3 AN ANOMALY AND A PREVIOUS SOLUTION

As discussed in [Quinlan, 1994], MDL can lead to poor choices among candidate categorical theories. One hypothetical illustration used in that paper supposes a dataset of 1000 cases of which 300 belong to the target class, with five candidate theories that give rise to various numbers of false positive and false negative errors as shown in Table 1. All five theories are further presumed to have the same theory cost, so that MDL will choose the theory with lowest exceptions cost. In this situation the uniform strategy will find an exact tie between T_1, with 47 errors on the training data, and T_5, with 953! The divided approach will chose T_4, with 74 errors, over the equally complex theory T_1 that makes far fewer errors. The choices made by MDL in this (admittedly contrived) example are clearly at odds with intuition.

The number of cases covered by a theory is given by

$$tp \ + fp \ - fn$$

where *tp* is the number of (true positive) cases belonging to the target class. In a categorical context, the proportion of cases covered by the theory can be interpreted as the predicted prior probability of the target class. Theories T_4 and T_5, which cover 37.4% and 70.9% of the cases respectively, are at marked variance with the data in which the prior probability of the target class is 30%.

In an attempt to force categorical theories to agree with the training data in this respect, [Quinlan, 1994] penalizes atypical theories. The details are unimportant here, but the idea is to multiply the description length of a theory by a factor based on the discrepancy between the predicted proportion of target cases and that observed in the data.

4 A NEW SOLUTION

Resorting to an ad-hoc penalty function is inherently unsatisfying, particularly since the principal attraction of MDL methods is their clean theoretical base. My justification for using it was an inability to find a method for coding theories that favors those whose predicted class distribution is similar to that observed in the data. I realized recently that I was concentrating on the wrong component of description length and that the method of encoding exceptions could be adapted to prefer such theories.

The proportions of target class cases predicted by a theory and observed in the training data are the same when the numbers of false positives and false negatives are equal. This suggests a new *biased* coding scheme as follows: Just as with the uniform scheme, the total number e of errors is sent to the receiver. Instead of transmitting the error messages for all the data, the sender first transmits the errors in the C cases covered by the theory and then those in the U uncovered cases. Under the assumption that false positives and false negatives are balanced, the probability of error in the covered cases is $e/2C$ and this probability is used to encode the error messages for covered cases. Once the false positives have been identified, the receiver can calculate the true number of false negatives as e-fp, so the probability of error on the uncovered cases is known to be fn/U. The total exceptions cost then becomes

$$\begin{aligned}&\log(|D|+1)\\&+ fp \times (-\log(\frac{e}{2C}))\\&+ (C - fp) \times (-\log(1 - \frac{e}{2C}))\\&+ fn \times (-\log(\frac{fn}{U}))\\&+ (U - fn) \times (-\log(1 - \frac{fn}{U})). \qquad (3)\end{aligned}$$

There is a slight complication: if the number C of covered cases is small, $e/2C$ may be greater than 1. To overcome this problem while retaining symmetry, the above scheme is followed when at least half the cases are covered by the theory; if less than half are covered, the (false negative) errors in the uncovered cases are transmitted first, using the probability $e/2U$, followed by the false positives using fp/C.

The final column of Table 1 shows the biased exceptions costs for the five theories of Section 3. These are smaller than either the uniform or the divided encoding costs when fp is close to fn, but larger when the assumption of balanced errors is grossly incorrect.

In this example, MDL would now place T_1 and T_2 well ahead of the other theories, an intuitively sensible outcome.

5 APPLYING THE SCHEME TO C4.5RULES

C4.5RULES is a program that generates rule-based classifiers from decision trees [Quinlan, 1993]. The algorithm proceeds in three phases:

1. A rule *if* $L_1\&L_2\&...\&L_k$ *then class* X is formulated for each leaf of the decision tree, where X is the majority class at the leaf. The left-hand side initially contains every condition L_i that appears along the path from the root of the tree to the leaf, but rules are usually generalized by dropping one or more of these conditions. As a result, the rules are no longer mutually disjoint.
2. For each class in turn, all rules for that class are examined and a subset of them selected.
3. An order for these class rule subsets is then determined and a default class chosen.

The second phase in which a subset of rules is selected for each class is guided by MDL. Although the learning task may have any number of classes, every subset selection is essentially a two-class problem in which the goal is to cover cases of the class in question while not covering cases belonging to any other class. The description length of each candidate subset is determined as before by calculating its theory cost (to encode the constituent rules) and exceptions cost (to identify misclassified cases). The subset with the lowest description length is then chosen.[2]

The use of MDL in C4.5RULES fits squarely in the context addressed by this paper, since a rule subset is a categorical theory that characterizes one class against all other classes. If the new encoding is doing its job, it should lead to a better choice of rules for each class and, ultimately, to a more accurate classifier.

To test this hypothesis, two versions of C4.5RULES were prepared that differ only in the method used to calculate exceptions costs. One version uses the uniform strategy as set out in (1) since this has been found to be generally more robust than the divided strategy [Quinlan, 1994]. The biased version employs the new strategy of (3); like the uniform strategy, this also transmits a single global error count, but uses the initial assumption of equal numbers of false positive and false negative errors to derive separate error probabilities for covered and uncovered cases.

A comprehensive collection containing 25 real-world datasets was assembled from the UCI Repository. The intention was to cover the spectrum of properties such as size, attribute numbers and types, number of classes and class distribution, with no attempt to favor either coding strategy. A summary of their main characteristics is given in the Appendix.

One hundred trials were carried out with each dataset. In each trial, the data were split randomly into a training set (90%) and a test set (10%). Rule-based classifiers were learned from the training data using both versions of C4.5RULES above, and these classifiers were evaluated on the test data. Table 2 shows, for each dataset, the average over 100 trials of the respective error rates on the test data and numbers of rules retained. The final columns record the numbers of trials in which the biased and uniform exceptions costs led to a more accurate classifier.

There are several ways in which these results can be used to compare the coding strategies:

- The biased strategy gives a lower average error that the uniform approach in 20 of the 25 domains, the same error rate in two domains, and a higher error rate in three domains (credit approval, horse colic, and sonar).
- If the performance of a strategy on a dataset is judged instead by the number of trials on which it is superior, the biased coding wins on 19 domains, ties on one, and loses on five domains.
- The biased approach gives a more accurate classifier on 593 of the 2500 trials, versus 318 trials on which the uniform strategy comes out ahead.
- For a particular domain, the ratio of the average error rate using the biased strategy to that obtained with the uniform approach measures the extent of the benefit (values less than 1) or loss (values greater than one) associated with using the former. The values of this ratio range from 0.94 (splice junction) to 1.02 (sonar), the average across all domains being 0.97. On a new domain, then, use of the biased strategy with C4.5RULES would be expected to lead to a lower error rate than if the uniform strategy were adopted.
- When the above ratio is computed for just the trials on which the strategies give different numbers of errors on the test data, the average ratio is 0.93. If coding strategy matters for a trial, therefore, the biased coding approach should give an error rate considerably lower than that obtained by the alternative.

[2] If there are more than a few rules, the consideration of subsets is not exhaustive. From Release 6, C4.5 now carries out a series of greedy searches, starting first with no rules, then a randomly-chosen 10% of the rules, then 20%, and so on; each search attempts to improve the current subset by adding or deleting a single rule until no further improvement is possible. The best subset found in any of these searches is retained. This differs from Release 5, described in [Quinlan, 1993], in which simulated annealing is used to search for the best subset.

Table 2: Comparison of biased and uniform exceptions coding strategies implemented in C4.5RULES.

Dataset	Biased Coding		Uniform Coding		Trials Superior	
	Error	Rules	Error	Rules	Biased	Uniform
	(%)		(%)			
audiology	22.8	20.6	23.1	21.2	10	5
auto insurance	25.0	19.1	26.1	18.9	16	7
breast cancer (Wi)	4.5	8.5	4.5	8.6	2	2
chess endgame	7.1	21.9	7.4	21.1	31	20
Congress voting	4.5	6.3	4.7	6.4	4	1
credit approval	15.9	15.0	15.8	15.6	14	21
glass identification	30.4	13.2	31.4	12.6	19	10
heart disease (Cl)	23.1	11.1	23.1	11.2	7	9
hepatitis	18.8	6.5	19.3	6.2	8	6
horse colic	15.8	9.5	15.7	9.9	8	13
hypothyroid	0.56	9.8	0.59	9.8	19	11
image regions	4.0	28.1	4.1	27.4	28	23
iris	4.7	4.1	4.9	4.1	1	0
led digits	32.0	12.4	33.3	11.6	31	13
lymphography	19.4	9.9	19.6	9.6	10	5
nettalk (phoneme)	22.9	335	24.2	353	87	7
nettalk (stress)	16.7	229	17.5	253	68	32
Pima diabetes	27.6	13.3	27.7	13.1	34	37
primary tumor	60.1	17.0	63.1	11.5	56	21
promoters	16.5	8.2	16.9	8.2	5	1
sick euthyroid	1.3	13.9	1.4	16.6	31	22
sonar	31.1	7.0	30.7	7.5	6	13
soybean disease	8.1	34.6	8.3	34.0	24	10
splice junction	6.6	72.0	7.0	73.2	56	22
tic-tac-toe	7.5	21.3	7.6	21.7	8	7

- The number of rules retained is a rough indicator of the complexity of the final theory. In this respect there is no systematic difference between the strategies: the biased coding approach leads to fewer rules in 12 domains, the same number of rules in three domains, and more rules in 10 domains.

By any of the accuracy metrics, the biased strategy defined in (3) emerges as clearly preferable to the uniform strategy over these trials.

6 RELATED RESEARCH

The anonymous reviewers drew my attention to two alternative approaches to selecting categorical theories, both of which resemble MDL in trading off the accuracy of a theory against its complexity. Both consider families of *loss functions*, or criteria used to judge the appropriateness of the selected theory.

Selecting a theory to minimize categorical error rate, under the title of the *pattern recognition problem*, is one of the tasks considered by Vapnik [1982]. He first derives an upper bound on the error rate of a selected theory such that, with confidence 1-η, the true error rate of the theory will not exceed the bound. Besides factors such as η, the amount to training data $|D|$, and the observed error rate of the theory, this bound also depends on the *capacity* of the set of candidate theories – roughly, the largest amount of data that can be partitioned into two subsets in all possible ways by the theories. This is the basis for *structural risk minimization*: candidate theories are first grouped into a sequence of subsets with increasing capacity (e.g., by placing all theories with similar complexity in one subset). The best candidate in each subset is found and a final theory selected by choosing one of the subsets, either by minimizing the upper bound on the error rate or by estimating the value of the loss function for each subset using a leave-one-out cross-validation.

Barron [1991] is also concerned about problems arising from the use of MDL with general loss functions and develops an alternative strategy of *complexity regularization.* A theory is chosen to minimize the sum of the error rate and a complexity component; for categorical loss functions, this is

$$\frac{e}{|D|} + \lambda\sqrt{\frac{m}{|D|}} \qquad (4)$$

where m is the cost of encoding the theory and e is its number of errors on the training data D. So long as the constant λ has a value greater than $1/\sqrt{2\log(2.718)}$ or approximately 0.6, Barron shows that the expected penalty for choosing this theory approaches zero as $|D|$ increases. When this criterion (using λ=0.6) was tried with c4.5rules, however, results were quite poor – for these datasets, the error rate component is dominated by the complexity component and very few rules are selected.

One reviewer also pointed out that exceptions coding costs can often be reduced by quantizing the transmitted number of errors e. If e is expressed in units of $\sqrt{|D|}$, rounded to the nearest integer, the number of bits needed to encode the error count is approximately halved. This gain is offset by the fact that the message probabilities are now known with lower accuracy. However, such quantization does not appear to be advantageous in the application discussed here, representing as it does a windfall benefit to values of e for which the message probabilities do not change appreciably. Among the theories of Table 1, T_3 has the lowest biased encoding cost if quantization is employed. Further, when the above quantization scheme was implemented in c4.5rules, performance was degraded in almost all of the 25 domains.

7 CONCLUSION

Like its predecessor, this paper focuses on the common learning scenario in which a theory induced from a training set is used to classify an unseen case by predicting its class, rather than by determining the posterior probabilities of all classes. The straightforward application of the Minimum Description Length Principle in such situations can lead to anomalous choices among contending theories. Better choices are obtained by the addition of a bias towards theories whose probability of predicting each class is similar to the relative frequency of that class in the training data. Instead of relying on an artificial penalty function to implement this bias, as was the case in [Quinlan, 1994], we have presented a biased exceptions coding strategy that achieves the same effect in a manner more in tune with the MDL Principle itself.

The new scheme has been tested in a rule learning program c4.5rules and has been shown to lead to greater predictive accuracy in most of the domains investigated. The improvement is not dramatic but could be described as "useful". The biased scheme involves no additional computation and will be incorporated in the next release of the C4.5 software.[3]

The biased exceptions cost has also been tested independently by William Cohen on 37 domains that include only seven of the datasets reported here. His ripper2 rule induction system [Cohen, 1995] previously used a uniform coding strategy; when this was altered to the biased strategy, the latter proved superior on 17 domains and inferior on 13. The average ratio of the error rate using the biased encoding to that using the uniform encoding is 0.96, but one domain in which the error rate dropped to zero has an undue impact on this average. Excluding the highest and lowest value of the ratio, we obtain an average over the remaining datasets of 0.98, a more modest gain.

Finally, the particular strategy described in (3) is not the only way to exploit an expected balance between false positive and false negative errors. For instance, we could transmit the number of false positive errors, then estimate the probability of false negatives under the assumption that there are the same number of errors in the uncovered cases. It will be interesting to see whether alternative biased encoding schemes might be more beneficial still.

Acknowledgements

This work was supported by a grant from the Australian Research Council and assisted by research agreements with Digital Equipment Corporation. The anonymous reviewers helped me greatly by identifying relevant research in Statistics. I am grateful to William Cohen for providing the experimental results for ripper2. Thanks to Patrick Murphy and David Aha for maintaining to the UCI Repository, and to the Ljubljana Oncology Institute, Slovenia, for providing the lymphography and primary tumor data.

References

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1984). *Classification and regression trees.* Belmont, CA: Wadsworth.

Barron, A.R. (1991). Complexity regularization with application to artificial neural networks. In *Nonparametric Functional Estimation and Related Topics* (G. Roussas, Ed), Boston: Kluwer Academic Publishers, 561-576.

Cohen, W.W. (1995). Fast effective rule induction. *Proceedings 12th International Conference on Machine Learning*, Tahoe City, in this volume.

Georgeff, M.P. and Wallace, C.S. (1985). A general selection criterion for inductive inference. Technical Note 372, SRI International, Menlo Park.

Hunt, E.B., Marin, J., and Stone, P.J. (1966). *Experiments in Induction.* New York: Academic Press.

[3] Anyone who has C4.5 Release 5 (published by Morgan Kaufmann) can obtain an update to the latest version via anonymous ftp from `ftp.cs.su.oz.au`, file `pub/ml/patch.tar.Z`. This compressed `tar` file contains replacements for those source code files that have been changed since Release 5. The more recent releases incorporate several changes that affect the system's performance, so retaining a copy of the old files is recommended!

Michalski, R.S. (1980). Pattern recognition as rule-guided inductive inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2, 349-361.

Pagallo, G. and Haussler, D. (1989). Two algorithms that learn DNF by discovering relevant features. *Proceedings 6th International Workshop on Machine Learning*, Ithaca, 119-123. San Mateo: Morgan Kaufmann.

Pednault, E.P.D. (1991). Minimal-length encoding and inductive inference. In *Knowledge Discovery in Databases* (G. Piatetsky-Shapiro and W.J. Frawley, Eds), Menlo Park: AAAI Press, 71-92.

Quinlan, J.R. (1987). Simplifying decision trees. *International Journal of Man-Machine Studies*, 12, 221-234.

Quinlan, J.R. (1993). *C4.5: Programs for Machine Learning.* San Mateo: Morgan Kaufmann.

Quinlan, J.R. (1994). The Minimum Description Length Principle and categorical theories. *Proceedings 11th International Conference on Machine Learning*, New Brunswick, 233-241. San Francisco: Morgan Kaufmann.

Quinlan, J.R. and Rivest, R.L. (1989). Inferring decision trees using the Minimum Description Length Principle. *Information and Computation*, 80, 227-248.

Rissanen, J. (1983). A universal prior for integers and estimation by minimum description length. *Annals of Statistics*, 11, 416-431.

Rivest, R.L. (1987). Learning decision lists. *Machine Learning*, 2, 229-246.

Schaffer, C. (1993). Overfitting avoidance as bias. *Machine Learning*, 10, 153-178.

Vapnik, V. (1983). *Estimation of Dependences Based on Empirical Data.* New York: Springer-Verlag.

Wallace, C.S. and Patrick, J.D. (1993). Coding decision trees. *Machine Learning*, 11, 7-22.

Appendix: Summary of Datasets

The following provides a brief description of the datasets used in these experiments in terms of

- *size*, the number of instances in the dataset;
- *attributes*, the number and types of attributes involved (c=continuous-valued, b=binary, n=nominal); and
- the number of distinct *classes*.

Dataset	Size	Attributes	Classes
audiology	226	60b+9n	24
auto insurance	205	15c+10n	6
breast cancer (Wi)	699	9c	2
chess endgame	551	39b	2
Congress voting	435	16n	2
credit approval	690	6c+3b+6n	2
glass identification	214	9c	6
heart disease (Cl)	303	8c+3b+2n	2
hepatitis	155	cc+12b+1n	2
horse colic	368	10c+1b+11n	2
hypothyroid	3772	7c+20b+2n	5
image regions	2310	19c	7
iris	150	4c	3
led digits	200	7b	9
lymphography	148	18n	4
nettalk (phoneme)	5438	7n	47
nettalk (stress)	5438	7n	5
Pima diabetes	768	8c	2
primary tumor	339	17n	22
promoters	106	57n	2
sick euthyroid	3772	7c+20b+2n	2
sonar	208	60c	2
soybean disease	683	2b+33n	19
splice junction	3190	60n	3
tic-tac-toe	346	9n	2