BENEFIT MAXIMIZATION IN CLASSIFICATION ON FEATURE PROJECTIONS

H. Altay Güvenir
Bilkent University
Computer Engineering Department, Ankara
Turkey

Abstract

In some domains, the cost of a wrong classification may be different for all pairs of predicted and actual classes. Also the benefit of a correct prediction is different for each class. In this paper, a new classification algorithm, called BCFP (for *Benefit Maximizing Classifier on Feature Projections*), is presented. The BCFP classifier learns a set of classification rules that will predict the class of a new instance with maximum benefit or minimum cost. BCFP represents a concept in the form of feature projections on each feature dimension separately. Classification in the BCFP algorithm is based on a voting among the individual predictions made on each feature. A genetic algorithm is used to select the relevant features. The performance of the BCFP algorithm is evaluated in terms of accuracy. As a case study, the BCFP algorithm is applied to the problem of diagnosis of gastric carcinoma. A lesion can be an indicator of one of nine different levels of gastric carcinoma. The benefit of correct classification of early levels is much more than that of late cases. Also, the cost of wrong classifications is different for all classes.

Key Words

Machine learning, feature projection, voting, benefit maximization

1. Introduction

Classical classification algorithms aim to maximize the number of correct classifications, or, in other words, minimize the number of incorrect classifications. However, in some domains, the cost of a wrong classification is different for each predicted/actual class pair. Also the benefit of correct prediction is different for each class. In this paper we propose an inductive classification learning algorithm, called Benefit Maximizing Classifier on Feature Projections (BCFP). BCFP is based on a knowledge representation technique, called *feature projections*, which has been successfully employed in CFP [1]. As a case study, we show its application to a medical dataset to diagnose the gastric tumors.

The input to the BCFP training algorithm is a set of training instances. Learning from the training examples, BCFP constructs a representation of the classification knowledge inherent in these examples. This knowledge is represented as the projections of the training dataset as feature intervals on each feature dimension separately. For each feature dimension, projection points with similar characteristics are grouped into intervals. Therefore, an interval is a generalization that represents a set of feature values that yield the same classifications. Classification in the BCFP algorithm is based on a voting mechanism among the individual predictions made on each feature. Since each feature participates independently of the others, both in learning and classification, BCFP enables an easy and natural way of handling missing feature values by simply ignoring them.

Other machine learning algorithms using feature projection based knowledge representation were successfully applied to medical domains. For example, an expert system named DES was implemented for differential diagnosis of erythemato-squamous diseases in dermatology [2] based on the VFI (Voting Feature Intervals) technique [3]. These classification systems, however, are not designed for cost-sensitive classification domains. Therefore they do not work on domains, where the benefit of correct classification is different for each class; also the cost of wrong classification is different for all pairs of predicted and actual classes.

The next section presents the BCFP algorithm. Section 3 describes the gastric carcinoma domain, and presents the results of the application of the BCFP algorithm to the gastric carcinoma domain. Also the BCFP algorithm is compared with the performance of the medical students specializing on gastroenterology. Finally, the last section concludes with some remarks and suggestions for feature work.

2. The BCFP algorithm

The BCFP algorithm is the classification cost sensitive version of the feature projection based classification algorithms family [1]. In the following subsections, the knowledge representation used in the BCFP algorithm, training, and classification algorithms will be explained through a simple example. Then, the feature selection using a genetic algorithm will be described.

2.1 Knowledge Representation

Each training example is represented by a vector of nominal (discrete) or linear (continuous) feature values plus the class label. The BCFP classification algorithm represents a concept description by a set of feature intervals. An interval is either a range or a point interval. A range interval is a set of consecutive values of a given feature, whereas a point interval is defined as a single feature value.

For range intervals, lower and upper bounds of the range value and the votes for each class are maintained. For point intervals, on the other hand, the lower and upper values are the same. Therefore, an interval is represented as a vector, whose first two elements store the lower and upper bounds and the remaining elements correspond to the votes for each class, as shown below:

$$\langle lb, ub, V_1, V_2, ... V_k \rangle .$$

Here, k is the number of classes in the domain, and V_i represents the vote of the interval for class C_i.

2.2 Training

The training process of the BCFP algorithm is shown in Fig. 1. For each feature *f*, all training instances are sorted with respect to their values for *f*, forming their projections on *f*. A point interval is constructed for each projection. The lower and upper bounds of the interval are equal to the value of feature *f* in the corresponding training instance. Given the normalized benefit table *NB*, the vote V_p of a class *p* is initialized as

$$V_p = \begin{cases} \frac{1}{N} \sum_{c=1}^{k} N_c \times NB[p,c] & \text{if } N_p > 0 \\ 0 & \text{otherwise} \end{cases} .$$

Here N is the total number of instances in the interval, N_c is the number of class c instances in the interval, and $NB[p,c]$ is the normalized benefit of classifying a class c instance as p. In other words, V_p is the average benefit to be gained by classifying all the instances in that interval as class p. If no instances of class p have been observed in that interval, then the vote for class p is 0. In order for an equal voting power for each interval, during querying, the votes of an interval are normalized later, so that

$$\sum_{p=1}^{k} V_p = 1 \cdot$$

If the *f* value of a training instance is unknown (represented by “?"), it is simply ignored for this feature *f*. Then, only for linear features, BCFP tries to generalize the point intervals. Consecutive point intervals whose highest votes are for the same class are joined, forming range intervals.

An indecisive interval, which distributes its vote among all classes evenly, is uninteresting and it should be removed. We call a rule *decisive* if the standard deviation of its votes is above a minimum threshold, called $s_{\min}$. The BCFP algorithm uses $s_{\min} = \frac{1}{k-1}\sqrt{\frac{1}{k}}$. This threshold is equal to the standard deviation when the interval casts 0 votes for one class, and distributes its vote evenly among all other classes.

```
train (TrainingSet):
begin
    for each feature f
        /* sort TrainingSet with respect to f */
        sort (f, TrainingSet)
        /* construct a list of point */
        interval_list ← make_intervals (f, TrainingSet)
        if f is linear
            /* join adjacent point intervals to form
               range intervals */
            interval_list ← join_interval(interval_list)
end.

join_interval (interval_list)
begin
    I = first interval in interval_list
    while I is not empty do
        I' is the interval following I
        if beneficial_class(I) = beneficial_class(I')
            /* beneficial_class of an interval is the class with the highest votes */
            merge I' into I
        else I ← I'
end.
```

Fig. 1. Training algorithm of BCFP.

An example training data set and the corresponding feature intervals constructed by the BCFP algorithms are shown in Fig. 2. The example domain consists of three features, namely f1, f2, and f3, the first two of which are linear and the last one is a nominal feature. The nominal feature f3 can take values from the set {A, B, C}. The class labels are C1, C2, and C3. There are seven training instances in this example. Training algorithm forms three intervals on the feature f1, two of which are range intervals. The first interval on f1, spans the value range [1,3], and it votes only for the C1 class.

2.3 Classification

The classification (querying) process in the BCFP algorithm is given in Fig. 3. The process starts by initializing the votes of each class to zero. The classification operation includes a separate pre-classification step on each feature.

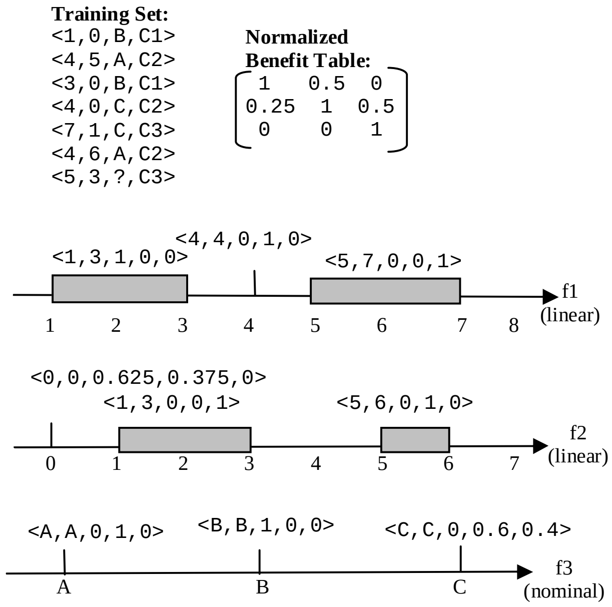

Fig. 2. Feature intervals formed for a sample training set.

The pre-classification on feature f involves a search for the interval on feature dimension f into which q_f falls, where q_f is the value of the query instance q for feature f. If that value is unknown (missing), then that feature does not participate in the voting. Hence, the features containing missing values are simply ignored.

If the q_f value is known, the interval I into which q_f falls is searched. If the q_f value does not fall in any interval on f, then again the feature f does not participate in the voting, which means that that value for the feature f has not been observed in the training set. If an interval I is found that includes the q_f value, then the votes of I are the votes that f casts in the voting. Since the sum of the votes of an interval is normalized to 1 during the training, each feature has an equal power in the voting.

Finally, the class that receives the highest amount of votes is returned as the predicted class of the query instance q. Although a single class returned as the prediction of the query instance, the votes received by each class are also made available to the user, so that the level of the confidence of this prediction can be measured. The *benefit accuracy* of the classification is obtained directly from the normalized benefit table. If the actual class of q is q_c, and the predicted class is p, then the accuracy is $NB[p, q_c]$. Note that the classical definition of predictive accuracy, as the ratio of number of correct classifications over the total number of test instances, is a special case obtained by setting all the diagonal entries of the *NB* table to 1, and all other entries to 0. Therefore, in the rest of the paper, the term accuracy will be used to refer the benefit accuracy.

```
classify(q): /* q: query instance to be classified */
begin
    /* initialize total votes */
    for each class c vote[c] = 0
    for each feature f
        if q_f value is known
            I = search_interval(f; q_f)
            if I is not empty
                for each class c
                vote[c] = vote[c] +  interval_vote(I ; c)
    return the class c, such that vote[c] is maximum.
end.
```

Fig. 3. Classification in the BCFP algorithm.

Continuing on the example in Fig. 2**,** consider the classification of a query instance q=<2,5,C>. The intervals corresponding to the query instance are shown in Fig. 4. The total votes for classes C1, C2 and C3 are 1, 1.6 and 0.4, respectively. The C2 class received the highest amount of votes. Therefore, C2 is the predicted class of that query instance. The confidence of this prediction is 1.6 / (1+1.6+0.4) = 53%.

```
Query <2, 5, C>
Feature: f1, q1=2, I1 = <1,3,1,0,0>
Feature: f2, q2=5, I2 = <5,6,0,1,0>
Feature: f3, q3=C, I3 = <C,C,0,0.6,0.4>
Total votes: <1, 1.6, 0.4>
Prediction: C2
```

Fig. 4. Classification example on the sample data set.

2.4 Feature Selection using a Genetic Algorithm

The performance and of classification is sensitive to the choice of the features used to construct the classifier. A natural and safe approach in inductive machine learning is to provide all available features, and let the machine learning system to determine and use only the relevant ones in classification. The problem of identifying the relevant subset of features in the data is called *feature subset selection*. Exhaustive evaluation of possible feature subsets is usually infeasible in practice because of the large amount of computational effort required. Genetic algorithms offer an attractive approach to find near-optimal solutions to such optimization problems [4].

A genetic algorithm attempts to find a good solution to the problem by genetically breeding a population of individuals over a series of generations. Each individual in the population represents a candidate solution to the given problem. The genetic algorithm transforms a population of individuals, each with an associated fitness value, into a new generation of the population using reproduction, crossover, and mutation [5].

We have coupled the BCFP algorithm with a genetic algorithm using the wrapper approach for feature subset selection [6]. A gene in the chromosome represents each feature. Therefore, the chromosome size is equal to the number of features. The genetic algorithm used with BCFP employed a two-way crossover operation. The fitness of a chromosome is computed as the 10-fold cross-validation accuracy of the BCFP algorithm. In the experiments, the population size was 500. The probability of crossover and probability of mutation were p_c = 0.9, and p_m = 0.001, respectively. The genetic algorithm was run for 500 generations.

3. A Case Study: The Gastric Carcinoma Domain

Cancer of the stomach, also called *gastric cancer*, is a disease in which cancer (malignant) cells are found in the tissues of the stomach. Sometimes cancer can be in the stomach for a long time and can grow very large before it causes any symptoms. In the early stages of the stomach cancer, a patient may have indigestion and stomach discomfort, a bloated feeling after eating, mild nausea, loss of appetite, or heartburn. In more advanced stages of cancer of the stomach, the patient may have blood in the stool, vomiting, weight loss, or pain in the stomach. Stomach cancer is difficult to detect in its early stages because its early symptoms are absent or mild. Unfortunately, this is a highly aggressive cancer and overall survival rate is very low. The chance of recovery (prognosis) and the choice of treatment depend on the stage of the cancer, whether it is just in the stomach or if it has spread to other places, and the patient's general state of health.

According to a report published by the Gastroenterology department of the Ankara University School of Medicine, the stomach cancer is the second most frequent type of cancer in men, and the third one in women [7].

3.1 The Stomach

The stomach is separated into upper, middle and lower portions. When the cancer infiltration (penetration) is limited in one of the three main portions, this is expressed by indicating C (Fundus, upper part), M (Body, middle part) and A (Antrum, lower part). The other possible locations are E (Esophagus) and D (duodenum). The cancer tumor placement also identified by the cross-sectional positioning.

3.2 Classification of Gastric Cancers

If there are symptoms of cancer, a physician will usually order an upper gastrointestinal x-ray or he may also look inside the stomach with a gastroscope. This procedure is called gastroscopy, and it is useful in the detection of most stomach cancers. According to the Japanese Gastroenterological Endoscopy Society, based on the visual inspection of the mucosal surface of the patient's stomach, gastric cancers are classified mainly into three categories as shown in Table 1. They are Early Gastric Cancers (EGC) and Advanced Gastric Cancers (AGC) and the remaining ones which cannot be included to these categories [8].

Table 1. Classification of Gastric Cancers.

Type	Classification
Type 0	Early Gastric Cancer (EGC)
Type 1	Advanced Gastric Cancer (AGC)
Type 2	
Type 3	
Type 4	
Type 5	The cancers that cannot be included under any of the above types

Early gastric cancer is defined as gastric cancer confined to the mucosa or submucosa, regardless of the presence or absence of lymph node metastasis as shown in Table 2 [9].

Table 2. Types of early gastric carcinoma.

Type	Properties
I	Exophytic, protruded
IIa	Superficially elevated
IIb	Even, flat
IIc	Superficially depressed
III	Excavated

On the other hand, in advanced gastric cancers, as defined by Bormann, the tumor is invaded into the proper muscle layer beyond the stomach [10]. Moreover, knowledge of these types permits a preliminary assessment of tumor spread. According to Bormann Classification AGC's are divided into four groups, Bormann I, Bormann II, Bormann III, and Bormann IV.

3.3 The Gastric Carcinoma Data Set

The Gastric Carcinoma data set used in this paper consists of 285 gastric cancer records. These recordings consist of 209 male and 67 female (9 missing sex information) patients with age ranging from 26 to 85.

3.3.1 Classes

The cancers that are classified in this domain are labeled as C1 through C9 as Early I (C1), Early IIa (C2), Early IIb (C3), Early IIc (C4), Early III (C5), BI (C6), BII (C7), BIII (C8), and BIV (C9). The data set contains 174 early and 111 advanced gastric cancer patients. The distribution of the record set among the diseases is shown in Table 3.

3.3.2 Features

Patient records collected for diagnosis and prognosis typically contain values of clinical and histopathological investigations. The features used in this domain are represented as a vector of 68 features. Seven of these features are linear valued and the others are categorical.

The data set contains 970 missing feature values, which means that 5% of the data set is missing.

Table 3. The distribution of classes in the data set.

Type	Class	Number of Patients
Early Gastric Cancers		174
Early I	C1	3
Early IIa	C2	55
Early IIb	C3	7
Early IIc	C4	103
Early III	C5	6
Advanced Gastric Cancers		111
BI	C6	6
BII	C7	17
BIII	C8	69
BIV	C9	19
TOTAL		285

3.3.3 Benefits and Costs

An important characteristic of the gastric carcinoma data set is that the benefit of correct classification depends on the class value. In this domain, benefit of correct classification of an early stage of a tumor is more than that of a later stage. For an incorrect classification, depending on the predicted and actual class values, a different cost is incurred. If the predicted class label is similar to the actual class, still a benefit is obtained. All this information is provided as a benefit table. The benefit table used in this experiment is given in Table 4. Positive values indicate benefits, while negative values indicate costs. The entry B[*p*,*a*] represents the benefit of predicting class *p* when the actual class is *a*. According to this table, classifying a C1 instance correctly provides 18 units of benefit, while classifying a C9 instance correctly provides only 5 units of benefit. On the other hand, predicting a C1 instance as C6 incurs 4 units of cost. However, incorrectly classifying a C7 instance as a similar class C6 still provides 2 units of benefit.

The benefit and cost values are difficult to measure and most of the time they are subjective. The amount of benefits and costs can be measured according to a combination of many criteria. In medical domains, the most important one is the possibility of saving the patient's life; the earlier the diagnosis, the longer survival. Other criteria may include the cost and the alternatives of the treatment procedure, which are inverse proportional with the benefit.

The entries of the benefit table can be set up using any measuring unit meaningful to the domain experts. In order to eliminate the effects of the measuring unit chosen, the BCFP algorithm initially normalizes the entries of the benefit table to the [0,1] range, so that the benefit of a correct classification is always 1, and the benefit of the most costly prediction is always 0.

Table 4. Benefits table for the gastric carcinoma domain. Negative values indicate costs.

	Actual Class								
Prediction	C1	C2	C3	C4	C5	C6	C7	C8	C9
C1	18	6	6	6	-1	-10	-12	-15	-20
C2	10	15	12	12	4	-8	-10	-13	-15
C3	10	12	15	12	4	-8	-10	-13	-15
C4	10	12	12	15	4	-8	-10	-13	-15
C5	5	7	7	7	10	-3	-8	-11	-13
C6	-4	-3	-3	-3	-1	8	2	1	-1
C7	-6	-5	-5	-5	-3	4	7	4	2
C8	-12	-10	-10	-10	-8	1	3	6	3
C9	-20	-15	-15	-15	-11	-6	1	3	5

3.4 Results

The BCFP algorithm and the accompanying genetic algorithm for feature selection have been implemented and experimented on the gastric carcinoma domain. In measuring the performance of the BCFP algorithm we used 10-fold cross-validation accuracy. This technique ensures that the training sets are disjoint, and each instance in the data set is classified exactly once.

Using all of the 68 features of the data set, the BCFP algorithm achieved 85.7% accuracy. However, the feature selection algorithm chose only 32 of the 68 features as relevant for a beneficial classification. With the selected set of features the BCFP algorithm achieved 94.3% accuracy. Some of the rules induced by the BCFP algorithm are shown in Fig. 5. The numbers following the class labels indicate the votes of each corresponding class.

```
If 1≤ depth ≤2 then
   C1/0.14 C2/0.18 C3/0.17 C4/0.18
   C5/0.14 C6/0.08 C7/0.07 C8/0.04 C9/0
If depth = 3 then
   C1/0 C2/0.13 C3/0 C4/0.14 C5/0 C6/0
   C7/0.37 C8/0.36 C9/0
If 4≤ depth ≤5 then C1/0 C2/0.04
   C3/0 C4/0.05 C5/0.07 C6/0.19 C7/0.22
   C8/0.23 C9/0.20
If flower bed app = Present then
   C1/0 C2/1 C3/0 C4/0 C5/0 C6/0 C7/0
   C8/0 C9/0
if infiltrated ulcer = Present then
   C1/0 C2/0 C3/0 C4/0.06 C5/0 C6/0
   C7/0.32 C8/0.33 C9/0.29
If erosion = Present then C1/0.24
   C2/0.32 C3/0 C4/0.29 C5/00 C6/0
   C7/0.15 C8/0 C9/0
```

Fig. 5. Sample rules induced by BCFP.

The rules constructed by the BCFP algorithm are easy to be verified by experts. According to these rules, if the depth of the lesion is 1 (mucosa) or 2 (sub mucosa), then it is more likely that the case is an early gastric cancer; while if the depth is 4 (subserosa) or 5 (serosa) then advanced gastric cancer is more certain. If the lesion has a flower bed appearance, then it is certainly Early IIa. On the other hand, if the infiltrated ulcer is present, then the case is either BII, BIII or BIV. The other rules can be interpreted in the similar manner.

In order to see how difficult it is to make a prediction with high benefits, we have conducted an experiment with 16 fellows on internal medicine. The students were shown only the data set that was used by the BCFP algorithm. As a group, the fellows' accuracy was 63%. This indicates that making accurate decision in the diagnosis of the gastric carcinoma is quite difficult.

4 .Conclusions

In this paper, a new classification algorithm, called BCFP, has been developed and applied to the diagnosis of gastric carcinoma tumors. The BCFP algorithm aims to maximize the benefit of classification, reducing the cost of possible misclassifications. It uses the feature projections based knowledge representation.

The feature projections based knowledge representation allows the BCFP algorithm to process each feature separately. The missing feature values of instances are simply ignored, and only the known values are used both in training and querying. The classification model is constructed only using the known feature values of the training instances. Also the class of a query instance is predicted by considering only the given values of the features. This feature of the BCFP algorithm makes it robust to missing feature values. Another advantage of using the feature projections as the knowledge representation is that the constructed rules are based on a single feature and an associated set of values. Therefore, the rules are simple and easy to be verified by a human expert. The rules constructed for the gastric carcinoma dataset have been verified and found to be correct by the expert gastro-enterologists.

The BCFP algorithm is applicable, in particular, to concepts where each feature, independent of the other features, can be used in the classification. One might think that this requirement may limit the applicability of BCFP, since in some domains the features might be dependent on each other. Holte has pointed out that the most real-world classification tasks are such that their attributes can be considered independently of each other [12].

The BCFP algorithm achieved very good accuracy on the gastric carcinoma dataset. The result was even better than the medical students specializing on internal medicine. This showed us that the differential diagnosis of gastric carcinoma classes is quite difficult even for medical doctors. We used a genetic algorithm for selecting the relevant features. With selected features the BCFP algorithm achieved excellent classification accuracy.

The BCFP algorithm constructs a rule for each interval formed by the projections of training instances on features. The votes of an interval to the class labels are based on the number of training instances with that class value falling in that interval, and the entries of the benefit table. A rule which gives similar votes to each class does not make any difference in the final classification of the query instance. Such a rule is usually uninteresting to the domain expert. Therefore it can be discarded from the model. As a future work we plan to develop a system that can measure the interestingness of a rule constructed by the BCFP algorithm. Selecting only the interesting rules may provide a domain expert with some pointers for further experiments and research ideas.

Acknowledgements

This project is supported by TUBITAK (Scientific and Technical Research Council of Turkey) under Grant 101E044.

References

[1] H.A. Güvenir, İ. Şirin, Classification by Feature Partitioning. *Machine Learning*, 23(1), 1996, 47-67.

[2] H.A. Güvenir, N. Emeksiz, An Expert System for the differential diagnosis of erythemato-squamous diseases. *Expert Systems with Applications*, 18(1), 2000, 43-49.

[3] H.A. Güvenir., G. Demiröz, N. Ilter, Learning Differential Diagnosis of Erythemato-Squamous Diseases using Voting Feature Intervals. *Artificial Intelligence in Medicine*, 13(3), 1998, 147-165.

[4] J. Yang, V. Honavar, Feature Subset Selection Using a Genetic Algorithm. *Feature Extraction, Construction, and Subset Selection: A Data Mining Perspective*. Motoda, H. and Liu, H. (Ed.) New York: Kluwer. 1998.

[5] D. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley, New York, 1989.

[6] G. John, R. Kohavi, and K. Pfleger, Irrelevant Features and the Subset Selection Problem. *Proc. of the Eleventh International Conference on Machine Learning*, New Brunswick, NJ, Morgan Kaufmann, 1994:121-129.

[7] Örmeci N, Demirci S, Tulunay Ö, Kuzu I, Akgül H, Uzunalimoğlu Ö, Yol S. Early Stomach Cancer in Turkey. Recent Advances in Management of Digestive Cancers. 1993:339-341.

[8] Kajitani, T., Clinical Classification. *Japanese Journal of Surgery*, 11, 1981, 127-139.

[9] Kurihara H. Detection of Early Gastric Cancer Outside the Mass Screening Program. Japanese Journal of Clinical Oncology, 1998:233-233.

[10] Yokota, T. Bormann's type IV gastric cancer: clinicopathologic analysis. Canadian Journal of Surgery, 1999;42:371-6.

[11] Demiröz G, Güvenir, HA, Classification by Voting Feature Intervals, Proc. of Ninth ECML, Springer-Verlag, LNAI 1224, 1997:85-92.

[12] Holte RC. Very simple classification rules perform well on most commonly used datasets. Machine Learning 1993;11:63-91.