Artificial Intelligence in Medicine (2004) **31**, 231—240

ARTIFICIAL INTELLIGENCE IN MEDICINE

http://www.intl.elsevierhealth.com/journals/aiim

Diagnosis of gastric carcinoma by classification on feature projections

H. Altay Güvenir[a,*], Narin Emeksiz[b], Nazlı İkizler[a], Necati Örmeci[c]

[a]*Computer Engineering Department, Bilkent University, 06800 Ankara, Turkey*
[b]*Central Bank of the Republic of Turkey, Ankara, Turkey*
[c]*Medical School, Ankara University, Ankara, Turkey*

Received 18 September 2002; received in revised form 23 October 2003; accepted 10 March 2004

KEYWORDS
Machine learning; Voting; Feature projection; Benefit maximization; Gastric carcinoma

Summary A new classification algorithm, called *benefit maximizing classifier on feature projections* (BCFP), is developed and applied to the problem of diagnosis of gastric carcinoma. The domain contains records of patients with known diagnosis through gastroscopy results. Given a training set of such records, the BCFP classifier learns how to differentiate a new case in the domain. 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. In the gastric carcinoma domain, a lesion can be an indicator of one of nine different levels of gastric carcinoma, from early to late stages. The benefit of correct classification of early levels is much more than that of late cases. Also, the costs of wrong classifications are not symmetric. In the training phase, the BCFP algorithm learns classification rules that maximize the benefit of classification. In the querying phase, using these rules, the BCFP algorithm tries to make a prediction maximizing the benefit. A genetic algorithm is applied to select the relevant features. The performance of the BCFP algorithm is evaluated in terms of accuracy and running time. The rules induced are verified by experts of the domain.

1. Introduction

In clinical medicine, reaching a conclusion about a patient's symptoms, when presented with complex and sometimes contradictory clinical information, is really difficult. A clinician usually makes decisions based on a set of measurements and observations about a patient. He evaluates all the factors subjectively in order to reach a diagnosis. However, it is obvious that clinicians may have great difficulty in analyzing enormous amount of clinical and histopathological data. Therefore, more sophisticated quantitative techniques are needed to help clinicians to consider all the data and make better diagnoses. Some sophisticated quantitative techniques are proposed to the doctors by computer scientists via machine learning techniques in order to help in this decision making process. In this paper we propose an inductive supervised learning algorithm called benefit maximizing classifier on feature projections (BCFP) applied to a medical data set in order to diagnose the gastroenterological tumors. BCFP is based on a technique, called *feature projections*, which has been successfully employed in classification by feature partitioning (CFP) [1].

The input to the BCFP training algorithm is a set of training instances that are the descriptions of subjects with known diagnoses. Learning from the

[*] Corresponding author. Tel.: +90-312-2901252; fax: +90-312-2664047.
E-mail address: guvenir@cs.bilkent.edu.tr (H. Altay Güvenir).

0933-3657/$ — see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.artmed.2004.03.003

training examples, BCFP constructs a representation of the classification knowledge inherent in these examples. This knowledge is represented as the projections of the training data set 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 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 voting feature intervals (VFI) technique [3]. CFI was applied to the diagnosis of cardiac arrhythmia from standard 12 lead Electrocardiograph (ECG) recordings [4]. These classification systems, however, are not designed for cost-sensitive classification domains. Therefore they do not work on the gastric carcinoma domain, where the benefit of correct classification of early stages is more than that of later stages; also the cost of wrong classification is different for all pairs of predicted and actual classes.

The next section describes the gastric carcinoma domain in detail. Section 3 explains the BCFP algorithm. Section 4 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 future work.

2. 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. The stomach is a J-shaped organ in the upper abdomen where the food is digested. Food reaches the stomach through a tube called the esophagus that connects the mouth to the stomach. After leaving the stomach, partially digested food passes into the duodenum then the small intestine and then into the large intestine called the colon.

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. Some factors that increase the chances of getting stomach cancer are a stomach disorder, called atrophic gastritis, disorder of the blood, called anemia, or a hereditary condition of growths, called polyps, in the large intestine. 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.

Gastric cancer is the seventh most frequent cause of cancer mortality in the United States. In 2000, approximately 21,500 citizens were diagnosed with gastric cancer and 13,000 of them died [5]. The disease is much more common in other countries, principally Japan, Central Europe, Scandinavia, Hong Kong, South and Central America, the Soviet Union, China, and Korea. In fact, it is a major cause of death worldwide especially in developing countries. 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; also, it has been encountered as the most common type of tumor in gastrointestinal system in Turkey [6]. The Japanese Research Society for Gastric Cancer, because of the high rates of stomach cancer, has carried out a very strict screening program. This has enabled them to identify the disease in very early stages.

2.1. 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 location of the tumor is essential for selecting the appropriate surgical procedure. The cancer tumor placement also identified by the cross-sectional positioning.

The cross-sectional circumference of the stomach is separated into four parts, the lesser curvature, the greater curvature, the anterior wall, and the posterior wall [7].

2.2. 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 thin, lighted tube called a gastroscope. This procedure is called gastroscopy, and it is useful in the detection of most stomach cancers. For this test, the gastroscope is inserted through the mouth and guided into the stomach and the stomach mucosa is examined. 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 [7].

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 [8].

On the other hand, in advanced gastric cancers, as defined by Bormann, the tumor is invaded into the proper muscle layer beyond the stomach [9]. Moreover, knowledge of these types permits a preliminary assessment of tumor spread. According to Bormann Classification AGCs are divided into four groups, Bormann I, Bormann II, Bormann III, and Bormann IV, as shown in Table 3.

Table 1 Classification of gastric cancers

Type	Classification
0	Early gastric cancer (EGC)
1—4	Advanced gastric cancer (AGC)
5	The cancers that cannot be included under any of the above types

Table 2 Types of early gastric carcinoma

Type	Properties
I	Exophytic, protruded
Iia	Superficially elevated
Iib	Even, flat
Iic	Superficially depressed
III	Excavated

Table 3 Advanced gastric cancer classification

Type	Properties
BI	Mainly exophytic growth, usually broad-based polypoid carcinomas with a protruding, papillary, cauliflower-shaped or villous surface
BII	Carcinoma with a central, bowl-shaped ulceration, elevated margins, the carcinoma being relatively sharply delineated from its surroundings
BIII	Centrally ulcerating carcinoma without ridged, elevated margins and indistinctly delineated from its surroundings
BIV	Diffuse tumor infiltration of the gastric wall

2.3. 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 (nine missing sex information) patients with age ranging from 26 to 85.

2.3.1. 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 actual class labels of the instances in the data set are determined by the pathology test results. 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 4.

2.3.2. 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

Table 4 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

features [6]. Seven of these features are linear valued and the others are categorical. The listing of the features is given in Table 5. The data set contains 970 missing feature values, which means that 5% of the data set is missing.

2.3.3. 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 6. 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 five units of benefit. On the other hand, predicting a C1 instance as C6 incurs four units of cost. However, incorrectly classifying a C7 instance as a similar class C6 still provides two units of benefit.

The benefit and cost values are difficult to measure and most of the time they are subjective [10]. 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, by subtracting the largest cost from all the values of a column of an actual class, therefore the benefit of the most costly prediction is always 0. This guarantees that all benefit values are positive. Such a benefit matrix is called, the *normalized benefit table*. The normalized benefit table used by the BCFP algorithm for this domain is given in Table 7.

3. The BCFP algorithm

The BCFP algorithm is the classification cost sensitive version of the feature projection based classification algorithms family [11]. 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.

3.1. 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:

$$\langle \mathrm{lb}, \mathrm{ub}, V_1, V_2, \ldots, V_k \rangle.$$

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

3.2. Training

The training in 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

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

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 to

Table 5 Features in the data set

Index	Name	Values	Index	Name	Values
F_1	Age	1—100	F_{35}	Surface mucosa nodularity (granular)	Absent, present
F_2	Sex	Male, female	F_{36}	Surface mucosa uneven	Absent, present
F_3	Blood type	A, B, 0, AB	F_{37}	Surface mucosa infiltrated	Absent, present
F_4	Rh factor (blood)	+, −	F_{38}	Erithematous changes of surface	Absent, present
F_5	Smoking habit (cigarette/day)	0—100	F_{39}	Surrounding mucosa	Normal, abnormal
F_6	Number of years smoking	0—100	F_{40}	Surrounding mucosa atrophy	Absent, present
F_7	Alcohol consumption (cc/day)	0—1000	F_{41}	Surrounding mucosa irregularity	Absent, present
F_8	How long years (drinking)	0—100	F_{42}	Surrounding mucosa infiltration	Uninfiltrated, infiltrated
F_9	Alcohol type	Beer, whisky, wine, brandy	F_{43}	Surrounding mucosa redness/hyperemic	Absent, present
F_{10}	Atrophy	C_1, C_2, C_3, O_1, O_2, O_3	F_{44}	Surrounding mucosa spotty redness	Absent, present
F_{11}	Intestinal metaplasia	+, −	F_{45}	Erithematous changes of surrounding mucosa	Absent, present
F_{12}	Section	E, C, M, A, D	F_{46}	Surrounding mucosa margin irregularity	Absent, present
F_{13}	Curvature	Lesser, greater, whole	F_{47}	Surrounding mucosa discoloration	Absent, present
F_{14}	Wall	Anterior, posterior, whole	F_{48}	Bleeding	Absent, present
F_{15}	Depth	1 (mucosa), 2 (submucosa), 3 (proper mucosa), 4 (subserosa), 5 (serosa)	F_{49}	Erosion	Absent, present
F_{16}	Ulcer scar present	0—5	F_{50}	Elevated lesion	Absent, present
F_{17}	Ulcerization	Absent, present	F_{51}	Mucosal elevation	Wooden map type elevation, slight elevation, no elevation
F_{18}	Ulcer present	Absent, present	F_{52}	Base granularity/nodularity	Absent, present
F_{19}	Cauliflower appearance	Absent, present	F_{53}	Base infiltrated	Absent, present
F_{20}	Flower bed appearance	Absent, present	F_{54}	Base irregularity	Absent, present
F_{21}	Deep ulcer	Absent, present	F_{55}	Base redness	Absent, present
F_{22}	Big ulcer	Absent, present	F_{56}	Base spotty redness	Absent, present
F_{23}	Irregular ulcer	Absent, present	F_{57}	Base smoothness (normal)	Absent, present
F_{24}	Infiltrated ulcer	Absent, present	F_{58}	Wide base	Absent, present
F_{25}	Irregular margin of ulcer	Absent, present	F_{59}	Protrusion	Absent, present
F_{26}	Mucosal fold	Absent, present	F_{60}	Polyp	Absent, present
F_{27}	Club or rod like thickening	Absent, present	F_{61}	Polypoid lesion	Absent, present
F_{28}	Abrupt disruption	Absent, present	F_{62}	Sessile	Absent, present
F_{29}	Shallow ulcer	Absent, present	F_{63}	Stomach deformation	Absent, present
F_{30}	Surface mucosa redness	Absent, present	F_{64}	Whole stomach infiltration	Absent, present
F_{31}	Surface mucosa spotty redness	Absent, present	F_{65}	Pylica gastrica sickness	Absent, present
F_{32}	Surface mucosa smoothness (normal)	Absent, present	F_{66}	Depressed area	Absent, present
F_{33}	Surface mucosa discoloration	Absent, present	F_{67}	Irregular margin of depressed area	Absent, present
F_{34}	Surface mucosa irregularity	Absent, present	F_{68}	Infiltrated mass	Absent, present

Table 6 Benefits table for the gastric carcinoma domain

Prediction	Actual class								
	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

Negative values indicate costs.

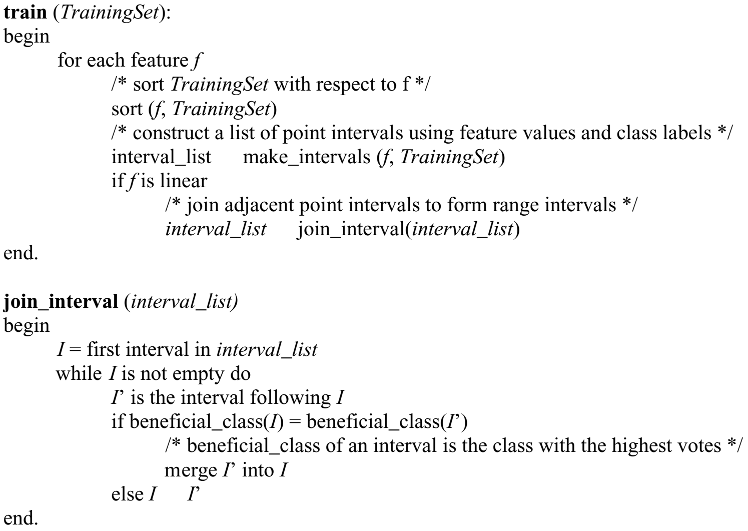
Table 7 Normalized benefit table for the gastric carcinoma domain

Prediction	Actual class								
	C1	C2	C3	C4	C5	C6	C7	C8	C9
C1	38	21	21	21	10	0	0	0	0
C2	30	30	27	27	15	2	2	2	5
C3	30	27	30	27	15	2	2	2	5
C4	30	27	27	30	15	2	2	2	5
C5	25	22	22	22	21	7	4	4	7
C6	16	12	12	12	10	18	14	16	19
C7	14	10	10	10	8	14	19	19	22
C8	8	5	5	5	3	11	15	21	23
C9	0	0	0	0	0	4	13	18	25

an equal voting power for each interval, during querying, the votes of an interval are normalized later, so that

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

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.



```
train (TrainingSet):
begin
    for each feature f
        /* sort TrainingSet with respect to f */
        sort (f, TrainingSet)
        /* construct a list of point intervals using feature values and class labels */
        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.
```

Figure 1 Training algorithm of BCFP.

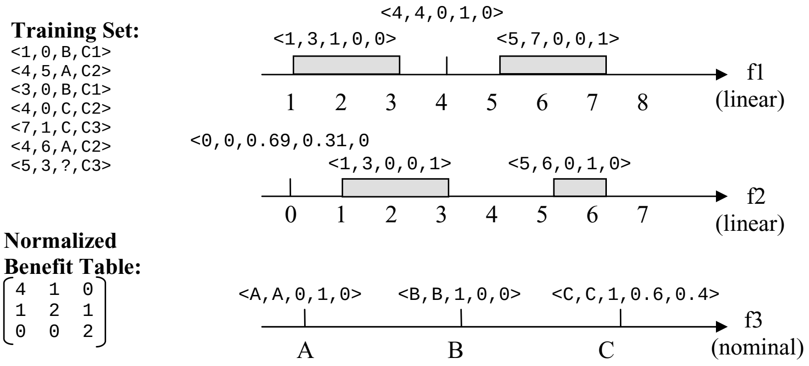


Figure 2 Feature intervals formed for a sample training set.

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 $s_{\min} = (1/k - 1)\sqrt{1/k}$. 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.[1] For the gastric carcinoma domain, $k = 9$ and $s_{\min} = 0.0417$.

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.

3.3. 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 preclassification step on each feature. The preclassification of 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 available to the user, enabling him/her to measure the level of the confidence of this prediction.

Note that in domains with imbalanced class distribution, the simple accuracy measure as the ratio of the number of correctly classified instances to the total number of instances fails to reflect the performance of the classification algorithm [12]. In benefit maximizing classification, we are interested in the total benefit achieved over all test instances, not in number of correct classifications. The *benefit accuracy* of the classification in BCFP is obtained directly from the normalized benefit table. The performance of the BCFP classifier on a test set T is measured as

$$\text{benefit accuracy} = \frac{\sum_i^T \text{NB}[p_i, a_i]}{\sum_i^T \text{NB}[a_i, a_i]}$$

where p_i and a_i are the predicted and actual class labels of the ith test instance from T. In the rest of

```
classify($q$): /* q: query instance to be classified */
begin
    for each class $c$ $vote[c]$ = 0 /* initialize total votes */
    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.
```

Figure 3 Classification in the BCFP algorithm.

[1] This threshold is meaningful if $k > 2$.

```
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
```

Figure 4 Classification example on the sample data set.

the paper, the term accuracy will be used to refer the benefit accuracy.

Continuing on the example in Fig. 2, consider the classification of a query instance $q = \langle 2, 5, \mathrm{C}\rangle$. 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 $1.6/(1+1.6+0.4) = 53\%$.

3.4. Feature selection using a genetic algorithm

The performance and the cost of classification are sensitive to the choice of the features used to construct the classifier. The natural and safe approach in inductive machine learning is to collect the values of all available features for instances, 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 [13].

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 [14].

We have coupled the BCFP algorithm with a genetic algorithm using the wrapper approach for feature subset selection [15]. Each feature is represented by a gene in the chromosome. Therefore, the chromosome size is equal to the number of features, which is 68 for the gastric carcinoma data set. 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. Experiments on the gastric carcinoma domain were conducted with one-point and two-point operations. For the probability of crossover $p_c = 0.7$, 0.8 and 0.9, and for the probability of mutation $p_m = 0.001$, 0.002 and 0.005 were tried. The genetic algorithm was run for 2000 generations. In all experiments, the same chromosome was found before the 800th generation. The chromosome found represented a subset of 30 features selected as relevant.

4. Results

The BCFP algorithm and the accompanying genetic algorithm for feature selection have been implemented in the C language to run the experiments. In measuring the performance of the BCFP algorithm we used 10-fold cross-validation accuracy, i.e., the whole data set is partitioned into 10 equal sized subsets. One of the subsets is used as the test set, and the other nine as the training set. This process is repeated 10 times, once for each subset as the test set. This technique ensures that the training and test sets are disjoint, and each instance in the data set is classified exactly once. Accuracy is the average of the accuracy values of these 10 runs. The execution of a 10-fold cross validation took 25 ms.

Using all of the 68 features of the data set, the BCFP algorithm achieved 83.5% accuracy. However, the feature selection algorithm chose only 30 of the 68 features as relevant for a beneficial classification. With the selected set of features the BCFP algorithm achieved 94.8% accuracy. The confusion matrix for 10-fold cross validation obtained using these 30 features is given in Table 8. The selected features are Rh factor, depth, flower bed appearance, big ulcer, infiltrated ulcer, mucosal fold, abrupt disruption, surface mucosa spotty redness, surface mucosa smooth/normal, surface mucosa discoloration, surface mucosa irregularity, surface mucosa uneven, surface mucosa infiltrated, surrounding mucosa irregularity, surrounding mucosa infiltration, surrounding mucosa redness/hyperemic, surrounding mucosa spotty redness, surrounding mucosa margin irregularity, bleeding, mucosal elevation, base infiltrated, base redness, base wide, polyp, sessile, stomach deformation, whole stomach infiltrated, pylica gastrica sickness, depressed area, and infiltrated mass. 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.

The rules constructed by the BCFP algorithm are easy to be verified by experts. According to these

Table 8 Confusion matrix

Prediction	Actual									
	1	2	3	4	5	6	7	8	9	Total
1	0	0	0	0	0	0	0	0	0	0
2	1	42	0	4	0	1	0	0	0	48
3	0	0	7	0	0	0	0	0	0	7
4	2	13	0	98	6	2	3	1	0	125
5	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	3	0	0	0	3
7	0	0	0	1	0	0	7	20	2	30
8	0	0	0	0	0	0	7	46	13	66
9	0	0	0	0	0	0	0	2	4	6
Total	3	55	7	103	6	6	17	69	19	285

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. This is because, in our data set, all the instances with flower bed appearance were an Early IIa case. On the other hand, if the surface mucosa discoloration is present, then the case is either Early IIa or Early IIc. If sessile is present, it is an Early I, with 30% certainty, Early IIa with 36% certainty, or Early IIc with 34% certainty. The other rules can be interpreted in the similar manner.

We have also compared the BCFP algorithm with the cost sensitive version of Naive Bayesian Classifier (NBC) [16]. We have used MetaCost [17] algorithm around NBC in order to make it sensitive to benefits. MetaCost, together with NBC as a base classifier has been taken from Weka [18]. Using all of the features, MetaCost with NBC achieved an accuracy of 90%, whereas using only the selected set of features, it achieved an accuracy of 93.7%.

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 and four faculty members in gastroenterology. The students and experts were shown only the data set that was used by the BCFP algorithm. The benefit accuracy of the students was 65%, and that of faculty members was 80%. This indicates that making an accurate decision in the diagnosis of gastric carcinoma is quite difficult.

```
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 C9/0.01
If depth = 3
   then C2/0.16 C4/0.17 C7/0.34 C8/0.33
If 4  depth  5
   then C2/0.05 C4/0.05 C5/0.07 C6/0.19 C7/0.22 C8/0.22 C9/0.22
If flower bed appearance = Present
   then C2/1
if infiltrated ulcer = Present
   then C4/0.08 C7/0.31 C8/0.33 C9/0.28
if surface mucosa discoloration = Present
   then C2/0.50 C4/0.50
if surrounding mucosa redness/hyperemic = Present
   then C2/0.26 C4/0.28 C5/0.24 C7/0.14 C8/0.08
if surrounding mucosa spotty redness = Present
   then C4/1
If bleeding = Present
   then C2/0.06 C4/0.06 C5/0.08 C6/0.19 C7/0.22 C8/0.21 C9/0.20
If sessile= Present
   then C1/0.30 C2/0.36 C4/0.34
If stomach deformation = Present
   then C8/0.49 C9/0.51
If depressed area = Present
   then C2/0.25 C3/0.24 C4/0.26 C5/0.20 C8/0.05
```

Figure 5 Sample rules induced by the BCFP algorithm.

5. 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.

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 data set 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. However, Holte has pointed out that the most real-world data sets for classification tasks are such that their attributes can be considered independently of each other [19]. Also, Kononenko claimed that in the data sets used by human experts there are no strong dependencies between features [20].

The BCFP algorithm achieved very good accuracy on the gastric carcinoma data set available. 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 an 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 and providing these interesting rules constructed from the data available 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. We thank the anonymous referees for their invaluable comments.

References

[1] Güvenir HA, Şirin I˙. Classification by feature partitioning. Machine Learning 1996;23:47—67.

[2] Güvenir HA, Emeksiz N. An expert system for the differential diagnosis of erythemato-squamous diseases. Expert Syst Appl 2000;18:43—9.

[3] Güvenir HA, Demiröz G, Ilter N. Learning differential diagnosis of erythemato-squamous diseases using voting feature intervals. Artif Intell Med 1998;13:147—65.

[4] Güvenir HA, Acar B, Demiröz G, Çekin A. A supervised machine learning algorithm for arrhythmia analysis. Computers Cardiol 1997;24:433—6.

[5] Torii A, Sakai M, Inoue K, Yamabe H, Ueda S, Okuma M. A clinicopathological analysis of early gastric cancer: retrospective study with special reference to lymph node metastasis. Cancer Detect Prevent 1994;18:437—42.

[6] Örmeci N, Demirci S, Tulunay Ö, Kuzu I, Akgül H, Uzunalimog˘lu Ö, et al. Early stomach cancer in Turkey. Recent Adv Manage Digest Cancers 1993:339—41.

[7] Kajitani T. Clinical classification. Jpn J Surg 1981;11:127—39.

[8] Kurihara H. Detection of early gastric cancer outside the mass screening program. Jpn J Clin Oncol 1998:233.

[9] Yokota T. Bormann's type IV gastric cancer: clinicopathologic analysis. Can J Surg 1999;42:371—6.

[10] Ting KM, Zheng Z. A study of adaboost with naïve Bayesian classifiers: weakness and improvement. Int J Comput Intell 2003;19:186—200.

[11] Demiröz G, Güvenir HA. Classification by voting feature intervals. In: Proceedings of the Ninth ECML. Springer-Verlag. LNAI 1997;1224:85—92.

[12] Kubat M, Holte RC, Matwin S. Machine learning for the detection of oil spills in satellite radar images. Machine Learning 1998;30:195—215.

[13] Yang J, Honavar V. Feature subset selection using a genetic algorithm. In: Motoda H, Liu H, editors. Feature extraction, construction, and subset selection: a data mining perspective. New York: Kluwer; 1998.

[14] Goldberg D. Genetic algorithms in search, optimization, and machine learning. New York: Addison-Wesley; 1989.

[15] John G, Kohavi R, Pfleger K. Irrelevant features and the subset selection problem. In: Proceedings of the Eleventh International Conference on Machine Learning. New Brunswick, NJ: Morgan Kaufmann; 1994. p. 121—9.

[16] Duda RO, Hart PE. Pattern classification and scene analysis. New York: Wiley & Sons; 1973.

[17] Domingos P. Metacost: a general method for making classifiers cost-sensitive. In: Proceedings of the International Conference on Knowledge Discovery and Data Mining, San Diego, CA; 1999. p. 155—64.

[18] Weka 3: Machine Learning Software in Java, The University of Waikato software documentation [http://www.cs.waikato.ac.nz/~ml/weka].

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

[20] Kononenko I. Inductive and Bayesian learning in medical diagnosis. Appl Artif Intell 1993;7:317—37.