Applied Intelligence 22, 61–75, 2005

Building a Medical Decision Support System for Colon Polyp Screening by Using Fuzzy Classification Trees

I-JEN CHIANG
Graduate Institute of Medical Informatics, Taipei Medical University Taipei, Taiwan
ijchiang@tmu.edu.tw

MING-JIUM SHIEH
Department of Biomedical Engineering, National Taiwan University Taipei, Taiwan

JANE YUNG-JEN HSU
Department of Computer Science and Information Engineering, National Taiwan University Taipei, Taiwan

JAU-MIN WONG
Department of Biomedical Engineering, National Taiwan University Taipei, Taiwan

Abstract. To deal with highly uncertain and noisy data, for example, biochemical laboratory examinations, a classifier is required to be able to classify an instance into all possible classes and each class is associated with a degree which shows how possible an instance is in that class. According to these degrees, we can discriminate the more possible classes from the less possible classes. The classifier or an expert can pick the most possible one to be the instance class. However, if their discrimination is not distinguishable, it is better that the classifier should not make any prediction, especially when there is incomplete or inadequate data. A *fuzzy classifier* is proposed to classify the data with noise and uncertainties. Instead of determining a single class for a given instance, *fuzzy classification* predicts the degree of *possibility* for every class.

Adenomatous polyps are widely accepted to be precancerous lesions and will degenerate into cancers ultimately. Therefore, it is important to generate a predictive method that can identify the patients who have obtained polyps and remove the lesions of them. Considering the uncertainties and noise in the biochemical laboratory examination data, *fuzzy classification trees*, which integrate decision tree techniques and fuzzy classifications, provide the efficient way to classify the data in order to generate the model for polyp screening.

Keywords: fuzzy classifications, polyp screening, fuzzy classification trees, fuzzy entropy

1. Introduction

Colorectal cancer (CRC) has become one of the leading causes of cancer death in Taiwan, with nearly 2900 new cases and 1900 deaths reported each year. Despite advances in treatment, early detection can probably reduce CRC mortality more than any other approaches. Therefore, it is important to develop a cost-effective cancer screening policy in the hopes of reducing CRC mortality by detecting lesions at any early, curable stage.

The prevalence of adenomatous polyp varies geographically in parallel with the incidence of colorectal cancer and an increasing risk of colorectal cancer [1–4]. The concept is now widely accepted that adenomas are precancerous lesions and will degenerate

into cancers ultimately. Nowadays, the majority of the pathogeneses of the colorectal cancer are attributed to the adenoma-adenocarcinoma sequence. Hence, the identification and removal of the precancerous lesion, an adenomatous polyp, have significant clinical implications and are now commonly recommended for the control of CRC. Endoscopy is considered the most sensitive diagnostic modality for detection of colorectal polyps. However, the effort and eventual cost involved based on this surveillance strategy are potentially enormous and not practical, except for high-risk groups. Owing to the shortage of medical resources at present, it is important to develop a most cost-effective and safe screening method to predict the existence of adenomatous polyps.

In order to determine the predictive value of the risk factors related to the existence of rectosigmoid colon polyps, physicians evaluate all putative risk factors obtained from checkup items. Bias inevitably occurs from this assumption, in that only factors that have been selected can be shown to have association. A collection of physical checkup data with the patients who underwent sigmoidoscopy enrolled for the polyp screening analysis.

Classification can be thought as the base of ability to knowledge acquisition [5]. Some classification techniques, e.g. decision trees [6–11], decision lists [12, 13] work well for pattern recognition and process control. Here, we choose these techniques to apply to colon polyp screening analysis [3]. Unfortunately, it is hard to clearly classify the data because of the uncertainties and noise. Obviously, a vague classification method is needed to deal with such problems. That is, a classifier is able to classify an instance into all possible ones and each class is associated with a degree which shows how possible an instance is in that class. According to these degrees, we can discriminate the more possible classes from the less possible classes.

As a result, a more reasonable answer from the classification system should present all probable conclusions, each of which is associated with a degree of possibility. When the number of variables to describe a process is not large, it models the process by (1) dividing the whole space into several subspaces, (2) representing each subspace by a simple linear function, and (3) interpolating several subspaces continuously. When a system is very complex, it is necessary to extract the relevant variables in the premises of fuzzy models. Sugeno and Kang [14] proposed to use a mathematical programming method to dealing with this problem. Schuermann and Doster [15] called it to be *hard-decision systems*. A large amount of calculation to identify premise parameters is unavoiding. Therefore, fuzzy classifications have proposed by Hsu and Chiang [16–18].

This paper introduces the use of the fuzzy classification approach to polyp screening. Section 2 gives the definition of classifications and problems of traditional classifiers. The definitions of *fuzzy classifications*, *fuzzy classification trees* and an example of fuzzy classification trees are presented in Section 3. The attribute selection measures are defined in Section 4. Section 5 describes the basic algorithm for constructing a Fuzzy Classification Tree (FCT) from a data set. The classification process is shown in Section 6. The empirical results compared FCT with C4.5 on polyp screening are shown in Section 6, followed by the conclusion.

2. Classifications

Given a set of instances, a classification problem is concerned with assigning each instance into a proper class based on its attribute values. An instance can be described in terms of its values corresponding to a given set of attributes. The most important issue in classification is to identify the key attributes of the instances. Classification is an important inductive technology that has been widely used to gain the information from a very large database. The data in the database are identified to belong to different classes. Classification methods are used to minimize the difference in a class and to maximize the difference among classes.

Consider an ordered set of attributes $A = (a_1, a_2, \ldots, a_n)$ for the instance description. An *attribute value vector* $\mathbf{x} = \langle x_1, x_2, \ldots, x_n \rangle$ consists of the value x_i for the corresponding attribute a_i. Each attribute may take either *ordered* or *categorical* values. Ordered values are typically numerical, either discrete or continuous, while categorical values are symbolic. For example, supposedly to decide whether one should play golf depends on attributes {`Outlook`, `Temperature`, ... }. The attribute value vector for the golf example may look like $\langle sunny, 85^\circ F, \ldots \rangle$, in which attribute `Outlook` takes a symbol as value and attribute `Temperature` is numerical. As another example, in a study on physical examinations performed at the National Taiwan University Hospital, there were more than 400 attributes associated with the examinations. The first attribute was gender, which had a symbolic value of either male or female; the second attribute,

age, had a discrete value ranging from 0 to 120; the third attribute, height, was defined over a continuous range from 0 to 200.

Through a classification method, a classifier can be constructed from a database. This classifier is able to predict which class a new instance is. Many techniques, such as Bayesian classifiers [19], decision trees [11], neural networks [20], rule based learners [21, 22], etc., have been applied to producing classifiers. A classifier is produced on a set of training instances and a decision is made automatically on each new instance based upon a forecast of the classification of the instance.

2.1. *Requirements of Classifiers*

Traditional classifiers, such as decision trees [11] and rule based learners [13, 21, 22], usually produce a classification of every new instance. Although those classifiers are generally efficient, they have serious problems in dealing with elaborate real-valued attributes [23, 24]. In some applications, it is advantageous not to produce a classification on every instance. In particular, when a learning program is being used to assist a person to perform some task, it might be desirable to have the machine automatically make some decisions while allowing others to be made by the person. It is necessary that a classifier can classify an instance into all possible classes and each class is associated with a degree which shows how possible an instance is in that class. According to these degrees, we can discriminate the more possible classes from the less possible classes. The classifier or an expert can pick the most possible one to be the instance class. However, if their discrimination is not distinguishable, it is better that the classifier should not make any prediction, especially when there is incomplete or inadequate data.

Bayesian classifiers estimate the probability that a test instance is a member of each class and thes test instance is assigned to the class with the highest probability. This kind of classification methods is based on Bayesian formula [25, 26]. The *a priori* probabilities are needed to explain the result of classifications. The data defined on classical statistics can classify into n mutually exclusive and equally likely outcomes. If n_A denotes those outcomes with property A, then the probability of A is the fraction n_A/n [27]. *A priori* probabilities are determined in accordance with the classical definition. The purpose of the learning task is to construct the rules from an amount of training samples. In order to make those rules sufficient to describe the universal world, it is necessary that the training samples are sufficient to be talked about by *a priori*. This seems a strong assumption for the training set. Ignoring some limitations in the classical, or a prior approach, for data analysis, we will find that it is difficult in the data that has high dimensions and is sparse.

Neural network classifiers that use a logistic activation function have units to produce an output ranged from 0 to 1 [20]. A test instance can be classified correctly according to the value of the activation on the unit with the maximum activation as a measure of the likelihood. A given threshold can be used to achieve making a significant discrimination on classification tasks. However, the complexity of defining the architecture of neural network classifiers and the low convergent rate of learning make this kind of classifiers improper for high-dimensional large databases.

Therefore, the concept of "*fuzzy classifications*" has been proposed [16–18]. Fuzzy classifications which satisfy the basic requirements addressed above are properly defined the classifiers that we need for data analysis. A fuzzy classifier, *fuzzy classification trees*, has been addressed instead of the other kinds of classifiers. Fuzzy classification trees are a kind of tree structure classifiers.

2.2. *Data Partition Problems*

Classification by decision trees has been successfully applied to problems in artificial intelligence, pattern recognition and statistics. However, as Quinlan [28] pointed out "the results of decision trees are categorical and so do not convey potential uncertainties in classifications." Missing or imprecise information may prevent a case from being classified at all. In the presence of uncertainties, what is preferred is an estimate of the degree of being in each class, e.g. in the medical domain.

CART [29], and C4.5 [11] choose a test for the root node to create its leaves, partition the training set into those nodes, and then apply the same algorithm recursively to each of the leaves. The test chosen is according to a goodness of split measured at each stage. According to the test, the data can be explicitly divided into a nested sequence of regions. The data partition can have favorable consequences for the bias of an estimator, but it generally increases the variance of the estimator [30]. Consider the linear regression model, for example, in which the variance of the estimates of the slope and intercept depends quadratically on the spread of data on

the projection axes of the corresponding independent variables. The points that are the most peripheral in the input space are those that have the maximal in decreasing the variance of the parameter estimates. Jordan and Jacobs called those algorithms variance-increasing algorithms [30].

Instead of classifying a case as belonging to exactly one class, and ruling out the others, one can estimate the relative probabilities of it belonging to each class. Casey and Nagy [31] designed a decision tree classifier using probabilistic model for the optical character recognition process. Breiman et al. [29] introduced the class probability estimate. Quinlan [22] proposed probabilistic decision trees to deal with uncertainties in data. Schuermann and Doster [15] also proposed using the probabilistic model to estimate the probability of each class. In addition, to deal with the search bias introduced in attribute selections and the hypotheses-space bias due to noisy data [25], Buntine [26] suggested the "averaging" method over multiple class probability trees. Unfortunately, even as class probability trees [29] that are used to produce accurate posterior class probabilities, rather than simply the label of the mostly likely class is also based on the data partition algorithm of the traditional decision tree classifier, such as CART or C4.5. That makes the class probability trees not able to avoid the variance-increasing problem.

Probabilistic approaches still assume there is only one decision node in the tree to which a case can be classified. A test instance falls down a single branch to arrive at a leaf labelled by a class and associated with the corresponding probability. Such classifications ignore the information at the other nodes. Several methods, including Buntine's classification trees [26], Rymon's *Set Enumeration* tree [32] have been addressed to solve this difficulty. However, their approaches are still inefficient in both time and space.

Due to the variance-increasing problem, a "soft" splitting method that allows the training instance to lie simultaneously in multiple regions is hence addressed. Jordan and Jacobs [30] proposed the variance-decreasing algorithm that is a hybrid method combined the bottom-up induction of decision trees and neural computing. Their method defined a complex decision tree architecture in which each node of the tree is a neural network. A complicated computation is involved in training this hybrid architecture. A simple and efficient method is necessary to overcome this problem.

Fuzzy decision trees [17, 33–35] which integrate decision tree techniques and fuzzy classifiers, provide a simple and efficient way to generate the classification model that can suffer from inadequately or improperly expressing and handling the vagueness and ambiguity associated with human thinking and perception [36]. Even by Quinlan's work [28], the types of uncertainties are not to be probabilistic, appearing as randomness or noise. Pedrycz and Sosnowski [37] pointed out that the concept of fuzzy granulation realized via context-based clustering is aimed at the discretization process. For the sake of vagueness, fuzzy classifications are issued. Through it, we can calculate the degree of possibility that the instance belongs to any of the classes. Using information-based measures, there is no need to generate multiple classification trees. Therefore, it requires less time and space than decision forests [38].

3. Fuzzy Classifications

Fuzzy classifications are proposed to overcome the difficulties that conventional classifiers cannot handle multiple instances with overlapping attribute values that belong to different classes, but keep the efficient as decision tree classifiers.

Definition 1. Given a *fuzzy classifier* $\mathbf{F}$ for a given classification problem $(\mathcal{X}, \mathcal{C})$ defines a total function

$$\mathbf{F} : \mathcal{X} \to \{\langle p_1, \ldots, p_n \rangle \mid p_i \in [0, 1]\}$$

where p_i is the *possibility* that a given instance $\mathbf{x}$ belongs to class C_i and n is the number of classes.

For ease of presentation, the function $\mathbf{F}$ is sometimes represented as a vector of functions

$$\langle \wp_1, \wp_2, \ldots, \wp_n \rangle,$$

where $\wp_i$ is a possibility function $\mathcal{X} \to [0, 1]$. For any given instance $\mathbf{x}$, the relation $\wp_i(\mathbf{x}) > \wp_j(\mathbf{x})$ indicates that it is more likely for the instance $\mathbf{x}$ to be in class C_i.

A fuzzy classifier can be readily implemented by a tree structure, such as fuzzy decision trees [33–36, 39]. In general, those methods can be separated into two types, pre-fuzzification and post-fuzzification. However, no matter what the type of fuzzy decision tree methods is, they all unavoid two phases processing to generate the decision rules. They either prefuzzify the data according to domain knowledge or postfuzzify the decision rules generated by the decision tree methods

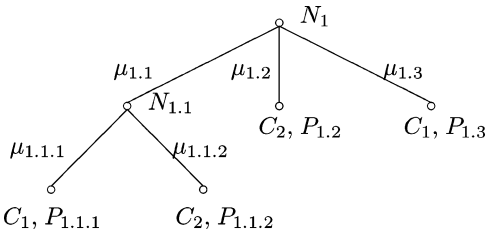

Figure 1. A sample FCT with $\mathcal{C} = \{C_1, C_2\}$

by some tuning methods. They do not concern the distribution of the data that can make improper classifications. Therefore, fuzzy classification trees [16–18] have been presented to solve those problems on pre-fuzzification and post-fuzzification.

This section briefly presents the basic definitions of *fuzzy classification trees* (FCTs). Figure 1 shows a sample FCT that classifies instances into two classes C_1 and C_2.

Let $\mathcal{L}$ be the set of all labels that is defined by a labeling function that uniquely assigns a label to each node and each branch.

Definition 2. Given an *FCT*, each node n in the tree $\mathcal{T}$ is given a label:

$$\text{Label}(n) = \begin{cases} 1 & \text{if } n \text{ is the root;} \\ \text{Label}(n').i & \text{if } n \text{ is the } i\text{th} \\ & \quad \text{child of node } n'. \end{cases}$$

where . is the concatenation operator.

N_L denote the node labeled by $L \in \mathcal{L}$, and B_L denote the branch leading into node N_L. Each non-terminal node in the tree is associated with a test, and the resulting branches, $B_{L.i}$, is associated with a membership function

$$\mu_{L.i} : \mathcal{X} \to [0, 1].$$

Intuitively, the membership defines the degree of possibility that an instance $\mathbf{x} \in \mathcal{X}$ should be propagated down the branch. In our implementation, each test at a node is tested on a single attribute. Therefore, the membership function is defined over the projection on that attribute, that is, $\texttt{projection}(\mathcal{X}, a_L)$, i.e. the domain of the testing attribute $a_L \in A$.

Suppose each node N_L is associated with a class C_L and a possibility function P_L.

Definition 3. Let the label for the parent node of N_L is denoted to be $\hat{L}$. The possibility function $P_L : \mathcal{X} \to [0, 1]$ is defined by composing the membership functions along the path from the root to node N_L. That is,

$$P_L = \begin{cases} 1 & \text{if } N_L \text{ is the root node;} \\ P_{\hat{L}} \otimes \mu_L & \text{if } N_{\hat{L}} \text{ is the parent of } N_L. \end{cases}$$

The composition operator $\otimes$ is defined in terms of some valid operation for combining two membership functions.

Several composition operators, e.g. fuzzy sum, fuzzy product, and fuzzy max, are supported in our implementation. For example,

$$P_L(\mathbf{x}) = P_{\hat{L}}(\mathbf{x}) + \mu_L(\mathbf{x})$$

when the fuzzy sum operator is applied.

Given any instance $\mathbf{x}$ at a terminal node N_L in an FCT, it is classified into class C_L with a possibility $P_L(\mathbf{x})$. As shown in Fig. 1, multiple terminal nodes may be associated with the same class. It follows that an FCT defines a unique fuzzy classifier

$$\mathbf{F} = \langle \wp_1, \ldots, \wp_n \rangle$$

such that the possibility for an instance belonging to class C_i is the *maximum* over all the possibility values at terminal nodes classified as C_i. That is, for $1 \le i \le n$,

$$\wp_i(\mathbf{x}) = \max\{P_L(\mathbf{x}) | N_L \text{ is a leaf } \wedge C_L = C_i\}.$$

Before going into detail about the tree construction algorithm, let us give a brief example to explain the fuzzy classification tree. Consider that the objects were human beings and the classification task involved the *hypertension*, the attributes might be

rsabp *systolic arterial blood pressure*
rdabp *diastolic arterial blood pressure*
fbs *fasting blood sugar*

The normal arterial blood pressure is 120/70 mm Hg. If a person whose resting systolic blood pressure is over 120 mm Hg and resting diastolic blood pressure is over 70 mm Hg, that can always increase this patient's risk for hypertension. Whatever the fasting blood sugar is will take no effect on the hypertension's diagnosis.

Table 1. A training set of hypertension.

No.	rsabp	rdabp	fbs	class
1.	118	65	110	normal
2.	114	68	120	normal
3.	130	75	100	hypertension
4.	122	76	140	hypertension
5.	108	60	99	normal
6.	120	78	102	hypertension
7.	115	73	121	normal
8.	125	70	110	hypertension
9.	124	69	108	hypertension
10.	113	61	122	normal
11.	109	62	98	normal
12.	135	80	104	hypertension
13.	116	67	103	normal
14.	133	83	117	hypertension
15.	119	75	112	hypertension

In our example, we assume that each object in the universe belongs to one of two mutually exclusive classes, *hypertension* or *normal*, which is shown in Table 1.

Our algorithm for constructing a fuzzy classification tree can automatically identify the characteristics of the attribute values in according to their appearances at different classes, which is able to eliminate the noise and adjust the classification of values. It also can distinguish which attributes are adequate or inadequate for the current classification. We say attributes are *inadequate* for the classification task if two objects are identical but belong to different classes, however, it is clearly impossible to differentiate between these objects with reference only to the given attributes.

The classification rule will be expressed as a fuzzy classification tree. A fuzzy classification tree that correctly classifies each object in the training set is given in Fig. 2. As mentioned above, leaves of a fuzzy classification tree are class name, other nodes represent attribute-based selections with a branch for each possible outcome. The tree is beginning from the root of the tree and proceeding down to its leaves. The tree constructing process continues until a leaf is encountered, then the object is asserted to belong to the class at the leaf.

As shown in Fig. 2, the fuzzy classification tree identifies the classification intervals for each attribute in different classes. The interval of the '*rsabp*' attribute for hypertension is (119, 135), and for normal is (108, 118). Since the intervals of the normal and the hypertension classes are nonoverlapped, so the membership function of each one is 1. The interval of the '*rdabp*' attribute for hypertension is (69, 83) with membership 1, and for normal is (60, 69.83) that is divided into two parts (60, 69) with membership 1 and (69, 69.83) with membership greater than 0 and less than 1. The intervals of the '*fbs*' attribute for these two classes as seen are almost overlapping. Of course, *fbs* is the most inadequate attribute for the hypertension classification.

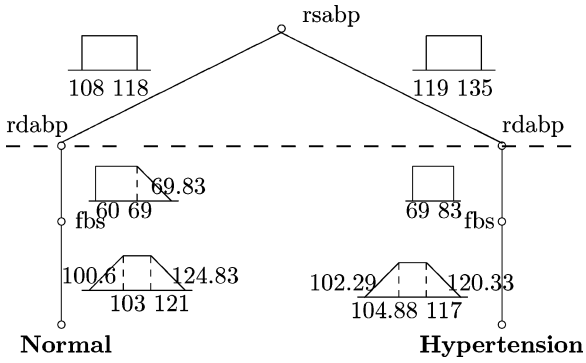

Figure 2. A fuzzy classification tree for the hypertension example.

From the root to the leaves of the fuzzy classification tree, it is easy to see that the *rsabp* is the most adequate attribute to construct a classification tree for the example. It can correctly classify each object in the training set. The essence of induction is to construct a fuzzy classification tree that correctly classifies not only objects from the training set but other unseen objects as well. The process of construction will be introduced in the succeeding section.

As we expect, the fuzzy classification tree can not only correctly classify the domain of each attribute but also identify the adequate attribute. In the Fig. 2, the attributes below the dash line are not needed to be considered further more. Both the *rdabp* and *fbs* attributes are not considered any more, especially the *fbs*. Because the *rsabp* attribute is enough to distinguish the objects which are hypertension or normal.

4. Information-Based Measure

At each node of a fuzzy classification tree, an attribute is used to calculate the membership that an instance should be split into a branch. This attribute is decided

at the learning time, that may create the best data clustering at the current node. The *goodness of split* is an important criterion for selecting attributes to expand a fuzzy classification tree. Some information-based measures have been widely applied to classifications for evaluating the goodness of split [11, 29, 40–42].

In order to evaluate the uncertainties in the data, Shannon has defined the information entropy function that refers to the Boltzmann's H theorem in statistical mechanics [43]. The foundation of Shannon's formula is based on probability theory. Quinlan [11], etc., have used such kind of uncertainty evaluation methods to construct tree classifiers. These information-based evaluation methods can be applied to the construction of probabilistic fuzzy classification trees. However, those methods are well-defined on probability.

According to the original probabilistic entropy defined by Shannon [43] and fuzzy entropy function defined by De. Luca and Termini [44], the information-based measure should satisfy the following criteria. Let the possibility $\wp_i$ for each i define the possibility of an instance, where $\wp_i \in [0, 1]$. Five criteria [16, 18] required for attribute selection in terms of an information-based measure of FCT are listed as follows.

Property 1. *Function $H(\wp_1, \wp_2, \ldots, \wp_n)$ should be continuous in $\wp_i$. This property prevents a situation in which a very small change in $\wp_i$ would produce a large (discontinuous) vibration.*

Property 2. *Function H must be* 0 *if and only if all the $\wp_i$ but one are zero. When all but one is possible, there exists no uncertainty in the data.*

Property 3. *Function H is the maximum value if and only if the $\wp_i$ are equal because there exists the most uncertainties in the data. That is, no matter what all the $\wp_i$ are, the largest uncertainties happened when all the $\wp_i$ are of the same value.*

Property 4. *Function H is a nonnegative valuation on the $\wp_i$.*

Property 5. *The purpose of an attribute selection measure is to reduce uncertainties in the data, so it is necessary that if a choice is broken down into several successive choices, the original H should be no less than the weighted sum of the individual values of H. This property prevents the data been classified to be worse than before.*

We can define our fuzzy entropy functions that follow the five criteria. Suppose we have a set of instances S_L at node N_L. Assume there are n classes associated with the possibilities of occurrences $\wp_1, \wp_2, \ldots, \wp_n$. Concerning about the measure of how much *choice* is involved in the selection of the instance in S_L or of how uncertain we are of the outcome, we choose the entropy function to evaluate that.

Definition 4. The entropy for the set of instances S_L at node N_L is defined by

$$\text{Info}(S_L) = -\sum_{\forall c \in \mathcal{C}} \frac{\mathcal{P}_L^c}{\mathcal{P}_L} \times \log_2 \frac{\mathcal{P}_L^c}{\mathcal{P}_L}.$$

where

$$\mathcal{P}_L = -\sum_{x \in S_L} P_L(\mathbf{x})$$

is the sum of the possibility value $P_L(\mathbf{x})$ of all instances at node N_L, and

$$\mathcal{P}_L^c = \sum_{\mathbf{x} \in S_L \wedge \text{Class}(\mathbf{x}) = c} P_L(\mathbf{x})$$

is the sum over instances belonging to class c.

The entropy of a set measures the average amount of information needed to identify the class of an instance in the set. It is minimized when the set of instances are homogeneous, and maximized when the set is perfectly balanced among the classes.

A similar measurement can be defined when the set is distributed into b_L subsets, one for each branch based on the test at node N_L. The expected information requirement is the weighted sum over the subsets.

$$\text{Info}_T(S_L) = \sum_{i=1}^{b_L} \frac{\mathcal{P}_{L.i}}{\mathcal{P}_L} \times \text{Info}(S_{L.i}).$$

To assess the "benefits" of a test, we need to consider the increase in entropy. The quality

$$\text{Gain}(\text{Test}_L) = \text{Info}(S_L) - \text{Info}_T(S_L).$$

measures the information gain due to the test Test_L. This gain criterion is used as the basis for attribute selection.

4.1. Choosing the Fuzzy Operations

Five criteria of fuzzy entropy limitate the fuzzy operators that can be used to calculate the possibility of each instance at a node. Here, the *fuzzy t-norm* operator is involved for the possibility evaluation because it can satisfy those criteria, especially, the fifth property.

Since the function, $\log_2$ is a continuous function, the fuzzy entropy defined by $\log_2$ is also a continuous function. It is easy to see that Info satisfies Property 1.

If S_L is the set of instances in N_L that has been purely classified into one class, that is all the $\wp_i$ of each instance but one are zero. Let $\wp_i \neq 0$ for some class C_i, then the possibility

$$\mathcal{P}_L = \sum_{x \in S_L} P_L(\mathbf{x}) = \sum_{x \in S_L} \wp_i(\mathbf{x}).$$

The possibilities $\mathcal{P}_L^c$ of the other classes are zero. Because

$$\mathcal{P}_L^c = \sum_{\mathbf{x} \in S_L \wedge \text{Class}(\mathbf{x})=c} P_L(\mathbf{x}) = 0$$

for $c \neq C_i$. The entropy value of $\text{Info}(S_L)$ will be zero when all the possibilities $\wp_i$ but one are zero.

Property 3 restricts that the entropy value is maximum when all the class possibilities are equal. According to that, it needs that $\sum_c \mathcal{P}_L^c$ should be no bigger than $\mathcal{P}_L$. Otherwise, this property will not be satisfied. Let $|\mathcal{C}|$ be the number of classes and $\mathcal{P}_L^{C_i} = \mathcal{P}_L^{C_j}$ for $i \neq j$. In the FCT algorithm, the sum operation $\sum$ is defined to be equal to the sum

$$\begin{aligned}\text{Info}(S_L) &= -\sum_{\forall c \in \mathcal{C}} \frac{\mathcal{P}_L^c}{\mathcal{P}_L} \times \log_2 \frac{\mathcal{P}_L^c}{\mathcal{P}_L} \\ &\leq -\sum_{i=1}^{|\mathcal{C}|} \frac{\mathcal{P}_L}{|\mathcal{C}|\mathcal{P}_L} \log_2 \frac{\mathcal{P}_L}{|\mathcal{C}|\mathcal{P}_L} \\ &= -\sum_{i=1}^{|\mathcal{C}|} \frac{1}{|\mathcal{C}|} \log_2 \frac{1}{|\mathcal{C}|}.\end{aligned}$$

operation in classical (crisp) set.

Since $0 \leq \mathcal{P}_L^c \leq \mathcal{P}_L$ for all class $c \in \mathcal{C}$, $\log_2 \frac{\mathcal{P}_L^c}{\mathcal{P}_L} \leq 0$ and $\text{Info}(S_L) \geq 0$. Therefore, it is no doubt that the fourth property is also satisfied.

The purpose of an attribute selection in FCTs is toward reducing the uncertainties in the data. After the fuzzy classification tree has been generating, the total entropy of the child nodes should be no greater than the entropy of their parent nodes. In the other word, the total entropy of child nodes from a node should be less than or equal to the entropy of that node before the tree expanded. That is,

$$\text{Info}(S_L) \geq \sum_{i=1}^{b_L} \frac{\mathcal{P}_{L.i}}{\mathcal{P}_L} \times \text{Info}(S_{L.i}).$$

This is what the fifth property gives, which is a strong constraint that restricts the kinds of fuzzy operations and the membership functions. It also limits the clustering methods to generate the membership function from a node.

The membership function is the kernel for fuzzy classifications. To determine the membership function from a data set, the method of clustering is used. Clustering is a well-used method in pattern recognition. It plays a key role in searching for structures in data. There may be different kinds of models simultaneously occurring in the data, that is called *multi-model* [26]. Data could be clustered into differential groups in accordance with their distribution models. The models construct the membership function of the data.

Fuzzy c-means clustering method [45], which satisfies the weaker requirement, is used to make a properly vague partition. The membership value of each datum defines how possible this datum is associated with a category. The membership gives a meaningful explanation on this vagueness. Therefore, to deal with the unavoidable observation and measurement uncertainties, fuzzy clustering is a very suitable choice applied to real world applications.

Theorem 1. *Let* $\otimes$ *be the* fuzzy t-norm *operator. If* $\sum_{i=1}^{b_L} \mu_{L.i}(\mathbf{x}) \leq 1$ *for every* $\mathbf{x} \in S_L$. *Definition* 3 *satisfies the fifth property of entropy. That is*

$$\text{Info}(S_L) \geq \sum_{i=1}^{b_L} \frac{\mathcal{P}_{L.i}}{\mathcal{P}_L} \text{Info}(S_{L.i})$$

Proof: Let α be the maximal membership value for all membership functions. Since $\sum_l \mu_l(\mathbf{x}) \leq 1$ and $\alpha \geq \mu_l(\mathbf{x}), \forall l, \mathbf{x}$ and $\sum_{c \in \mathcal{C}} \mathcal{P}_L^c \leq \mathcal{P}_L$, the right-hand-side of the inequality is derived as follows.

$$\begin{aligned}\sum_{i=1}^{b_L} \frac{\mathcal{P}_{L.i}}{\mathcal{P}_L} \text{Info}(S_{L.i}) &= -\sum_{i=1}^{b_L} \frac{\mathcal{P}_{L.i}}{\mathcal{P}_L} \sum_{c \in \mathcal{C}} \frac{\mathcal{P}_{L.i}^c}{\mathcal{P}_{L.i}} \log_2 \frac{\mathcal{P}_{L.i}^c}{\mathcal{P}_{L.i}} \\ &= -\sum_{i=1}^{b_L} \frac{\mathcal{P}_{L.i}}{\mathcal{P}_L} \sum_{c \in \mathcal{C}} \frac{\mathcal{P}_L^c \otimes \mu_{L.i}}{\mathcal{P}_L \otimes \mu_{L.i}}\end{aligned}$$

$$\times \log_2 \frac{\mathcal{P}_L^c \otimes \mu_{L.i}}{\mathcal{P}_L \otimes \mu_{L.i}}$$
$$\leq -\sum_{c \in \mathcal{C}} \frac{\mathcal{P}_L^c \otimes \alpha}{\mathcal{P}_L \otimes \alpha} \log_2 \frac{\mathcal{P}_L^c \otimes \alpha}{\mathcal{P}_L \otimes \alpha}$$
$$\leq -\sum_{c \in \mathcal{C}} \frac{\mathcal{P}_L^c}{\mathcal{P}_L} \log_2 \frac{\mathcal{P}_L^c}{\mathcal{P}_L}$$
$$= \text{Info}(S_L).$$

□

5. Algorithms

This section presents the learning algorithm for constructing a fuzzy classification tree from a set of training instances containing real-valued attributes. Previous approaches to this problem usually fuzzify the data before they are used to construct a decision tree [36]. The linguistic variables have to be defined ahead of time based on existing domain knowledge.

The main algorithm for FCT construction as shown in Fig. 3 takes an input a set S_0 of instances, and starts by creating a root node N_1, adding its label to $\mathcal{L}$, and initializing S_1 to be S_0.

The fuzzy information gain evaluation is based on the algorithm in Fig. 4. The procedure Spawn_New_Tree(N_L, a_i) that expands the tree from node N_L according to some attribute a_i is shown in Fig. 5. Fuzzy c-means clustering algorithm is then taken to derive the membership of data based on the selected attribute a_i [18].

Suppose there are n attributes with m possible classes, and each attribute (no matter what is category of numerical) is associated with at most v values. Considering the worse case, which is similar to conventional decision tree algorithms, the *Build_FCT* algorithm will use all the attribute values to generate classification rules for each classes. Therefore, its worse case time complexity is $O(n \times m \times v)$.

6. Experiments

The dataset selected is from a general population who were admitted for two-day physical checkups at National Taiwan University Hospital (NTUH) from November 1, 1993 to October 31, 1994. All the subjects had no prior history of any colorectal pathology. During this one-year period, 2987 patients were admitted for physical checkup. A total of 2746 patients who underwent sigmoidoscopy enrolled for the polyp screening analysis. There were 264 patients (9.5%) found to have rectosigmoid polyps by 60 cm-flexible sigmoidoscopy. Since the national health insurance system did not cover the fee of physical checkup, most cases were considered from upper and middle socioeconomic classes.

The purpose of this study was to determine the prevalence of distal large bowel polyps, both adenomatous and hyperplastic. At NTUH, there are about 500 checkup records for each patient in a two-day physical checkup. Sigmoidoscopy using 60cm flexible endoscope without sedation was administered by

Algorithm *Build_FCT*
[Input] A set of training instances S_0
[Output] An FCT
1. $L \leftarrow 1$
/* Initialize L to be 1 which is the label at the root node. */
2. $\mathcal{L} \leftarrow \{1\}$
/* Let $\mathcal{L}$ be the set of labels represented the nodes that have not been expanded. */
3. $S_1 \longleftarrow S_0$
/* S_1 at the root node is set to be the original set S_0. */
4. **loop** until $\mathcal{L} = \phi$
5. $L \leftarrow \text{random}(\mathcal{L})$
/* Random select one of the label from $\mathcal{L}$. */
6. $\mathcal{L} \leftarrow \mathcal{L} \setminus \{L\}$
7. $\forall a_i,\ \tau_i \leftarrow \text{Spawn\_New\_Tree}(N_L, a_i)$
8. Find τ_k s.t. $\text{Info}(\tau_k) = \max_j \text{Info}(\tau_j)$
9. $\text{Gain} \leftarrow \text{Info}(\mathcal{T}_L) - \text{Info}(\tau_k)$
10. **if** $\text{Gain} > \epsilon$ **then**
$\mathcal{L} \leftarrow \mathcal{L} \cup \text{leaf}(\tau_k)$
Assign subsets of S_L into $S_{L.1}, \ldots, S_{L.k}$

Figure 3. The algorithm to construct FCTs.

Algorithm *Evaluate_Entropy*
[**Input**] An FCT with root node N_L
[**Output**] The entropy value of $\mathcal{T}_L$
1. $\forall l \in \mathcal{L}$, s.t. N_l is any node in $\mathcal{T}_L$,
   $\text{Info}(S_l) \leftarrow -1$ /* Initialization */
   /* $\text{Info}(S_l)$ is nonnegative, and therefore set a negative value to it first. */
2. $\forall l \in \mathcal{L}$, s.t. N_l is a leaf node,
   $\text{Info}(S_l) \leftarrow -\sum_{c \in \mathcal{C}} \frac{\mathcal{P}_l^c}{\mathcal{P}_l} \times \ln \frac{\mathcal{P}_l^c}{\mathcal{P}_l}$
3. **loop** until $\text{Info}(S_L) \geq 0$
   if $\forall i, 1 \leq i \leq b_l$ $\text{Info}(S_{l.i}) \geq 0$ **then**
   $\text{Info}(S_l) \leftarrow \sum_{i=1}^{b_l} \frac{\mathcal{P}_{l.i}}{\mathcal{P}_l} \times \text{Info}(S_{l.i})$
end
4. **return** $\text{Info}(S_L)$.

Figure 4. The gain ratio evaluation algorithm.

Algorithm *Spawn_New_Tree*
[**Input**] An unexpanded node N
An attribute a
[**Output**] An expanded tree rooted at node N

$\forall i$, $1 \leq i \leq n$ do the following:
1. *Project* instances at node N of class C_i onto attribute a
2. *Smooth* the resulting histogram using k-median method
3. *Partition* the smoothed histogram into clusters
4. *Create* a new branch from N_L for each cluster
5. *Define* the membership function for each branch

Figure 5. The algorithm to expand the fuzzy classification trees at each node.

experienced endoscopists on all patients except those who gave up this procedure. If polyps were detected, the endoscopists should describe the size, number and location in detail. According to the endoscopic appearance, submucosal tumor, such as leiomyoma, lymphoid follicle, lipoma, and normal mucosa excrescences, were considered as negative findings for this study. Although biopsies might be done at the screening site, it was not mandatory to this study at this stage.

Twenty one attributes, such as blood type, sex, age, body mass index, serum cholesterol, triglyceride, total protein, albumin/globulin, albumin, Zinc Turbit Test, direct bilirubin, total bilirubin, alkaline phosphotatase, acid phosphotatase, alanine aminotransferase, asparate aminotransferase, mean corpuscle volume, hemoglobin, hemoglobin A1C, alcohol consumption, and smoking, were selected for discovering the knowledge about the patients who will get polyp.

6.1. Cross Validation Estimates

A three-fold cross validation for the polyp screening data set was performed. The original data set is randomly split into two parts. One (2/3) is for training, and the other (1/3) is used for testing. FCT and C4.5 methods have been compared across a variety of learning tasks in each experiment.

A frequent application of Bayes' theorem is to evaluate the performance of a diagnostic test intended for use in screening program. Let B denote the event that a person has disease in question; $\bar{B}$ the event that the person does not have the disease; A the event that the person gives a positive response to the test; and $\bar{A}$ the event that the patient gives a negative response.

The results of this trial of the screening test may be represented by the two conditional probabilities $P(A \mid B)$ and $P(A \mid \bar{B})$. The probability $P(A \mid B)$ is the conditional probability of a positive response given that the person has the disease; the larger $P(A \mid B)$ is, the more *sensitive* the test is. The probability $P(A \mid \bar{B})$ is the conditional probability of a positive response given that the person is free of disease; the smaller $P(A \mid \bar{B})$ is (equivalently, the larger $P(\bar{A} \mid \bar{B})$ is), the more *specific* the test is [46].

Let *False Negative* denote the conditional probability that the person who has polyps but obtains negative response, and *False Positive* denote the conditional probability that the person who does not have polyps but gets positive response. Therefore, the less the value of the *False Negative* is, the more sensitive the screening method is. The less the value of the *False Positive* is, the more specific the screening method is. These two values identify two error types. In medical examinations, *False Negative* is the important factor for evaluating the effect of screening method. We need to develop a method with the lowest *False Negative* ratio. On the other hand, we compare these two kinds of error rates instead of the accuracy rates and list in Table 2. There are 211 runs of cross validation tests have been

Table 2. The error rates of the NTUH checkup data set for polyp screening (1).

	Error rate	
Method	*False negative*	*False positive*
FCT	0.226478 ± 0.087654	0.010175 ± 0.007056
C4.5	0.971804 ± 0.020626	0.007173 ± 0.001265

performed. These results were obtained according to the F-test under the confident level of 68%. According to Table 2, the error rate on *False Negative* of C4.5 is 0.971804 ± 0.020626 which is higher than FCT's 0.226478 ± 0.087654. Since 1 minus the value of *False Negative* is the value for sensitivity, we can conclude that the sensitivity of C4.5 is 0.092827 ± 0.001265 and the sensitivity of FCT is 0.989825 ± 0.007056. It means that FCT is more adapted than C4.5 for polyp screening. About 78% patients who have polyps will get positive response without taking colonscope examinations. However, about 1% patients who do not have polyps will be detected to have polyps by FCT, that is less specific than C4.5. The detail is shown in the following section.

In another experiment, five attributes, albumin/globulin, albumin, alanine aminotransferase, asparate aminotransferase, and mean corpuscle volume, are substituted by uric acid, Na^+, K^+, Ca^{++}, and Cl^-. After we performing the three-fold cross validation 200 runs, the error rates of FCT and C4.5 are listed in Table 3. Those substituted attributes are not important in the polyp screening dataset because they seldom occur in a fuzzy classification tree or a C4.5's decision tree, even as they occurred as the tests in the trees are far beneath the root of the tree. However, those five attributes, uric acid, Na^+, K^+, Ca^{++}, and Cl^- are less important than albumin/globulin, albumin, alanine aminotransferase, asparate aminotransferase, and mean corpuscle volume for polyp screening because of the increased error rates.

Table 3. The error rates of the NTUH checkup data set for polyp screening (2).

	Error rate	
Method	*False negative*	*False Positive*
FCT	0.251768 ± 0.092644	0.176667 ± 0.02075
C4.5	0.971804 ± 0.021626	0.007173 ± 0.001746

Table 4. The error rates of the NTUH checkup data set for polyp screening (3).

	Error rate	
Method	*False negative*	*False positive*
FCT	0.352234 ± 0.107644	0.182367 ± 0.120444
C4.5	0.986231 ± 0.010325	0.003242 ± 0.002241

In most of the fuzzy classification trees for polyp screening, we found that age, body mass index, triglyceride, Zinc Turbit Test and hemoglobin A1C were at the important locations (root or near the root as possible) for constructing the classification trees. It seems that these five attributes are the key features for polyp screening. If we substituted uric acid, Na^+, K^+, Ca^{++}, and Cl^- for age, body mass index, triglyceride, Zinc Turbit Test and hemoglobin A1C, the error was increased. Table 4 lists the error rates. Comparing the error rates in Table 4 with Table 3, we can come to the conclusion that some of those attributes, age, body mass index, triglyceride, Zinc Turbit Test and hemoglobin A1C are important for polyp screening.

From those empirical results on polyp screening, we find that FCT is more suitable than C4.5 for polyp screening. Not only FCT is able to make more precise decision for polyp screening, but also FCT is able to properly reflect the effects of features. C4.5 is not capable of doing them.

7. Discussion

In some applications, the classifier is advantageous not to produce a classification on every instance. The classifier is needed to produce the reasonable classification to assist a person to perform the final decision. When there is incomplete or inadequate data, a system that makes no prediction may provide more useful information than a system that makes its best guess on every case. In addition, for disease screening, the classifier should satisfy the following criteria.

- Due to the limitation of medical resources, the classifier needs to identify the patients who do not get the disease and do not need to take any further diagnosis.
- The classifier is able to distinguish the patients who should take a further diagnosis. That is, the classifier can identify the patients who are at the risk of getting the disease.

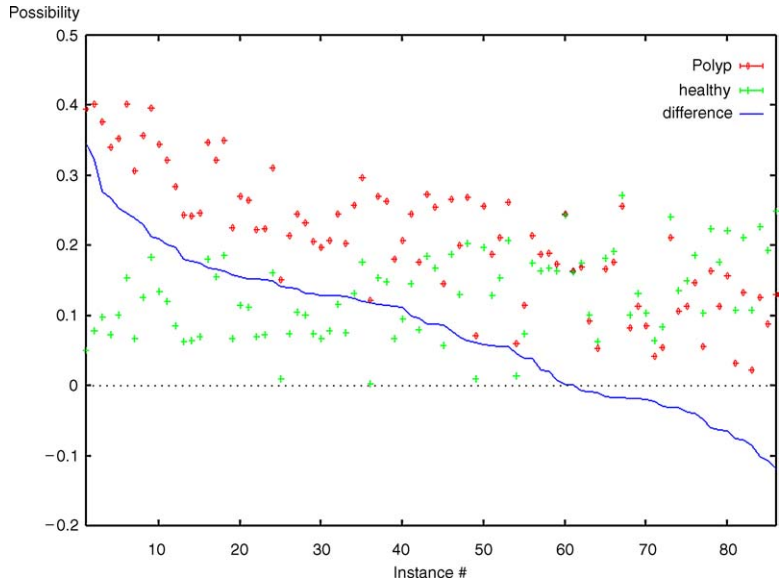

Figure 6. The differences of the predicted possibilities between every "polyp" instances belonging to class `polyp` and class `healthy`.

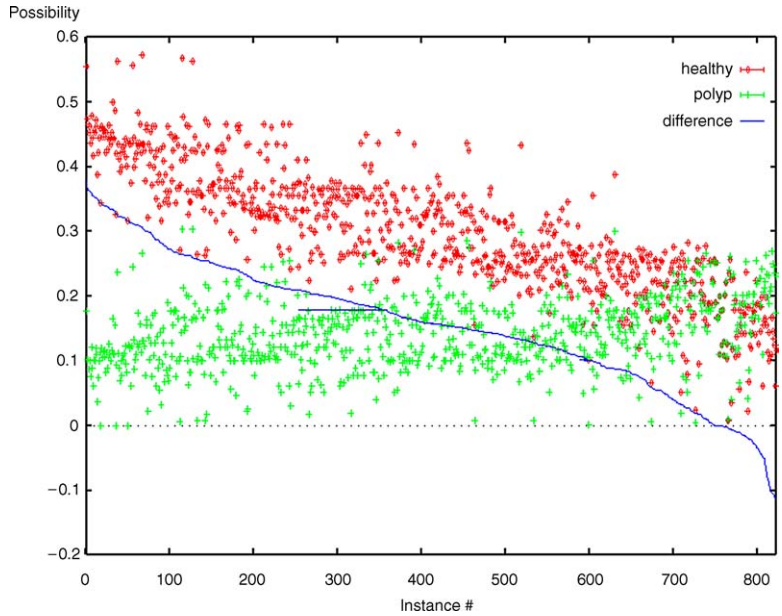

Figure 7. The differences of the predicted possibilities between every health instances belonging to class `polyp` and class `healthy`.

A one-run cross validation test is depicted in Figs. 6 and 7 to check if the FCT system satisfies the requirements of classifiers. Under the condition of the original attributes, a one run cross validation testing result is collected for illustrating the effect of FCT on polyp screening. The data set is divide into two groups: one is the set that all patients have rectosigmoid polyp which is called `polyp` group; the other is the set of the healthy persons which is called `healthy` group. Of course, the `healthy` group contains the persons who were not detected to have rectosigmoid polyp by 60 cm-flexible sigmoidoscopy.

Figure 6 shows the sensitivities of FCT. It shows the predicted possibilities and the possibility differences

Table 5. The ratio of the difference of the predicted possibilities of two classes that is less than a threshold in the NTUH checkup data set for FCT polyp screening.

	Difference between two classes	
Criterion	*False negative*	*False positive*
≤0.15	0.0513002	0.001686
≤0.1	0.226478	0.010175

of an instance between two classes. There are eighty-six patients detected to have rectosigmoid polyps by colonscopic examination. According to the twenty one attributes, FCT and C4.5 can construct classifiers for polyp screening respectively. Basically, a requirement for disease screening strategies is that few false negative results should be determined. Awfully, C4.5 always makes wrong decisions for the patients who have polyps. In this run, only four patients have been detected to have polyps. The decisions of C4.5 are biased to the majority, if only a small proportion of population will get the disease. It can be thought as an *unbalanced learning*. In medical and financial applications, the classifiers should avoid classifying an instance into only one class. It is better to give a possibility result of each class. Then we can make further judgment from the other diagnoses and information.

Figure 7 shows the specificity of FCT. It shows the predicted possibilities and the possibility differences of an instance between two classes.

A useful data mining tool is not expected to substitute human being. The most important is that the tool can help people filter some impossible results. FCT gives each patient the possibility of being in each class. As seen in the Fig. 7, although almost a quarter of patients whose possibilities of being `healthy` are higher than possibilities of being `polyp`, the difference between these two predicted possibilities is always smaller than 0.1. According to our experiments, there are almost no possibility differences greater than 0.15.

7.1. A Reduced Dataset

The 400 patients are extracted from the original polyp dataset to be a reduced subset. It includes 200 polyp patients and 200 healthy persons. We have performed the three-fold, five-fold, and ten-fold cross validation respectively on this subset based on the original twenty one attributes. The empirical results are listed in Table 6

Table 6. The error rates of the three cross validation tests.

		Error rate	
Validation	Method	*False Negative*	*False Positive*
Three-fold	FCT	0.2108 ± 0.06501	0.2306 ± 0.05546
Three-fold	C4.5	0.4243 ± 0.10652	0.3863 ± 0.11224
Five-fold	FCT	0.2017 ± 0.05923	0.2009 ± 0.04944
Five-fold	C4.5	0.3724 ± 0.10736	0.3372 ± 0.10962
Ten-fold	FCT	0.2008 ± 0.03612	0.1874 ± 0.04017
Ten-fold	C4.5	0.3008 ± 0.08716	0.2918 ± 0.08193

where each cross validation test has been performed ten times.

The ten-fold cross validation test is always considered to be a unbiased test [47, 48] in machine learning, because the accuracy rate of the ten-fold cross validation test is more precise than the accuracy rate of the others (three-fold and five-fold) for most of classifiers. In this reduced subset, the change of the error rates of FCT is less than the change of the error rates of C4.5 from the three-fold cross validation test to the ten-fold cross validation test. Comparing the results of the reduced set with the results of original set, we can make a further conclusion that those attributes can not provide a distinct classification for polyp screening. The reason may be due to the data is incomplete or inadequate. Therefore, we should collect enough information for classifications and collect the patient data of real total colon examinations.

8. Conclusion

The uncertainties and noise make classification difficult. Missing or imprecise information may prevent a case from being classified at all. It is occurred in the boundaries of the data in two more different classes [49, 50]. In the presence of uncertainties, it is often desirable to have an estimate of the degree that an instance is in each class.

Probabilistic tree classifiers [15, 22, 26, 29, 31] have been proposed to deal with uncertainties and noise. However, the *a priori* probabilities are needed to explain the result of classifications. In addition, probabilistic tree classifiers do not give a good solution for data partition. For numerical attributes, discretization [11, 51] makes the data in the overlapped region be classified into only one branch. A test instance falls down a single branch to arrive at a leaf where a

probability is associated with each class. Such classifications ignore the possibility that instance could be classified into the other branches.

In a fuzzy classification tree, an instance has a membership value at each node. Instead of determining a single class for any given instance, fuzzy classification trees can predict the degree of *possibility* for every class. Using information-based measures, there is no need to generate multiple classification trees. Therefore, it requires less time and space than decision forests [38].

C4.5 is totally useless for polyp screening. All the patients who have polyps are almost classified into the `healthy` class. Basically, a requirement for disease screening strategies is that few false negative results should be determined. Unfortunately, C4.5 always makes wrong decisions for the patients who have polyps. Only few instances can be clearly classified. The testing result of the checkup dataset is formally under the consideration of F-test at the confident level of 95%. Using the three-fold cross validation testing, we will see that the error rate on false negative of FCT is less than the error rates on false negative of C4.5. That is, FCT is more sensitive than C4.5. The decisions of C4.5 are always biased to the majority, if only a small proportion of population will get the disease. In medical and financial applications, it is important that a classifier should give the estimate degrees of all potential classes. The classifiers should avoid classifying an instance into only one class. The fuzzy classifier, fuzzy classification trees, can estimate the possible degrees of all classes. According to these possibilities, even if we pick the class with the high possibility to be the patient's class, a much better prediction can be made by FCT than by C4.5.

References

1. E. Sato, A. Ouchi, and T. Ishidate, "Polyps and diverticulosis of large bowel in autopsy, population of Akita prefecture, compared with Miyagi: High rate of colorectal cancer in Japan," *Cancer*, vol. 37, pp. 1316–1321, 1976.
2. J. Sauar, G. Hoff, and T. Hausken, "Colonoscopic screening examination of relatives of patients with colorectal cancer," *Scandinavian Journal of Gastroenterology*, vol. 27, pp. 667–672, 1992.
3. M. Shieh, I. Chiang, J. Wong, C. Huang, S. Huang, and C. Wang, "Prevalence of colorectal polyps in Taiwan: 60cm-sigmoidoscopic findings," *Biomedical Engineering-Application, Basis, Communication*, vol. 7, no. 3, pp. 50–55, 1995.
4. A.R. Williams, B.A. Balasooriya, and D.W. Day, "Polyps and cancer of the large bowel: A necropsy study in Liverpool," *Gut*, vol. 23, pp. 835–842, 1982.
5. Z. Pawlak, "*Rough Sets*," Kluwer Academic, Dordrecht, 1991.
6. D. Heath, S. Kasif, and S. Salzberg, "Learning oblique decision trees," in *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*, Chambery, France, 1993, pp. 1002–1007.
7. S.K. Murthy, "*On growing better decision trees from data*," PhD dissertation, The Johns Hopkins University, Baltimore, Maryland, 1995.
8. S.K. Murthy, S. Kasif, and S. Salzberg, "A system for induction of oblique decision trees," *Journal of Artificial Intelligence Research*, vol. 2, pp. 1–32, 1994.
9. S.K. Murthy, S. Kasif, S. Salzberg, and R. Beigel, "OC1: Randomized induction of oblique decision trees," in *Proceedings of the Eleventh National Conference on Artificial Intelligence*, Washington, DC, 1993, pp. 322–327.
10. J.R. Quinlan, "Decision trees and decision making," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 20, pp. 339–346, 1990.
11. J.R. Quinlan, *C*4.5*: Programs for Machine Learning*, Morgan Kaufmann: Los Altos, CA, 1993.
12. P. Clark and T. Niblett, "The CN2 induction algorithm," *Machine Learning*, vol. 3, pp. 261–283, 1989.
13. R. Rivest, "Learning decision lists," *Machine Learning*, vol. 2, pp. 229–246, 1987.
14. M. Sugeno and G.T. Kang, "Structure identification of fuzzy model," *Fuzzy Sets and Systems*, vol. 28, pp. 15–33, 1988.
15. J. Schuermann and W. Doster, "A decision theoretic approach to hierarchical classifier design," *Pattern Recognition*, vol. 17, no. 3, pp. 359–369, 1984.
16. I. Chiang and J. Hsu, "Integration of fuzzy classifiers with decision trees," in *Proceedings of Asian Fuzzy Systems Symposium*, Kenting, Taiwan, 1996, pp. 65–78.
17. J.Y. Hsu and I. Chiang, "Fuzzy classification trees," in *Proceedings of the Ninth International Symposium on Artificial Intelligence*, Cancun, Mexico, 1996, pp. 431–438.
18. I. Chiang and J. Hsu, "Fuzzy classification trees for data analysis," *Fuzzy Sets and Systems*, vol. 13, no. 1, pp. 87–99, 2002.
19. R. Duda and P. Hart, *Pattern Classification and Scene Analysis*, John Wiley and Sons: New York, 1973.
20. D. Rumelhart, G. Hinton, and R. Williams, "Learning internal representations by error prpagation," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations*, edited by D. Rumelhart and J. McClelland, MIT Press, Cambridge, MA, 1986, pp. 318–362.
21. M. Pazzani and D. Kibler, "The utility of knowledge in inductive learning," *Machine Learning*, vol. 9, no. 1, pp. 57–94, 1991.
22. J.R. Quinlan, "Learning logical definitions from relations," *Machine Learning*, vol. 5, pp. 239–266, 1990.
23. J. Dougherty, R. Kohavi, and M. Sahami. "Supervised and unsupervised discretization of continuous features," in *Proceedings of the Twelfth International Conference on Machine Learning*, San Mateo, CA, 1995, pp. 194–202.
24. U.M. Fayyad and K.B. Irani, "On the handling of continuous-valued attributes in decision tree generation," *Machine Learning*, vol. 8, pp. 87–102, 1992.
25. W. Buntine, Myths and legends in learning classification rules, in *Proceedings of the Eighth National Conference on Artificial Intelligence*, Boston, MA, 1990, pp. 736–742.

26. W. Buntine, "Learning classification trees," *Statistics and Computing*, vol. 2, pp. 63–73, 1992.
27. A.M. Mood, F.A. Graybill, and D.C. Boes, *Introduction to the Theory of Statistics, 3rd edition*, McGraw-Hill: New York, 1975.
28. J.R. Quinlan, "Probabilistic decision trees," in *Proceedings of the Fourth International Workshop on Machine Learning*, edited by P. Langley, Los Altos, CA, 1987.
29. L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*, Chapman and Hall: London, 1984.
30. M.I. Jordan and R.A. Jacobs, "Supervised learning and divide-and-conquer: A statistical approach," in *Proceedings of the Ninth International Conference on Machine Learning*, Aberdeen, Scotland, 1992, pp. 159–166.
31. R.G. Casey and G. Nagy, "Decision tree design using a probabilistic model," *IEEE Transactions on Information Theory*, vol. 30, no. 1, pp. 93–99, 1984.
32. R. Rymon, "An SE-tree based characterization of the induction problem," in *Proceedings of the Tenth International Conference on Machine Learning*, Amherst, MA, 1993, pp. 268–275.
33. X. Boyen and L. Wehenkel, "Automatic induction of fuzzy decision trees and its application to power system security assessment," *Fuzzy Sets and Systems*, vol. 102, pp. 3–19, 1999.
34. K.J. Cios and L.M. Sztandera, "Continuous ID3 algorithm with fuzzy entropy measures," in *Proceedings of the International Conference on Fuzzy Systems*, San Diego, CA, 1992, pp. 469–476.
35. A. Suárez and J.F. Lutsko, "Globally optimal fuzzy decision trees for classification and regression," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 21, no. 12, pp. 1297–1311, 1999.
36. Y. Yuan and M.J. Shaw, "Induction of fuzzy decision trees," *Fuzzy Sets and Systems*, vol. 69, pp. 125–139, 1995.
37. W. Pedrycz and Z.A. Sosnowski, "The design of decision trees in framework of granular data and their application to software quality models," *Fuzzy Sets and Systems*, vol. 123, pp. 271–290, 2001.
38. P.M. Murphy and M.J. Pazzani, "Exploring the decision forest: An empirical investigation of Occam's razor in decision tree induction," *Journal of Artifical Intelligence Research*, vol. 1, pp. 257–275, 1994.
39. C.Z. Janickow, "Fuzzy decision trees: Issues and methods," *IEEE Trans. on System, Man, and Cybernetics B: Cybernetics*, vol. 28, no, 1, pp. 1–14, 1998.
40. P.W. Baim, "A method for attribute selection in inductive learning system," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 10, no. 9, pp. 888–896, 1988.
41. J. Mingers, "An empirical comparison of selection measures for decision-tree induction," *Machine Learning*, vol. 3, pp. 319–342, 1989.
42. J.R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, pp. 81–106, 1986.
43. C.E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, 1948.
44. A. De Luca and S. Termini, "A definition of a nonprobabilistic entropy in the setting of fuzzy sets theory," *Information and Control*, vol. 20, pp. 301–312, 1976.
45. J.C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press: New York, 1981.
46. J. Yerushalmy, "Statistical problems in assessing methods of medical diagnosis, with special reference to X-ray techniques," *Public Health Reports*, vol. 62, pp. 1432–1449, 1947.
47. P. Murphy and D. Aha, "UCI repository of machine learning databases," Department of Information and Computer Science, University of California at Irvine, 1992.
48. A.P. White and W.Z. Liu, "Bias in information-based measures in decision tree induction," *Machine Learning*, vol. 15, pp. 321–329, 1994.
49. R.S. Michalski, "Learning flexible concepts: Fundamental ideas and method based on two-tiered representation," in *Machine Learning: An Artificial Intelligence Approach*, vol. III, edited by Y. Kodratoff and R.S. Michalski, Morgan Kaufmann, Los Altos, CA, 1990.
50. L. Rendell and H. Cho, "Empirical learning as a function of concept character," *Machine Learning*, vol. 5, no. 3, pp. 267–298, 1990.
51. R. Kerber, "ChiMerge: Discretization of numeric attributes," in *Proceedings of the Tenth National Conference on Artificial Intelligence*, San Jose, CA, 1992, pp. 123–128.