Journal of Automated Reasoning **18:** 5–24, 1997.

A New Probabilistic Induction Method

RONG-HUEI HOU
Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan, R.O.C.

TZUNG-PEI HONG*
Department of Information Management, Kaohsiung Polytechnic Institute, Kaohsiung, Taiwan, R.O.C.

SHIAN-SHYONG TSENG and SY-YEN KUO
Department of Computer and Information Science, National Chiao-Tung University, Hsin-Chu, Taiwan, R.O.C.

(Received: 12 June 1993)

Abstract. Knowledge acquisition by interviewing a domain expert is one of the most problematic aspects of the development of expert systems. As an alternative, methods for inducing concept descriptions from examples have proven useful in eliminating this bottleneck. In this paper, we propose a probabilistic induction method (PIM), which is an improvement of the Chan and Wong method, for detecting relevant patterns implicit in a given data set. PIM uses the technique of residual analysis and several heuristics to effectively detect complex relevant patterns and to avoid the problem of combinatorial explosion. A reasonable trade-off between the induction time and the classification ratio is achieved. Moreover, PIM quickly classifies unknown objects using classification rules converted from the positively relevant patterns detected. Three experiments are conducted to confirm the validity of PIM.

Key words: adjusted residual, induction, probabilistic, relevant pattern.

1. Introduction

Much progress has been made in the field of artificial intelligence in recent years. Among the most significant areas of progress has been the development of powerful computer systems known as 'expert' or 'knowledge-based' systems. These systems are designed to represent and to apply factual knowledge from specific areas of expertise to solve problems [5–7]. Expert systems are knowledge intensive, and the process of acquiring the necessary knowledge by interviewing domain specialists is tedious and difficult, since the experts are usually unaware of how to express their knowledge effectively. Moreover, the knowledge acquired is often incomplete, inconsistent, or irrelevant. Methods for inducing concept descriptions from examples, on the other hand, have proved useful in facilitating the acquisition of knowledge in constructing expert systems [21–23].

The field of machine learning has been studied by many researchers over the past two decades, and many approaches toward machine learning have been pro-

* Responsible for correspondence.

posed [13–15]. Some research on machine learning has focused on techniques for developing concept descriptions from training examples. Given a set of examples and counterexamples, the learning program tries to induce general concept descriptions that describe the positive training instances and exclude the negative ones.

Many factors are involved in designing a good learning system [10]. One of the most important of these factors is reducing the influence of noise [9, 18]. Noise means incorrect values or incorrect classes in the training instances. Incorrect training instances may originate from unreliable or incorrect information or may arise from input error. Two cases exist for wrong classification of a training instance. An originally positive training instance that is wrongly classified into the negative class is called a '*false negative*' training instance; an originally negative training instance that is wrongly classified into the positive class is called a '*false positive*' training instance. Attribute values of training instances may also be given incorrectly. For example, '*size* = 20' may be wrongly given as '*size* = 10'.

The validity and relevance of the concepts eventually learned depend heavily on the accuracy of the chosen training instances. In real applications, data provided to learning systems by experts, teachers, or users usually contain noise. Noise can be expected to affect the formation and use of the learned concepts in two ways [18]:

(1) Learning strategies usually must apply some form of generalization to derive the desired concepts. This is usually accomplished by identifying subsets of the given training instances that share common properties. Noise in the training instances will tend to confuse a generalization mechanism of this type.

(2) Problems arise when the concepts formed from the noisy training instances are used to classify other objects. Since the given training instances contain noise, the learned concepts may contain errors, and the results obtained by using the incorrect concepts to classify the objects in question might well be incorrect also.

Many induction methods for managing noisy data sets have been proposed [4, 16, 19, 22]. One of these, the probabilistic induction system proposed by Chan and Wong in 1990, reduced the effect of noise by using a statistical approach [2, 3]. This approach first detects the relevant classification patterns of a single attribute-value pair based on residual analysis. It then constructs rules with individual weights based on the relevant patterns detected. Finally, it uses the rules constructed to classify unknown data. Although Chan and Wong's method indeed works in noisy domains, it is still subject to several disadvantages:

(1) It considers only the relevant relations between a single attribute-value pair and a class and thus misses more complex patterns.

(2) It allows negatively relevant patterns to be converted into rules, thus causing erroneous classification when no positively relevant patterns are matched.

(3) It requires all of the rules constructed to be checked each time an unknown object is classified, thus making the classification process very inefficient.

This paper proposes a new probabilistic induction method (PIM) [11] that avoids the above three problems. The new method proposed here uses heuristics to find complex relevant patterns and to avoid the problem of combinatorial explosion. It also uses the information implicit in the negatively relevant patterns to find other positively relevant patterns so as to help raise the classification ratio. In addition, when classifying an object, PIM stops checking the remaining rules once the first rule matched is found.

Three experiments, involving fitting contact lenses domain [1], simulated artificial data [1], and brain tumor diagnosis [22], were conducted to evaluate the performance of PIM. The results reveal that PIM indeed is more accurate than Chan and Wong's method.

The remaining parts of this paper are organized as follows. The terms and notations used in this paper are introduced in Section 2. Chan and Wong's method is reviewed in Section 3, and some problems with their method are then discussed in Section 4. In Section 5, a new probabilistic induction method (PIM) and a corresponding reasoning method are proposed to improve on Chan and Wong's method, and an example is provided to illustrate their use. Experiments conducted to verify the proposed methods are described in Section 6. The results of the paper are summarized in Section 7.

2. Terms and Notations

The following notation will be employed in this paper:

- N, the number of training instances,
- M, the number of available attributes,
- A_j, the jth attribute, where $1 \leqslant j \leqslant M$,
- a_j, the number of possible values of A_j,
- v_{ji}, the ith value of A_j, where $1 \leqslant i \leqslant a_j$,
- C, the number of possible classes to be classified,
- p, the pth class, where $1 \leqslant p \leqslant C$,
- o_p, the total number of objects in the training set that belong to class p,
- o_{ji}, the total number of objects with characteristic v_{ji},
- o_{pji}, the total number of objects in the training set that belong to class p and are characterized by v_{ji},
- e_{pji}, the expected number of objects in the training set that belong to class p and are characterized by v_{ji} under the assumption that values of attribute A_j are uniformly distributed in class p,
- r_{pji}, the adjusted residual between the characterization v_{ji} and the class p.

Each instance is represented by one or several discrete attributes and its class. If an attribute value of a training instance is originally continuous, it is first

transformed into a discrete value by a discretization technique. Each instance is then represented as follows:

$$A_1 = v_{1i_1}, \quad A_2 = v_{2i_2}, \ldots, A_m = v_{mi_m}, \quad \text{class} = p.$$

The adjusted residual r_{pji} is defined as follows [8]:

$$r_{pji} = z_{pji}/\sqrt{y_{pji}}, \tag{1}$$

where

$$z_{pji} = (o_{pji} - e_{pji})/\sqrt{e_{pji}},$$

and

$$y_{pji} = (1 - o_p/N)(1 - o_{ji}/N).$$

If the absolute value of r_{pji} is greater than 1.96, or 95% of the normal distribution, the attribute-value pair $A_j =$ the ith value (v_{ji}) is considered an important feature for class p. The sign of r_{pji} is also important: $r_{pji} \geqslant +1.96$ indicates that v_{ji} is a positively relevant characteristic of class p, whereas $r_{pji} \leqslant -1.96$ indicates that v_{ji} is a negatively relevant characteristic of class p (meaning an object with characteristic v_{ji} has a low probability of belonging to class p).

An elementary pattern is a pattern with only one attribute-value pair. A complex pattern is a pattern with more than one attribute-value pair. A positively relevant pattern is a pattern for which the value of adjusted residual is greater than or equal to $+1.96$. A negatively relevant pattern is a pattern for which the value of the adjusted residual is less than or equal to -1.96. A pattern is irrelevant iff it is not relevant. A dominated pattern is a complex pattern composed of an elementary positively relevant pattern and one or more positively irrelevant patterns. A nondominated pattern is a complex pattern composed of only positively irrelevant patterns.

3. Review of Chan and Wong's Method

To effectively handle uncertainty and noise in classification tasks, Chan and Wong developed a probabilistic induction system (PIS) based on the idea of analyzing residuals in statistic and weighting evidence in information theory. PIS is capable of detecting and extracting statistically significant patterns from a set of data, and is thus effective in dealing with uncertainty and noise. Chan and Wong's method works well even when (1) the data contains inaccurate, incomplete, and inconsistent values, or (2) the training sample size is relatively small [2, 3]. Chan and Wong's method consists of three main phases:

1. detecting relevant patterns of a single attribute-value pair through residual analysis;

2. constructing classification rules with individual weights based on the relevant patterns detected; and

3. using the classification rules to classify the objects in question.

3.1. DETECTING RELEVANT PATTERNS

Attributes important to classification usually need to be identified in order to form efficient rules. In the past, some systems use the chi-square test to find attributes that are statistically dependent on classes. The chi-square test, however, only evaluates each attribute as a whole even though some of the possible values may be irrelevant to the classification. Instead of using the chi-square test to find relevant attributes, Chan and Wong's method uses the adjusted residual to find relevant attribute-value pairs [2, 3]. If the absolute value of the adjusted residual of an attribute-value pair for a class is greater than 1.96 (this value can be arbitrarily assigned by the user) or considered 95% of the normal distribution, then the attribute-value pair is an important feature for that class. The sign of the adjusted residual is also important. A positive value indicates that an instance with the attribute-value pair very likely belongs to the class; a negative value indicates that an instance with the attribute-value pair is very unlikely to belong to the class.

3.2. CONSTRUCTING CLASSIFICATION RULES

The relevant patterns detected are used to construct the classification rules. If an attribute-value pair $A_j = v_{ji}$ is relevant (positive or negative) to a certain class p, a rule is formed as follows:

$$\text{If} \quad A_j = v_{ji}, \quad \text{then class} = p$$

$$\text{with} \quad W = \log \frac{\text{Probability(attribute } A_j = v_{ji} \mid \text{class} = p)}{\text{Probability(attribute } A_j = v_{ji} \mid \text{class} \neq p)} .$$

This type of rules probabilistically describes the relation between an attribute-value pair and a class.

3.3. CLASSIFICATION OF UNKNOWN OBJECTS

Suppose that an unknown object *obj* is described by N attributes. The set of classification rules (constructed by Step 2) is searched to determine which rules match *obj*, and the object *obj* belongs to a class with weight equal to the sum of the weights of the matched rules for that class. Among all possible classes, *obj* is assigned to the class with the maximum weight as the classification result.

Although Chan and Wong's method indeed works well in noisy domains, it also has several problems, as discussed below.

4. Some Problems with Chan and Wong's Method

The first problem with Chan and Wong's method is that only relevant patterns of a single attribute-value pair are detected. In real-world application domains,

data with dependent attribute-value pairs are very commonly seen. For example, attribute-value pair $A1$ $(A = 1)$ may show no correlation with class $\delta 1$ and attribute-value pair $B1$ $(B = 1)$ may also show no correlation with class $\delta 1$. However, $A1$ and $B1$ together may be relevant to Class $\delta 1$. This kind of complex relevant pattern cannot be detected by Chan and Wong's method, and thus the rules constructed include no preconditions for $A1$ and $B1$. Relevant patterns with more than one attribute-value pair must be found to raise the classification ratio.

The second problem is that negatively relevant patterns are also converted into rules with negative weights, thus causing wrong classification when no positively relevant patterns are matched. For example, assume that by Chan and Wong's method, attribute-value pair $A1$ is only negatively relevant to Class $\delta 1$, and $B1$ and $C1$ are irrelevant to any class. Also assume a training instance with description $(A1, B1, C1)$ is to be classified. According to the rule of classification, the class with the maximum weight is to be assigned to the instance. In this example, only $\delta 1$ is matched, and it is then assigned to the instance, causing a classification error. As another example, assume that by Chan and Wong's method, $A1$ is only negatively relevant to Class $\delta 1$, $B1$ is only negatively relevant to Class $\delta 2$, and $C1$ is only negatively relevant to Class $\delta 3$. Again, assume a training instance with description $(A1, B1, C1)$ is to be classified. In this case, all three classes have negative weights, representing negative evidence. The class with the maximum weight is then quite unlikely to be the class of the instance. Rather than merely acting as rules with negative weights, negatively relevant patterns should provide a clue to finding complex positively relevant patterns that can be used to raise the classification ratio.

The third problem is that all rules constructed must be checked in order to classify an unknown object. As mentioned before, when an unknown object is presented, all rules are searched, the weights of the matched rules for each class are added, and the class with the maximum total weight is chosen as the desired class. All of this computation must be done for each new object. As a better alternative, rules with complex patterns can be found from the elementary relevant patterns and then sorted according to their weights. In this approach, the first rule that matches an unknown object determines the class to which the object belongs, and the remaining rules need not be checked. This is a form of '*knowledge compilation*'.

In the next section, a new probabilistic induction method and a corresponding reasoning method are proposed that remove the disadvantages of Chan and Wong's method.

5. A New Probabilistic Induction Method

The probabilistic induction method (PIM) proposed here can detect complex relevant patterns instead of patterns of only a single attribute-value pair. It uses heuristics to avoid the problem of combinatorial explosion. Moreover, only posi-

tively relevant patterns are converted into classification rules. Negatively relevant patterns are not converted into rules, but rather are used to derive other positively relevant classification rules. Several heuristics are used in our algorithm to help find complex relevant patterns more efficiently and thus alleviate the problem of combinatorial explosion.

5.1. USE OF HEURISTICS

The heuristics are based on the following four observations.

Observation 1

Suppose the elementary positively relevant patterns $(A1,\ \delta 1)$ and $(B1,\ \delta 1)$ have been detected. Since $A1$ and $B1$ both show a positive correlation with class $\delta 1$, $A1$ and $B1$ together will intuitively show more correlation with $\delta 1$ than either only $A1$ or only $B1$. The first heuristic used in the algorithm is that a complex pattern composed of two elementary positively relevant patterns is also most likely to be positively relevant.

Observation 2

Suppose $(A1,\ \delta 1)$ is an elementary positively relevant pattern and $(B1,\ \delta 1)$ is irrelevant. Since $B1$ shows no correlation with class $\delta 1$, $A1$ and $B1$ together will intuitively show no more correlation with $\delta 1$ than only $A1$ alone. Furthermore, a rule constructed using only $A1$ will be more general than a rule based on $A1$ and $B1$ together. The second heuristic used in the algorithm then stipulates that a dominated pattern should not be added to the rule set.

Observation 3

Suppose $(A1,\ \delta 1)$ is an elementary negatively relevant pattern, meaning any instance with attribute value $A1$ has a very low probability of being classified as $\delta 1$. Since classes are mutually exclusive, the implicit meaning of this rule is that any instance with attribute value $A1$ has a high probability of being classified as $\delta 2$ or $\delta 3$ (assuming only three classes may exist). In other words, an elementary negatively relevant pattern $(A1,\ \delta 1)$ implies $A1$ may have some correlation with $\delta 2$ or $\delta 3$. Assume $(A1,\ \delta 2)$ and $(A1,\ \delta 3)$ are not positively relevant. The complex patterns composed of $(A1,\ \delta 2)$ or $(A1,\ \delta 3)$ with other patterns detected in this way are more likely to be positively relevant than patterns composed of any two randomly chosen irrelevant patterns. So, the third heuristic used in the algorithm is that elementary negatively relevant patterns should be utilized to find some other complex positively relevant patterns.

Observation 4

Suppose both $(A1,\ \delta 1)$ and $(B1,\ \delta 1)$ are irrelevant patterns. Since neither $A1$ nor $B1$ shows a correlation with class $\delta 1$, it is difficult to determine whether $A1$ and $B1$ together will be relevant to class $\delta 1$. The adjusted residual must be calculated to answer this question. So, the fourth heuristic used in the algorithm is that irrelevant patterns should be utilized as late as possible to find some other complex positively relevant patterns.

5.2. THE LEARNING ALGORITHM

The above heuristics are the basis for the induction algorithm, which is outlined below.

The New Probabilistic Induction Algorithm

STEP 1. Find all elementary relevant patterns (both positively and negatively relevant);
STEP 2. Retain elementary positively relevant patterns in the candidate set;
STEP 3. Form complex positively relevant patterns with K attribute-value pairs (initially $K = 2$) by combining elementary positively relevant patterns. Retain them in the candidate set;
STEP 4. Detect nondominated positively relevant patterns with K attribute-value pairs by utilizing the elementary negatively relevant patterns. Retain them in the candidate set;
STEP 5. Construct classification rules with weights from the positively relevant patterns in the candidate set;
STEP 6. Sort the classification rules by the weights;
STEP 7. If the percentage of the undetermined training examples is smaller than a predefined threshold, stop the induction algorithm and add a default rule by setting the most common class in the undetermined training examples as the default class; otherwise, do Step 8;
STEP 8. Find all other nondominated positively relevant patterns each with K attribute-value pairs by the residual analysis. If no such relevant patterns are detected, stop the induction algorithm and add a default rule; otherwise, go to Step 9;
STEP 9. Convert the new positively relevant patterns into classification rules. If the percentage of the undetermined training examples is smaller than a predefined threshold, stop the induction algorithm and add a default rule; otherwise, set $K = K + 1$ and go to Step 3.

A flowchart of the learning algorithm is shown in Figure 1.

The training set in Table I is used as an example through this section to illustrate the operation of the algorithm. This training set is composed of four attributes, $a,\ b,\ c$, and d, and three classes, $\delta_1,\ \delta_2$, and δ_3.

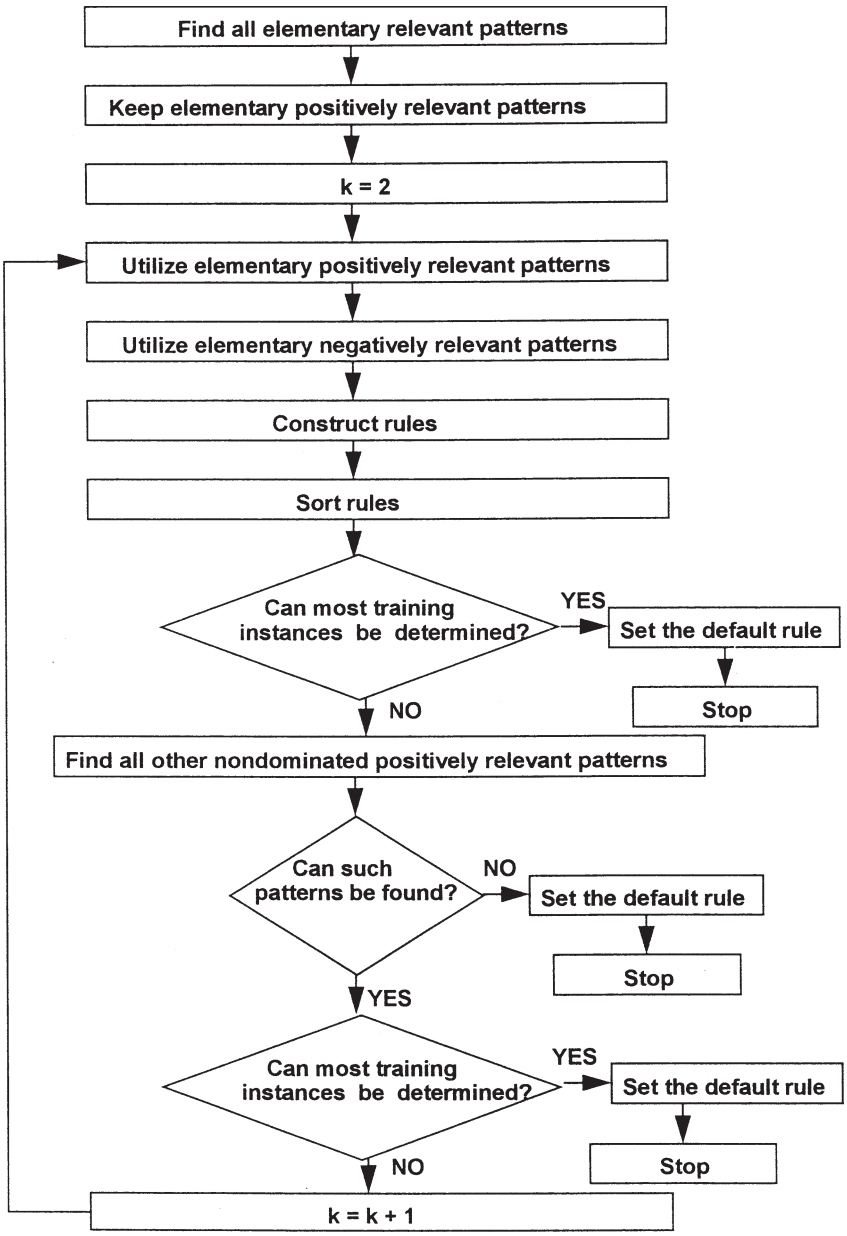

Figure 1. System flowchart.

STEP 1. Find all elementary relevant patterns

An elementary pattern is a pattern with only one attribute-value pair. The adjusted residual of all elementary patterns are first calculated by Formula (1). Take

Table I. Data table for the example

	Value of attribute				class		Value of attribute				class		Value of attribute				class
	a	b	c	d	δ		a	b	c	d	δ		a	b	c	d	δ
1	1	1	1	1	2	9	2	1	1	1	1	17	3	1	1	1	1
2	1	1	1	2	1	10	2	1	1	2	3	18	3	1	1	2	3
3	1	1	2	1	2	11	2	1	2	1	1	19	3	1	2	1	1
4	1	1	2	2	1	12	2	1	2	2	3	20	3	1	2	2	3
5	1	2	1	1	2	13	2	2	1	1	1	21	3	2	1	1	2
6	1	2	1	2	1	14	2	2	1	2	3	22	3	2	1	2	2
7	1	2	2	1	2	15	2	2	2	1	1	23	3	2	2	1	2
8	1	2	2	2	1	16	2	2	2	2	3	24	3	2	2	2	2

Table II. Adjusted residuals of all elementary patterns (items with * represent relevant patterns)

	δ_1	δ_2	δ_3
a_1	0.586	1.225	$^{*}-2.000$
a_2	0.586	$^{*}-2.449$	$^{*}\ 2.000$
a_3	-1.171	1.225	0.000
b_1	0.828	-1.732	0.943
b_2	-0.828	1.732	-0.943
c_1	0.000	0.000	0.000
c_2	0.000	0.000	0.000
d_1	0.828	1.732	$^{*}-2.828$
d_2	-0.828	-1.732	$^{*}\ 2.828$

(a_1, δ_1) as an example to illustrate the finding of the adjusted residuals (here attribute a is considered the first attribute). In this example, $p = 1,\ j = 1$, and $i = 1$, and thus the adjusted residual r_{111} is to be found. From Table I, o_{111} is equal to 4 (the second, the fourth, the sixth, and the eighth training instances), and e_{111} is equal to $10/3$ (since ten training instances belong to class δ_1 and attribute a has three possible values). Moreover, o_1 is equal to 10, o_{11} is equal to 8, and N is equal to 24. By Formula (1), r_{111} is then calculated as 0.586, meaning that the attribute a_1 is not a relevant characteristic of the class δ_1. The other adjusted residuals are calculated in the same way. Table II shows the results for the training set in Table I.

When the absolute values of the adjusted residuals are compared with 1.96, 95% of the normal distribution, the elementary relevant patterns are found to be as follows:

$$(a_2\ \delta_3\ +), (d_2\ \delta_3\ +);$$
$$(a_1\ \delta_3\ -), (a_2\ \delta_2\ -), (d_1\ \delta_3\ -),$$

where $+$ indicates a positively relevant pattern and $-$ indicates a negatively relevant one.

STEP 2. Retain elementary positively relevant patterns in the candidate set

Only the positively relevant patterns are retained here to construct rules. For this example, only the following elementary positively relevant patterns are considered:

$$(a_2\ \delta_3\ +), (d_2\ \delta_3\ +).$$

STEP 3. Form complex positively relevant patterns with K attribute-value pairs ($K = 2$ initially) by combining elementary positively relevant patterns. Retain them in the candidate set

The first heuristic is used here to detect complex positively relevant patterns by combining elementary positively relevant patterns. Suppose both $(a_2\ \delta_3\ +)$ and $(d_2\ \delta_3\ +)$ are elementary positively relevant patterns. PIM presumes the combinatorial pattern $(a_2\ d_2\ \delta_3\ +)$ is also a positively relevant pattern. For the training set in Table I, the following positively relevant patterns are detected:

$$(a_2\ \delta_3\ +), (d_2\ \delta_3\ +), (a_2\ d_2\ \delta_3\ +).$$

STEP 4. Detect nondominated positively relevant patterns with K attribute-value pairs by utilizing the elementary negatively relevant patterns. Retain them in the candidate set

According to the second heuristic, a dominated pattern will not be retained in the candidate set, since the elementary positively relevant pattern in the dominated pattern is enough to represent the correlation. Therefore only nondominated patterns are checked here.

In Step 1, all the elementary positively and negatively relevant patterns are detected. In Step 2 and Step 3, only the elementary positively relevant patterns and the combinations of such patterns are considered. In this step, elementary negatively relevant patterns are used to find further complex positively relevant patterns (the third heuristic).

For example, $(a_2\ \delta_2\ -)$ is an elementary negatively relevant pattern in the data set. This rule indicates that any instance with attribute value a_2 can be classified as δ_2 with low probability. According to the third heuristic, an elementary negatively relevant pattern $(a_2\ \delta_2\ -)$ implies a_2 may have some correlation with δ_1 or δ_3. Since the elementary positively relevant pattern $(a_2\ \delta_3\ +)$ was detected in Step 1, $(a_2\ \delta_3\ +)$ will not be considered again, because it is relevant. Only the

combinations of the pattern $(a_2\ \delta_1)$ with other possible patterns are tested by the residuals.

For the data set in Table I, the elementary negatively relevant patterns are processed as follows:

$(a_1\ \delta_3\ -) \rightarrow (a_1\ \delta_1)$: considered and $(a_1\ \delta_2)$: considered,
$(a_2\ \delta_2\ -) \rightarrow (a_2\ \delta_1)$: considered and $(a_2\ \delta_3)$: not considered,
$(d_1\ \delta_3\ -) \rightarrow (d_1\ \delta_1)$: considered and $(d_1\ \delta_2)$: considered.

Therefore, the following three possible nondominated positively relevant patterns are formed:

$$(a_1\ d_1\ \delta_1), (a_1\ d_1\ \delta_2), (a_2\ d_1\ \delta_1).$$

The adjusted residuals of these three nondominated patterns are then calculated by Formula (1). Results are shown as follows:

$(a_1\ d_1\ \delta_1) \rightarrow -1.852$,
$(a_1\ d_1\ \delta_2) \rightarrow 3.098$,
$(a_2\ d_1\ \delta_1) \rightarrow 2.592$.

$(a_1\ d_1\ \delta_2\ +)$ and $(a_2\ d_1\ \delta_1\ +)$ are then put into the candidate set.

STEP 5. Construct classification rules with weights from the positively relevant patterns in the candidate set

Each elementary positively relevant pattern (attribute $A_j = v_{ji}$ class $= p\ +$) is now converted into a classification rule:

$$\text{If} \quad A_j = v_{ji}, \text{ then class} = p \ (\text{or } A_j = v_{ji} \rightarrow p).$$

Next, based on the information theory, the weight of this rule is calculated by the following formula:

$$W = \log \frac{\text{Probability(attribute } A_j = v_{ji} \mid \text{class} = p)}{\text{Probability(attribute } A_j = v_{ji} \mid \text{class} \neq p)}.$$

For a complex positively relevant pattern with two attribute-value pairs (attribute $A_{j_1} = v_{j_1 i_1}$ attribute $A_{j_2} = v_{j_2 i_2}$ class $= p\ +$), the rule is

$$\text{If} \quad A_{j_1} = v_{j_1 i_1} \quad \text{and} \quad A_{j_2} = v_{j_2 i_2}, \text{ then class} = p$$

$$(\text{or } A_{j_1} = v_{j_1 i_1} \wedge A_{j_2} = v_{j_2 i_2} \rightarrow p).$$

The weight of the rule is

$$W = \log \frac{\text{Probability(attribute } A_{j_1} = v_{j_1 i_1}, \text{ attribute } A_{j_2} = v_{j_2 i_2} \mid \text{class} = p)}{\text{Probability(attribute } A_{j_1} = v_{j_1 i_1}, \text{ attribute } A_{j_2} = v_{j_2 i_2} \mid \text{class} \neq p)}.$$

For more complex patterns, the rules and the weights can be formed and calculated in the same way.

For the data set in Table I, the following positively relevant patterns are detected after execution of Step 4:

$(a_2\ \delta_3\ +), (d_2\ \delta_3\ +), (a_2\ d_2\ \delta_3\ +), (a_1\ d_1\ \delta_2\ +)$ and $(a_2\ d_1\ \delta_1\ +)$.

The following classification rules are then constructed:

1. $a_2 \rightarrow \delta_3 \quad W = 0.477$,
2. $d_2 \rightarrow \delta_3 \quad W = 0.477$,
3. $a_2 \wedge d_2 \rightarrow \delta_3 \quad W = \infty$,
4. $a_1 \wedge d_1 \rightarrow \delta_2 \quad W = \infty$,
5. $a_2 \wedge d_1 \rightarrow \delta_1 \quad W = \infty$.

STEP 6. Sort the classification rules by the weights

To efficiently match the relevant rules and classify an unknown object, the classification rules constructed in Step 5 are sorted by their weights, with rules with higher weights placed first. For this example, the classification rules are arranged in the following order:

1. $a_2 \wedge d_2 \rightarrow \delta_3 \quad W = \infty$,
2. $a_1 \wedge d_1 \rightarrow \delta_2 \quad W = \infty$,
3. $a_2 \wedge d_1 \rightarrow \delta_1 \quad W = \infty$,
4. $a_2 \rightarrow \delta_3 \quad W = 0.477$,
5. $d_2 \rightarrow \delta_3 \quad W = 0.477$.

STEP 7. If the percentage of the undetermined training examples is smaller than a predefined threshold, stop the induction algorithm and add a default rule by setting the most common class in the undetermined examples as the default class; otherwise, do Step 8

A training example is said to be determined if it can match the condition part of at least one classification rule. If an object matches more than one rule, then the rule with the highest weight will be used to determine the class. As an example, if the object $(a_2\ b_2\ c_2\ d_2)$ is to be classified, the following three rules will all be matched:

1. $a_2 \wedge d_2 \rightarrow \delta_3 \quad W = \infty$,
4. $a_2 \rightarrow \delta_3 \quad W = 0.48$,
5. $d_2 \rightarrow \delta_3 \quad W = 0.48$.

Since Rule 1 has the highest weight, the object $(a_2\ b_2\ c_2\ d_2)$ is classified as δ_3. Thus the matching process can stop once the first rule matched is found, because the list of the classification rules is sorted in order of decreasing weight. This strategy greatly reduces the amount of time needed for matching.

If the percentage of undetermined training examples is smaller than a predefined threshold, the set of classification rules is then enough to ensure good performance. For the data set in Table I, the seventeenth, the nineteenth, the twenty-first, and the twenty-third cannot be determined by the classification rules. If the percentage of the undetermined training instances is set at, for example, 10%, then STEP 8 must be executed to determine these four instances.

If certain training instances cannot be determined by the classification rules, one of two factors may be the cause:

(1) other complex positively relevant patterns have not yet been found; or
(2) noise is present in the training set.

The next step is intended to cope with this situation.

STEP 8. Find all other nondominated positively relevant patterns each with K attribute-value pairs by the residual analysis. If no such relevant patterns are detected, stop the induction algorithm and add a default rule; otherwise, go to Step 9

For this step to be executed, there must be some training instances left undetermined by the existing classification rules. As mentioned before, there are two possible reasons for this situation. A simple heuristic is used to determine which is applicable. Nondominated positively relevant patterns with K attribute-value pairs are detected. If no such positively relevant patterns are detected, the first value of K that yields a positively relevant pattern could be sought by gradually increasing the value of K. However, the first nondominated positively relevant pattern may contain $K+1,\ K+2$, or even more attribute-value pairs. When no positively relevant patterns of K attribute-value pairs are detected, it is hard to decide what value of K is enough to yield another positively relevant pattern, so instead we then stop the learning algorithm and consider the undetermined training instances to be noise. This policy allows us to avoid the problem of combinatorial explosion. By means of this heuristic, a reasonable trade-off between induction time and the classification ratio can be achieved.

For the above example, the following nondominated positively relevant patterns with two attribute-value pairs are detected:

$$(a_3\ b_2\ \delta_2) \rightarrow 3.098,$$
$$(a_1\ d_2\ \delta_1) \rightarrow 2.592,$$
$$(b_2\ d_1\ \delta_2) \rightarrow 2.000.$$

STEP 9. Convert the new positively relevant patterns into classification rules. If the percentage of the undetermined training examples is smaller than a predefined threshold, stop the induction algorithm and add a default rule; otherwise, set $K = K + 1$ *and go Step 3*

If positively relevant patterns are detected, they are converted into classification rules. If the ratio of the undetermined training samples is smaller than a predefined threshold, the set of classification rules is then enough to achieve good performance. Otherwise, more complex relevant patterns are possible (since the current value of K can cause a positively relevant pattern), and the algorithm returns to Step 3 for another run. This heuristic is called the '*property of continuity*'.

For the above example, the following classification rules are then constructed:

1. $a_3 \wedge b_2 \rightarrow \delta_2 \quad W = \infty$,
2. $a_1 \wedge d_2 \rightarrow \delta_1 \quad W = \infty$,
3. $b_2 \wedge d_1 \rightarrow \delta_2 \quad W = 0.602$.

For the data set in Table I, the seventeenth and the nineteenth still cannot be determined by the classification rules. If the percentage of the undetermined training instances is set at, for example, 10%, the following default rule is added into the rule set:

Other conditions $\rightarrow \delta_1$.

The learning process then stops, with the following set of classification rules output as the result:

1. $a_2 \wedge d_2 \rightarrow \delta_3 \quad W = \infty$,
2. $a_1 \wedge d_1 \rightarrow \delta_2 \quad W = \infty$,
3. $a_2 \wedge d_1 \rightarrow \delta_1 \quad W = \infty$,
4. $a_3 \wedge b_2 \rightarrow \delta_2 \quad W = \infty$,
5. $a_1 \wedge d_2 \rightarrow \delta_1 \quad W = \infty$,
6. $b_2 \wedge d_1 \rightarrow \delta_2 \quad W = 0.602$,
7. $a_2 \rightarrow \delta_3 \quad W = 0.477$,
8. $d_2 \rightarrow \delta_3 \quad W = 0.477$,
9. Other conditions $\rightarrow \delta_1$.

Based on the induction method, an easy and effective approximate reasoning method can be presented here. When an object is to be classified, it matches the condition part of each classification rule to determine its class. If an object matches more than one rule, then the rule with the highest weight will be used to determine the class. If an object matches no rule, the default rule is used to

determine its class. By this reasoning method, the classification rules derived in the above example can classify the data in Table I with 100% accuracy.

Some related remarks about PIM are discussed below.

1. In STEP 8, only nondominated positively relevant patterns are detected. Negative patterns are not detected. As a good alternative, negatively relevant patterns (with K attribute-value pairs) can also be detected to act as negative constraints for the next STEP 4 $(K+1)$. However, the maximum possible attribute number of positively relevant patterns derived from the negative constraints is $2 * K$.

2. In STEPs 7–9, the termination criterion is evaluated by the number of undetermined training instances. As an alternative, the termination criterion could be evaluated by the number of wrongly classified training instances. This alternative, however, will take more time and get more specific rules than the original algorithm.

3. When the numbers of instances for each class are not comparable, it is possible that the patterns correctly classifying certain classes are very specific. The implication is that no positively relevant patterns for these classes could be found when the termination criterion is reached. In this situation, the algorithm may regard these classes of instances as noise (they could actually be noise) or may set a default rule to classify them (STEPs 7–9). As an alternative, the termination criterion evaluated by the wrongly classified instances can be used to get more specific rules of these classes (see Remark 2). It is hard to decide which alternative is better, since these classes of instances could actually be noise.

4. Traditional symbolic learning is most suitable for deriving linear boundaries (represented by attributes). It has less power in deriving nonlinear boundaries. For example, the learning results by PIM for the XOR problem are four rules, with each rule covering only one training instance. No generalization is achieved. But actually, only two rules are enough for the XOR problem (e.g., $x^*y = 1 \rightarrow$ class 1, and $x^*y = 0 \rightarrow$ class 2). For this kind of problems, numerical learning methods (such as neural network learning) are good candidates.

6. Experiments

The proposed algorithm was applied to three problem domains – fitting contact lenses [1] (a noise-free domain with a small number of instances), a simulated data set [1] (a noise-free domain with a larger number of instances), and brain tumor diagnosis [22] (a noisy domain) – to demonstrate its effectiveness. The classification accuracy of an induction algorithm is usually evaluated by the following steps. The data set is first split into a training set and a test set; the induction algorithm is run on the training set to induce concept descriptions; and the concept descriptions are then tested on the test set to measure the percentage of correct predictions. In each of the following experiments, 75% of the cases

Table III. Accuracy for the contact lens fitting domain

Method	Accuracy
PRISM	99.40%
ID3	96.89%
PIM	94.70%
Chan and Wong	90.60%

were selected at random for training, and the remaining 25% were used for testing.

6.1. FITTING CONTACT LENSES

Consider an adult spectacle wearer who consults an optician for purchasing contact lenses [1]. Assume that, from the optician's point of view, this is a three-category problem with four factors $a, b, c,$ and d that must be taken into consideration. Table III compares the accuracy averaged over 100 runs on this problem of PRISM [1], ID3 [12, 17, 20], PIM, and Chan and Wong's method [2, 3].

Although all methods have high accuracy on this problem, PIM performs slightly better than Chan and Wong's. As mentioned before, an induction algorithm for noisy domains may cause classification errors when applied to noise-free domains. Thus it is not surprising that ID3 and PRISM (two noise-free learning algorithms) perform better than PIM.

6.2. ARTIFICIAL DATA

Suppose there are four attributes, $a, b, c,$ and d. Attribute a has five possible values (1, 2, 3, 4, 5); Attributes b and c have four possible values (1, 2, 3, 4); and Attribute d has three possible values (1, 2, 3). Thus a complete training set would consist of $5 \times 4 \times 4 \times 3 = 240$ instances. Also suppose that all instances are generated according to the following rules:

rule 1: $a_4 \wedge d_2 \rightarrow \delta_1$
rule 2: $c_1 \wedge d_1 \rightarrow \delta_1$
rule 3: $a_2 \wedge c_4 \wedge d_2 \rightarrow \delta_1$
rule 4: $a_5 \wedge c_4 \wedge d_2 \rightarrow \delta_1$
rule 5: all others are of class δ_2.

Table IV compares the accuracy of the four learning methods on this problem, averaged over 100 runs. The results are similar to those for Experiment 1.

Table IV. Accuracy for the artificial data domain

Method	Accuracy
PRISM	89.30%
ID3	86.80%
PIM	85.20%
Chan and Wong	75.90%

6.3. BRAIN TUMOR DIAGNOSIS

The brain tumor diagnosis domain is quite interesting and challenging because the brain is very complex and the cause of many brain tumors is still unclear. The most reliable technique for diagnosing brain tumors is generally considered computer tomography (CT). Nearly all intracranial lesions can be detected with CT. The usual examination involves scanning the neurocranium in a series of transverse slices lying in parallel. The head is bent forward so that the sectional plane lies at an angle of 12° to the orbitomeatal lines. Each slice is 8 millimeters thick, so that 8–15 slices are usually sufficient to visualize the intracranial structures which are to be examined. Normally, the process of diagnosing a brain tumor comprises several stages. First, the CT pictures of a patient's brain are analyzed and compared in order to determine the location and density of the lesion. Next, the CT pictures are further analyzed to obtain edema, shape of edema, degree of enhancement, appearance of enhancement, general appearance, size of mass, mass effect, and bone change. Finally, the brain tumor can be identified as one of several possible types of tumors.

Data on 204 cases of patients with the brain tumors were provided by the Department of Radiology of Veterans General Hospital, Taipei, Taiwan [22]. Each case is described using 12 attributes and six types of brain tumors may be distinguished. A training instance, for example, is shown as follows:

1. Sex ..... M.
2. Location ..... Sellar.
3. Precontrast ..... High.
4. Calcification ..... Marginal.
5. Edema ..... No.
6. Shape_edema ..... Smooth and regular.
7. Degree of enhancement ..... Less than vessel.
8. Appearance of enhancement ..... Homogeneous with lucency inside.
9. General of appearance ..... Solid with small cyst/cysts.

10.	Bone_change	Sellar enlargement.
11.	Mass effect	With mass effect.
12.	Hydro	No hydrocephalus.
Pathology		Pituitary adenoma.

A total of 58 rules are derived. A rule, for example, is shown as follows:

If $(Location = Sellar)$ *and* $(Precontrast = High)$ *and* $(Hydro = No)$
Then Class is Pituitary adenoma.

Table V compares the accuracy of the four learning methods on this problem, averaged over 100 runs. The table shows that PIM has the highest accuracy among the four learning methods.

Table V. Accuracy for the brain tumor diagnosis domain

Method	Accuracy
PRISM	82.30%
ID3	74.80%
PIM	87.10%
Chan and Wong	80.60%

7. Conclusion

We have proposed a new probabilistic induction method for obtaining relevant patterns implicit in data sets. Compared with Chan and Wong's method, PIM has the following advantages:

(1) It can detect complex positively relevant patterns and avoid the problem of combinatorial explosion. A reasonable trade-off between the induction time and the classification ratio can be achieved.

(2) It converts only positively relevant patterns into classification rules. Negatively relevant patterns are not converted, but rather used as information for inducing other positively relevant patterns. PIM thus avoids erroneous classification of unknown objects when no positively relevant patterns are matched.

(3) It lists the rules by order of weight, with the rules of highest weights at the front. The matching process can then stop when the first rule matched is found, thus decreasing the matching time.

Three experiments were conducted to compare the performance of PIM with that of three other learning methods. PIM outperformed Chan and Wong's in all three experiments. PIM also outperformed the ID3 and PRISM learning methods in the brain tumor diagnostic domain (a noisy domain). PIM is clearly a good candidate for application in noisy learning environments.

Acknowledgement

The authors thank the Department of Radiology of Veterans General Hospital, Taipei, Taiwan, for providing the patient data on the brain tumor domain. The authors also thank the anonymous referees for their very constructive comments.

References

1. Cendrowska, J.: PRISM: An algorithm for inducing modular rules, *Int. J. Man–Machine Studies* **27** (1987), 349–370.
2. Chan, K. C. C. and Wong, A. K. C.: Automatic construction of expert systems from data: A statistical approach, *Proc. IJCAI'89 Workshop on Knowledge Discovery in Databases*, 1989.
3. Chan, K. C. C. and Wong, A. K. C.: Performance analysis of a probabilistic inductive learning system, in *Proc. 5th Int. Conf. Machine Learning*, 1990, pp. 16–23.
4. Clark, P. and Niblett, T.: The CN2 induction algorithm, *Machine Learning* **3** (1989), 261–283.
5. Fillmer, J. F., Mellichamp, J. M., Miller, D. M., and Narayanan, S.: An expert system for wide area network component configuration, *Expert Systems* **9**(1) (1992), 3–9.
6. Gisolfi, A. and Balzano, W.: Constructing and consulting the knowledge base of an expert system shell, *Expert Systems* **10**(1) (1993), 29–34.
7. Grimm, F. and Bunke, H.: An expert for the selection and application of image processing subroutines, *Expert Systems* **10**(2) (1993), 61–71.
8. Haberman, S. J.: The analysis of residuals in cross-classified tables, *Biometrics* **29** (1973), 205–220.
9. Hong, T. P.: A Study of Parallel Processing and Noise Management on Machine Learning, Ph.D. dissertation, National Chiao-Tung University, Taiwan, R.O.C., Jan. 1992.
10. Hong, T. P. and Tseng, S. S.: Learning concept in parallel based upon the strategy of version space, *IEEE Transactions on Knowledge and Data Engineering*, **6**(6) (1994), 857–867.
11. Hou, R. H.: A New Probabilistic Inductive Learning Method, Master's thesis, National Chiao-Tung University, Taiwan, R.O.C., June 1992.
12. Jackson, A. H.: Machine learning, *Expert Systems* **5**(2) (1988), 132–149.
13. Kodratoff, Y. and Michalski, R. S.: *Machine Learning: An Artificial Intelligence Approach*, Vol. 3, Toiga, Palo Alto, CA, 1990.
14. Michalski, R. S., Carbonell, J. G., and Mitchell, T. M.: *Machine Learning: An Artificial Intelligence Approach*, Vol. 1, Toiga, Palo Alto, CA, 1983.
15. Michalski, R. S., Carbonell, J. G., and Mitchell, T. M.: *Machine Learning: An Artificial Intelligence Approach*, Vol. 2, Toiga, Palo Alto, CA, 1984.
16. Mingers, J.: An empirical comparison of pruning methods for decision tree induction, *Machine Learning* **4** (1989), 319–342.
17. Quinlan, J. R.: Learning efficient classification procedures and their application to chess end games, in *Machine Learning: An Artificial Intelligence Approach*, Vol. 1, Toiga, Palo Alto, CA, 1983, pp. 463–482.
18. Quinlan, J. R.: The effect of noise on concept learning, in *Machine Learning: An Artificial Intelligence Approach*, Vol. 2, Toiga, Palo Alto, CA, 1984.
19. Quinlan, J. R.: Simplifying decision trees, *Int. J. Man–Machine Studies* **27** (1987), 221–234.
20. Safavian, S. R. and Landgrebe, D.: A survey of decision tree classifier methodology, *IEEE Transactions on Systems, Man, and Cybernetics* **21**(3) (1991), 660–674.
21. Smyth, P. and Goodman, R. M.: An information theoretic approach to rule induction from database, *IEEE Transactions on Knowledge and Data Engineering* **4**(21) (1992), 301–316.
22. Wang, C. H. and Tseng, S. S.: A brain tumor diagnostic system with automatic learning abilities, in *3rd IEEE Symp. Computer-Based Medical Systems Conf.*, 1990, pp. 313–320.
23. Witten, L. H. and Macdonald, B. A.: Using concept learning for knowledge acquisition, *Int. J. Man–Machine Studies* **29** (1988), 171–196.