Survey paper on improved methods of ID3 decision tree classification

Shikha Chourasia

Computer Engineering Department, 23 Park Road, SGSITS Indore 452003 MP India

Abstract- Decision tree classification technique is one of the most popular techniques in the emerging field of data mining. There are various methods for constructing decision tree. Induced Decision tree (ID3) is the basic algorithm for constructing decision trees. After ID3 various algorithms were proposed by different researchers and authors those are extensions of ID3 algorithm. This paper contains a survey about the improved methods of ID3 decision tree classification and those are FID3 (fixed induced decision tree) and VPRSFID3 (variable precision rough set fixed induced decision tree). In this short survey we will investigate which method is best among all the other methods.

Index Terms- Decision tree, ID3, FID3, VPRSFID3

I. Introduction

Classification is the prediction approach in data mining techniques. There are many algorithms based on classification that is Instance based, neural networks, Bayesian networks, support vector machine, and decision tree, Decision tree classify instances by sorting them down the tree from the root to some leaf node, which provides the classification of instance. Each node in the tree specifies a test of some attribute of the instance and each branch descending from that node corresponds to one of the possible values for this attribute.

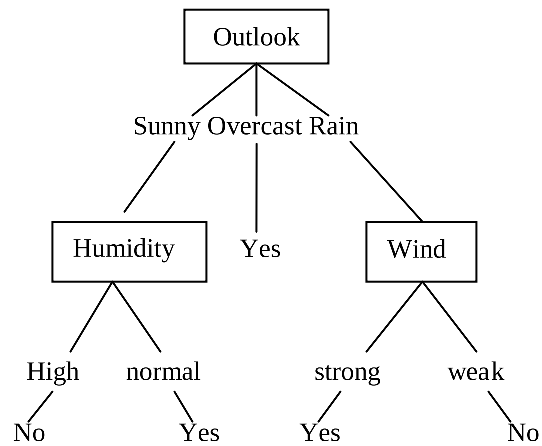
Decision Tree Classifiers (DTC's) are used successfully in many diverse areas such as radar signal classification, character recognition, remote sensing, medical diagnosis, expert systems, and speech recognition, to name only a few. Perhaps, the most important feature of DTC's is their capability to break down a complex decision-making process into a collection of simpler decisions, thus providing a solution which is often easier to interpret.

A. *Decision tree representation*

Figure illustrates a typical decision tree [2]. This decision tree classifies the situation of playing tennis according to the weather. For example, the instance

$$\{Outlook = Sunny, Temperature = Hot$$
$$Humidity = High\ , Wind = Strong\}$$

Would be sorted down the left most branch of the decision tree and would therefore be classified as a negative instance it means the tree predicts that play tennis = no.



II. BACKGROUND STUDY

A. *Rough set theory*

Rough set theory was proposed [7] by Poland in 1982 is a mathematical tool to deal with vagueness and uncertainty [3] was introduced to process the uncertainty and imprecise information. Here are concepts of rough set theory.

1. *Indescernibility relation*

Let S= (U, C, D, V, f) be a decision table and A=C U D, then with any B∈A there is an equivalence relation $IND_A(B)$:

$$IND_A(B) = \{(X, X') \epsilon U^2 : \forall a \epsilon B, a(X) = a(X')\}$$

$IND_A(B)$ is called the B-indescernibility relation, its classes are denoted by $[x]_B$.

2. *Set approximation*

Let T = (U, A) and let and We can approximate *X* using only the information contained in B by construction the β-lower and β-upper approximations of *X*, denoted and respectively, where

$$\underline{BX} = \{x | [x]_B \subseteq X\}$$
$$\overline{BX} = \{x | [x]_B \cap X \neq \emptyset\}$$

3. *Positive Region*

Suppose P and Q are the equivalent relationship in U, then the positive region P of Q can be marked as $POS_p(Q)$, and $POS_p(Q) = UP^*(Q)$

If $POSp(Q)=POS_{p-\{r\}}(Q)$, r∈P, we say that attribute r is omissible in P with respect to Q. Otherwise, r is necessary.

4. Reduct and Core

Keep only those attributes that preserve the indescernibility relation and, consequently, set approximation. There are usually several such subsets of attributes and those which are minimal are called reducts.

The set of attributes is called a reduct of *C*, if *S'= (U, R, D)* is independent and

The set of all the condition attributes indispensable in *S* is denoted by CORE(C).

Where RED(C) is the set of all reducts of *C*.

B. Variable precision rough set theory

VPRST [4] was proposed Wojciech Ziarko Variable precision rough sets (VPRS) attempts to improve upon rough set theory by relaxing the subset operator. Let $X, Y \subseteq U$, the relative classification error is defined by:

$$C(X,Y) = \begin{cases} 1 - \dfrac{|X \cap Y|}{|x|} & |X| > 0 \\ 0 & |X| = 0 \end{cases}$$

Where $|X|$ is the cardinality of that set. Observe that c(X, Y) =0 if and only if X $\subsetneq$ Y degree of inclusion can be achieved by allowing a certain level of error, *β*, in classification:

iff $c(X,Y) \leq \beta$, $X \subseteq_\beta Y$, $0 \leq \beta < 0.5$

VPRS generalization is aimed at handling uncertain information. Ziarko has introduced 0≤β≥0.5 (subset operator)

1. β-lower and β- upper Approximation of Set

Suppose (*U*, *R*) is an approximation space. *U/R ={X1, X2…, Xn}* where *Xi* is an equivalence class of *R*. For any subset *X* ⊆*U*, lower approximation $\underline{R_\beta X}$ and upper $\overline{R_\beta X}$ approximation of X with precision level *β* respect to *R* is respectively defined as:

$$\underline{R_\beta X} = \cup\{Y \epsilon\, U/R \mid \frac{X \cap Y}{x} \leq \beta$$

$$\overline{R_\beta X} = \cup\{Y \epsilon\, U/R \mid \frac{X \cap Y}{x} < 1 - \beta$$

Where the domain of *β is 0 ≤β < 0.5,* $\underline{R_\beta X}$ is also called *β*-positive region (POS (*X*, *Y*)).

III. Induced Decision Tree (ID3) Classification

ID3 [5] [6] is an algorithm for building decision tree. The key idea of id3 is to choose attributes with the biggest information gain based on entropy as current classification attribute and then recursively expand the branches of decision tree until whole tree has been built completely.

A. Algorithm of ID3 decision tree

In decision tree approach ID3 algorithm is the most popular algorithm. Suppose S is the set of example set, and the number of equivalence class constructed by indescernibility relation is n then entropy is defined as:

$$Entropy = -\sum_{i=1}^{n} p_i \log_2 p_i$$

Where $p_i = \frac{S_i}{|S|}$, $|S|$ is the number of example set S.

Given an attribute A∊C is the set of condition attribute the domain of A is denoted as V_A, then the expected information of the entropy is given as follows:

$$info_A(S) = \sum_{i-1}^{n} \frac{|S_i|}{|S|} Entropy(S_i)$$

Hence the information gain on A∊C is defined as:

$$Gain(S,A) = Entropy(s) - Info(s)$$

We compute the information gain of each condition attribute, and the attribute with the maximum information gain is the most informative attribute.

IV. Fixed Induced Decision Tree (FID3) Classification Algorithm

FID3 algorithm makes use of rough set theory [1]. And removes drawback of ID3 algorithm it was proposed by Baoshi Ding, Yongqing Zheng, Shaoyu Zang in 2009.

A. Dependency of attribute(k)

The algorithm FID3 introduces Dependency of attribute (k). If all the values of attribute from Q are uniquely determined by values of attribute from p, then we say that a set of attribute Q depends totally on asset of attribute p, denoted p => Q. we can define a functional dependency as follows:

Suppose P and Q are subsets of C U D. where C is conditional attribute and D is decision attribute we say that Q depends on p in a degree k (0 ≤ k ≤ 1) denoted by $P \Rightarrow_k Q$, if

$$k = \frac{|pos_P Q|}{|U|} \quad \overline{R_\beta X}$$

Considering the value of k:

If 0 <k <1, Q depends partially on P.

If k=1, Q depends totally on P.

If k=0, it means there is no dependency between P and Q.

Fixed Information gain

FID3 proposes fixed information gain $Gain_{fix}$ as the new standard for selecting splitting attributes, it is defined as:

$$Gain_{fix} = \sqrt{k \cdot \frac{Gain}{m}}$$

Where k is the dependency of attribute Q on P where Q ϵ D and P ϵ C, m means number of different values of attribute A and A ϵ C.

A. Algorithm of FID3[1]

Input: A decision table S= (U, C, D, V, F) as the training sets.

Output: A decision tree.

Step1:

Create a node as the initial node of the tree. Check that whether the samples are all of the same class. If they are, then the node turns into a leaf and return the leaf labeled with that class.

Step2:
For each attribute A in C, we begin to calculate $k = \frac{|POS_A(D)|}{|U|}$ and then we can calculate all the values of k to A.

Step3:
For each attribute C_i in C' where $C'=\{X(x\epsilon C)\wedge(k_x\neq 0)\}$, calculate its $gain_{fix}$, choose the attribute c_i so that the samples are partitioned accordingly.

Step4:
Attribute that has maximum $gain_{fix}$ choose it as root attribute and grow the branches according to the different values of attribute c_i so that the samples are partitioned accordingly.

Step5:
If it reaches a node where all the samples belong to the same class, turn it into a leaf and label it with that class.

Step6:
Otherwise we continue to construct the decision tree recursively for the samples at each partition. Once an attribute has occurred at a node, it needs not to be considered in any of the node's descendents.

Step7:
Partitioning the samples repeatedly from top to down until one of the conditions below is satisfied.

- All samples for a given node belong to the same class, return a leaf labled with that class.
- If there are no more there are no more samples to be classified, we can create a leaf belong to the same class, return a leaf labeled with that class.
- If it reaches the node where all the attributes have been chosen in the path from root to current node, return a leaf node labeled with the most common class.

V. VPRSFID3 CLASSIFICATION

This algorithm [5] [8] uses the concept of variable precision rough set theory. It removes drawback of FID3. This algorithm is proposed by Rajkumar Sharma, Pranita Jain, Shailendra K.Shrivastava in 2012. It introduces subset operator β and provides relaxation to the inclusion of attribute and uses significance of attribute (σ).

A. β-Dependency of attribute k_β

The degree of dependency of condition attributes on decision attribute $\gamma_\beta(X,Y)$ is defined as:

$$k_\beta = \gamma_\beta(X,Y) = \frac{|POS_\beta(X,Y)|}{|U|}$$

$\gamma_\beta(X,Y)$ implies the proportion that objects in the domain *U* can be correctly classified for a given value of *β*, and it evaluates the ability of classification to object.

B. Significance of Attributes

The higher the change in dependency, the more significant the attribute is. If the significance is 0, then the attribute is dispensable without losing information .More formally, given *X*, *Y* and an attribute $a \in X$, the significance of attribute *a* upon *Y is* defined by:

$$\sigma X(Y,a) = \gamma X(Y) - \{a\}(Y)$$

Attributes that has sigma (σ) value zero will be discarded and attribute that has value greater then
zero will be taken for further classification.

C. Enhanced Information gain

VPRSFID3 proposes a new selection criterion to select attribute $gain_{enh}$.

$$gain_{enh} = \sqrt{k_\beta * \frac{gain_{enh}}{m}}$$

Algorithm VPRSFID3-

The steps of algorithm are as follows:
Input: An information systems S=(U, C ∪ D, V, f), the training sets, the threshold parameter β, (0≤ β<0.5)
Output: A decision tree T.

Step1:
Create a node as the initial node of the tree. Judge that whether the samples are all of the same class. If they are, then turns the node into a leaf and return the leaf labelled with that class.

Step2:
For each attribute in C_i , calculate $Gain_{enh}$, choose the attribute Ai with the maximum value of $Gain_{enh}$ as the root node. where C_i is the set of β- reducts.

Step3:
Construct the branches according to different values of attribute C_i so that the samples are partitioned accordingly.

Step 4:
If
It reaches a node where all the samples belong to the same class, turn it into a leaf and label it with that class.

Else
Continue to construct the decision tree recursively for the samples at each partition. Once an attribute has fixed to a node, it needs not to be considered in any of the node's descendents.

Step 5:
Partitioning the samples repeatedly from top to down until one of the conditions below is satisfied.

- $k_\beta \leq \beta$

- All samples for a given node belong to the same class, return a leaf labelled with that class.
- There are no more training samples to be classified, we can create a leaf belong to the class in majority among samples.

Step 6:

Output the decision tree T.

VI. RESULTS

This section shows the experimental result and comparison among ID3, FID3, VPRSFID3.

Comparison between ID3, FID3, VPRSID3 in terms of accuracy and number of leaves

Data sets		Weather Nominal	Iris	Wine
Instances		64	5	
Attributes				
ID3	Accuracy	85.7%	64.1%	69.6%
	No of leaves	5	7	15
FID3	Accuracy	84.2%	66.5%	74.7%
	No of leaves	5	5	11
VPRSFID3	Accuracy	92.18%	94.6%	72.4%
	No of leaves	5	4	7

REFERENCES

[1] Baoshi Ding, Yongqing Zheng, Shaoyu Zang, "A New Decision Tree Algorithm Based on Rough Set Theory," Asia-Pacific Conference on Information Processing, IEEE, pp. 326-329, 2009.

[2] http://www.cs.princeton.edu/courses/archive/spr07/cos424/papers/mitchell-dectrees.pdf

[3] Baowei Song Chunxue Wei, " Algorithm of Constructing Decision Tree Based on Rough Set Theory, " International Conference on Computer and Communication Technologies Agriculture Engineering, IEEE 2010.

[4] Ziarko, "Variable precision rough sets model," Journal of Computer and System Sciences, vol. 46, pp. 39–59, 1993.

[5] Rajkumar Sharma, Pranita Jain, Shailendra K.Shrivastava "An Optimize Decision Tree Algorithm Based on Variable Precision Rough Set Theory Using Degree of β-dependency and Significance of Attributes" (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 3 (3) , 2012,3942-3947.

[6] R. Quinlan, "Induction of decision trees," Machine Learning, vol.1, No. 1, pp. 81–106, 1986.

[7] Z. Pawlak, "Rough sets," International Journal of Computer and Information Sciences, vol. 11, pp. 341–356, 1982.

[8] Weiguo Yi, Jing Duan, Mingyu Lu "Optimization of Decision Tree Based on Variable Precision Rough Set" International Conference on Artificial Intelligence and Computational Intelligence IEEE 2010.

AUTHORS

First Author – Shikha Chourasia, Computer Engineering Department, 23 Park Road, SGSITS Indore 452003 MP India, Shikhachourasia754@gmail.com