The Estimation of Entropy Variation in Decision Tree

Maiga Chang, Tzong-Tsann Wu and Jia-Sheng Heh
Dept. of Information and Computer Engineering,
Chung-Yuan Christian University, 32023, Chungli, Taiwan

Abstract. Decision tree learning system is widely used in machine learning. Many researchers, such as Quilan and Utgoff, have made the learning system to be more efficient, including making it incremental. This paper tries to modify the learning sequence of ID5R by analyzing the upper and lower bounds of entropy variation to estimate the tolerance limit of incremental learning. Based on the entropy variation measurement, the whole tree does not need retrain within the tolerance limit. Therefore, the cost of learning time will decrease. Besides the upper and lower bounds are analyzed, the number of acceptable new arrival data can also be estimated in advance. Experiment system is also implemented on World-Wide Web. Several rules could be extracted from the decision tree which is constructed by our system, and new training data may be inserted.

Keywords: decision tree, incremental learning, entropy, information gain

1. Introduction

Quilan proposed a famous learning method, named ID3, which is good for constructing a decision tree. [1] And an incremental learning algorithm of decision tree, called ID5R, which is an extension of ID3 proposed by Utgoff.[2][3] Although the revised decision tree made by ID5R can be used to retrain the decision tree dynamically when new training data arrive, it recalculates the gains of all nodes on classification path for deciding whether modify or not. This means, ID5R takes long processing time for only one arrival data.

The upper and the lower bounds of entropy variation for the incorporating new data is proposed in this paper. Hence, the classification attribute of nodes in decision tree have not be changed in the tolerance limit based on the difference between the upper and lower bounds. This paper also makes a comparison with TDIDT [4] to verify the advantage of the entropy variation utilizing. A prototype system is accomplished on WWW.

In Section 2, the relation between decision tree and entropy is described. And the entropy variation used in the incremental learning system is then analyzed in Section 3. Besides estimating how many new training data is tolerated without retraining based on the estimated value, Section 4 also makes a comparison to TDIDT. An experiment incremental learning system is then implemented on WWW in Section 5. Section 6 is a conclusion.

2. Decision Tree and Entropy

Given a set of training data S, it is possible to generate several kinds of decision trees. One example with four records are shown below. In Figure 1, the left decision tree is well-classified with only attribute A, but the right one needs more attributes for classification. This means, the most important issue for constructing a decision tree is to select an appropriate attribute. Based on information theory (Shannon, 1948), Quilan (1986) selects suitable attributes by finding the information hiding behind data, so-called *entropy*, in ID3 algorithm. [1][5] ID3 searches all attributes which can be used in classification and then selects the best one. This methodology belongs to hill-climbing search and the *information gain* is chosen as a measure for selecting classification attributes.

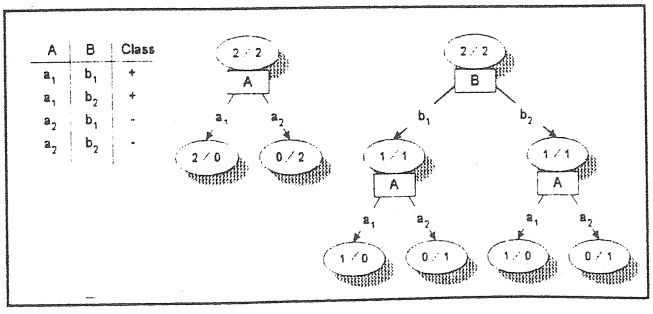


Figure 1. Different attribute selection in decision tree

Supposed that the training set S contains h classes C_λ, $\lambda = 1,\ldots,h$, and the probability for each class can be noted as $\frac{freq(C_\lambda, S)}{|S|}$. According to information theory, the *information* of C_λ is $-\log_2\left(\frac{freq(C_\lambda, S)}{|S|}\right)$ and then the *average information* (*entropy*) *in* S is

$$info(S) = -\sum_{\lambda=1}^{k} \frac{freq(C_\lambda, S)}{|S|} \log_2\left(\frac{freq(C_\lambda, S)}{|S|}\right). \quad (1)$$

When S is classified by some specific attribute A, its information of S becomes

$$info_A(S) = \sum_{i=1}^{m} \frac{|S_i|}{|S|} * info(S_i), \quad (2)$$

and the *information gain* derived from such classification is

$$gain(A) = info(S) - info_A(S). \quad (3)$$

With Eq.(3), ID3 computes all the information gain of candidate attributes and selects the attribute with the maximal information gain. Then the training set S will be divided into several sub-set, corresponding to each sub-tree of the classified decision tree. Such process will be repeated until no classification is possible.

ID3 selects classifying attributes with information gain, which always makes too many branches and affects the classification efficiency. Quilan (1993) solves such problem with C4.5 algorithm.[6] First, C4.5 defines one kind of *split information* of all subset classified with attribute A

$$split\ info(A) = -\sum_{i=1}^{m} \frac{|S_i|}{|S|} * \log_2\left(\frac{|S_i|}{|S|}\right). \quad (4)$$

In order to avoid too many splitting subsets, *gain ratio* is proposed as the measure for attribute selection

$$gain\ ratio(A) = gain(A) / split\ info(A) \tag{5}$$

After a decision tree is constructed, some pruning works is necessary with criteria such as *predicted error rate.* Several important and simple rules than can be generated by finding paths from root to different leaves. All classification attributes in the path are antecedence, and the data of leaf node is consequence. Rule formula is usually used to express the knowledge in expert systems, such as a causal relationship is always can be simply expressed as the "If-Then" form. Unfortunately, even decision tree can be used to generate rule simply, it takes large amount of time to process the new arrival data. There are batch learning and incremental learning to solving this kinds of problems. This paper focuses on the incremental ways, proposes a method which can be used in decision tree for incremental learning based on the relation analysis between entropy and incremental learning.

3. Entropy and Entropy Variation in Incremental Learning

Consider a *two-class problem* in which the training data set S is divided into a positive instance set S^+ and a negative instance set S^-. The *entropy of this decision node* is then computed as

$$\varepsilon = I(p, n) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}, \tag{6}$$

where p and n indicate the number of positive and negative instances respectively, and $\frac{p}{p+n}$ and $\frac{n}{p+n}$ are the probabilities of S^+ and S^- respectively.

In the incremental learning operation of decision tree, when a new training data is arrived, the entropy variation of the decision node should be analyzed. To make entropy variation be explicitly described, let $R = \frac{p}{n}$ then Eq.(6) becomes

$$\varepsilon = I(R) = -\frac{R}{1+R}\log_2\frac{R}{1+R} - \frac{1}{1+R}\log_2\frac{1}{1+R}. \tag{7}$$

Figure 2 illustrates that entropy increases fast for $R \in [0,1]$, and then decreases slowly. The entropy variation speed can then be obtained through differentiation:

$$I'(R) = \frac{-\log_2 R}{(1+R)^2}. \tag{8}$$

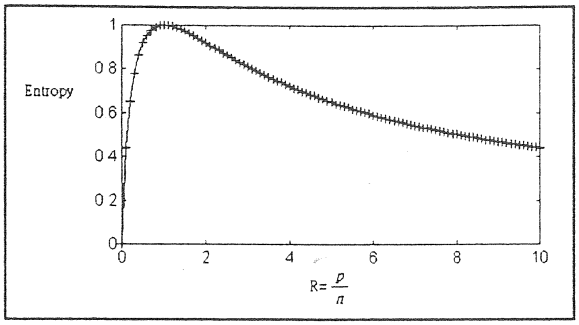


Figure 2. Entropy versus R

Furthermore, because both positive and negative training instances are possibly be the new arrival record, R will become R^+ or R^- if a new positive or negative instance arrival. According to mean-value theorem, $\Delta\varepsilon = I'(r) * \Delta R$, where the ratio r falls between R^+ and R^-, the entropy variation of a new positive or negative instance arrival is then computed within

$$\Delta\tilde{\varepsilon}^+ = \frac{-\log_2 r^+}{(1+r^+)^2} \cdot \frac{1}{n} \quad \text{and} \quad \Delta\tilde{\varepsilon}^+ = \frac{-\log_2 r^-}{(1+r^-)^2} \cdot \frac{-p}{n(n+1)}, \tag{9}$$

where the average ratio

$$r^+ = \frac{\frac{p+1}{n} + \frac{p}{n}}{2} = \frac{p+0.5}{n} \tag{10}$$

is for a new positive instance arrival and the average ratio

$$r^- = \frac{\frac{p}{n+1} + \frac{p}{n}}{2} = \frac{p(2n+1)}{2n(n+1)} \tag{11}$$

is for a new negative instance arrival. The estimation error of such entropy variation will be decreased as p or n increase.

Therefore, we can find the lower and upper bound of entropy variation (ΔE). Assuming m subsets are classified based on specific attribute and the entropy of this attribute is

$$E=\sum_{i=1}^{m}\frac{p_i+n_i}{p+n}\left(-\frac{p_i}{p_i+n_i}\log_2\frac{p_i}{p_i+n_i}-\frac{n_i}{p_i+n_i}\log_2\frac{n_i}{p_i+n_i}\right). \quad (12)$$

When a new instance arrives, the new entropy is changed into E', therefore the entropy variation $\Delta E=E'-E$ would be

$$E'=\sum_{i=1,i\neq j}^{m}\left(\frac{p_i+n_i}{p+n+1}\varepsilon_i-\frac{p_i+n_i}{p+n}\varepsilon_i\right)+\left(\frac{p_j+n_j+1}{p+n+1}\varepsilon_j-\frac{p_j+n_j}{p+n}\varepsilon_j\right). \quad (13)$$

According to Eqs.(9) and (13), the upper and lower bound of entropy variation, $\Delta E_{\max}$ and $\Delta E_{\min}$ are

$$\Delta E^{+}_{\max}=\frac{p_j+n_j}{p+n}\cdot\frac{\left(-\log_2 r^{+}\right)}{\left(1+r^{+}\right)^2}\cdot\frac{1}{n_j}+\frac{1}{p+n},\quad r^{+}=\frac{p_j+0.5}{n_j}, \quad (14)$$

$$\Delta E^{-}_{\max}=\frac{p_j+n_j}{p+n}\cdot\frac{\left(-\log_2 r^{-}\right)}{\left(1+r^{-}\right)^2}\cdot\frac{-p_j}{n_j\left(n_j+1\right)}+\frac{1}{p+n},$$

$$r^{-}=\frac{p_j\left(2n_j+1\right)}{2n_j\left(n_j+1\right)}, \quad (15)$$

$$\Delta E_{\min}=\frac{-1}{p+n+1}. \quad (16)$$

4. Incremental Learning for Decision Tree

ID3 classifies the attribute with maximal information gain in the training set. In other words, the minimal entropy value of classification attribute is selected. Assuming A_s is the selected attribute with maximal information gain and A_c is its competitor attribute. Let E_o be the entropy of original training data, and E_s and E_c correspond the entropy of the decision tree classified by A_s or A_c respectively. Therefore, we have the following relationship:

$$gain(A_s) \geq gain(A_c) \Leftrightarrow E_o - E_s \geq E_o - E_c \Leftrightarrow E_s \leq E_c \quad (17)$$

For incremental learning, the values of E_s and E_c change with the ratio r_i. As for A_s, a new data arrival always makes the entropy variation smaller than ΔE^+_{max} or ΔE^-_{max}. As contrasted with A_s, the entropy variation of A_c would be always larger than ΔE_{min}. Consider the worst case that E_s increases and E_c decreases. Then A_s will not be replaced by A_c until *the re-training condition* occurs:

$E_s + \Delta E_{max} \leq E_c + \Delta E_{min}$, where ΔE_{max} may be ΔE^+_{max} or ΔE^-_{max}. (18)

In Eq.(12), there can be k new instance arrival, then

$$\Delta E_{max.k} < k\ \Delta E_{max}, \quad (19)$$

and

$$\Delta E_{min.k} > k\ \Delta E_{min}. \quad (20)$$

And the retraining condition Eq.(18) and the corresponding *retraining index* can be estimated by:

$$\begin{cases} E_s + \Delta E_{max.k} \leq E_c + k\ \Delta E_{min.k} \\ E_s + k\ \Delta E_{max} \leq E_c + k\ \Delta E_{min} \end{cases}, \quad k \leq \left\lfloor \frac{E_c - E_s}{\Delta E_{max} - \Delta E_{min}} \right\rfloor \quad (21)$$

Kalles and Morris (1996) developed an incremental decision tree algorithm, called TDIDT. [4] TDIDT defines the upper and lower bounds of entropy variation for specific classification attribute:

$$\Delta E'_{max} = \frac{\log_2(p_j + n_j + 1) + \log_2 e}{p + n + 1}, \quad (22)$$

$$\Delta E'_{min} = \frac{-1}{p + n + 1}. \quad (23)$$

After insert "a" new data ($a \leq (p_j + n_j)$), the total entropy variation of upper and lower bound are changed:

$$\Delta E'_{\max.a} \le \sum_{i=1}^{a} \frac{\log_2(p_j + n_j + 1) + \log_2 e}{p+n+1} \le a \cdot \frac{\log_2(p_j + n_j) + 1 + \log_2 e}{p+n+1}, \quad (24)$$

$$\Delta E'_{\min.a} \ge a \cdot \frac{-1}{p+n+1}. \quad (25)$$

Eq.(25) indicates the lower bound of TDIDT is same as our Eq.(16), but the upper bound differs. Figure 3 illustrates their relationships.

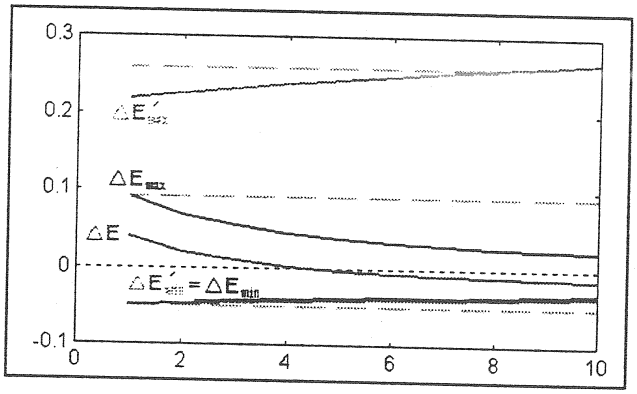


Figure 3. Entropy variations of TDIDT and our method

5. System Implementation and Evaluation

The proposed incremental decision-tree learning method is implemented on WWW (World-Wide Web) as a prototype system, as Figure 4 shown. Decision tree is not only built from database in the experiment system, but also modified the tree according to new training data. The test database is Mushroom, adopted from Machine Learning Repository. (Murphy and Aha, 1995) From the complete 5644 records in Mushroom, 4926 records are chosen for batch learning to build tree and other 718 ones for incremental learning to modify tree.

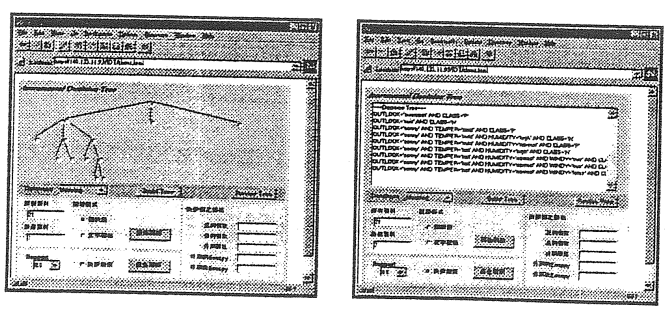
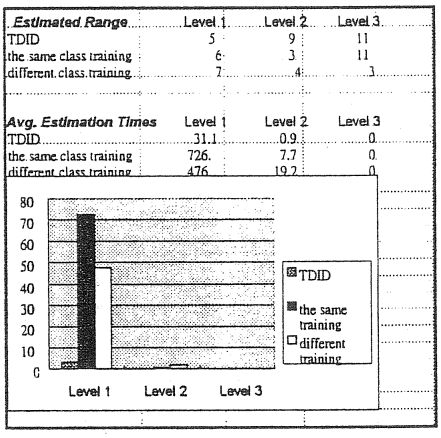


Figure 4. Resultant decision tree and corresponding rules

The method TDIDT is used to compare with to our proposed method in this

paper. In Table 1, *estimated range* is the times of estimated index k decreasing to 0, whereas *average estimation times* is the sum of estimation value k divided by estimated range. The estimated count in Table 1 means that the attribute in each node still keeps the best when the new data arrival. Thus we re-compute the range of entropy of each classifying attribute. As the result in Table 1 shown, when insert new data, our estimated range possesses larger value than TDIDT, then has higher value of the tolerance limit.

Table 1. Comparison of TDIDT and our method



Both resultant trees are equal. Table 2 compares incremented learning with batch one. The platform is IBM-compatible PC, with Pentium-133 CPU, 32MB memory, and Windows 95 operating system.

Table 2. Comparison of batch learning and incremental learning.

Learning Time	Batch Learning	Incremental learning
First: 4926 records	About 18 min	
Second:4926+718 records	About 21 min	About 11 min
Total	39 min	29 min

6. Conclusion

Decision tree learning system is widely used in machine learning. For this reason, many researchers are devoted to make this method more efficient, including making it incremental. This paper, modifying from TDIDT and ID5R, analyzes the upper and lower bounds of entropy variation to estimate the range for incremental learning. Experiment system is also implemented on World-Wide Web. Several rules could be extracted from the decision tree which is constructed by our system,

and new training data may be inserted.

According to the analysis made above, the whole tree does not need retrain within the tolerance limit. Therefore, the learning time will be less than that of batch learning system. As new training data arrive, besides the entropy variation is applied to estimate the best classification attribute, upper and lower bounds is calculated in advance for the criteria of the retraining timing.

References

1. J. R. Quinlan, "Induction of Decision Tree," Nacgube Kearbubg, 1, pp.81-106, 1986
2. P. E. Utgoff, "ID5: An incremental ID3," Proceedings of the 5th International Conference on Machine Learning, pp. 107-120, 1998
3. P. E. Utgoff, "Incremental induction of decision tree," Machine Learning, 4(2), pp.161-186, 1989
4. Dimitrios Kalles and Tim Morris, "Efficient Incremental Induction of Decision Trees," Machine Learning, 24, pp.231-242, 1996
5. R. J. Schalkoff, "Pattern Recognition : Statistical, Structural and Neural Approaches," Fing & Sons Printers Pte Ltd., 1992
6. J. R. Quinlan, "C4.5: Programs for Machine Learning," Morgan Laufman, 1993