Enhancing the Performance of Entropy Algorithm using Minimum Tree in Decision Tree Classifier

Khalaf Khatatneh[1]& Ibrahiem M. M. El Emary[2]
1 Al Balqa Applied University, 2 Al Ahliyya Amman University
Amman, Jordan
E-mail: doctor_ebrahim@yahoo.com

Abstract

Classification builds a model based on historical data (training data set). Once the model is built, it is used to predict the class for a new instance. Many methods have been proposed to solve the classification problem (referred to as classifiers); one of the most popular and best classifiers proposed so far is the decision tree classifier. Multiple trees can be generated from the same dataset, all the trees yields the same outcome for a given new instance to be classified. The possible trees for a dataset vary in their size where the size of the tree depends on the sequence in which the dataset attributes is used to build the tree. However, we prefer the minimum tree because the minimum tree needs the shortest time to figure out the outcome of the model. One of the best algorithms that have been proposed to find the sequence that yield the minimum tree if used is the entropy algorithm. We proposed in this article a new algorithm (enhanced entropy algorithm) that reduces complexity and execution time of the original entropy algorithm and at the same time yields the same sequence that can be found by applying entropy algorithm.

Keywords

Enhanced Entropy Algorithm (EEA), Entropy Algorithm (EA), Data Mining (DM), Gain information measure (GINI).

1. Introduction

Data mining (DM) is defined as “the knowledge discovery in database” [1, 2 ,3]; it attempts to apply machine learning, statistics, artificial intelligent concepts and techniques on databases in order to discover useful patterns and trends hidden in vast amount of data stored in form of files and database records.

Classification [4] is a supervised data mining technique [5, 6]; in which we predict categorical class values for new instances based on the training data collected over time. One of the most popular classifiers proposed so far is the decision tree classifier where decision tree classifier builds a model from the training data in the form of a tree. Once the model is built and verified, it can be used to predict the class for new given instances. Decision tree classifier is one of the best classifiers proposed for many reasons, these reasons include:

1) Can handle numeric and categorical data.
2) Needs reasonable time to construct the model.
3) It’s an eager classifier.

Multiple trees [7] can be generated from the same dataset where all trees yield the same outcome (class value) for a given new instance to be classified. Possible trees for a dataset vary in their size. The size of the tree depends on the sequence in which the dataset attributes is used to build the tree. However, we prefer the tree with the smallest size because such tree needs the shortest time to figure out the outcome of the model. Two algorithms have been proposed to find the sequence that yield the smallest tree if used; entropy algorithm and GINI algorithm. Entropy algorithm is usually used because it's suitable for almost all types of datasets types while GINI algorithm is suitable for a small portion of datasets types.

2. Entropy Algorithm

Entropy algorithm computes the information gain value for each attribute in the dataset [8] excluding the class attribute. After that, the attributes are sorted in descending order based on their information gain values, the sorted attributes

compromise the sequence that if used by the decision tree classifier construction process, the smallest tree will be generated.
The information gain measure (abbreviated as iGain) for attribute A in dataset S is measured by the following function:

$$\text{iGain(S, A)} = \text{Entropy (S)} - \sum \frac{|S_v|}{|S|} * \text{Entropy}\ (S_v)$$

Where S_v is the subset of instances of S which have "v" as attribute "A" value, and $|S_v|$ is the number of instances in the set S_v, |S| is the number of instances in the dataset S.

The entropy of a set S of (positive class, negative class) training instances is:
Entropy(S)$- p \log_2(p) - n \log_2(n)$

Where p the number of instances with a positive class is divided by the number of instances in S, and n is the number of instances with a negative class divided by the number of instances in S.

3. Enhanced Entropy Algorithm

The new algorithm I proposed in this paper (Enhanced Entropy Algorithm) finds the sequence that yield the smallest tree using the following steps:

1) For each attribute "A" in the dataset, determine the classes distribution to find the classes distribution:

   a) Determine the attribute distinct values.
   b) For each distinct value V and class value C, determine the number of instances that have A=V and class value=C.

2) Sort the attributes in ascending order according to the number of zeros in the class distribution of every attribute.

3) In case of two attributes with the same number of zeros, find the difference between the maximum value and the minimum value in each attribute class distribution and sort them in ascending order according to the difference found.

4) In case of two attributes with the same number of zeros with the same difference, compute the information gain measure for both attributes using entropy algorithm and sort them in descending order according to the Information gain.

4. Results

4.1. Entropy algorithm Performance

When you think about implementing the entropy algorithm concept using a programming language, you will find that it requires a number of nested loops to get the information gain value for each attribute.

The complexity of any code written to implement the entropy algorithm will never be less than O(n^3). The reason behind the n^3 complexity of entropy algorithm is that in order to find the information gain measure for an attribute, you have to scan all attributes in the dataset and for each attribute "A" you have to scan each distinct value "V" within the attribute "A" and for each distinct value you have to scan all distinct classes values "C" in order to find the number of instances that satisfy "A"="V" and class value="C".

One of the major disadvantages of entropy algorithm is that there is no best case or worst case when it's given a dataset to find the sequence of the minimum tree from, the information gain measure for all attributes must be computed and you always have to iterate through the three loops to find the sequence of the minimum tree.

4.2. Enhanced Entropy Algorithm Performance

Enhanced entropy algorithm unlike the entropy algorithm that always has a constant case has best and worst case. In general, the best case happens when all attributes have different number of zeros. The attribute will be sorted according to the number of zeros and the sorted attributes list compromise the final sequence, or when there are some attributes that have the same number of zeros but the difference between the maximum value and the minimum value in their classes distribution is different, so that the attributes with the same number of zeros can be sorted according to the difference and the final sequence is reached.

In the best case, EEA needs two nested loops to find the number of zeros for each attribute. The same number of loops is required to find the difference between the maximum value and the minimum value in each attribute class distribution. The two nested loops indicate that the complexity of EEA in the best case is O(n^2).

In contrast, the worst case happens when all attributes have the same number of zeros and the difference between the maximum value and the minimum value in all attributes classes distribution is equivalent. In this case, EEA become the same as entropy algorithm. The information gain measure for all the dataset attributes must be measured, then the attributes is sorted according to the information gain value. The complexity of enhanced entropy algorithm in the worst case is the same as the complexity of entropy algorithm which is O(n^3)[9].
.

5. Conclusion and Future works

The major drawback of entropy algorithm is that it needs to perform complex computations to find out the sequence of the minimum tree. The complexity of computations required grows exponentially when the number of attributes in a given dataset increases.

The proposed algorithm in this paper is enhancing entropy algorithm (EEA) which finds the sequence of the minimum tree in a dataset using few logical and mathematical operations that are simple compared to the complex operations needed by entropy algorithm to find the same sequence for the same dataset.

Comparing EEA to entropy algorithm in terms of algorithms complexity yield that entropy algorithm always need O(n^3) to find the sequence of the minimum tree. There is no worst case nor best case, while EEA complexity at the worst case is O(n^3) which is the same as entropy algorithm complexity, but in the best case it needs a complexity of O(n^2).
When we enhance EEA, there would be no worst case by eliminating the need to compute the information gain measure for two or more attributes if they have the same number of zeros and their differences is equivalent.

References

[1] Mendelson and Smola [2003] Mendelson and A.Smola, “Advanced Lectures on Machine learning”.

[2] Michie and Spiegelhalter and Taylor [1994] D.Michie, D. Spiegelhalte, and C.Taylor, “Machine learning, Neural and Statistical Classification”.

[3] Mitchell [1997] Mitchell, “Machine Learning”.

[4] Duda and Hart and Stork [2001] R.Duda, P.Hart, and D.Stork, “Pattern classification”.

[5] Adriaans and Zantinge [1996] P.Adriaans and D.Zantinge, “Data Mining”.

[6] Almasri and Navathe [2004] R,Elmasri and S.B.Navathe, “Fundamentals of database systems”, fourth edition,(2004).

[7] White and Liu [1997] A.White and W.Liu, “The Importance of attribute selection measures in descision tree induction”.

[8] Mingers [1989] J.Mingers, “An empirical comparison of selection measures for decision tree induction”.

[9] Naps and Pothering [1986] T.naps and G.Pothering, “Introduction to data structure and algorithm analysis with Pascal”.