Asian Journal of Convergence in Technology **Volume IV Issue I**
ISSN NO.:2350-1146 I.F-5.71

Decision Tree for data with Known Hierarchical Class Labels

Sachin Gavankar, *Department of Computer Engineering, Datta Meghe College of Engineering, Mumbai University,*
Dr. Sudhirkumar Sawarkar, *Department of Computer Engineering, Datta Meghe College of Engineering, Mumbai University.*
sachingavankar@yahoo.co.in

***Abstract*— Decision Tree classifier builds a classification model using training data. It consists of records having attribute values and corresponding class label. Very few algorithms deal with class labels that are organized as hierarchical structure. In this paper we propose a framework for decision trees which considering the prior knowledge of the class hierarchies in training data. Classification algorithm is applied multiple time to predict class label from higher level (coarse-grain) to lower level (fine-grain) by using respective training records.**

Keywords*— *data mining; classification; decision tree; hierarchical class label

I. INTRODUCTION

Data Mining extracts patterns in the data. Decision Tree is one of the classification method where classification model is created using training data. Training data consists of records with attributes values and class label associated with it. Once the model is created it could be used for classification of unclassified data i.e. based on the values are attributes class label is predicted. Commonly used algorithm for decision tree is C4.5 proposed by Quinlan [4].

Most of the existing approaches do not consider the hierarchical structure within class. In this paper we propose a framework of decision trees which considers the prior knowledge of hierarchies in class label. It applies decision tree algorithm initially using higher level class label (coarse- grain) and recursively moves to the lower levels (fine-grain). As it progresses from higher level to lower level, it utilizes the relevant records from training set.

This paper is organized as follows: Section 2 -formalizes the problem, Section 3 – proposed framework, Section 4 – presents the performance evaluation and finally in Section 5 – conclusion and future work.

II. PROBLEM DEFINITION

The following training data (Table 1) consists of 16 customer records. Class label (Preferred product) is hierarchical in nature. Higher Level represents more general classification i.e. Desktop or Laptop, Lower Level is more specific i.e. subcategories of Desktops or Laptops.

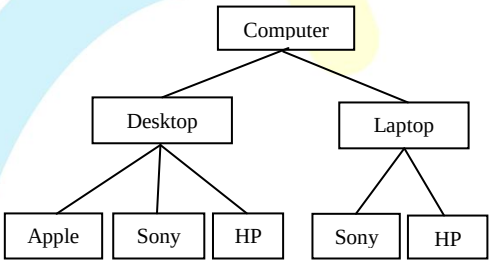


Figure 1. A hierarchy tree of product categories

The figure 1 shows the hierarchy within class labels. Class labels at higher level represent more general concept i.e. Desktop or Laptop and class labels at lower level represent specific concept i.e. Apple Desktop, HP Desktop, etc.

TABLE I. TABLE 1. TRAINING SET FOR CUSTOMERS

GENDER	CUSTOMER'S CAREER	CUSTOMER'S INCOME	PREFERRED PRODUCT
F	Business	1324	Apple
F	Business	1176	Apple
F	Engineering	1412	Apple
F	Engineering	1471	Apple
F	Education	1324	Apple
F	Education	1265	Apple
F	Engineering	1176	Apple
F	Education	1324	HP Laptop
F	Engineering	1912	Acer Desktop
F	Engineering	2059	Sony Desktop
M	Business	2353	Sony Laptop
M	Business	1412	Sony Laptop
M	Business	2206	HP Laptop
M	Business	1471	HP Laptop
M	Engineering	2206	HP Desktop
M	Education	1471	Acer Desktop

In standard decision tree algorithms like C4.5, all 7 class labels are considered at single level without any correlation between them.

The C4.5 tree structure for the data product categories as follows –

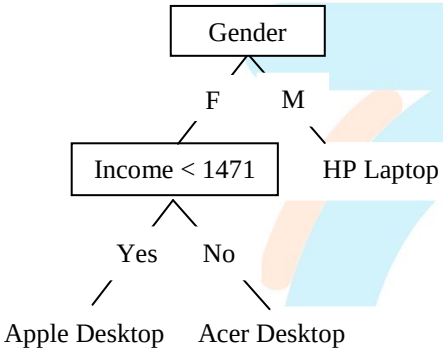


Figure 2. C4.5 tree for product categories

However, here, we would like to use existing knowledge of class hierarchies i.e. these 7 class labels can be categorized under class label – Desktop or Laptop. Our goal is to construct decision tree which will use known hierarchies for class labels to increase the accuracy.

III. KHCL ALGORITHM

In order to make use of known hierarchies in class attribute, we propose hierarchical classification framework, a form of ensemble classifier. Classifiers at the leaves conduct fine grained classification and classifiers at non-leaf nodes further up the hierarchy conduct coarse-grained classification, categorizing records using groups of labels.

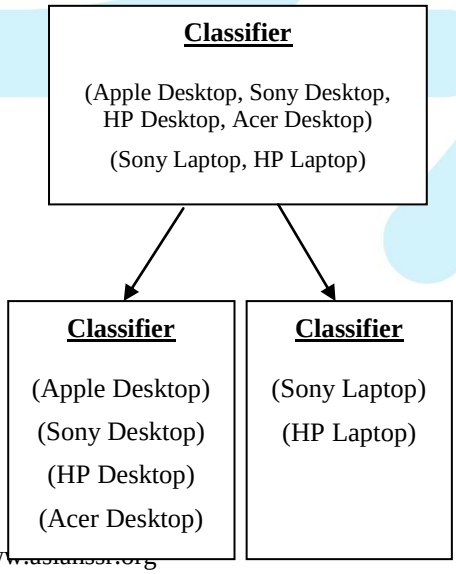


Figure 3. Classification hierarchy example

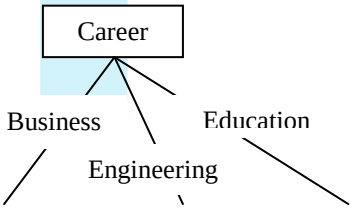
Framework:

1. Identify class label hierarchy nodes using prior knowledge.
2. Create individual classifiers at every node-
   a. Replace class value in training set by its next lower level group label
   b. Select only those training records which belongs to the group class label.
3. Classification –
   a. Replace the class label by it's highest level grouping (coarse-grain) and run classifier
   b. Based on the result of the above classifier, select next level classifier
   c. Repeat the process till we reach to the leaf node of class label hierarchy (fine-grain).

In the given example, 3 different trees to be constructed in advance.

1. Classifier with class values – Desktop & Laptop (All 16 training records to be used)
2. Classifier with class values - Apple Desktop, Sony Desktop and HP Desktop (11 respective training records)
3. Classifier with class values – Sony Laptop, HP Laptop (5 respective training records)

At the time of classification, the record to be classified using classification tree at root node (coarse-grained). Based on the output of classification, it will select appropriate next level classification tree. In this example first classification will decide whether the instance belongs to Desktop or Laptop category. Once it is predicted, the appropriate classification tree will be selected for next level prediction.



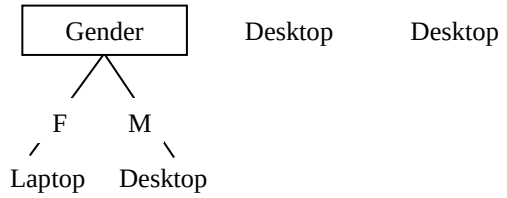


Figure 4. Classifier at root

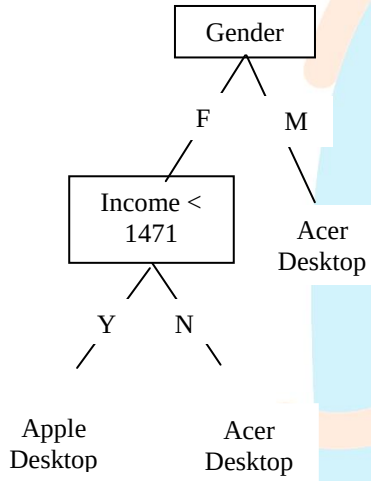


Figure 5. Classifier for group 'Desktop'

Figure 4 and 5 shows the classifiers at Root and Desktop nodes. Classifier at for Laptop is only single layer and predicts class as 'HP Laptop'.

Another advantage of this method is classification results are also available in 'coarse-grain' to 'fine-grain' fashion to end user providing more insight into classification process.

Many of the times, users many not be interested in the final class value, even higher level visibility of class value is very useful in decision making process.

IV. Performance Evaluation

We used J-48 algorithm from WEKA [7] machine learning software. In the given example of product categories for C4.5 algorithm the prediction accuracy is 62.50%. It correctly classified 10 out of 16 examples. However, our proposed framework correctly classified 13 of 16 examples providing classification accuracy to 75.00%. Algorithm at root correctly classified 15 examples (out of 16) for level class i.e. Desktop or Laptop. 'Desktop tree' and 'Laptop tree' further correctly classified 10 (out of 11) and 2 (out of 4) examples resulting into total 12 out of 16 correctly classified examples.

V. Conclusions and Future Work

The framework mentioned in this paper helps to improve the classification accuracy of decision tree classifier in case of hierarchical class labels. Classification results initially gives more general class label and at the end more specific class.

References

[1] Jiawei Han, Michline Kamber, 'Data Mining Concepts and Technique' , Kaufmann Pulications, 2001

[2] David Hand, Heiki Mannila, Padhraic Smyth, 'Principles of Data Mining', The MIT Press.

[3] Tom Mitchell, 'Machine Learning', Mc-GrawHill Publications

[4] J.R.Quinlan, 'C4.5 Programs for Machine Learning', Morgan Kaufmann Publications, San Mateo, CA, 1993.

[5] Yen-Liang Chen, Hsiao-Wei Hu, Kewi Tang, 'Consturcting a decision tree from data with hierarchical class labels', Expert Systems with Applications 36 (2009) 4838-4847.

[6] Alshdaifat E., Coenen F., Dures K. (2013) Hierarchical Single Label Classification: An Alternative Approach. In: Bramer M., Petridis M. (eds) Research and Development in Intelligent Systems XXX. pp-39-52 Springer, Cham.

[7] I. H. Witten, E. Frank. Nuts and bolts: Machine Learning alogrithms in java. In:Data Mining: Practical Machine Learning Tools and Techniques with Java Implementation, pp.265-320. Morgan-Kaufmann, 2000.

[8] Sachin Gavankar, Sudhir Sawarkar, "Decision Tree: Review of Techniques for Missing Values at Training, Testing and Compatibility", Proc. of the 3rd Int'l Conf on Artificial Intelligence, Modelling and Simulation, AIMS2015, Malaysia, 2015.