Customizing the WEKA for Visual Improvements in the Output of J48 Decision Tree

M.A.S.M. Perera, W.D.S.S. Appuhami, A.N.P.Bandara, D.S.Deegalla
Department of Computer Engineering,
Faculty of Engineering,
University of Peradeniya

Abstract - **The integration of communication and computing has dawned an enriched era of society which feeds on information. Yet, a huge amount of potentially useful and implicit information is locked up in databases as raw data. Data Mining (DM) is an active research area which deals in finding patterns, anomalies and trends in this data and summarizing them with quantitative and qualitative models to predict future aspects. In this paper, we describe the visual improvement of the output produced by the J48 decision tree algorithm in the WEKA toolkit, which is an open source DM toolkit widely used in teaching as well as research. We improved the visual representation of decision trees by coloring nodes according to the class distribution of the instances reached at each node, displaying values of information gain and gain ratio of intermediate nodes and displaying decision rules of leaf nodes accordingly. These improvements enhance the model's user friendliness and understandability.**

1. Introduction

Large amount of data could be collected and stored with recent advances in communication and computing technologies. This data can be utilized for betterment of human beings when it is analyzed and some dependencies and correlations are detected. Data Mining (DM) [1] is the field of study which involves in finding patterns, anomalies and trends in these data and summarizing them with satisfactory models to predict behavior of future occurrences of data of same type. The strength of DM lies in advanced tools, techniques and algorithms developed by different individuals and groups. WEKA [2] is one such open source toolkit available for DM tasks, which is popular among academic and research personnel.

In this paper we present improvement of the visual representation of the decision tree produced by J48 algorithm of WEKA tool. J48 is an open source java implementation of C4.5 [3] algorithm. Figure 1 illustrates the decision tree produced by this algorithm for the dataset in Table 1. If we compare this dataset and the decision tree building methodology against the output decision tree, it is an obvious fact that this output is a minimal representation of the result, because it can be made more comprehensive and informative by including more details in the output. Since a better visual representation is a dominant feature of a good toolkit, we targeted on accomplishing three objectives to make WEKA an outstanding DM tool among its competitors.

Our objectives are

- Coloring nodes according to the class distribution of instances reached at each node.
- Displaying values of information gain and gain ratio of intermediate nodes.
- Displaying decision rules of leaf nodes accordingly.

Information gain and gain ratio are important information measuring criterions used in J48 algorithm.

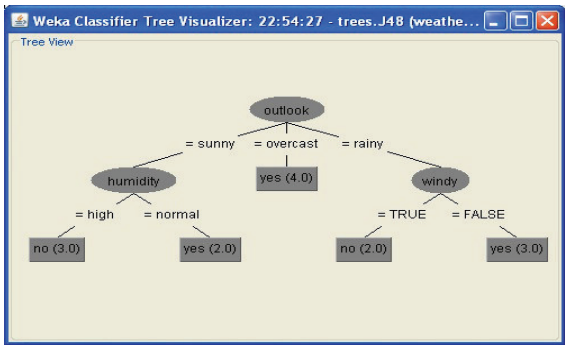

Figure 1. Decision Tree Produced by J48 Algorithm

This will cater DM teachers, students and researches to understand what features are dominant in the tree model.

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

Table 1. Weather Data set

2. Background

WEKA (Waikato Environment for Knowledge Analysis) is developed at the University of Waikato, New Zealand. It contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to this functionality.

There is a number of development projects done related with WEKA tool. However there is a less number of development projects done particularly on J48 algorithm. Most of the projects related with J48 algorithm are research projects that use it as the classification learning method. In Schistosomiasis risk mapping tool [4] the state of Minas Gerais, Brazil has customized J48 decision tree by colouring leaf nodes. They have coloured each leaf node according to class attribute value of that node with different colours for different class values. In our project, apart from leaf nodes, we have coloured the intermediate nodes also according to the distribution of class values at each node.

In TADA-Ed [5] by Wake Forest University, they have coloured the nodes of the decision tree by considering the dominant attribute value at each node. But it does not reflect the distribution of attribute values at each node as in our implementation.

3. Implementation

The most challenging part was to understand the relevant source code areas to be modified. A reasonable amount of time was spent on studying and hacking into the code. We divided the development of the project into the following three phases.

Modifying WEKA native dot format.
Generating decision rules.
Modifying the decision tree drawing

Modifying WEKA native dot format

The final J48 decision tree produced by WEKA is drawn by successively translating a native dot format defined as a string into a java drawing. We had to modify this format in order to include the additional information of class distribution, information gain and gain ratio. Figure 3 and figure 4 display the previous dot format and modified dot format for weather data set in Table 1.

Generating Decision Rules

WEKA produces the hierarchy of the decision tree as a string in its output. So we divided this output using java string manipulation techniques into a set of decision rules. Then each rule was assigned to its corresponding leaf. Figure 4 displays the hierarchical string produced by WEKA for the dataset in Table 1. Figure 5 displays the set of rules generated from this string.

Modifying the Decision Tree Drawing

After completing phase 1 and phase 2, source code related with drawing the tree was modified in order to translate the dot format into the decision tree. Here a user can visualize both the default decision tree and customized decision tree. Also a legend

was included in the output to display colours relevant to each class.

```
digraph J48Tree {

N0 [label="outlook" ]
N0->N1 [label="= sunny"]
N1 [label="humidity" ]
N1->N2 [label="= high"]
N2 [label="no (3.0)" shape=box style=filled
]
N1->N3 [label="= normal"]
N3 [label="yes (2.0)" shape=box
style=filled ]
N0->N4 [label="= overcast"]
N4 [label="yes (4.0)" shape=box
style=filled ]
N0->N5 [label="= rainy"]
N5 [label="windy" ]
N5->N6 [label="= TRUE"]
N6 [label="no (2.0)" shape=box style=filled
]
N5->N7 [label="= FALSE"]
N7 [label="yes (3.0)" shape=box
style=filled ]
}
```

Figure 2. WEKA Native Dot Format

```
digraph J48Tree {

N0 [label="outlook #2 yes(9) no(5)
info(0.247) ratio(0.156)"]

N0->N1 [label="= sunny"]
N1 [label="humidity yes(2.0) no(3.0)
info(0.971) ratio(1.0)"]
N1->N2 [label="= high"]
N2 [label="no (3.0)" shape=box style=filled
]
N1->N3 [label="= normal"]
N3 [label="yes (2.0)" shape=box
style=filled ]
N0->N4 [label="= overcast"]
N4 [label="yes (4.0)" shape=box
style=filled ]
N0->N5 [label="= rainy"]
N5 [label="windy yes(3.0) no(2.0)
info(0.971) ratio(1.0)"]
N5->N6 [label="= TRUE"]
N6 [label="no (2.0)" shape=box style=filled
]
N5->N7 [label="= FALSE"]
N7 [label="yes (3.0)" shape=box
style=filled ]
}
```

Figure 3. Modified Dot Format

```
outlook = sunny

|   humidity = high: no (3.0)
|   humidity = normal: yes (2.0)
outlook = overcast: yes (4.0)
outlook = rainy
|   windy = TRUE: no (2.0)
|   windy = FALSE: yes (3.0)
```

Figure 4. Hierarchical String Produced by WEKA

```
If outlook = sunny and humidity = high then
play=no
If outlook = sunny and humidity = normal
then play=yes
If outlook = overcast then play=yes
If outlook = rainy and windy = TRUE then
play=no
If outlook = rainy and windy = FALSE then
play=yes
```

Figure 5. Set of Rules

4. Results and analysis

At the implementation time, the latest version of WEKA was version 3.7.2 and its source came as an eclipse project. Therefore we configured the source code as a project in eclipse IDE Galileo version 1.2.2. Apache Ant 1.7.1 was used as the java building tool. Our development was tested for all the datasets available in WEKA for J48 algorithm. In the following sections we summarize the output of the project.

1. Customized output of J48 Algorthm

Figure 6 illustrates the customized output produced for weather dataset after the completion of development. In the modified version of WEKA tool one can visualize both the customized decision tree and the normal decision tree. When this customized output is compared with the normal output it is evident that a decision tree learner can get more information about how the result has been built and the class distribution at each node. This will be useful in many aspects. Especially this kind of representations can be used as a teaching aid in DM subjects to give the students a better understanding about this learning method. This is one of the main targets of our work.

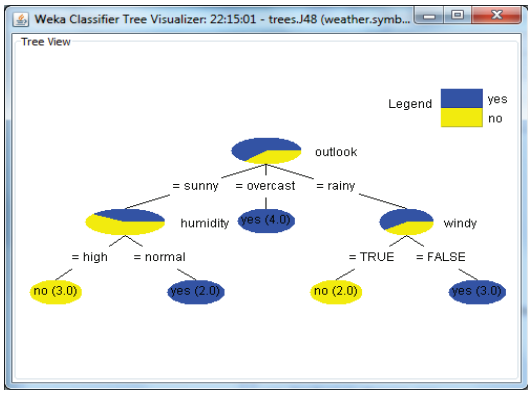

Figure 6.Customized Output of J48 Decision Tree

2. Information at intermediate nodes

One of the objectives in our project was to display the values of information gain and gain ratio of each intermediate node. These two values are important information measuring criterions used in building the decision tree output. In the normal output this was not displayed as an output. So we added those functionalities to the output. When the user clicks on an intermediate node a figure will be displayed as follows. Figure 8 displays the details about humidity intermediate node. This output was obtained using JFreeChart [6] library. In this representation the distribution can be obtained with the number of instances of each class with their percentage values. The 'info' stands for information gain and 'ratio' stands for gain ratio. We used JFreeChart version 1.0.13 in our project.

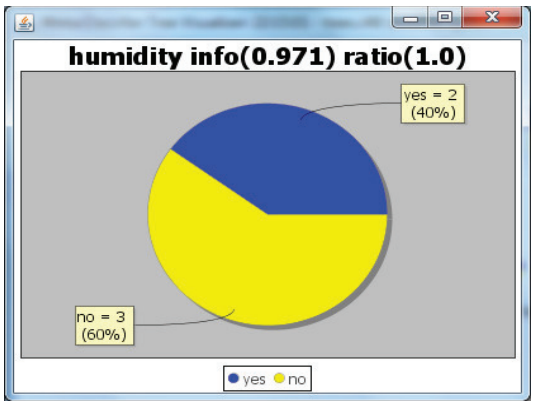

Figure 7.Details of Humidity Intermediate Node

3. Decision Rules

In classification learning, technique of decision rules is a popular method. Decision tree model could be used to create a set of decision rules.

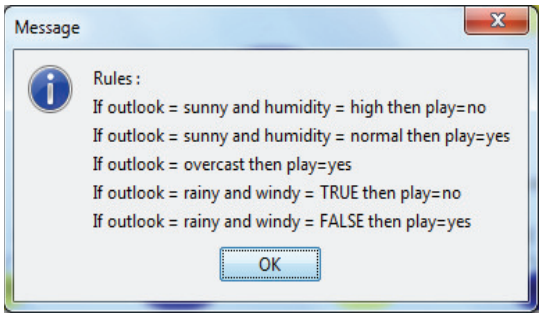

Figure 8.Rule Set for Weather Data set

Therefore one can generate decision rules by assigning one rule for each path from the root node to a leaf node in a decision tree. Following the same, we facilitate the output with decision rules. When user clicks on the tree and selects the 'Get Rule Set' option the final rule set will be displayed. Also if he wants to identify for which leaf node it belongs to he can click on the relevant leaf node to get the right rule. Figure 8 displays the rule set. Figure 9 displays the result when user clicks on the leftmost leaf node.

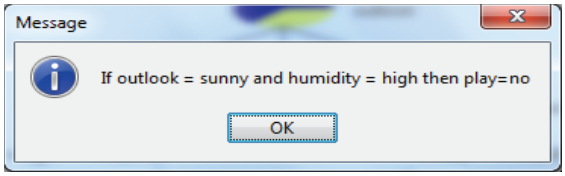

Figure 9. Decision Rule for Leftmost Leaf Node

5. Conclusions

In this paper, we describe the enhancements made to the output of J48 decision tree learning algorithm of the WEKA tool to improve its readability, user friendliness and understandability. We have identified three targets to achieve above objectives in order to cover the most basic requirements in decision tree learning.

When one compares the modified version of the output of J48 decision tree algorithm with the default version, it is obvious that modified tree visualization is more illustrative than the original output. This modified version could be used by DM teachers to provide better understanding on decision tree learning algorithm to their students. Further, it could also be useful for other research and academic purposes.

As a further improvement one can modify this tree to illustrate the values of information gain and

gain ratio of each attribute at each node which will provide the user a better understanding on selection of an attribute as the node attribute.

References

[1] Ian H. Witten and Eibe Frank. Data mining, Practical Machine Learning Tools and Techniques.Morgan Kaufmann Publishers, second edition edition, 2005

[2] http://www.cs.waikato.ac.nz/ml/weka/ [1-October-2011]

[3] Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993.

[4] Flavia T Martins-Bede et.al. Schistosomiasis, Risk mapping in the state of minas gerais, Brazil. http://www.ar-tracking.de/Markers.141.0.html, 2010. [21-March-2011].

[5] http://imej.wfu.edu/articles/2005/1/03/. [1-October-2011].

[6] http://www.jfree.org/jfreechart/[1-October-2011]