Interactive Exploration of Decision Tree Results

Khang N. Pham[1] and Nghi T. Do[2]

[1] IRISA
Campus de Beaulieu
F35042 Rennes Cedex, France
(email: pnguyenk,amorin@irisa.fr)
[2] INRIA Futurs
L.R.I., University Paris-Sud
F91405 ORSAY Cedex, France
(e-mail: dtnghi@lri.fr)

Abstract. Our investigation aims at interactively exploring the decision tree results obtained by the machine-learning algorithms like C4.5. We propose an interactive graphical environment using the new radial tree layout, zoom/pan techniques and some existing visualization methods like explorer-like, hierarchical visualization, interactive techniques to represent large decision trees in a graphical mode more intuitive than the results in output of usual decision tree algorithms. The interactive exploration system on one hand can preserve the global view in a large representation of radial layout, zoom/pan techniques and on the other hand, it also provides a very good performance for an interesting sub-tree in the explorer-like view with simplicity, speed of task completion, ease of use and user understanding. The user can easily extract inductive rules and prune the tree in the post-processing stage. He has a better understanding of the obtained decision tree models. The numerical test results with real datasets show that the proposed methods have given an insight into decision tree results.
Keywords: Post-processing decision trees, Interactive exploration, Visual data mining.

1 Introduction

Real-world databases have increased rapidly [Fayyad *et al.*, 2004] in recent years so that Data-mining [Fayyad *et al.*, 1996] is necessary to extract hidden useful knowledge from large datasets in a given application. This usefulness relates to the user goal, in other words only the user can determine whether the resulting knowledge answers his goal. Therefore, data mining tools should be highly interactive and participatory. The idea here is to increase the human involvement through interactive visualization techniques in a data-mining environment.

Over the last decade, a large number of visualization methods developed in different domains have been used in data exploration and knowledge discovery process [Fayyad *et al.*, 2001], [Keim, 2002]. The visualization

methods are used for data selection (pre-processing step) and viewing mining results (post-processing step). Some recent visual data mining methods [Ankerst *et al.*, 2000], [Do and Poulet, 2004] try to involve more intensively the user in the data-mining step through visualization. The effective cooperation can bring out some advantages such as: using the domain knowledge during the model construction, improving the confidence and comprehensibility of the obtained model, using the human pattern recognition capabilities in model exploration and construction.

Our investigation aims at interactively exploring the decision tree results obtained by decision trees algorithms like C4.5 [Quinlan, 1993]. Decision trees are powerful and popular tools [Kdnuggets, 2006] for classification and prediction. The attractiveness of decision trees is due to the fact that, in contrast to neural networks, decision trees represent rules in an easy way to understand for the user. In real-world classification task, the user is not able to explore a large tree in text mode in output of decision tree algorithms.

Some tree visualization systems [Lamping and Rao, 1996], [Munzner, 1997], [Brunk *et al.*, 1997], [Yee *et al.*, 2001] try to enhance decision trees with interactive tools. In data mining context, the important requirements of the user are: facility for using, large scale capacity, easy interpretation, extraction of interesting rules and pruning useless tree branches in a post-processing stage. In this aim, we propose the new radial tree layout and zoom/pan techniques for supporting interactive exploration of decision trees. An interactive graphical environment uses them and some existing visualization methods like explorer-like view, hierarchical visualization, linking, brushing, expand, unexpand techniques to represent large decision trees of over one thousand nodes. The interactive exploration system on one hand can preserve the global view in context of the entire hierarchy with a large representation volume of radial layout, focus-context techniques and on the other hand also provides a very good performance for viewing an interesting sub-tree in the explorer-like view with simplicity, speed of task completion, ease of use and user understanding. The user can interactively explore the decision tree result, easily extract interesting inductive rules and prune tree branches in a post-processing stage. He has a better understanding of the obtained decision tree models. The numerical test results with real datasets show that the proposed methods have given an insight into decision tree results.

This paper is organized as follows. In section 2, we present an explorer-like representation of decision trees. We show how to combine explorer-like, interactive tools and our radial layout, zoom/pan techniques into an hierarchical visualization for exploring ing large decision trees in section 3. We present some test results before the conclusion and future work.

2 Explorer-like representation

Explorer-like view maps the root of decision trees to the left top corner of the screen. A node is represented by a graphical point whose color denotes the majority class. The leaf size is with respect to the number of datapoints in the decision tree leaf node. The tree is hierarchically grown up. Figure 1 is explorer-like visualization of the decision tree with the Shuttle dataset [Michie *et al.*, 1994] (58000 datapoints in 9 dimensional input space, 7 classes).

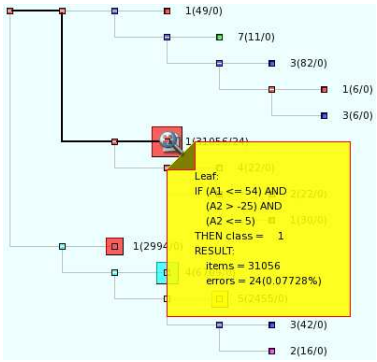

Fig. 1. Inductive rule extraction for classifying the Shuttle dataset

The user can easily explore the decision tree result in a graphical mode more intuitive than the results in output of decision tree algorithms. The node size helps the user determine if the decision rules create meaningful or spurious groups. Usually, the small nodes are less interesting than larges nodes because it is more cost effective and statistically reliable. Therefore the graphical leaf size and the hierarchical level can helps the user to find some important rules and prune the decision tree in a post-processing stage.

The user can extract inductive rules: clicking on a graphical leaf will highlight the path from the root to this leaf and then an inductive rule (IF-THEN) created from the root to the leaf is displayed. Figure 1 is an example of interactive rule extraction from the decision tree with the Shuttle dataset. With some interactive techniques like expand, unexpand, focus, zoom in/out, the user can navigate into the tree for exploring the information associated with the current node e.g. the number of errors and datapoints. With these interactive tools, level, size, color and information of the current node, the user has a good knowledge for pruning by himself the decision tree in a post-processing step. It is necessary to improve the estimated accuracy of the obtained model on new data (avoid overfitting in a learning task). For example, the user can unexpand the branch of the decision tree in figure 2 because he gets good information of this sub-tree like the level, the color, the size

and the number of errors and datapoints. Thus, the error rate (∼0.07%) in this case is not significant. Therefore, the user decides, with this meaningful information, to prune the branch as shown in figure 2. Our investigation also aims at trying to involve more intensively the user in the post-processing step through interactive visualization.

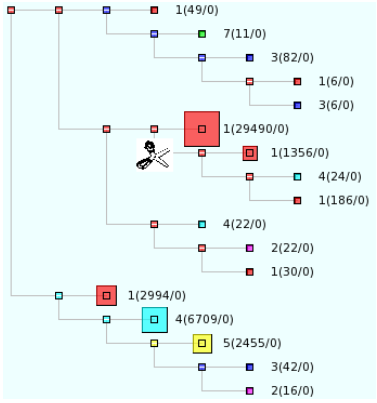

Fig. 2. Pruning the branch

The explorer-like view shows a very good performance for medium trees, both with simplicity, speed of task completion, easy of use and user understanding.

3 Radial layout, zoom/pan and hierarchical visualization techniques

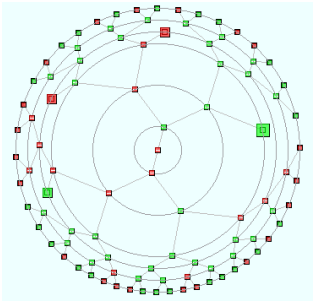

Fig. 3. Focus on the interesting level

For large trees, we try to improve the representation surface by using radial layout, zoom/pan techniques regarding the global view in context of

the entire hierarchy and hierarchical visualization techniques. With the radial layout algorithm, the root of decision trees is mapped on the screen center, multiple generations of children are mapped out towards arc of circle. The children are allocated space based on number of nodes descended from them.

We propose to use the fisheye technique that can help the user to focus on the levels relative to the current node. He can explore the interesting nodes of the decision tree. Figure 3 is the example of radial layout and fisheye representation for navigating the levels interesting for the current node.

We also propose a new interactive technique, zoom/pan that can help the user not only to focus on interesting nodes but also to see the global view. The idea is to zoom-in interesting nodes and zoom-out the other ones.

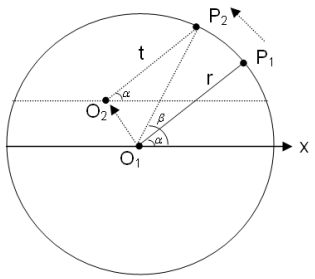

Fig. 4. Zoom/pan technique

Let us consider a root O mapped at a position O_1 and a node P mapped at a position P_1, as depicted in figure 4. The root O is moved from $O_1(x_1, y_1)$ to $O_2(x_2, y_2)$ so that the angle $\alpha(\overrightarrow{OX}, \overrightarrow{OP})$ is preserved. Thus the node P is also moved from P_1 to P_2. So we obtain the angle $(\overrightarrow{O_1X}, \overrightarrow{O_1P_2}) >$ the angle $(\overrightarrow{O_1X}, \overrightarrow{O_1P_1})$. The position P_2 with given (x_2, y_2) and α only requires the largest value of t^* solution of a quadratic equation (1):

$$t^2 + 2(\Delta x * cos(\alpha) + \Delta y * sin(\alpha))t + \Delta^2 x + \Delta^2 y - r^2 = 0 \qquad (1)$$

$$\text{where } \Delta x = x_2 - x_1 \text{ and } \Delta y = y_2 - y_1$$

Our zoom/pan technique is very simple and fast to project the tree nodes. The user easily uses our interactive technique to focus on interesting nodes as shown in figure 5 by dragging operations. He explores the information in the interesting nodes of decision trees with zoom/pan, fisheye and rotation techniques.

User experiments with tree visualization systems [Barlow and Neville, 2001] show that no single technique is the best for large tree exploration both with simplicity, speed, focus, global view, easy of use and user understanding. We would like to combine different techniques to overcome the single one. We propose a hierarchical visualization using explorer-like, radial layout, fisheye

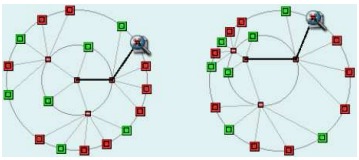

Fig. 5. Focus on interesting nodes with the zoom/pan technique

and zoom/pan techniques providing more useful information to the user. The idea is to divide the screen into two view panels. The first one on the left is used for visualizing the full decision tree with radial layout, fisheye and focus techniques. The user can select an interesting sub-tree and it is visualized by the explorer-like representation in the panel on the right. Thus, the user can interactively explore large decision trees of over one thousand nodes with the hierarchical technique. The exploration system on one hand can preserve the global view of the entire hierarchy with large representation volume of radial layout, fisheye, zoom/pan techniques and on the other hand, also provides a very good detail of an interesting sub-tree in the explorer-like view with simplicity, speed of task completion, ease of use and user understanding.

We have proposed to measure the important requirements of the user in a data mining with decision tree context like ease of use, large scale capacity, easy of interpretation, extraction of interesting rules and pruning useless tree branches in a post-processing stage.

We study the spam classification task (ref. http://www.ics.uci.edu/∼m-learn/databases/spambase). The spam email database was created by G. Forman and his colleagues at Hewlett-Packard Labs. The spambase has 4601 instances (1813 Spam = 39.4%) and 58 columns (attributes). Most of the attributes indicate whether a particular word or character was frequently occurring in the email. There are 48 continuous real [0, 100] attributes of type word_freq_WORD, 2 run-length attributes measuring the length of sequences of consecutive capital letters and the statistical measures of each attribute.

The decision tree algorithm C4.5 using decentred entropy proposed by [Lallich *et al.*, 2007] has classified the spam database with 92.24 % correctness. The resulting tree with 148 nodes is shown in figure 6. The user is able to explore the result with some interactive techniques like rotation, expand, unexpand, fisheye, zoom/pan. He easily navigates into the information associated with the current node e.g. the number of errors and datapoints. With these tools, level, size, color and information on current node, the user has a good knowledge for extracting inductive rules or pruning himself the decision tree in a intuitive graphical mode. As shown in figure 6, the user can focus on an interesting branch of the decision tree in the radial layout and then he easily explores in the graphical mode this sub-tree with the explorer-like view, i.e. intuitive rule extraction (in the radial layout or the explorer-like view) based on level, size and color of nodes. The user has an insight into the decision tree result of the spambase classification task.

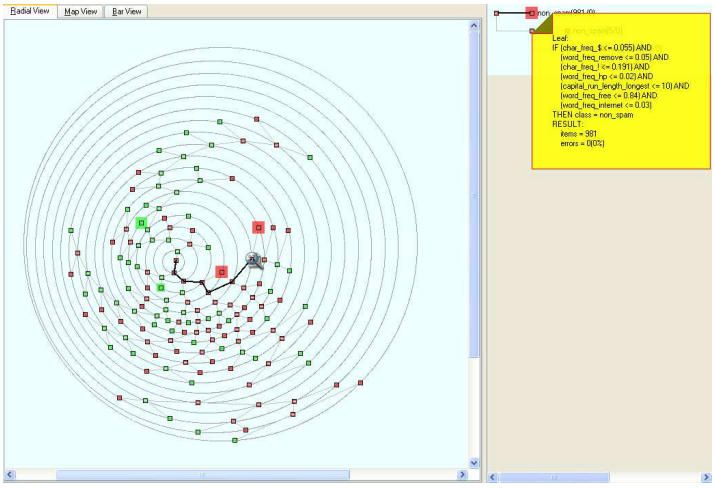

Fig. 6. Graphical extraction of inductives rules for classifying the Spambase

4 Conclusion and future work

We present in this paper a new graphical environment for exploring the decision tree results obtained by decision tree algorithms. Our investigation aims at improving the user's important requirements of decision trees in a data mining context: easy interpretation, extraction of interesting rules and pruning useless tree branches in a post-processing stage. We have proposed an interactive graphical toolkit using the new radial tree layout and zoom/pan techniques and some visualization methods exist like explorer-like, hierarchical visualization, interactive techniques to represent large unbalanced decision trees in a graphical mode more intuitive than the results in output of decision tree algorithms. The system can easily handle with large trees of over one thousand nodes. The user can interactively explore the decision tree result, easily extract interesting inductive rules and prune tree branches in a post-processing stage. The interactive exploration system provides a global view of the entire hierarchy in large representation volume of radial layout, focus-context techniques and a good detail of interesting sub-tree in an explorer-like view with simplicity, speed of task completion, ease of use and user understanding. The numerical test results with real datasets show that the proposed methods have given an insight into large, unbalanced tree in a graphical mode more intuitive than the results in output of decision tree algorithms. With some interactive techniques like expand, unexpand, fisheye, zoom/pan, rotation, the user is able to explore decision tree results in a post-processing stage. He easily navigates into information associated with tree nodes e.g. the number of errors and datapoints. The user has also a good knowledge for extracting inductive rules or pruning the decision tree in a intuitive graphical mode based on level, size, color and information of

tree nodes. The user has a better understanding of the obtained decision tree models.

A forthcoming improvement will be to find appropriate abstractions for exploring extremely large trees in a complex classification task.

References

[Ankerst *et al.*, 2000]M. Ankerst, M. Ester, and H-P. Kriegel. Towards an effective cooperation of the computer and the user for classification. In *Proceeding of KDD'00, 6th ACM SIGKDD*, pages 179–188, 2000.

[Barlow and Neville, 2001]T. Barlow and P. Neville. Comparison of 2-d visualizations of hierarchies. In *Proceeding of IEEE InfoVis*, pages 131–138, 2001.

[Brunk *et al.*, 1997]C. Brunk, J. Kelly, and R. Kohavi. Mineset: An integrated system for data access, visual data mining, and analytical data mining. In *Proceedings of KDD'97, AAAI Press*, pages 135–138, 1997.

[Do and Poulet, 2004]T-N. Do and F. Poulet. Enhancing svm with visualization. In *Discovery Science 2004, E. Suzuki et S. Arikawa Eds.*, pages 183–194, 2004.

[Fayyad *et al.*, 1996]U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, 17(3):37–54, 1996.

[Fayyad *et al.*, 2001]U. Fayyad, G. Grinstein, and A. Wierse. *Information Visualization in Data Mining and Knowledge Discovery.* Morgan Kaufmann Publishers, 2001.

[Fayyad *et al.*, 2004]U. Fayyad, G. Piatetsky-Shapiro, and R. Uthurusamy. Summary from the kdd-03 panel - data mining: The next 10 years. *SIGKDD Explorations*, 5(2):191–196, 2004.

[Kdnuggets, 2006]Kdnuggets. Data mining methods. In *KDnuggets Polls*, 2006.

[Keim, 2002]D. Keim. Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):1–8, 2002.

[Lallich *et al.*, 2007]S. Lallich, P. Lenca, and Vaillant B. Construction dune entropie décentrée pour l'apprentissage supervisé. In *3ème Atelier Qualité des Données et des Connaissances*, pages 45–54, 2007.

[Lamping and Rao, 1996]J. Lamping and R. Rao. The hyperbolic browser: A focus + context technique for visualizing large hierarchies. *Journal of Visual Languages and Computing*, 7(1):33–55, 1996.

[Michie *et al.*, 1994]D. Michie, D.J. Spiegelhalter, and C.C. Taylor. *Machine Learning, Neural and Statistical Classification.* Ellis Horwood, 1994.

[Munzner, 1997]T. Munzner. H3: laying out large directed graphs in 3d hyperbolic space. In *IEEE InfoVis*, pages 2–10, 1997.

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

[Yee *et al.*, 2001]K-P. Yee, D. Fisher, R. Dhamija, and M. Hearst. Animated exploration of dynamic graphs with radial layout. In *IEEE InfoVis*, pages 43–50, 2001.