Human Involvement and Interactivity of the Next Generation's Data Mining Tools

Mihael Ankerst

The Boeing Company,
P.O. Box 3707 MC 7L-70, Seattle, WA 98124
mihael.ankerst@boeing.com

ABSTRACT
The basic task of the knowledge discovery and data mining (KDD) process is to extract knowledge from data such that the resulting knowledge (pattern) is useful in a given application. Obviously, only the user can determine whether the resulting knowledge satisfies this requirement. Moreover, what one user may find useful is not necessarily useful to another user.
Instead of allowing an automated data mining process to iterate in a trial-and-error manner, a natural but neglected way to enhance the process is to support human involvement. To achieve the goal that the user steers and monitors the information flow without burdening him performing tasks that can be done automatically, an interface for human involvement has to be well designed and integrated in the KDD process. As additional benefits from this approach, the user better understands and trusts the resulting patterns. *Visual Classification* which is a recently introduced approach has shown the benefits of this new direction for decision tree classifiers.

1. INTRODUCTION

Knowledge discovery [FPS 96] can be defined to be the non-trivial process of identifying patterns in data that are valid, novel, potentially useful and understandable. In [BA 96], the authors stress the interactive nature of the knowledge discovery process consisting of steps like selection, preprocessing, transformation, data mining and evaluation. Practice has shown that the process is virtually a loop which is iterated many times until good results are obtained. Major vendors of data mining software like SPSS Clementine or SAS EM have taken this fact into account and provide graphical views to model and visualize this complex process.
Even though data mining is the core analytical step, the quality of the results heavily rely on data preparation which usually takes at least 80-90% of the total time [Pyle 99]. During data preparation, domain knowledge is used to prepare the data and make it suitable for a data mining algorithm. If the evaluation of the patterns is not satisfactory there are two possibilities. Either the user can just reiterate the data mining step or he reiterates the whole process returning to the data preparation phase. The first case seems to be more natural and efficient, however, most state-of-the-art tools just facilitate a careful tuning of various algorithm-specific parameters in a trial-and-error manner [FPS 96b], [AEK 00]. Therefore, a costly return to the data preparation phase is required to incorporate new domain knowledge which has been acquired in a previous iteration.

2. KNOWLEDGE TRANSFER IN THE DATA MINING STEP

Historically, the exploding amount of available data has led researchers to the area of knowledge discovery and data mining. The main motivation is that humans are not capable to analyze the current size of the available data neither manually nor with basic statistical methods. As a result, the technological challenge of performing *everything* automatically has dominated the awareness of researchers and developers of commercial tools up to the present. However, the knowledge discovery process is not meant to exclude the human since the discovered knowledge addresses the human! Therefore, the roles of the computer and the human have to be properly identified.

There are two ways to enable human involvement in the data mining step. Either the user specifies constraints in some textual form or a visualization provides an interface where the user acquires knowledge about the current state of the KDD process and is enabled to manipulate a mining algorithm through interaction.

Though text-based human involvement has been shown to be effective for particular tasks, see e.g. [WHH 00][THL+ 01], we will focus on a discussion of potential benefits of interaction based on visualization techniques since this approach entails powerful data and knowledge representations. Several approaches have been proposed recently (commonly classified by the term 'visual data mining'), however, almost all of them can be classified into one of the following two groups. The first group comes from the research field 'information visualization'. It is based upon a data visualization providing an overview of the data but not supporting a data mining algorithm explicitly. It is typically used in the preprocessing step or directly before the algorithm is invoked. The second group addresses knowledge representations, visualizing the patterns produced by an algorithm. Thus it is applied after the mining algorithm has terminated and basically supports the evaluation of the patterns.

We think that the potential benefits of human involvement are even better exploited if the visualization and interaction facilities are more tightly coupled with a mining algorithm, see also [AEEK 99][AEK 00][WFH+ 00][Wong 99]. To tightly couple these two worlds, mining algorithms have to be well understood to design appropriate visualization techniques supporting human involvement during the run of a mining algorithm and to identify key interaction points without burdening the user. The advantages of such a tightly coupled approach include:

1. **Transfer of domain knowledge in both directions**.
2. **Data mining tools are becoming more effective**.
3. **Less KDD steps are involved in each iteration**.

First, domain knowledge can be transferred from the human to the computer and vice versa. The user can incorporate his knowledge like e.g. focus on certain patterns, constraints, relations already known or knowledge about attributes and attribute values. On the other hand, data and knowledge representations can increase both the trust of the user in the discovered patterns and the user's domain knowledge about the specific application.

Second, human involvement can make data mining tools more effective if the user can specify how to focus the search. A run of a mining algorithm typically includes searches in large search spaces which cannot be performed exhaustively. At such points, involvement of the user can narrow down the search space significantly facilitating the mining algorithm to conduct a more accurate search. The visualization of the latter situation may even enable the user to make better decisions solely based on his perception than a mining algorithm.

Third, less KDD steps are involved in each iteration making the whole process more efficient. If the evaluation of some patterns are not satisfactory the user may just reiterate the data mining step if he is enabled to incorporate his domain knowledge reflecting his discontent with the current patterns.

To summarize, a tightly coupled visual data mining system enables the user to supervise the run of an algorithm and to intervene during the run by either using his domain knowledge or his perception.

3. AN EXAMPLE FOR HUMAN INVOLVEMENT IN THE DATA MINING STEP: VISUAL CLASSIFICATION

Based on a recently proposed approach called *visual classification* [AEK 00][AEEK 99], we show how a mining algorithm can be interleaved by human involvement such that the cooperation of the human and the computer yields the advantages described above. The visual classification approach decomposes the construction of a decision tree classifier into the steps depicted in

figure 1. The system is initialized with a decision tree consisting of the root node which corresponds to the whole training data set. A visualization is generated representing the data objects of the current node.

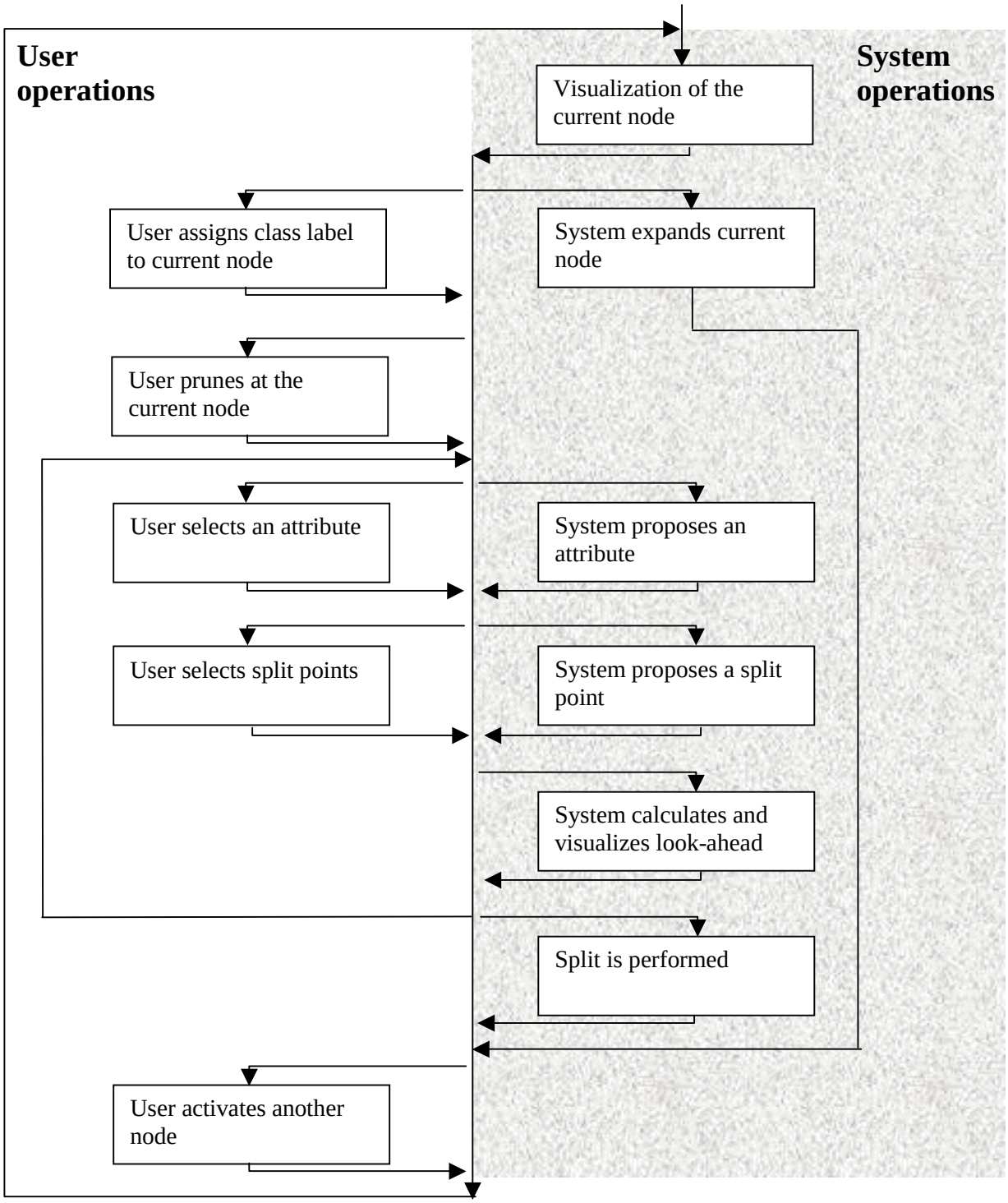

Figure 1: Cooperative decision tree construction

The task that is performed by a (univariate) decision tree algorithm is the search for the best split points in an attribute with respect to some goodness measure. To accomplish this task within a reasonable time several simplification are made by state-of-the-art algorithms, e.g. just the single best split point is evaluated and the evaluation is just based upon class distributions of the resulting partitions. At this point the visual classification approach sets an example of the benefits of human involvement since the task of split point selection can be performed by the user either by his perception, e.g. identifying multiple split points in an attribute or by using his domain knowledge, e.g. favoring an attribute or certain split points.

4. CONCLUSIONS AND OPEN ISSUES

Current data mining tools and algorithms provide very limited possibilities to incorporate domain knowledge if any at all. In our opinion, developers of commercial tools and most researchers still underestimate or do not recognize the potential benefit of human involvement in the data mining step to accelerate the whole KDD process and to improve the results. As pointed out in this paper, the benefits of human involvement in the data mining step can be the transfer of domain knowledge, more effective data mining tools and as a result less and shorter iterations within the knowledge discovery process loop. The approach of visual classification has shown this benefits in the context of decision tree classifiers.

Open issues in the next future are a) to design interfaces for human involvement in various data mining methods like text mining, clustering or association rules and b) to address scalability issues not just to main memory restrictions but also to visualization techniques.

If advances will be made in these fields then we think that next generation's data mining tools will improve knowledge acquisition and the trust and understanding of the patterns by the human.

5. REFERENCES

[AEEK 99] Ankerst M., Ester M., Kriegel H.-P.: "Visual Classification: An Interactive Approach to Decision Tree Construction", Proc. Int. Conf. on Knowledge Discovery and Data Mining, pp. 392-397, 1999.

[AEK 00] Ankerst M., Ester M., Kriegel H.-P.: "Towards an Effective Cooperation of the User and the Computer for Classification", Proc. Int. Conf. on Knowledge Discovery and Data Mining, pp. 178-188, 2000.

[BA 96] Brachmann R., Anand T.: "The Process of Knowledge Discovery in Databases: A Human-Centered Approach", Advances in Knowledge Discovery and Data Mining, AAAI Press, Menlo Park, CA, pp. 37-58.

[FPS 96] Fayyad U., Piatetsky-Shapiro G., Smyth P.: "From Data Mining to Knowledge Discovery: An Overview", Advances in Knowledge Discovery and Data Mining, AAAI Press, Menlo Park, CA, pp.1-30.

[FPS 96b] Fayyad U., Piatetsky-Shapiro G., Smyth P.: "The KDD process for extracting useful knowledge from volumes of data", Communications of the ACM, 39(1).

[Pyle 99] Pyle, Dorian: "Data Preparation for Data Mining", Morgan Kaufmann, 1999.

[THL+ 01] Tung A.K.H., Han J., Lakshmanan L.V.S., Ng R.T.: "Constraint-based Clustering in Large Databases", Proc. Int. Conf. on Database Theory, pp. 405-419, 2001.

[WFH+ 00] Ware M., Frank E., Holmes G., Hall M., Witten I.H.: "Interactive Machine Learning – Letting Users Build Classifiers", http://www.cs.waikato.ac.nz/ml/publications.html

[WHH 00] Wang K., He Y., Han J.:"Mining Frequent Itemsets Using Support Constraints", Proc. 26th Int. Conf. On very Large Data Bases, pp. 43-52, 2000.

[Wong 99] Wong P.C.:"Visual Data Mining", IEEE Computer Graphics and Applications, Vol. 19(5), pp. 20-12, 1999.