BLUE: An Interactive Visualization System for Categorical Data
Technical Note

Robert W. Kerbs
Computer Science Department
California State Polytechnic University, Pomona
rwkerbs@csupomona.edu

Abstract

The visual data mining field is facing a difficult challenge; namely, how to build models from databases that possess primarily categorical attributes. Global models that incorporate categorical attributes are often created non- interactively resulting in non-intuitive and difficult to interpret representations. This paper introduces a prototype data visualization system, BLUE, to help induce meaningful decision trees from databases that contain primarily categorical attributes. BLUE is an interactive model creation and exploration system that provides a localized view of categorical attribute pairs within the global context of a decision tree. This approach facilitates the visual construction of globally relevant models in both a bottoms-up and tops-down manner while promoting its use by domain and non-domain experts alike.

1. Introduction

Data storage capacity has been doubling every nine months for over ten years [1]. As a result, companies and organizations are facing difficulties in extracting useful models from their data [2]. Timely construction of representational models should identify patterns and regularities that facilitate the prediction of future events based upon past performance. Consequently, models are often constructed without the benefit of domain expert knowledge or by automated processes alone. In an effort to bridge this gap, an interactive visualization framework is introduced to help construct domain relevant models.

1.2 Approach

Various data visualization techniques have been developed to aid with decision-making processes [3]. Many techniques provide a global view of data at the expense of local detail [4]. Other approaches use methods that allow for local detailed views while also maintaining a global context of the data [5], [6]. BLUE, the interactive model creation and exploration system introduced in this paper, employs the latter approach.

Two data visualization techniques are used in order to construct relevant models. A new visualization technique, called Blue’s Independence DiagramS (BIDS), facilitates the study of relationships between attribute pairs in relational databases. BIDS are a modification to the independence diagrams found in [7]. A decision tree is used to represent globally induced models from the analysis of individual BIDS. The overall framework is consolidated into a prototype data visualization system called BLUE.

Models are constructed interactively. Users possess the ability to backtrack to previously visited nodes of the induced decision tree. This capability allows the expert user to explore the data set extensively and provides for the possibility of discovering previously unknown knowledge. Basic visual construction guidelines are demonstrated showing how the non-domain expert may also build models.

BLUE is a unified environment in which data sets possessing small to a moderate number of attributes can be explored. This paper introduces the techniques and guidelines for exploring data sets and extracting global models using BLUE.

1.3 Related Work

The analysis of relationships in databases has been studied for many years [8]. Most analyses are based upon information theory or statistical metrics [9]. The overriding goal has been to discover useful, otherwise unknown relationships in the data, referred to as Knowledge Discovery in Databases [10], [11], [12], [13], [14].

Many visualization techniques exist to analyze local detail of data sets. However, most of these techniques analyze only numeric data. One explanation for this preference is that categorical data must use a different metaphor than that which is used with numeric data. Another explanation is that visualization techniques that focus on categorical data tend to be application-specific [15]. BLUE has been designed for exploration of databases that contain primarily categorical attributes in a non-application specific manner.
An outline of this paper is as follows: section 2 introduces BIDS and the theoretical foundations upon which BLUE is constructed; section 3 discusses the application of BLUE to a data set adapted from [16] and presented in [17] as well as recommendations for future research; finally, section 4 presents conclusions.

2. BLUE

The impetus for creating BLUE is the desire to produce a visual data mining approach that relies on users’ visual selectivity preferences and knowledge of a data set to guide the induction process. A directed acyclic graph (DAG), otherwise known as a decision tree, is used to represent global models induced from BLUE. A new visualization technique, called BLUE’s Independence DiagramS (BIDS), is introduced for viewing local detail in the data. Based upon proper analysis of BIDS the decision tree can be interactively constructed in either a tops-down or bottoms-up manner. BLUE supports both single and multiple class supervised learning.

2.1 BIDS

Independence Diagrams [7] are two-dimensional images that compare attributes from a data set. The image format is a combination equi-depth and equi-width histogram, referred to as a equi-slice histogram. These images facilitate user detection of interesting patterns in data sets that would otherwise go without detection. BIDS differ from independence diagrams in a number of ways.

First, the approach taken in [7] suggests a viewing of independence diagrams for all combinations of attributes (i.e. \( n(n-1)/2 \), where \( n \) represents the number of attributes in the relation of interest). While useful for viewing all possible dependencies between attribute-pairs, it is inefficient for the problem undertaken in this investigation, namely; classification. Consequently, it is only necessary to construct BIDS between the classification attribute and each attribute. This reduces the number of BIDS needed for analysis to \((n-1)\).

Second, preliminary experiments using grayscale, as in [7], resulted in difficulties in determining if two attributes were fully correlated or not. Consequently, two colors were added to better discern similarities and differences between correlation rectangles (i.e. the rectangles comprising a BIDS). Green is used to show correlation rectangles that possess the highest tuple population while red is used to show correlation rectangles possessing no tuple population. Correlation rectangles between the two extremes are linearly interpolated with grayscale.

Finally, BIDS are focused on categorical data necessitating a different form of tuple density measure. BIDS construct correlation rectangles in direct proportion (i.e. size) to the tuple population found in each correlation rectangle. When combined with green/red correlation rectangle color coding, users can very quickly see what categories are clearly correlated to specific classification attribute classes.

As BIDS are constructed, users select those correlation rectangles deemed relevant to grow the decision tree. The tuples corresponding to the induced decision tree are removed from the data set so that other relationships between attribute pairs can be made. In addition, as leaves are created in the decision tree classification rules are simultaneously extracted.

2.2 Software Prototype

BLUE is shown in Figure 1. The top section of BLUE contains four BIDS with corresponding attribute labels. The y-axis in each BIDS represents an attribute’s categorical relationship with the x-axis classification attribute (i.e. the x-axis is always the classification attribute). The large BIDS in the center-left of the screen is a zoomed version of the first BIDS. This zooming ability allows users to better analyze individual BIDS. The reporting window, to the right of the zoomed BIDS, gives detail about the zoomed BIDS. The bottom left pane displays the decision tree as it is grown. The lower right pane displays the extracted classification rules. At any time users can backtrack up the decision tree and reanalyze BIDS and grow a different decision tree.

3. Application

BLUE is demonstrated with a data set adapted from [16] and presented in [17]. The data set contains five attributes and fourteen records. The attributes are: outlook, temperature, humidity, windy, and the classification attribute play. Based upon the values found in the records the goal is to construct a global model of the data set in decision tree form that will direct a user whether s/he should play an outdoor activity like golf or tennis. For instance, one record has the following values: the outlook is sunny, the temperature is hot, there is high humidity, it is not windy, and the classification attribute directs the user to don’t play the activity. Similarly, other records have different values for these attribute instances.

First, through the use of BIDS, each attribute is correlated individually with the classification attribute play. The corresponding four BIDS are displayed near the top of BLUE’s main display window in Figure 1. Note how each respective attribute is identified at the top of each BIDS and the x-axis represents the classification attribute play. It can be immediately determined that there is a correlation occurring in the first BIDS. This can be ascertained by seeing the two horizontally located red and green correlation rectangles (i.e. the red rectangle indicates no tuple density, the green rectangle indicates maximum tuple density). The left mouse button is clicked over this
Figure 1: The Saturday Morning Data Set as explored with BLUE.

BIDS to place it in the larger zoomed window (towards the left-center of the screen). The BIDS is zoomed and detail about the image is now presented in the reporting window adjacent to the zoomed window. It can now be seen that cat2 corresponds to the outlook category OVERCAST and class2 corresponds to the play attribute category PLAY. Since there is a green and red correlation rectangle-pair, it can be determined that all four OVERCAST tuples correlate directly to the classification category PLAY. Regardless of other attribute values (i.e. humidity, windy, temperature), if the outlook is OVERCAST the user will always play the outdoor activity.

At this stage the decision tree leaf node is grown as shown in the lower left decision tree window. If this induced leaf is acceptable to the user, then the corresponding classification rule is extracted and displayed in the lower right window. Consequently, the four tuples corresponding to this rule are removed from the data set and new BIDS are displayed that no longer contain these respective tuples.

3.1 Variations

The sample shown above demonstrates a simple case where selection of a particular BIDS, in lieu of another, is based upon a clear relationship of attribute categories. In particular, the red and green correlation rectangles highlight a fully correlated relationship between the

Extracted Classification Rules

Overcast (outlook) --> PLAY
attribute outlook and the classification attribute play. The question then arises, what should be done if multiple BIDS possess correlations or none of the BIDS contain any correlations? The answer is that the user will need to rely upon his/her knowledge of the data set or a set of visualization guidelines will need to be followed in order to advance induction of a representative model. Visual guidelines that fall in the latter category are introduced in [18].

3.2 Future Research

BLUE has been utilized for data sets containing up to 17 attributes. It would be useful to experiment with data sets possessing more attributes. It would also be useful to explore the management of attributes that contain many categories. It has been found that when an attribute contains more than about ten categories the corresponding BIDS become visually complex and difficult to interpret. Finally, it would be desirable to come up with alternatives for displaying attributes with continuously valued numeric data.

4. Conclusions

Representing categorical attributes in a global model often results in knowledge structures that are non-intuitive and difficult to interpret. This paper introduced BLUE, a prototype data visualization system that allows for local detail between categorical attribute pairs to be examined from within the global context of a decision tree. A new visualization diagram was introduced, BIDS, to help clearly represent statistical relationships of categorical attribute pair ranges. Finally, BLUE was interactively demonstrated with a simple example.

References


