Abstract:

Interactive analysis introduces dynamic changes in Visualization. On another hand, advanced visualization can provide different perspectives of the data to the user, hence, provide effective way of data mining. This paper discusses new ideas for interactive data mining tool based on R through HCI techniques. Also demonstrates the purposed features through data mining and visualization examples, such as ensemble method and tree map. Lastly, explore some possibilities and difficulties from the view of implement.

Because of the fast rate of increasing in data complexity, existing efficiency of data mining is facing great challenge. There are emerging different data mining languages and related tool set. R language is one of the popular languages, which is specially welcomed by scientists and other professionals in education. They have accumulated thousands of data mining packages for different algorithms and also a lot of examples are available. It’s very helpful for coming researchers and novice learners like us to focus on R language and related packages currently.

However, current data analysis tools for R like R-Studio simply integrate some visualization tool without considering the significance of human interactions. Traditional machine learning techniques don’t emphasize the user’s involvement. Such tools tend to rely on statistical analysis and plotting, and are not open enough to combine other cutting edge visualization techniques in time. Other commercial statistic analysis software like SPSS without openness is also not fit for our research purpose. To implement effective interactive data mining, this paper discusses a new idea and proposes for data mining tool combining human computer interaction techniques based on new visualization methods.
1 Related data mining concepts and techniques

This section only introduces related concepts and techniques that we will discuss about in later sections.

1.1. KDD Model

Data mining has another popular term, knowledge discovery from data, or KDD, which shows the emphasis on mining from huge amounts of data. It follows certain process including data preprocessing, establish models or build data patterns based on Data mining algorithms and perform predictions or extract knowledge. Lastly, we need to effectively present the knowledge to users.

Figure 1 below shows the general process that Data Mining involved with.

Data mining algorithms include classification, clustering, semantic annotation, etc. Among these, classification has two-step process in general. First, a classification model is built based on training data. Second, if the model’s accuracy is acceptable, we will apply the model to classify new data.

There are also many new algorithms to improve the accuracy of classification. One common method is ensemble method, which combines a series of individual classifier models and then learns the new data. Adaptive boosting\(^1\)\(^{,8}\) is one new researching area based on ensemble methods. The basic idea is that a series of \(k\) weak classifiers is learned separately, and the weights assigned to each training tuple are updated adaptively to allow the subsequent classifier, \(M_{t+1}\) to give more or less attention to the
training tuples that were previously classified by $M_i$. The final integrated classifier combines the votes to improve the accuracy.

![Diagram of ensemble individual classifiers](image)

**Figure 2** Ensemble individual classifiers

(Adopted from “Data mining: concepts and techniques” by Han et al. 2011)

1.2. Visualization techniques
1.2.1. General Visualization techniques

Visualization turns the abstract data into graphic through computer. Through this way, data is conveyed to users effectively and clearly. It helps people to recognize the hidden data relationships. Hence, it is regarded as one visual interface between users and the data. There are different Data Visualization techniques for variant data. It includes pixel-oriented techniques, geometric projection techniques, icon-based techniques, and hierarchical and graph-based techniques.

Visualization technique involves traditional statically scatter-plot matrices mapping two attributes to 2-D grids, to configurable sophisticated new methods such as tree-maps, which display hierarchical partitioning of the screen.

1.2.2. Tree-maps

Tree-maps are good at handling hierarchical data. One example is to visualize Google news headline stories (Han et al., 2011). Tree-maps categorize all news stories into certain number of groups, each shown in a large rectangle with different color. Within each large rectangle, the news stories are further separated into smaller subcategories, figure as below.
1.2.3. Parallel coordinates

Parallel coordinates is a common way to handle high-dimensional geometry and analyze multivariate data.

It draws $n$ equally spaced axes, one for each dimension, parallel to one of the display axes. This visualization is closely related to time series visualization.
1.3. General Interactive technique
1.3.1. Perception and cognition process

Interactive technique allows user to access to different level information from the data set through managing and developing the data interactively. This transforms data presentation from static visualization to dynamic visualization. Through this approach, user has more chance to adjust or control the visualization process, such as scaling, rotating to fit for the data mining purpose.

When talking about interaction design, it is also important to present information that human can readily perceive, like seeing, hearing and touch. Such as seeing, when we use Icon or color, it’s important to make them easily distinguishable.

When People are dealing with complex data, they tend to have limited capabilities in the processes like thinking, decision making, as well as seeing, remembering, learning (Rogers et al., 2011). However, visualization options presented through graphic user interface can extend human recognition capabilities when people interact with computer. Novice researchers in data mining can benefit from exploring through the highly visualized data environment until they recognize or decide the correct tasks they want to perform for the complex data flow.

These techniques are especially important for data mining, due to the high demand for recognition and memory.

1.3.2. Virtual reality

Virtual reality is one of the new Interactive techniques, associated with interactive, artificiality, highly visual, immersive 3D environments. It creates a virtual environment with multiple perceptions. Through manipulating the data as multi-dimensional object, users can interact with the object in the virtual environment and get more intuitive data understanding and then analyze the data. It provides one of the strongest interactive functions.

As one famous example, Dr. Hans Rosling uses Gapminder bubbles to present global health and economics in CNN Global Public Squares. It’s quite amazing that when he touches the bubbles or even simple gestures, all the 3D data objects and related economic information updates accordingly.

---

Virtual data mining is an important research direction based on this technique.

2 Traditional Data Mining Tools Analysis

Traditional data mining tools might integrate certain visualization kit or simple graphic user interface. However, these tools focus on static analysis of graph and table, often neglecting or simply overlooking the dynamic profiles hidden in the data set without enough involvement of user’s interaction.

In another side, the visualization techniques in the tools are limited, like not allowing many configurations or customization, so that user interactions are restricted. Or, due to the restriction of the R language, such as memory and programmable ability, it tends to use simple and unitary technique, such as pixel-oriented techniques or geometric projection techniques, however, not good for spatial distribution of the data, or not able to providing cooperated multi-view visualization.

Users feel easy to get confused or lost when there are huge amount of data set, metadata, iterative processes or cross verifications. They also feel difficulty in understand the visualized result if interactions are not allowed. The traditional R tool like R-Studio only integrates the command line interface, script editor, and passive visualization of the result. As a novice user and beginner in researching data mining algorithm, I feel very difficult to learn and develop my own methods on some complex data mining algorithm, like previously mentioned Adaptive boosting. So that’s why I come up the idea to build a new interactive data mining Tool.

3 Features analysis for the new Interactive data mining tool

Human interaction can be integrated in different phases of previous mentioned KDD process, which empower user’s perception of information when visually exploring the data set, data patterns and rules, and assist users in thinking and decision making.

Below figure is the updated KDD process based on above idea.
Visual exploration through different visualization Techniques that assists perception and cognition can help to view the information from different perspectives to avoid overlooks. Users can look in to more details and more importantly recognize the key points through interactively changing different viewpoints, so that a fresh idea popping up as early as possible might save a huge amount of time in the whole data mining phase.

- Virtual data objects for highly interactive

Visualization of the information can be broken down to virtual data objects in the software, so that they are controllable in the way like virtual reality through specific devices allowing different senses or as simple as IPad for user to touch. Such as simulating the decision tree as real tree, user touches one interesting leaf, related possibility, accuracy or even related sampling data tuples will come up flowingly. It can zoom in and out with your finger gesture. This helps users exploring the rationale behind the data visualization. This is especially useful when you have been thinking several hours sitting before the computer. Now you can switch to play the data mining like a game but still working on it.

When the virtual data mining concept is put in to software, we can migrate easily to the simple version in App software like on android or Mac, which syncs and switch all the data from or to the workspace of main workstation smoothly as needed, so that users can easily enjoy exploring the visualized data set through different view with more perceptions.

- Incremental data mining

We can implement incremental data mining through saving primary data sets including the result from data preprocessing, sampling results including training set and test set, previously generated model including decision trees, rules, previously
used filters, and predictions, etc. We can also deduce to some templates for certain approach. So that when we change the data sets, we can reuse the approach by only modifying the templates. The GUI should support user to configure the templates with clues of frequent choices, such that the previous user experiences can be repeated.

In our learning phase, we can practice this algorithm as fully supervised boosting. We can manually adjust the weights from previous classifier result like assign the highest weights to the weakest classifier and then check the effect to the ensemble.

Below is the picture for above idea, it can be used to develop certain Ensemble technique to classify a complex problem, like adaptive boosting.

![Diagram of Ensemble Method]

**Figure 6** Ensemble Method
Cooperated multi-view and configurable Visualization

Components can be dragged in to one multi-view canvas. It will support interactive operation, such as component highlight, overview and detail, panning and zooming, drag and drop. Task component can be further broken down to elements, as showing in above figure for adaptive boosting.

The new visualization technique like Hierarchical Visualization and Parallel Coordinate Technique need to be embedded in our tool as separate components.

Tree-map has four distribution algorithms for different data attributes. They have their own advantages and weakness. Such as square distribution can achieve best visualization effect for rectangular size order. Slice-and-dice and strip distribution are fit for multi-variables. When geometric information is concerned, spatially-ordered distribution are preferred. The hierarchical structure and depth of Tree-map can be interactively changed. The size, color, and distribution pattern can be interactively configured based on dimension characters and requirement from analyzing. For the hierarchal structure, brushing and linking should be supported for the node choice on different level.

We can visualize related parameters and make them configurable. The visualization component will communicate to the Multi-View controller so that different views can be coordinated.

For data mining component, the separate graphic result like Scatter diagram, Box plot, Decision tree from each single R step can be combined in the component view.

Comparing similar data set by putting them together gives users an intuitive view and then analysis. Five-number summary let user choose different color, size or icon to differentiate the different result. This technique is useful especially for preprocessing phase. You can have an intuitive for the data to be data mining.
4 Implementation Analysis for the Interactive data mining tool

This section will examine the possible solutions and difficulties to implement above ideas for the new tool.

The general framework is to build up a Java platform with script language Python which further integrates R toolset. Python will deal with the Data Mining workflow including data preprocessing, virtual data objects handling, templates operation, the interactions with R. Detailed Data Mining algorithms are still executed through R code by reusing the existing packages. But we still need to further evaluate the efficiency of executing R in python through high complex data sets and consider the detailed implementation techniques to improve.

JAVA provides basic graphic user interface including the canvas and visualization components. And also it is easy to be deployed among different systems. Through java software, the mining results or data sets can be easily demonstrated or explored from synchronized APP version. However, we may meet some difficulties in the component of visualization, such as adaptive Tree-map and co-operated multi-view implementation. We still need more explorations in these areas.
As for Python, it is known for gluing other language, which combines different functions in an integrated platform. Rpy2 package in Python allows R language to be used in Python easily. Python script is also good at building data mining template or executes certain routings performing data mining tasks. The core Data mining tasks can still be finished by R because of its abundance of Data mining packages. From another side, Python has higher efficiency than R, and also Python can build a stronger command based interface, which allows interaction and can effectively control the data mining process (See references [3][4]). So some data preprocessing and time consuming tasks can be executed by Python, whereas, in R-Studio, it’s often hard to interrupt the execution of the long looping R code.

R program will return the data object to Python interface. Python script can store the result, visualize it or feed to next work flow. Frequent work flows with certain patterns such as classification, regression, can be saved as templates. So for next project, it’s easy to reuse previous work effort.

From above analysis, we can see this new tool idea is feasible, although expecting many challenges.

Reference

[1] Jiawei Han, Micheline Kamber, Jian Pei. Data mining: concepts and techniques. 3rd ed. 2011.


