TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 06/11 pp429-439
Volume 17, Number 4, August 2012

Evaluation of a Pointwise Local Visual Pattern Exploration Method*

Zhenyu Guo**, Matthew O. Ward, Elke A. Rundensteiner, Carolina Ruiz

Department of Computer Science, Worcester Polytechnic Institute, Worcester, MA 01609, USA

Abstract: Sensitivity analysis is a powerful method for discovering the significant factors that contribute to understanding the interaction between variables in multivariate datasets. A number of sensitivity analysis methods fall into the class of local analysis, in which the sensitivity is defined as the partial derivatives of a target variable with respect to a group of independent variables. In a recent paper, we presented a novel pointwise local pattern exploration system for visual sensitivity analysis. Using this system, analysts are able to explore local patterns and the sensitivity at individual data points, which reveals the relationships between a focal point and its neighbors. In this paper we present several evaluations of the system, including case studies with real datasets, user studies on the effectiveness of the visualizations and interactions, and a detailed description of the experience of a user.

Key words: knowledge discovery; sensitivity analysis; local pattern visualization; evaluation

Introduction

Sensitivity analysis[1] is the study of the variation of the output of a model as the input of the model changes. When we study the correlation between a target (response) variable Y and a set of independent variables $\{X_1, X_2, \cdots, X_n\}$, sensitivity analysis tells analysts the change rate of Y as X_i varies. Analysts can also discover which input parameters are significant for influencing the output variable. Sensitivity analysis has been widely applied for understanding multivariate behavior and model construction in analyzing quantitative relationships among variables[2]. For example, it can be applied to car engine designs; fuel consumption is dependent on the relationships among the design choices, such as fuel injection timing, as well as operation-varied conditions, such as engine speed[3]. The analysis results are important in helping engineers tune the parameters in designing an engine.

Sensitivity analysis is essential for decision making and system understanding, as well as model construction. Numerous approaches have been proposed to calculate the sensitivity coefficients. In this paper, we focus on differential analysis[4], where sensitivity is defined as the partial derivative of a target variable with respect to a set of independent variables. Because the sensitivity using partial derivatives is extracted in a small neighborhood of the data, it is usually called local analysis. Generally, any information extracted around a single focal point can be viewed as a local pattern, such as neighbor count, distances to neighbors, and partial derivatives. Local analysis is performed using the extracted local patterns, and sensitivity information is one important type of local pattern.

Although many visual analytics systems for sensitivity analysis follow this local analysis method, there are few that allow analysts to explore the local pattern in a pointwise manner, i.e., the relationship between a focal point and its neighbors is not visually conveyed. This pointwise exploration is helpful when a user wants to understand the relationship between the focal point and its neighbors, such as the distance

Received: 2012-05-01; revised: 2012-07-05

∗Supported by the National Science Foundation of USA (No. IIS-0812027)

∗∗To whom correspondence should be addressed.
E-mail: zyguo@cs.wpi.edu; Tel: 1-774-670-7250

and direction. The analysis result can assist analysts in understanding which neighbors do not conform to the local pattern. This discovery can be used to detect local anomalies and find potentially interesting neighbors.

We have been developing a system focused on pointwise analysis of sensitivity. In a recent paper[5], we presented a prototype system with the following attributes:

- **A novel pointwise exploration environment** It supports users in browsing a multivariate dataset from a pointwise perspective.
- **A novel visualization approach for sensitivity analysis** The designed local pattern exploration view indicates the relationships between the focal point and its neighbors, and whether the relationship conforms to the local pattern or not.
- **Adjustable sensitivity** We allow users to interactively adjust the sensitivity coefficients, which gives users flexibility to customize their local patterns based on their domain knowledge and goals.

In this paper, after a brief overview of our system, we focus on several ways that we have evaluated the technology, including:

- **Feature evaluation using a real-world dataset** We evaluated the effectiveness and utility of our system's features using a database of diamonds and their attributes.
- **User studies on visualization design options** We performed a formal user study to evaluate the designed local pattern display. Two aspects are examined: the layout and the glyph type.
- **Detailed usage case** We analyzed the behavior of a user of the system and solicited feedback as the subject performed real tasks on the diamond dataset.

1 Related Work

Sensitivity analysis has been studied in the scope of multivariate data analysis[6]. Sensitivity analysis is the analysis of the variation of the output from a model based on small changes in their inputs. A variety of approaches have been proposed in recent years. A number of methods fall into the class of local analysis, such as adjoint analysis[7] and automated differentiation[8], where the sensitivity parameters are found by simply taking the derivatives of the output with respect to the input. Because this is usually done in a small neighborhood of the data, they are usually called local methods. Our approach is based upon partial derivatives calculated using numerical differentiation. There are many ways to calculate partial derivatives[4,9]. We obtain the partial derivatives using the local linear regression model coefficients.

Subgroup pattern mining is a very popular and simple form of knowledge extraction and representation[10]. An advanced subgroup mining system called "SubgroupMiner"[11] allows the analyst to discover spatial subgroups of interest and visualize the mining results in a Geographic Information System (GIS). It has been shown that subgroup discovery methods benefit from the utilization of user background knowledge[12]. In this paper, we assume each group of local neighbors is a subgroup, and thus the anomalous local patterns can be discovered using subgroup pattern mining techniques. Our system allows users to detect interesting local patterns and compare the local pattern with the global one both visually and statistically.

In recent years, many visual analytics approaches have been proposed that allow analysts to visually perform sensitivity analysis. Barlowe et al.[13] proposed a system called Multivariate Visual Explanation (MVE). This system allows users to interactively discover correlations among multiple variables and use histograms to visualize the partial derivatives of the dependent variable. The histograms reveal the correlations, positive or negative, between the output and the coefficients. Correa et al.[14] presented a framework to support uncertainty in the visual analytics process through statistical methods such as uncertainty modeling, propagation, and aggregation. It has been shown that the proposed framework leads to better visualizations that improve the decision-making process and help analysts gain insight into the analytical process itself. Chan et al.[15] proposed a flow-based Scatterplot system, which extended two-dimensional (2-D) scatterplots using sensitivity coefficients to highlight local variation of one variable with respect to another. In their system, a number of operations, based on flow-field analysis, are supported to help users navigate, select, and cluster points in an efficient manner. In this paper, we also propose a visual solution for sensitivity analysis. However, inspired by the street view in Google maps, we allow users to explore the correlations among variables from a new perspective, i.e., pointwise examination of relationships among variables, and the relations between the focal point and its neighbors. The main difference between our work and previous work is that the local information about each data point is visually conveyed.

2 Local Pattern Extraction for Sensitivity Analysis

2.1 Neighbor definition

For each data point, the local pattern is extracted based on its vicinity. We compute the neighborhood of a point as a region around that point. The shape of its neighborhood could be sphere-shaped or box-shaped. For a sphere-shaped area, a radius is specified by the user and all the data points whose distances (usually the Euclidean distance after normalization) to the focal point are less than the specified radius are considered that point's neighbors. For a box-shaped area, the user can specify the box size on each dimension.

Our system allows users to perform this neighborhood definition in a parallel coordinate view by dragging and resizing a box-shaped region. The neighbors are all the data points in the hyper-box, taking the focal point as the box center.

2.2 Calculating the sensitivities

There are many ways to compute the sensitivity of one dependent variable with respect to an independent variable. We follow a variational approach, where the sensitivity can be calculated by the partial derivative of one variable with respect to another. The derivative of a target variable, y, as the independent variable, x, changes is approximated as $\partial y/\partial x$. The relationship is geometrically interpreted as a local slope of the function of $y(x)$. Since we do not know the closed form of the function $y(x)$ between variables in the general case, we approximate the partial derivatives using linear regression. The regression analysis is performed in neighborhoods around each point. A tangent hyperplane for each focus point is calculated based on its neighbors using linear regression. This linear function represents how the independent variables influence the target variable, assuming a constant local changing rate for all independent variables. The representation enables the model to predict the target value given the independent variables, as well as to assess the error between the predicted value and the observed value.

2.3 Local pattern extraction

Generally speaking, any local information that can assist analysts in performing local pattern analysis can be extracted and visually represented for examination, such as the neighbor count, distances to neighbors, and orientation to neighbors. In our work, we focus on the orientation from the focus point to the neighbors. We chose this pattern for two reasons. First, this pattern tells users the relationship between the focus point and its neighbors, i.e., the direction to move from the focus point to its neighbors. Second, and more importantly, since our system is designed for sensitivity analysis and we extract a linear regression model, this direction reveals whether the relationship conforms with the local trend. Knowing this can assist analysts in performing sensitivity analysis in this neighborhood region. Here "conforms with the local trend" means the vector between the focal point and a neighbor is approximately parallel to the local trend.

Due to the unit differences, the extracted local linear trend may be dominated by some attributes. For example, a linear pattern of $y = 10\ 000x + 5$ is dominated by y. In this case, all the connecting lines between the focal point to other neighbors are nearly parallel with each other (when reduced to one dimension y). To remove the unit differences, we assign a weight using the regression coefficient for each independent attribute, so that the changing rates are the same between each independent variable and the target variable. This step can be considered a normalization. After the normalization, the slopes of the linear trend are all $\pi/4$ in all dimensions, and the angle θ is between 0 and π.

To sum up, in our system, the extracted local pattern for a single point is a list of values V, in which each value is an angle between the focal point and a neighbor, relative to the local regression line. The size of V is the same as the neighbor count.

2.4 Anomaly detection

Our system allows users to detect anomalous local patterns that deviate from others. In general, we follow the idea of subgroup discovery to identify interesting subgroups from the dataset.

Since each local pattern is extracted from a small subset, we can take each local pattern as a subgroup. Thus subgroup discovery can be applied to discover the local patterns of certain special significance, such as the ones different from the others, i.e. anomalies. The word "anomalous" implies that there is something basic to which each subgroup can be compared, i.e., there is some notion of "background" or "expected" pattern. For example, the angles from the trend normal to the neighbors mentioned before are expected to be π/2. We know this is because the analysts have knowledge of regression analysis. In general, however, users may not have this prior knowledge.

As a general solution, we assume each subgroup is one extracted sample (a subset of individuals). All the samples could be integrated as a population to simulate the underlying model that generates the individuals. We use the term “global pattern” to represent the integrated pattern. Each local pattern is compared with this global one to decide whether it is different from it.

As a statistical method, the significance value of each local pattern is evaluated by a quality function. The value of this function is an outlier factor showing how likely it is that this local pattern is an anomaly. As a standard quality function, the binomial test is used to examine if the sample is significantly different from the rest of the population[11]. The z-score is calculated as

$$\frac{\mu - \mu_0}{\sigma_0}\sqrt{n}\sqrt{\frac{N}{N-n}}.$$

μ is the mean of the sample, μ_0 is the population mean, and σ_0 is the standard deviation of the population. N and n are data sizes of the population and the sample, respectively.

3 System Overview

In this section, we introduce the proposed local pattern exploration method and our system design. Several coordinated views assist users in exploring the local patterns. Three are described in this section (see Ref. [5] for a complete description).

3.1 Global space exploration

The global view is designed to give users a global sense of the whole dataset. Basically, any multivariate data visualization technique, such as scatterplot matrices, parallel coordinates, pixel oriented techniques[16], or glyphs, can be used to display and explore the data globally. Of these methods, only glyphs show each data point individually and completely as an entity without overlapping. We use a star glyph[17] because the analyst can easily specify which individual data point he/she wants to examine, thus leading to an easy exploration of the local patterns around that data item. A star glyph for a multivariate data point with D dimensions consists of D equally spaced rays, each proportional in length to the value of the corresponding dimension, emanating from a center point. The endpoints of the rays are connected by a closed polygon.

To assist analysts in discovering anomalous local patterns, i.e., a subgroup of neighbor data points that are different from the global pattern, we encode the statistical results using color. As shown in Fig. 1, gray color means there is no significant difference between the sample and the population (p-value is larger than 0.05), suggesting the local pattern is not an anomaly. Red and blue colors mean that a significant difference is detected (p-value is less than 0.05). Red means the z-score is less than zero (the critical value is -1.96 for a 0.05 level), which means the local pattern has a significantly lower mean value than that of the global pattern. Similarly, blue means the z-score is larger than zero (the critical value is 1.96 for 0.05 level), indicating a higher mean value compared to the global pattern.

3.2 Local pattern examination

During the interactive exploration in the global view, when the user moves the cursor onto a data item, another view displaying all its neighbors and the selected point are drawn. This new view is called the local pattern view. The main purpose for this view is to illustrate the relationships between the focal point and all its neighbors.

The focal point is shown in the center of the display using a star glyph. The two cross lines (vertical and horizontal) create four quadrants with the focal point as the origin. As a layout strategy, we map the difference in target values between a neighbor and the focal point as Y, meaning for each neighbor, if its target value is higher than the focal point’s target value, it is located in the upper half. Contrariwise, if the target value is lower than the focal point, it is located in the lower half. The higher the absolute difference is, the further away the neighbor is placed.

As discussed before, the local pattern we use is the orientation angle θ. The angle is mapped to X in this view. The angle of the focal point is π/2, assuming the direction conforms with the local trend. When the angle between a connecting vector and the normal vector of the local trend is less than π/2, the corresponding neighbor is placed in the left half of the view. If θ is smaller (larger) than π/2 it means the neighbor’s target value Y is smaller (larger) than the estimate. Notice that for this view, X and Y positions have different meanings. For Y values, higher means the neighbor’s target value is higher than the focal point’s target value (compared to the focal point); for X values, a neighbor

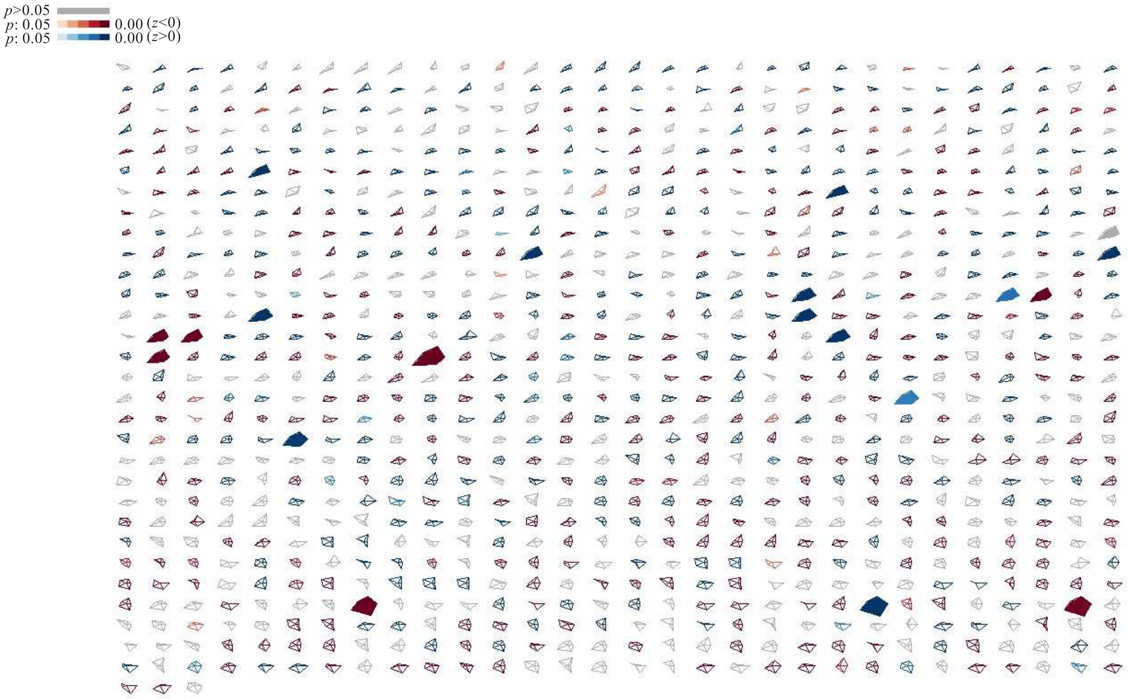


Fig. 1 The global display using star glyphs (903 records from the diamond dataset). The color represents whether the data item is an anomalous local pattern or not. The red records and the blue records are two kinds of anomalous local pattern. The filled star glyphs are selected local pattern neighbors.

in the right means this neighbor's target value is higher than its (this neighbor's) estimate value based on the local trend (compared to itself).

For each neighbor, we support two display methods. The first one is the original value display, which means that for each neighbor, the attribute values in the original dataset are shown. In this case, we again use the star glyphs to represent each neighbor, so that users can connect this view with the global view. The second display method is a comparative display, in which the focal point is the base line, represented as m dashes, where m is the number of attributes. For each neighbor, there are m bars corresponding to its m attributes, where an upward (downward) bar for an attribute indicates that the neighbor's value in that dimension is higher (lower) than that of the focal point. This piece of information is also redundantly represented using colors: blue means higher and red means lower. The larger the difference is, the darker the color is.

3.3 Adjusting the local pattern

The local partial derivative values reflect how the independent variables influence the target variable in the local area. However, the derivative values may not necessarily meet the user's requirements when they want to find interesting neighbors. Thus we allow users to interactively adjust the weight of each variable. The local pattern adjusting view uses parallel coordinates to convey the weightings. The partial derivatives of the local pattern are drawn as a poly-line. The last dimension is the constant (intercept) of the linear trend. The user can interactively change the coefficient values, i.e., the slope of the trend line, by dragging the poly-line on each axis. During the adjustment, the local pattern view is also dynamically changed to reflect the new relationships among the focal point and its neighbors, using the new coefficients.

4 Case Studies

In this section, we discuss some case studies used to evaluate our approach and show the effectiveness of our system. The dataset is a diamond dataset obtained from an online jewelry store[18]. Each data item is one diamond. The target attribute is price. There are 4 different independent attributes that influence the price of a diamond: weight (carat), color, clarity, and cut. The goal is to assist customers in choosing a diamond. The discovery can also tell the retailer whether the price of a certain diamond is set appropriately. We use a subset of the diamonds with a certain price range ($5000 - $8000), since we assume that customers have a budget range for shopping, rather than caring about the whole dataset. The whole dataset has 13 298 data items and the subset has 903 data items.

For easier understanding, we start from a single independent attribute weight. The user begins by defining an appropriate neighborhood: two diamonds are neighbors when they have similar weight and price, as well as when they are of the same color, clarity, and cut. The extracted local pattern is the orientations to the neighbors. Figure 1 shows the global star glyph display. The color indicates whether the diamond is an anomalous one.

To understand the normal and abnormal data items in detail, we show three local pattern views for gray, red, and blue data points. Figure 2 shows the local pattern view of a gray data point. All the neighbors of this data point are in the center of the view (x position), indicating that the directions to the neighbors are all about $\pi/2$. This means that all the data points in the local area fit the regression hyperplane, which is very common in the dataset. To assist the analyst in performing the sensitivity analysis, i.e., the change rate of the target as an independent attribute value varies, we show the local regression model in the bottom bar. It is shown that in this local area, as the weight increases, the price increases, which means a positive influencing factor.

Figure 3 shows the local pattern view for a diamond that is blue in Fig. 1, suggesting that it is an anomaly and the test result shows the mean of this local pattern is significantly higher than the global pattern. The user can see that all the neighbors are in the right half of the view. This means for each neighbor, the direction is larger than $\pi/2$. In particular, the local sensitivity shows that for every 0.01 carat increase, the price increases \$118. However, the price of the local neighbors are higher than estimated considering this changing trend. The ones above the focus have a greater rate, but the price increases more than the local trend. The ones below have lower weight, but the price decrease is less than expected. Thus for blue diamonds, generally most of the neighbors are in the right half side of the view, which means they are worse deals compared with the focal point. Thus, the blue diamonds should be preferable to the customers.

Finally, we give an example of a diamond mapped to red in Fig. 1. Similar to the discussion for blue diamonds, a red diamond means there are many better deals compared with this one. Figure 4 shows the local pattern view of a red diamond. It is shown that locally for every 0.01 carat increases, the price increases \$332. The neighbors to the left of the focus are better deals. Some have a higher weight with the same price, while others have a lower price for the same weight.

5 User Studies

In this section, we discuss a user study for evaluating the effectiveness of the visual representations of the local pattern. We focused on two visual factors: different types of glyph representations and different layout strategies. To remove the interaction effects among the

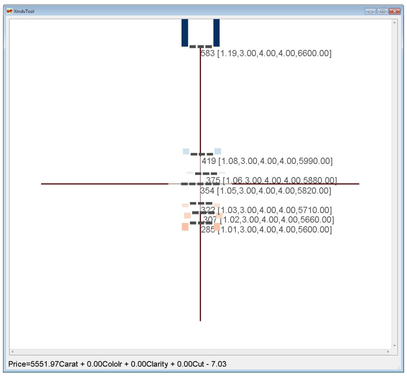


Fig. 2 The local pattern view of a gray data item. The orientations from the focal point to all its neighbors are π/2, which is common in the dataset.

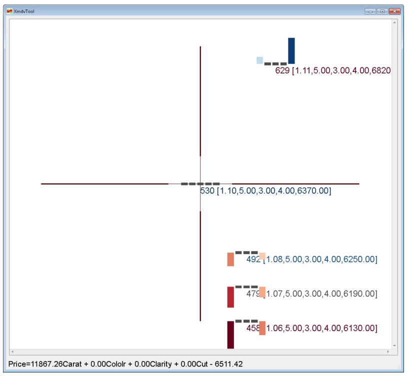


Fig. 3 The local pattern view of a blue data item. The orientations from the focal point to most of its neighbors are larger than π/2, which means the neighbors' target values are higher than estimated. In other words, the focal point is a "good deal".

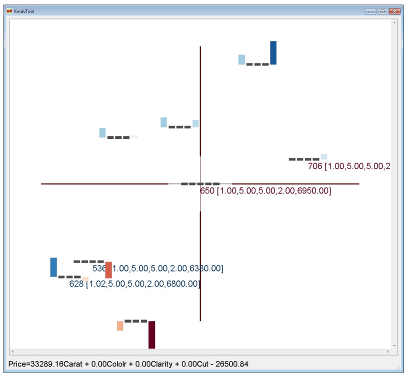


Fig. 4 The local pattern view of a red data item. The orientations from the focal point to most of its neighbors are lower than π/2, which means the neighbors' target values are lower than estimated. In other words, the focal point is a "bad deal".

two factors, we evaluate the two factors independently.

For the glyph type, our goal was to examine the effectiveness of the comparative display, i.e., using upward and downward bars to represent the relationship between the focal point and its neighbors. To compare with other methods, we implemented two other types of commonly used glyph representations: profile glyphs (Fig. 5) and star glyphs (Fig. 6). To make the comparison fair, we also integrated the same color strategy into these two glyph types. Our hypothesis was that the comparative glyph method better reveals the relationships between the selected focal point and its neighbors. A sample question was "Compared to the focal diamond, how many neighbors have both lower color and lower clarity?"

For the layout strategy, our goal was to examine the effectiveness of the local pattern view layout, namely, placing the selected focal point in the center and placing the neighbors in the four quadrants according to the interestingness (such as diamond price). For comparison, we implemented a scatterplot display which maps the attribute values to the x and y locations. The focal point was differentiated by both size and color. Our hypothesis was that the centered layout can better help analysts locate interesting neighbors. A sample question was "How many more dollars are needed to buy a diamond with both higher color and higher clarity?" The dataset we used is the same as the dataset mentioned in the case study which had 4 independent attributes and 1 target.

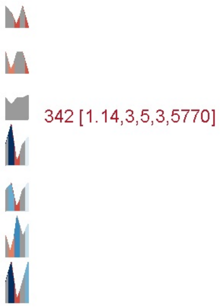


Fig. 5 The profile glyph display.

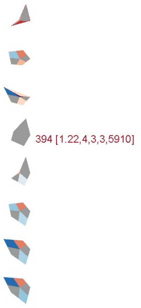


Fig. 6 The star glyph display.

We invited students to be the subjects (21 in total) in the user study. The subjects were asked to answer 8 questions about local pattern based on visual representations. In this user study, we didn't ask the subjects to use our system because the main goal was to evaluate the local pattern design method. In Section 6, we describe in detail how a user explored a dataset using our system. The subjects answered the questions based on screen-copied figures printed out on paper. Note that any single question could be answered based on different visual representation methods of the same local pattern, such as different glyph types or different layout strategies. Subjects were randomly assigned a visual representation method to answer a given question. Take the evaluation of the layout strategy for example. We designed two questions (question Q_a and question Q_b) to compare the two layout methods. We generated two groups of questions, group G_A and group G_B, as follows. Each question group had both questions Q_a and Q_b. In group G_A, question Q_a would be answered based on the designed local pattern layout strategy, while question Q_b would be answered based on the scatterplot layout. In group G_B, the questions are the same, but we exchanged the layout strategies: question Q_a was represented using the scatterplot and question Q_b was represented using our local pattern layout method. In the study, we randomly assigned half of the subjects to question group G_A and the other half to question group G_B. Similarly, we generated three groups of questions to evaluate the glyph types because there are three different glyph representations.

Before the study, the subjects signed a consent form. Then each subject was shown a brief explanation of the study using examples and sample questions, such as which dataset we used and how to read the figures. The subjects finished the study by answering several questions. We recorded the time each subject spent on each question for further analysis.

Figure 7 uses error bars with a 0.95 confidence interval to show the accuracy for the three glyph types. We found that the comparative glyph and the profile glyph were very similar in terms of accuracy. It is clear that both the comparative glyph and the profile glyph are much better than the star glyph: the p-values are 0.017 and 0.023, respectively.

We also examined the time spent for each glyph

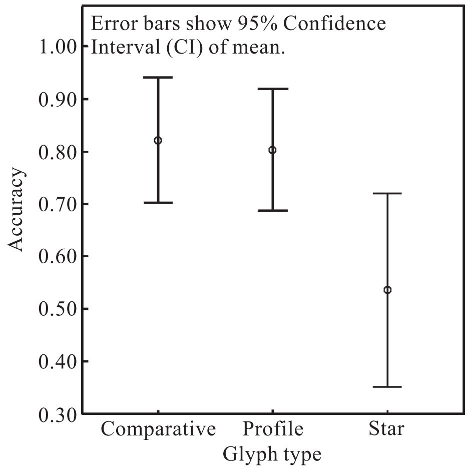


Fig. 7 The comparison of accuracy for different glyph types.

type and the results are shown in Fig. 8. Similarly, the comparative glyph and profile glyph are better than the star glyph. The difference between comparative glyph and star glyph is significant (p-value=0.026). Although there is no significant difference between comparative and profile glyphs (p-value=0.232), the time subjects spent on the comparative glyph was much lower than for the profile glyph. To conclude, we found comparative glyphs and profile glyphs were better than the star glyphs for both accuracy and time. The accuracy for comparative glyphs and profile glyphs are very similar, but they spent more time on profile glyphs.

Lastly, we compared the two layout strategies. In terms of accuracy, the two strategies are almost the same (nearly 80%). However, in terms of task completion time, we noticed that the subjects spent much more time when using the scatterplot layout. The average time for the centered layout was 62 s, while for the scatterplot layout it was 87 s. This is a statistically significant difference (p-value=0.038). We also noticed that the time variance of the centered layout is large. We believe this is because of different learning rates for this new layout method. Some subjects seemed to learn and get used to this layout very quickly, while others had difficulties and spent more time getting used to it. In a future evaluation, we will try to confirm this difference in learning rates and repeat the tests with trained subjects.

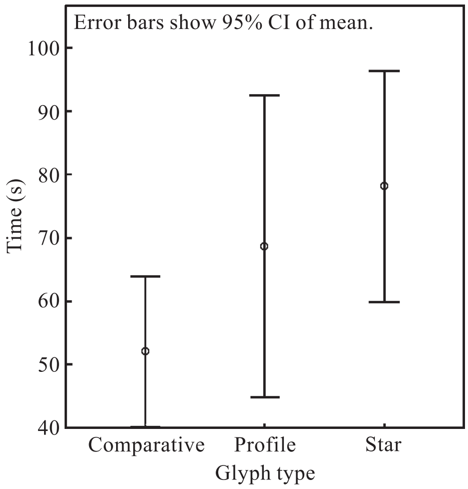


Fig. 8 The comparison of time for different glyph types.

6 Usage Session

We now demonstrate how our visual exploration method could be used for solving real life problems. Our usage session was again based on the diamond dataset. We invited a user who was trying to make a decision on buying a diamond to test our system.

Before using our system, he first browsed some online diamond selling websites on the internet to get familiar with the diamond purchasing task. There were two reasons for this activity prior to using our system. The first was to help him understand which attributes are important to him, i.e., to develop a personal preference. The second reason was that he could determine the minimum requirements and the price range he'd like to choose from. He told us his preferred price range was roughly between $6000 and $7000. In terms of the importance of different attributes, he thought weight (size) was the most important one. The second attribute important to him was color. The other two attributes, clarity and cut, were not very important to him. He said this was because he thought the latter two attributes were not as noticeable as weight and color for him. He also indicated minimum requirements on these attributes: weight needed to be at least 1.1; color needed to be at least H (the required value was 4 where the best color value is 8); clarity needed to be at least SI1 (the required value was 3 where the best clarity value is 8); he did not have any requirements on the attribute cut.

With these requirements and preferences, he started using our visual exploration system to perform the task. The first step was to define the local neighborhood range. After being given some explanation on this step, he decided to define two diamonds as neighbors when they have similar weight (within 0.15), color (plus or minus 1) and price (within $500). He did not care about the other two attributes, clarity and cut, so he decided to remove their influence at this step.

He then explored in the data in the global view (the star glyph display) by hovering the cursor over the glyphs (data items). The data attributes are shown when the cursor is on that data item. The glyphs are ordered based on price, so he roughly picked some interesting candidates within his preferred price range. He had two criteria when choosing the candidates. For the first one, since he considered weight the most important attribute, he picked several heavy (large) diamonds. The second criterion was to focus more on the blue data items. This is because we told him that generally glyphs colored blue are usually better deals. After this initial rough selection, he chose three candidates as shown in Table 1. These three diamonds are all blue, i.e., they are better than most of their neighbors.

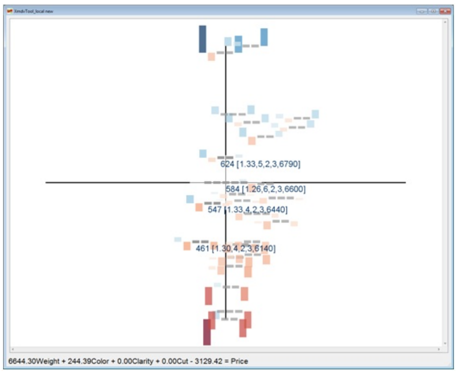
Then he decided to refine his selection by examining each candidate in the local pattern view. He opened the local pattern view and compared the pre-selected candidates with their similar local neighbors. The three local pattern views of these candidates are shown in Figs. 9-11. The attributes are in the same order as introduced in Section 4: the first attribute is weight and the last attribute is price. He wanted to find more interesting candidates on the left hand side in this view.

When he viewed the local neighbors of diamond 584, he noticed that diamond 624 was also a good choice because its weight is higher. Although the price is a little higher, since it is on the left hand side, it may still be worth buying. The second neighbor he was interested in was diamond 547. This diamond has the same weight as diamond 624, but it is much cheaper. Another interesting neighbor was diamond 461, whose weight is higher than candidate 584, but much cheaper. All three interesting neighbors are on the left hand side, indicating they are worth buying compared to the candidate diamond 584. Therefore, at this point, he removed diamond 584 from the candidate list and added the three newly found diamonds onto the list.

Then he opened the local pattern view for diamond 567. He noticed that the neighbor diamond 400 was a much better choice. The weight and color are both better than those of the previous chosen diamond 567, yet with a lower price. So he removed 567 from the candidate list and added diamond 400 to it. He didn't find any interesting neighbors for candidate diamond 544.

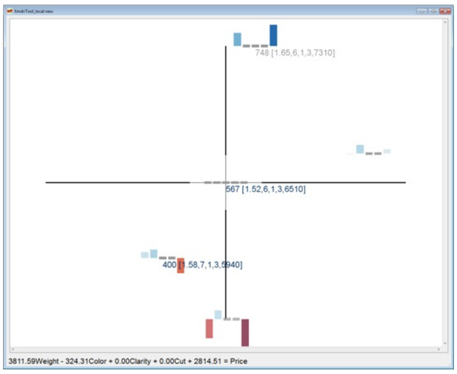
Table 1 Candidate diamonds after a rough exploration in the global star glyph view.

ID	Weight (carat)	Color	Clarity	Cut	Price ($)
584	1.26	6	2	3	6600
567	1.52	6	1	3	6510
544	1.51	2	3	2	6420



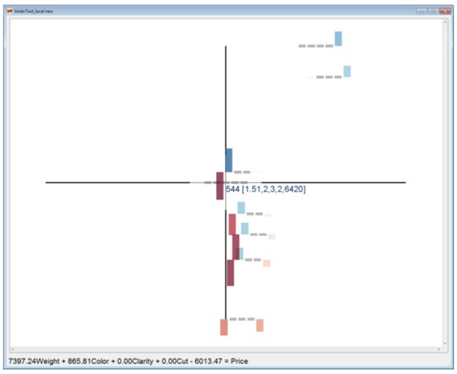
6644.30Weight + 244.39Color + 0.00Clarity + 0.00Cut - 3129.42 = Price

Fig. 9 The local pattern view of diamond 584.



3811.59Weight - 324.31Color + 0.00Clarity + 0.00Cut + 2814.51 = Price

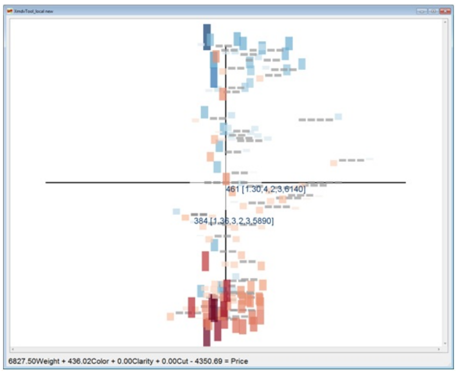
Fig. 10 The local pattern view of diamond 567.



7397.24Weight + 865.81Color + 0.00Clarity + 0.00Cut - 6013.47 = Price

Fig. 11 The local pattern view of diamond 544.

Next, he wanted to view the local patterns of the newly added candidates to further enlarge the candidate list with more choices. He didn't find any better choices for diamonds 624, 547, and 400. When he viewed the local pattern of diamond 461 (Fig. 12), he found an interesting neighbor, diamond 384. Because its weight is higher and its price is lower, he added it to the list. The



6827.50Weight + 436.02Color + 0.00Clarity + 0.00Cut - 4350.69 = Price

Fig. 12 The local pattern view of diamond 461.

candidate list at this point is shown in Table 2. Notice that after refinement, he had found several additional interesting candidates and only one pre-chosen diamond survived after examining the neighbors.

He then made a final decision among candidates on this list. He first removed diamonds 544 and 384 because their color, an important attribute, did not satisfy his minimum requirement. He then removed diamond 400 because its clarity was lower than that of the rest. After this, he noticed that all the candidates' clarity were lower than his initial requirement. Since he cared about weight and color much more, he decided to make a compromise on clarity, i.e., reduce the minimum requirement from SI1 (value 3) to SI2 (value 2). Now he narrowed his choices down to three similar diamonds: ID 461, 624, and 547. He decided to remove diamond 624 because the price was high compared to the other two. After a careful comparison between diamonds 461 and 547, he finally decided to purchase diamond 461. This is because diamond 461's weight is only slightly smaller than diamond 547, which is probably not noticeable, but the price is $300 cheaper.

After the study, he said that overall this system was very helpful. The local pattern view helped him compare similar data items, find more interesting candidates, and guide him to make a more comprehensive decision. He mentioned that the system was easy to use and helped him finish the task very quickly.

Table 2 Candidate diamonds after examining each local pattern of the pre-selected diamonds.

ID	Weight (carat)	Color	Clarity	Cut	Price ($)
384	1.36	3	2	3	5890
461	1.30	4	2	3	6140
624	1.33	5	2	3	6790
400	1.58	7	1	3	5940
544	1.51	2	3	2	6420
547	1.33	4	2	3	6440

We asked him whether he had some suggestions for improving our system. He pointed out some limitations and gave us some useful suggestions. He said the neighbor definition in the parallel coordinate view is somewhat confusing and he had difficulty understanding it. He said sometimes given a candidate, he only wanted to examine the neighbors with higher weight or color. He suggested we could add a function so that the user can dynamically change the neighbor definition and give him greater flexibility in defining neighbors not only centered in the focal diamond, but also can take the focal diamond's value as maximum or minimum, such as only cheaper neighbors.

Another suggestion was a sorting functionality. He said he might want to sort the star glyphs in the global view during exploration. This functionality is not currently supported but would not be difficult to add. A filtering functionality was also mentioned. He told us that a range query filter would be useful. It could be used to hide the less interesting diamonds which don't satisfy the minimum requirement. This functionality could be effective, especially in the case when a large number of local neighbors exist. The last comment was to have a comparative view for the selected candidates. The view could provide him an overall comparison, where he could select any of the candidates as the focal diamond.

7 Conclusions

This paper presents an overview and multi-pronged evaluation of a novel pointwise visualization and exploration technique for visual multivariate analysis. Generally, any local pattern extracted using the neighborhood around a focal point can be explored in a pointwise manner using our system. In particular, we focus on model construction and sensitivity analysis, where each local pattern is extracted based on a regression model and the relationships between the focal point and its neighbors. Using this system, analysts can explore the sensitivity information at individual data points. The layout strategy of local patterns can reveal which neighbors are of potential interest. Therefore, our system can be used as a recommendation system. We discussed case studies with a real dataset to investigate the effectiveness and usefulness of our approach, performed comparative evaluations to confirm our glyph design and layout decisions, and described the experience of a user performing a real task with the system.

Based on our evaluations, future work we are actively pursuing includes:

- **Supporting other types of local patterns** We plan to expand our system to support more types of local patterns, such as distances to neighbors and errors of the neighbors, in terms of the extracted local model.
- **Customize the local pattern view** When multiple types of local patterns are extracted, users should be able to specify how to visually map the values using color, size, and position.
- **Interactions** Users should be able to interactively discover interesting local patterns using brushing techniques in the local pattern view. This can save display space in the global display view since only interesting local patterns would be shown.
- **Performance and scalability** We plan to incorporate efficient neighbor finding techniques in our system in the future, such as binning the space or using k-d tree search[19]. For datasets with a large number of attributes, a user-driven attribute selection technique or a dimension reduction technique based on influencing factors could be applied[20].

References

[1] Cruz J B. System Sensitivity Analysis. Stroudsburg, PA, USA: Dowden, Hutchinson and Ross, 1973.

[2] Saltelli A, Ratto M, Andres T, et al. Global Sensitivity Analysis: The Primer. New York, NY, USA: John Wiley and Sons, 2008.

[3] Jelovič M, Jurić J, Konyha Z, et al. Interactive visual analysis and exploration of injection systems simulations. In: Proc. IEEE Conference on Visualization. Minneapolis, MN, USA, 2005: 391-398.

[4] Cain G, Herod J. Multivariable Calculus (on-line textbook). Georgia Tech, 1997.

[5] Guo Z, Ward M O, Rundensteiner E A, et al. Pointwise local pattern exploration for sensitivity analysis. In: IEEE Conference on Visual Analytics Science and Technology. Providence, RI, USA, 2011: 129-138.

[6] Tanaka Y. Recent advance in sensitivity analysis in multivariate statistical methods. *Journal of the Japanese Society of Computational Statistics*, 1994, **7**(1): 1-25.

[7] Cacuci D. Sensitivity and Uncertainty Analysis: Theory Vol. 1. London, UK: Chapman and Hall, 2003.

[8] Griewank A. Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2000.

[9] Mcclave J, Sincich T. Statistics (10th edition). Upper Saddle River, NJ, USA: Prentice Hall. Inc., 2003.

[10] Klösgen W, Zytkow J M. Handbook of Data Mining and Knowledge Discovery, Chapter 16.3: Subgroup Discovery. Oxford, UK: Oxford University Press, Inc., 2002.

[11] Klösgen W, May M. Spatial subgroup mining integrated in an object-relational spatial database. Principles of Data Mining and Knowledge Discovery (PKDD). Berlin, Germany: Springer-Verlag, 2002: 275-286.

[12] Atzmueller M. Exploiting background knowledge for knowledge-intensive subgroup discovery. In: Proc. 19th Intl. Joint Conference on Artificial Intelligence (IJCAI-05). Edinburgh, Scotland, 2005: 647-652.

[13] Barlowe S, Zhang T, Liu Y, et al. Multivariate visual explanation for high dimensional datasets. In: Proc. IEEE Symposium on VAST'08. Columbus, OH, USA, 2008: 147-154.

[14] Correa C D, Chan Y H, Ma K L. A framework for uncertainty-aware visual analytics. In: Proc. IEEE Symposium on VAST'09. Atlantic City, NJ, USA, 2009: 51-58.

[15] Chan Y H, Correa C, Ma K L. Flow-based scatterplots for sensitivity analysis. In: Proc. IEEE Symposium on VAST'10, 2010: 43-50.

[16] Keim D A, Kriegel H P. Visdb: Database exploration using multidimensional visualization. *IEEE Computer Graphics and Applications*, 1994, **14**(5): 40-49.

[17] Siegel J H, Farrell E J, Goldwyn R M, et al. The surgical implication of physiologic patterns in myocardial infarction shock. *Surgery*, 1972, (72): 126-141.

[18] James Allen Jewelry Online Store. http://www.jamesallen.com, May 19, 2010.

[19] Panigrahy R. An improved algorithm finding nearest neighbor using kd-trees. In: Proc. LATIN 2008. Búzios, Brazil, 2008: 387-398.

[20] Yang J, Patro A, Huang S, et al. Value and relation display for interactive exploration of high dimensional datasets. In: Proc. IEEE Symposium on Information Visualization. Austin, TX, USA, 2004: 73-80.