Pointwise Local Pattern Exploration for Sensitivity Analysis

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

Computer Science Department
Worcester Polytechnic Institute
{zyguo,matt,rundenst, ruiz}@cs.wpi.edu [*]

Abstract

Sensitivity analysis is a powerful method for discovering the significant factors that contribute to targets and 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. Incorporating sensitivity analysis in visual analytic tools is essential for multivariate phenomena analysis. However, most current multivariate visualization techniques do not allow users to explore local patterns individually for understanding the sensitivity from a pointwise view. In this paper, we present 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. During exploration, users are able to interactively change the derivative coefficients to perform sensitivity analysis based on different requirements as well as their domain knowledge. Each local pattern is assigned an outlier factor, so that users can quickly identify anomalous local patterns that do not conform with the global pattern. Users can also compare the local pattern with the global pattern both visually and statistically. Finally, the local pattern is integrated into the original attribute space using color mapping and jittering, which reveals the distribution of the partial derivatives. Case studies with real datasets are used to investigate the effectiveness of the visualizations and interactions.

Keywords: Knowledge Discovery, sensitivity analysis, local pattern visualization.

Index Terms: H.5.2 [Information Interfaces and Presentation]: User Interfaces—Graphical user interfaces

1 Introduction

Sensitivity analysis 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 $\{X_1,X_2,\ldots,X_n\}$, sensitivity analysis can tell 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 for analyzing quantitative relationships among variables [24]. 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 [18]. The analysis results are important in helping engineers tune the parameters in designing an engine.

Sensitivity analysis is essential for decision making, system understanding, as well as model constructing. Numerous approaches have been proposed to calculate the sensitivity coefficients. In this paper, we focus on differential analysis, where sensitivities are defined as the partial derivatives 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 generally not visually conveyed. The key idea behind this paper is analogous to the street view in a Google map [12], where the user can stand in a position (focal point) in the global map (the attribute space) to browse the vicinity (neighbors and local patterns), such as who are the neighbors and what are the distances to the neighbors.

This pointwise exploration is helpful when a user wants to understand the relationship between the focal point and its neighbors, such as the distances and directions. 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.

To better understand the usefulness of pointwise sensitivity analysis, we discuss an application scenario for selecting an apartment near a campus. The target variable is the price and the independent variables are several apartment attributes that influence the target, such as room size, bedroom number, distance to campus, and so on. The local sensitivity analysis can tell users (students) how the price is influenced by an independent variable, either positively or negatively, as well as which variables are important for choosing an apartment. However, users often cannot easily decide which apartment is worth renting. Given a particular apartment or the one in which they currently reside, it is not always clear whether there are any better choices compared to this one. Specifically, can the student pay a little more to get a much better apartment, or find a similar one that is much cheaper. Finally, if users have domain knowledge or certain requirements, they should be able to use this to change this apartment finding task. For example, if the students know that distance is much more important, i.e., they prefer a closer one rather than a bigger one (assume both choices increase costs the same amount), they should increase the influencing factor for distance, or similarly decrease the influencing factor of size.

We seek to develop a system focusing on these problems and challenges. In this paper, we present a novel pointwise local pattern visual exploration method that can be used for sensitivity analysis and, as a general exploration method, for studying any local patterns of multidimensional data. Specifically, our system allows users to interactively select any single data instance for browsing the local patterns. Each instance is assigned a factor using statistical means to reveal outliers that do not conform to the global distribution. In the local pattern view, the layout strategy reveals the relationships between the focal point and its neighbors, in terms

[*] This work is supported under NSF grant IIS-0812027.

IEEE Symposium on Visual Analytics Science and Technology
October 23 - 28, Providence, Rhode Island, USA
978-1-4673-0013-1/11/$26.00 ©2011 IEEE

of the sensitivity weighting factors. Users can interactively change the sensitivity information, i.e., the partial derivative coefficients, based on their requirements. Users can also compare the local pattern with the global pattern both visually and statistically.

The primary contributions of this paper include:

- *A novel pointwise exploration environment*: It supports users in browsing a multivariate dataset from a *pointwise* perspective view. This exploration assists users in understanding the vicinity of a focal point and reveals the relationships between the focal point and its neighbors.

- *A novel visualization approach for sensitivity analysis*: Sensitivity analysis is one important local analysis method, thus is well suited for our pointwise exploration. 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. This helps the user find potentially interesting neighbors around the focal point, and thus acts as a recommendation system.

- *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.

- *System evaluation using real-world dataset*: We evaluate the effectiveness of our system based on a real-world dataset.

2 Related work

Multivariate analysis (MVA) involves observation and analysis of more than one statistical variable at a time, which is an important data analysis method and widely applied in many domains. There exist many automatic techniques for multivariate analysis. For example, regression analysis [9] establishes a linear relationship between independent variables and a target (response) variable. Generalized additive models [15] describe more complex relationships, such as nonlinear relationships. Response surface analysis [3] explores the relationships between several explanatory variables and response variables, using a sequence of designed experiments to obtain an optimal response. Our approach is different from the automatic techniques in that our system takes advantage of multivariate analysis methods and interactive visual exploration, which enables users to intuitively examine and adjust the analysis results.

Model construction and selection is another important research topic in MVA. First, many multivariate analysis algorithms heavily depend on the underlying models used. Second, the construction and selection of an appropriate model can help analysts predict the target or class attribute value, as well as explain and describe the multivariate phenomena using explanatory models. Numerous automated methods have been discussed for model construction and selection [21, 10, 4]. In recent years, many user-centered and semi-automated model selection and construction approaches have been proposed, such as D2MS [16]. These approaches give users the ability to easily examine various alternatives and to compare the competing models quantitatively using effective visualizations. Guo et al. [14] presented a model space visualization system that assists users in discovering linear patterns in a dataset. The proposed system uses linear models as predictive models and to explain the relationships among variables. The advantage is that our new system allows users to interactively adjust the local model based on their prior knowledge and the expected model.

Sensitivity analysis has been studied in the scope of multivariate data analysis [25]. Sensitivity analysis is the analysis of the variation of the output in a model based on small changes of 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 [5] and automated differentiation [13], 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 [11, 22]. 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 [20]. In [19], an advanced subgroup mining system called “SubgroupMiner” was proposed, which allows the analyst to discover spatial subgroups of interest and visualize the mining results in a Geographic Information System (GIS). In [1], it has been shown that subgroup discovery methods benefit from the utilization of user background knowledge. 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. [2] 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 over the independent variables. The histograms reveal the correlations, positive or negative, between the output and the coefficients. Correa et al. [8] 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. [6] proposed a flow-based Scatterplot system, which extended 2D 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 so as 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.

3 Local Pattern Extraction for Sensitivity Analysis

3.1 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 vicinity region 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 is 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. This gives users flexibility to define the neighborhood based on different applications and requirements. For example, for categorical independent attributes, such as the country of origin or manufacturer of a car, the coefficients of the sensitivity analysis are meaningless, since the attribute values are not ordinal. However, for different *origins* or *manufacturers*, the coefficients may be different and it is useful to compare them. In this case, the user can specify the box size on the categorical attributes so that the cars of the same origin and manufacturer are neighbors. 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. For the numerical independent attributes, the box size on these dimensions controls how local the pattern is.

3.2 Calculating the Sensitivities

As mentioned earlier, there are many ways to compute the sensitivity of one dependent variable with respect to an independent variable. In this paper, 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 different 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, considering a constant local changing rate for all independent variables. Also, 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. In a sense, analysts assume that the local neighbors fit this trend since the sum of the square errors to the regression line is minimized.

3.3 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 neighbor count, distances to neighbors, and orientation to neighbors. In this paper, in particular, we focus on the orientations from the focus point to the neighbors. We choose this pattern for two reasons. First, this pattern tells users the relationships between the focus point and its neighbors, i.e., the directions 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 or not, which 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, such as the blue point shown in Fig. 1.

Similar to the street view in Google Map, when a user stands at a single point (the focal point) to examine the neighborhood, the orientations to the neighbors tell users which direction they should move from the standing point (the origin) to reach each of the neighbors. In the map coordinate system, this direction is usually described using an angle between a standard direction vector, such as north, and a connecting vector, from the focal point to a neighbor point. In our system, to assist users in performing sensitivity analysis, we take the normal vector of the regression hyperplane as the standard direction. Since there are two normal vectors of one plane, without any loss of generality, we take the one directed to the negative side of the target variable as the standard normal direction. For each neighbor of the focal point, we calculate an angle θ between the normal vector of the regression hyperplane and the connecting vector between the focal point and that neighbor, as shown in Figure 1. $Cos(\theta)$ is the dot product of the two unit vectors.

Due to the unit differences, the extracted local linear trend may be dominated by some attributes. For example, a linear pattern of $y = 10000x + 5$ is dominated by y. In this case, all the connecting lines between the focal point to other neighbors are nearly parallel with each other (reduced to one dimension y). To remove the unit differences among the different attributes, 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 (as shown in Fig. 1), and the angle θ is between 0 and π. The direction of the normal vector is orthogonal to the local gradient, taking the focal point as the starting position. Therefore, the angle θ for one neighbor represents whether the relationship between the focal point and this neighbor conforms with the local linear trend or not. The expectation of this angle is $\pi/2$, assuming all the local points fit the extracted linear model very well. If the angle is $\pi/2$, it means that the vector from the focal point to this neighbor is the same as the local trend (the blue point in Fig. 1). If the angle is less than $\pi/2$ (the green point in Fig. 1), it indicates that the neighbor's target attribute value is smaller than the estimate using the extracted model. Note that when we say the predicted value, we do not mean it is the predicted value using the local regression plane (the solid red line in Fig. 1). Since we care about the relationships between the focus point and its neighbors, the predicted value is based on the regression plane that is moved to the focal point in parallel (the dotted red line in Fig. 1). In contrast, if the angle is larger than $\pi/2$ (the yellow point in Fig. 1), it means that the neighbor's target attribute value is larger than the estimate, taking the origin as the focal point.

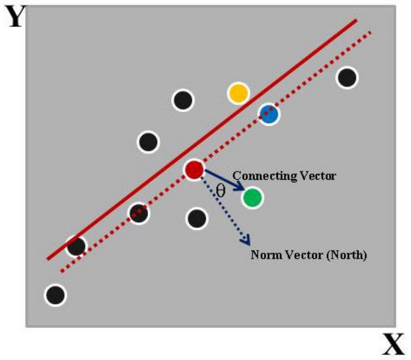


Figure 1: The extracted local pattern. The red point is the focal point. The three colored points indicate neighbors with different directions from the focal point. The angle θ shows the direction to the green point.

To sum up, in our system, the extracted local pattern for a single point is a vector V, in which each value is an angle introduced as before. The size of V is the same as the neighbor count.

3.4 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, i.e., neighbors of a single data point, 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 $\pi/2$. We know this is because the analysts have knowledge of regression analysis. In general, however, users may not have this prior knowledge.

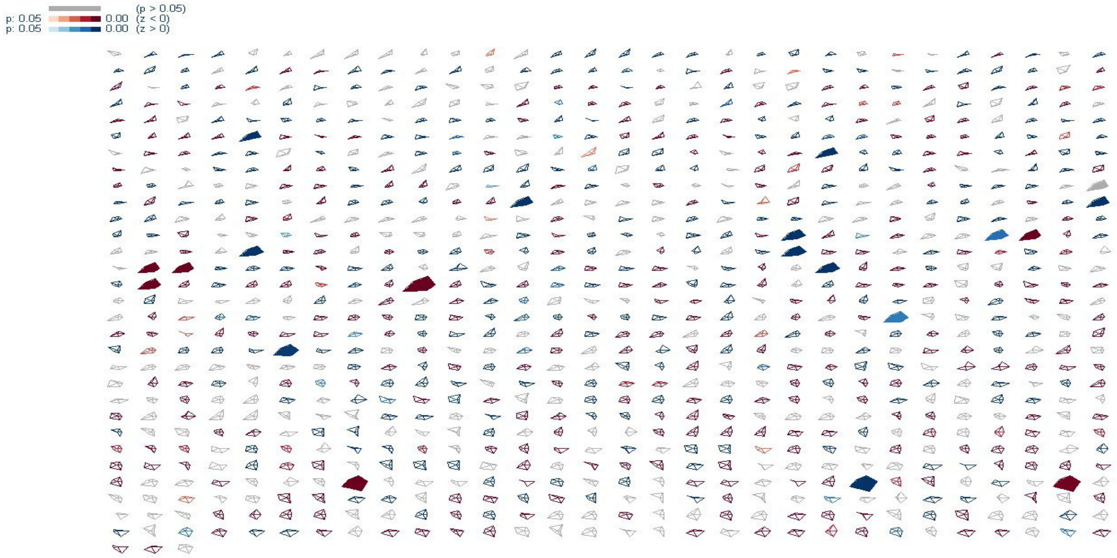


Figure 2: The global display using star glyphs (903 records from the diamond dataset). The color represents whether the data item is a anomalous local pattern or not. The filled star glyphs are selected local pattern neighbors.

As a general solution, we assume each subgroup (a local pattern) 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. To better understand this idea, we give an example for searching for anomalous patterns on a map. In this example, the extracted pattern is the percentage of water coverage around each sample position, and the goal is to detect anomalous areas in terms of this pattern. Since we assume users do not know the expected pattern, we integrate all the local patterns (percentages of water coverage) together and use a statistical test to detect anomalies. It is not hard to understand that for a map of mainland, areas near lakes and shores are anomalies; for a map of the ocean, islands are anomalies.

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 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 [19]. 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. The μ_0 and σ_0 are the mean and standard deviation values of the population. N and n are data sizes of the population and the sample, respectively. Although this is a general way to detect anomalies, visual exploration on each single pattern is still often needed. This is because this approach is based on the assumption that the population is normally distributed, which does not always hold for all applications. In our system, we support users examining each local pattern and comparing it with the global one both statistically and visually.

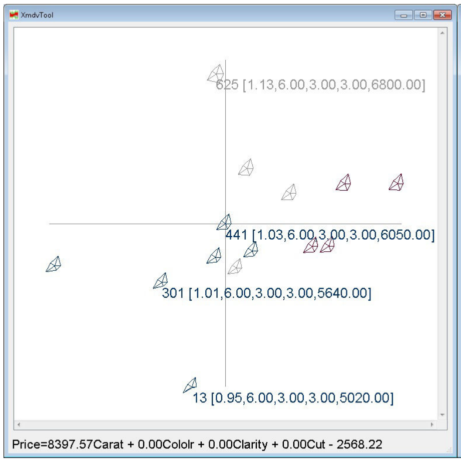
4 System Introduction

In this section, we introduce the proposed local pattern exploration method and our system design. In our system, we provide 5 different coordinated views to assist users in exploring the local patterns.

4.1 Global Space Exploration

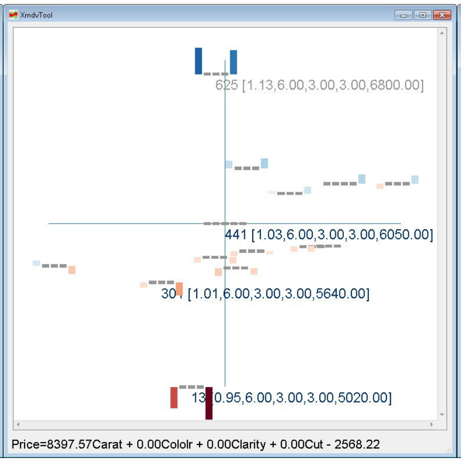
The *global view* is designed to give users a global sense of the whole dataset. Basically, any multivariate data visualization techniques, such as scatterplot matrices, parallel coordinates, pixel oriented techniques, or glyphs, can be used to display and explore the data globally. Of these methods, only glyphs show each data point individually and completely as an entity without overlapping. We use a star glyph because the analyst can easily specify which individual data point he/she wants to examine, thus leading to an easy exploration of the local pattern of that data point. A major drawback for the glyph display method is the scalability issue. When the data size is very large, each glyph is very small and it is difficult to recognize and specify a single data item. A solution is to use brushing and filtering techniques to hide uninteresting local patterns to save the display space. Another solution is clustering similar local patterns and displaying different clusters in separate views. We discuss these in the future work section.

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. 2, 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 0.05 level), which means the local pattern has 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. We use a diverging color strategy for two colors from ColorBrewer [7]; this strategy is also used in the local pattern view for comparative neighbor representation. The darker the red and blue colors are, the higher the significance is (i.e., a smaller p-value is obtained). When users examine each individual local pattern, red and blue items are generally of users' interests. Though we use 0.05 as the default significant level, if users only want to focus on the data items that are extremely different from the global pattern, they can change the significant level to a smaller value, such as 0.01



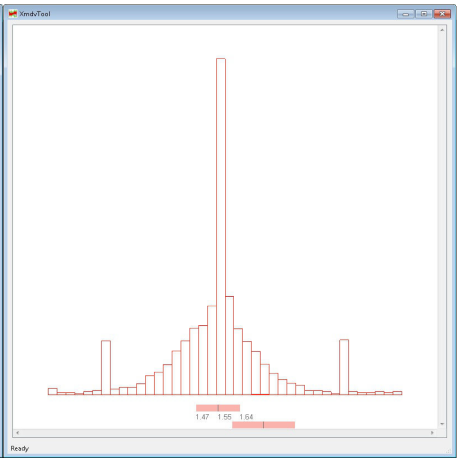
Price=8397.57Carat + 0.00Cololr + 0.00Clarity + 0.00Cut - 2568.22

Figure 3: Neighbor representation using original values.



Price=8397.57Carat + 0.00Cololr + 0.00Clarity + 0.00Cut - 2568.22

Figure 4: Neighbor representation using comparative values.



Ready

Figure 5: The comparison view. The two pink bars at the bottom represent the confidence interval of the global pattern (upper) and the selected local pattern (lower).

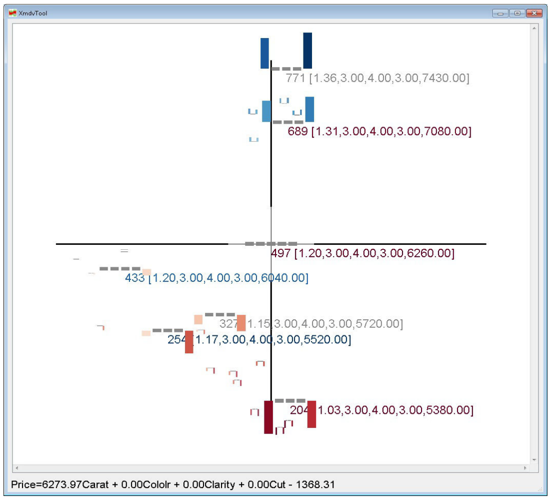
or 0.001, to reduce the number of anomalous local patterns.

When the user moves the cursor onto a single data item, its neighbors and the item itself are highlighted using larger filled glyphs to draw the user's attention. Meanwhile, the basic statistical information is shown in the bottom bar, such as neighbor count, mean value, z-score, and p-value.

4.2 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, called the *local pattern view*. The main purpose for this view is to illustrate the relationships between the focal point and all its neighbors. As a general solution, assume that the focal point is placed in the center of this view; all the neighbors' positions should be designated to reflect their relationships, according to different types of extracted local patterns and the users' requirements.

In particular, in our system, the focal point is shown in the center of the display using a star glyph, which allows the user to easily recognize the connection between the local pattern view and the global view. The two cross lines (vertical and horizontal) passing the center create four quadrants, using 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. This layout strategy tells users where to find an interesting neighbor when the goal is to discover a neighbor with different target attribute values, such as looking for a more/less expensive apartment.



Price=6273.97Carat + 0.00Cololr + 0.00Clarity + 0.00Cut - 1368.31

Figure 6: Users can use a scale factor to shrink the size of data items for reducing overlapping and visual clutter. Some data items are shown in the original size by hovering and clicking the cursor.

As discussed before, the local pattern in this paper is the orientation angle θ. The angle is mapped to X in this view. The angle of the focal point is $\pi/2$, 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 $\pi/2$, the corresponding neighbor is placed in the left half of the view. If θ is smaller (larger) than $\pi/2$ it means the neighbor's target value is smaller (larger) than the estimate. The user can use this piece of information to discover interesting neighbors. For instance, taking the example of the apartment finding problem, given a focal apartment, the students should have more interest in the neighbor apartments shown on the left side, as those neighbors are detected by our system as having lower prices than predicted comparing with the focal point.

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 (Fig. 3). The second display method is a comparative display (Fig. 4), 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 a 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. Note that the height of a bar represents the difference between the neighbor and the focal point in the normalized space, so that when the relationship between the neighbor and the focal

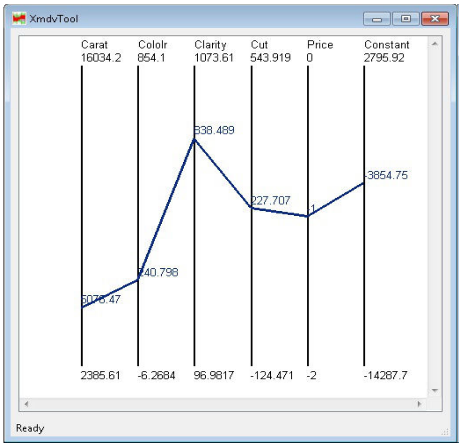
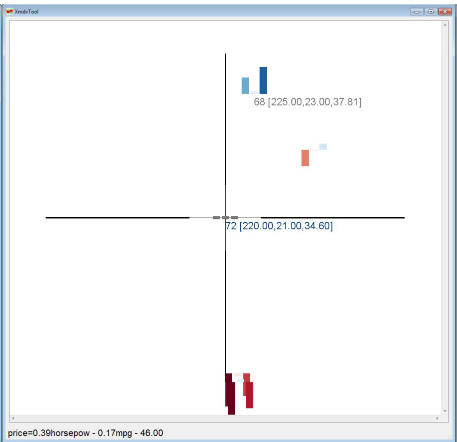
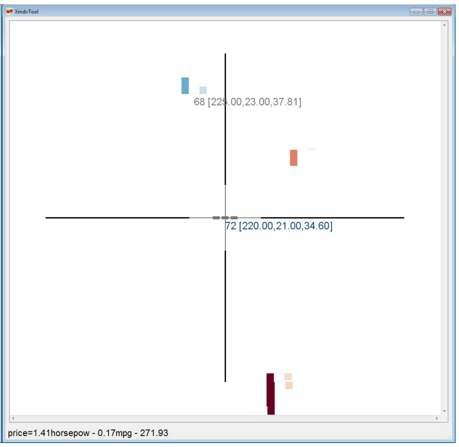


Figure 7: The local pattern adjusting view. The poly-line represents the adjustable coefficients.



price=0.39horsepow - 0.17mpg - 46.00

Figure 8: The local pattern view before adjusting the horsepower coefficient. The neighbor (ID 68) is a worse deal.



price=1.41horsepow - 0.17mpg - 271.93

Figure 9: The local pattern view after adjusting the horsepower coefficient. The neighbor (ID 68) became a better deal.

point conforms with the local trend, the sum of the bar heights of the independent attributes are the same as the bar height of the target for that neighbor. To reduce overlapping when there is a large number of neighbors, we allow users to interactively change a scale factor to reduce the size of each data item. A data item will be enlarged to its original size when the user moves the cursor onto it. Fig. 6 shows a local pattern view after scaling the data items to reduce overlapping. Some data items are shown in the original size after being clicked. In terms of the scalability for the comparative display, when there is a large number of attributes, a dimension reduction or selection technique could be applied before analysis. This issue is discussed in the future work section.

The local regression line in an equation form is shown in the bottom bar to assist the analyst in performing sensitivity analysis. For the interactions in this view, when the user moves the cursor on the focal point or one of the neighbors, the data item ID and its attribute values are displayed next to it (in the form of ID[attribute 1, attribute 2, ..., attribute n]). The user can click any data point to show or hide the attribute values.

4.3 Compare the Local Pattern with the Global Pattern

The color of each data point in the global view represents the statistical test results, i.e., an outlier factor indicates how likely the local subgroup is an anomaly. However, knowing the statistical test results is often insufficient. For example, some insignificant results may also be interesting due to a large deviation. Therefore, a visual comparison of the local with the global is still needed. To allow the user to compare the local pattern with the global pattern both statistically and visually, we provide users a *comparison view*, showing the global distribution (directions to neighbors) using a histogram. The mean values and confidence intervals for both the global and local pattern are also shown in the bottom (Figure 5). The use of this view is shown in the case study section.

4.4 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 expectations and requirements when they want to find interesting neighbors. For instance, assume that the students want to move to another apartment from the current one and are willing to increase their payments, e.g., they would be willing to pay around $100 more for one more bedroom, or pay $100 more for moving a mile closer to the campus. In this case, one more bedroom is the same as 1 mile closer, in terms of influencing ability on the target. For different users, the requirements are likely different. Students with cars may prefer a larger apartment, while ones without cars prefer a closer apartment. In the first case they would like to increase the influencing factor of size on price, while in the second case, they would like to increase the influencing factor of distance. It means that different users have different ways to define "better" when they want to find "better" neighbors around the focal point.

In our system, we provide users a *local pattern adjusting view*, using parallel coordinates (Fig. 7). The partial derivatives 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 in the new "environment", i.e., using the new trend. This is because we calculate the relationships among the focal point and its neighbors based on the normal vector of the hyperplane. Since we define the standard direction using the normal vector, we can understand this tuning as equivalent to changing the definition of north in a map.

Figures 8 and 9 show the local pattern view before and after changing the coefficients. The dataset is a car sales dataset (from the SPSS sample datasets). For easier understanding, only two independent attributes are considered: *horsepower* and *MPG*. The target is the price of the car. The goal is to compare a neighbor car, whose ID is 68 (the upper one with attribute values) with the focal one (ID is 72). It is shown that locally horsepower influences the price positively. Before adjusting, this neighbor is in the right hand side, which means a worse deal since the price is higher than estimated. We can recognize this by the comparative display of the neighbor; the sum of the height of the independent attribute bars is less than the target bar height (a lower bar for horsepower than the bar for price), which means the price is higher than estimated. After changing the weight (coefficient) of horsepower to a higher value, this neighbor become a better deal (in the left side). This is because the customer considers horsepower as an important attribute. After changing, the sum of the bar heights for independent attributes increases and exceeds the target bar height. This example shows users can change the coefficients according to their priorities.

4.5 Integrate the Local Pattern into the Global Space

Generally, a local pattern is a value (e.g., neighbor count) or a vector (e.g., distances to neighbors). Thus, the local pattern can be treated the same as the attribute values in the original data space. Assume there are n independent attributes and 1 target attribute, we can create n new dimensions taking the derivative values as derived dimensions and integrate them into the original attributes, thus resulting in a new dataset with $2n+1$ dimensions. Any multivariate data visualization technique can be used to display this new dataset, such as scatterplot matrices and parallel coordinates. This visualization enables users to discover the relationships among the derived dimensions and the original dimensions.

Fig. 10 shows an example of the *integrated view*. In this example, each data point is a child's basic body information: age, gender, height and weight. The age range is between 5 and 11. We use weight as the target and the goal is to discover for children of different ages and genders, how height influences weight. The neighbors are defined as children with the same age and gender, and similar height and weight. The figure shows the distribution of the derivatives ($\partial weight/\partial height$) in the original space (age and gender). The derivative values are color-coded (darker color means higher value) and the points are jittered to avoid overlaps. We can discover that the derivatives increase as age increases. Analysts can also compare the derivatives for different genders to answer questions, such as for 8-years-old children, which gender has larger derivative values (the answer is female).

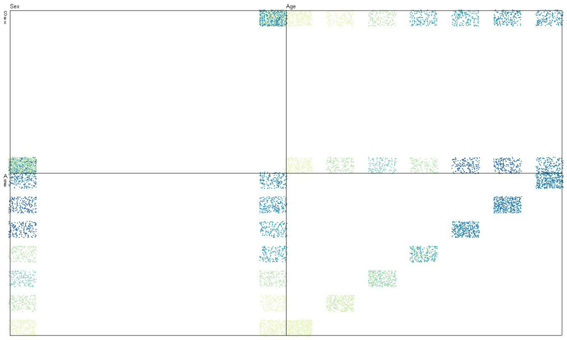


Figure 10: The view for integrating derivatives into global space. The jittered points with different colors indicate the coefficient of $\partial weight/\partial height$. As age increases, the coefficient increases. For the same age, the coefficient values are different for different genders.

5 Case Study

In this section, we discuss case studies to evaluate our approach and show the effectiveness of our system. The dataset is a diamond dataset obtained from an online jewelry store [17]. 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 13298 data items and the subset has 903 data items.

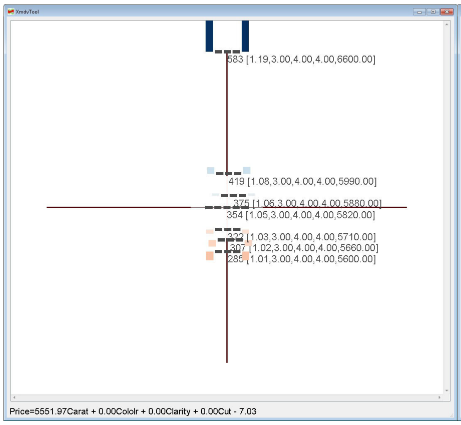
The main computational bottleneck is in the calculations involved in finding neighbors, which would be performed in a $O(n^2)$ time cost without any index data structure, assuming the data size is n. After the neighbors for each data item are found, the least square linear regression cost is $O(Km^2)$, where K is the average neighbor count and m is the dimension number. During the exploration of each local pattern, there is no computational cost since the neighbor index is already created. Another cost in our system is in the local pattern adjusting period, which is $O(k)$ (k is the neighbor count of the examined focal point). On a 3 Ghz dual core desktop PC with 4 GB of RAM and an ATI Radeon X1550 graphics card, we ran our system both for the whole dataset and the subset of the diamond dataset (neighbor range is defined as 0.1 of the entire range of each attribute). For the subset, the time for finding neighbors and regression calculating took less than 2 seconds. For the whole dataset, the time required is about 6 minutes. The huge difference is mainly due to the quadratic time complexity for finding neighbors. For both datasets, the exploration of local patterns, as well as local pattern adjustment, can be performed and updated in real time. We discuss improvements in finding neighbors in the future work section.

5.1 Where are the Good Deals

For easier understanding, we start from a single independent attribute *weight*. The user of our system can achieve this by defining an appropriate neighborhood: two diamonds are neighbors when they have similar *weight* and *price*, as well as they are of the same *color*, *clarity* and *cut*. The extracted local pattern is the orientations to the neighbors. Fig. 2 shows the global star glyph display. The color indicates whether the diamond is an anomalous one. To examine the global distribution, the user can open the comparison view (Fig. 5). The global distribution is similar to a normal distribution, except for that there are two peaks on each side. We will show later this is due to some anomalies, i.e., some diamonds whose prices are not set appropriately. The mean of the distribution is about $\pi/2$, which is the same as we discussed before, assuming the neighbors fit the local linear trend.

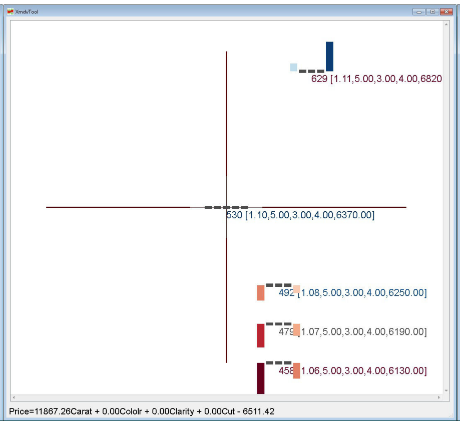
To understand the normal and abnormal data items in detail, we show three local pattern views for gray, red, and blue data points. Figure 11 shows the local pattern view of a gray data point. All the neighors of this data point are in the center of the view (x position), indicating that the directions 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. We can also recognize this local fitting by the comparative representation of all neighbors: the height of the first bar (*weight*) is almost the same as the height of the last bar (*price*). This indicates the price difference, between the focal point and one neighbor, is proportional to the weight difference. To assist the analyst in performing the sensitivity analysis, i.e., what is 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. The changing rate of price is $55, as the weight increases 0.01 carat. The influencing factors of the other independent attributes are all 0, since all neighbors have the same values.

Fig. 12 shows the local pattern view for a diamond that is blue in Fig. 2, 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 as weight increases 0.01 carat, the price increases $118. However, the price of the local neighbors are higher than estimated considering this changing trend. Take the upper diamond for example. The upper half means a higher target value based on our local pattern layout strategy.For this neighbor, the weight is 0.01 carat higher than the focal point, while the price is $450 higher than the focal point, which is a larger change rate, compared with the local trend. The user can also read this from the comparative representation of this neighbor: a higher and darker bar for price than the bar for weight, which means the price change rate is higher than weight. This tells users that this neighbor is a worse deal compared with the focal point. Similarly, we can consider another neighbor whose price is lower than the focal point, i.e., in the bottom half of the display (the



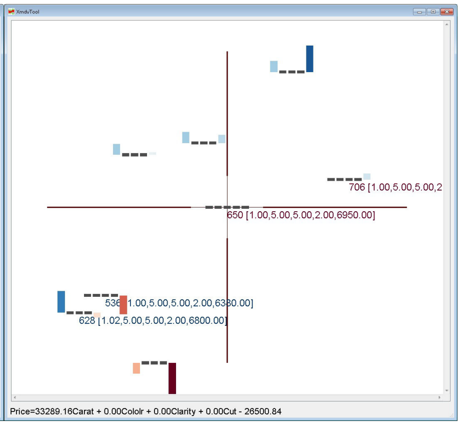
Price=5551.97Carat + 0.00Cololr + 0.00Clarity + 0.00Cut - 7.03

Figure 11: The local pattern view of a gray data item. The orientation from the focal point to all its neighbors are $\pi/2$, which is common in the dataset.



Price=11867.26Carat + 0.00Cololr + 0.00Clarity + 0.00Cut - 6511.42

Figure 12: The local pattern view of a blue data item. The orientations from the focal point to most of its neighbors are larger than $\pi/2$, which means the neighbors' target values are higher than estimated. In other words, the focal point is a "good deal".



Price=33289.16Carat + 0.00Cololr + 0.00Clarity + 0.00Cut - 26500.84

Figure 13: The local pattern view of a red data item. The orientations from the focal point to most of its neighbors are lower than $\pi/2$, which means the neighbors' target values are lower than estimated. In other words, the focal point is a "bad deal".

nearest one to the focal point). The neighbor's weight is 0.02 lower than the focal point. If this neighbor fits the local trend, the price would be \$118*2=\$236 lower than the focal diamond. However, the price is only \$120 lower than the focal diamond, which also means this neighbor is not a good deal compared with the focal diamond. From these discussions, we know that for blue diamonds, generally most of neighbors are in the right half side of the view, which means there are worse deal compared with this one. Thus, the blue diamonds should be preferable for the customers.

Finally, we give an example of a diamond mapped to red in Fig. 2. Similar with the discussion for blue diamonds, a red diamond means there are many better deals compared with this one. Fig. 13 shows the local pattern view of a red diamond. It is shown that locally as the weight increases 0.01 carat, the price increases \$332. The two neighbors (with attribute values) are better than this one (left part). For the upper neighbor, the weight is the same as the focal point, while the price is \$570 lower than the focal point (a downward red bar). For the lower neighbor, the weight is higher than the focal point, while the price is \$150 lower than the focal diamond. For the focal diamond, the neighbors in the left half are better recommendations. Since there are many blue and red diamonds (anomalies), the distribution of the global pattern has two peaks in each side. From the retailer side, it should consider increasing (decreasing) the prices of the blue (red) diamonds.

This method of discovering good and bad deals in this dataset is also suitable for more than one independent attribute. We choose only one independent attribute just because it is easy to verify whether the diamonds are worth buying.

5.2 Display the Local Pattern in the Global View

It is shown that for different local patterns (subsets of neighbors), the price increases differently as the weight increases. This means the coefficients ($\partial price/\partial weight$) are different in the whole space. It is useful to give users a global sense in terms of how the sensitivity derivatives are distributed in the original space. To assist users in better understanding this, we use the whole dataset rather than a subset of a certain range. Fig. 14 shows a scatter plot view of the dataset. We use color to represent the derivative values: dark blue means high and dark orange means low. The color strategy is again diverging. The points are jittered to reduce overlapping.

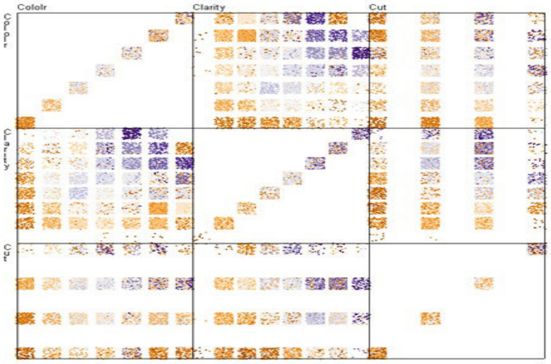
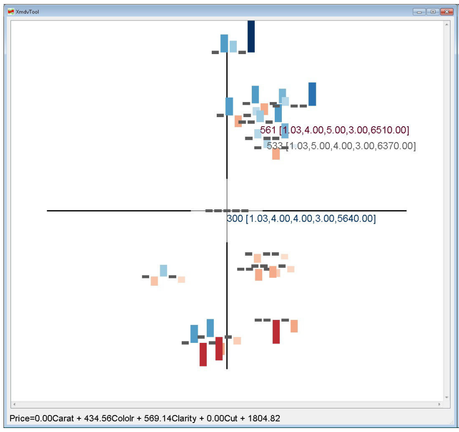


Figure 14: The coefficients of $\partial price/\partial weight$ are color-mapped and displayed in a scatterplot matrix of original attribute space.

Users can discover that the derivatives are pretty consistent for diamonds of the same color, clarity and cut. This means that for different subset of neighbors, although their weights and prices are of different range, the influencing factors of weight on price are very similar. Another discovery is that as color, clarity and cut increase, the derivatives generally increase (from dark orange to dark blue). This means that for diamonds of higher quality, the weight is more important for price, i.e., the price is very sensitive with changing weight for the subspace of higher color, clarity and cut. When customers notice that, they could consider changing their choices based on this discovery. For the blue region, they can consider choosing a diamond of lower weight, since it will save them a lot of money. In contrast, for the orange region, they can consider choosing a diamond of higher weight, since it won't increase their costs too much. We can also notice that in the upper right of the plot of clarity vs. color, there is a dark orange block in the blue area. A possible explanation for this divergence from the main pattern is that there are not enough diamonds in this region, whose color and clarity values are both very high. The low price variance results in low coefficient values.

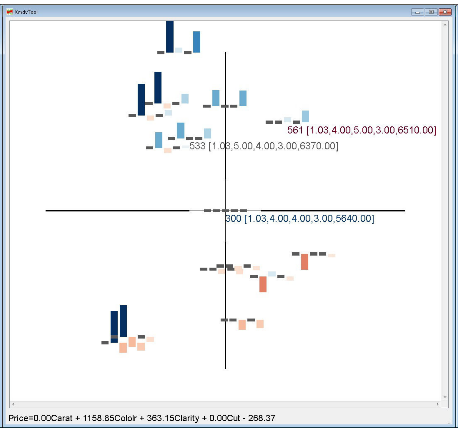
5.3 Customize the Local Pattern

Given budget limits, customers have to find a trade-off when considering the diamond attributes. We show an example to illustrate how customers can customize their requirements. Assume that a



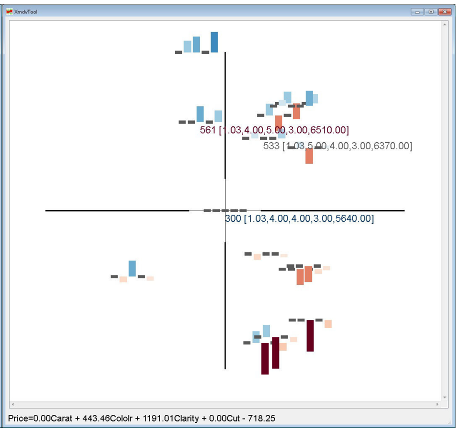
Price=0.00Carat + 434.56Cololr + 569.14Clarity + 0.00Cut + 1804.82

Figure 15: The local pattern view before tuning the coefficients. One neighbor (ID 533) has higher *color* and the other neighbor (ID 561) has higher *clarity*.



Price=0.00Carat + 1158.85Cololr + 363.15Clarity + 0.00Cut - 268.37

Figure 16: The local pattern view after increasing the coefficient of *color* and decreasing the coefficient of *clarity*. The neighbor with higher *color* became a "good" deal.



Price=0.00Carat + 443.46Cololr + 1191.01Clarity + 0.00Cut - 718.25

Figure 17: The local pattern view after decreasing the coefficient of *color* and increasing the coefficient of *clarity*. The neighbor with higher *clarity* became a "good" deal.

customer has decided the weight and cut of the selection, and is struggling with higher color or higher clarity. In this case, the neighborhood is defined as diamonds of the same weight and cut. For color and clarity, the neighborhood region covers three levels of each, indicating lower, current, and higher values. Fig. 15 shows the local pattern view of a preferable diamond before adjusting the coefficients. The two neighbors, shown with attribute values, are two alternative options compared with the focal one. Both of them are more expensive than the focal one: one has higher (better) color and one has higher (better) clarity. Before tuning the coefficients, none of them are better deals (in the left half). If the customer knows that she prefers higher color (clarity), she can accordingly increase the coefficient for color (clarity) and/or decrease that for clarity (color). Fig. 16 and Fig. 17 show the local pattern views after adjusting the coefficients. In Fig. 16, the coefficient for color is increased and the coefficient for clarity is decreased. It is clear that the neighbor with high color became a good deal. These two neighbors can be easily differentiated and the customer can tell which one is worthy purchasing in this circumstance. A similar result is shown in Fig. 17. In this case, the coefficient for clarity is increased and the coefficient for color is decreased. We can discover that the two neighbors shift in the opposite directions compared with Fig. 16. According to this example, we can see that customers can define "good" when selecting a diamond. Generally speaking, for any other type of local patterns, users can customize the definition of "interestingness" and the system is able to provide users different recommendations of neighbors.

6 Conclusion

This paper presents 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 are able to 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. During exploration, analysts can interactively change the local pattern, i.e., the derivative coefficients, to perform sensitivity analysis based on different requirements. Following the idea of subgroup mining, we employ a statistical method to assign each local pattern an outlier factor, so that users can quickly identify anomalous local patterns that deviate from the global pattern. Users can also compare the local pattern with the global pattern both visually and statistically. We integrated the local pattern into the original attribute space using color mapping and jittering to reveal the distribution of the partial derivatives. We discuss case studies with real datasets to investigate the effectiveness and usefulness of our approach.

Some future work we are actively pursuing are as follows:

- *Supporting other types of local patterns*: we plan to expand our system to support more types of local patterns, such as distances to the neighbors, and errors of the neighbors, in terms of the extracted local model.
- *Solving other tasks*: besides the regression analysis and sensitivity analysis, we plan to expand our system to solve other multivariate analysis and data mining tasks, such as classification based on nearest neighbors.
- *Local pattern management*: in addition to visually examining the local patterns, we plan to allow users to efficiently manage the extracted local patterns, such as finding similar local patterns, connecting similar local patterns, and clustering similar local patterns.
- *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 and queries*: user can interactively submit a query to discover interesting local patterns using brushing techniques in the local pattern view. This can also save the display space in the global display view since only interesting local patterns are shown.
- *Evaluation*: we plan to perform formal user studies and an expert study to better evaluate our approach in the future.
- *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 [23]. For datasets with large number of attributes, a user-driven attribute selection technique, or a dimension reduction technique based on influencing factors, could be applied [26].

References

[1] M. Atzmueller. Exploiting background knowledge for knowledge-intensive subgroup discovery. In *Proc. 19th Intl. Joint Conference on Artificial Intelligence (IJCAI-05)*, pages 647–652, 2005.

[2] S. Barlowe, T. Zhang, Y. Liu, J. Yang, and D. Jacobs. Multivariate visual explanation for high dimensional datasets. *IEEE Symposium on Proc. VAST '08.*, pages 147–154, 2008.

[3] G. Box and N. Draper. *Empirical Model-Building and Response Surfaces*. John Wiley and Son, New York, NY, USA, 1987.

[4] K. P. Burnham and D. Anderson. *Model Selection and Multi-Model Inference (2nd edition)*. Springer, New York, NY, USA, 2002.

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

[6] Y.-H. Chan, C. Correa, and K.-L. Ma. Flow-based scatterplots for sensitivity analysis. *IEEE Symposium on Proc. VAST '10.*, 2010.

[7] Color brewer. http://colorbrewer2.org.

[8] C. D. Correa, Y.-H. Chan, and K.-L. Ma. A framework for uncertainty-aware visual analytics. *IEEE Symposium on Proc. VAST '09.*, pages 51–58, 2009.

[9] N. Draper and H. Smith. *Applied Regression Analysis*. John Wiley an Sons, New York, NY, USA, 1998.

[10] B. S. Everitt. *Multivariable Modeling and Multivariate Analysis for the Behavioral Sciences*. Chapman and Hall Statistics in the Social and Behavioral Science, London, 2009.

[11] G.Cain and J. Herod. *Multivariable Calculus*. Georgia Tech, 1997.

[12] Google maps. http://maps.google.com.

[13] A. Griewank. *Evaluating derivatives: principles and techniques of algorithmic differentiation*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2000.

[14] Z. Guo, M. O. Ward, and E. A. Rundensteiner. Model space visualization for multivariate linear trend discovery. *IEEE Symposium on Proc. VAST '09.*, pages 75–82, 2009.

[15] T. Hastie and R. Tibshirani. *Generalized Additive Models*. Chapman and Hall, London, 1990.

[16] T. B. Ho and T. D. Nguyen. Visualization support for user-centered model selection in knowledge discovery in databases. *Tools with Artificial Intelligence, IEEE International Conference*, 0:228, 2001.

[17] James allen jewelry online store, obtained on may 19 2010. http://www.jamesallen.com.

[18] M. Jeloviȅ, J. Jurić, Z. Konyha, and D. Gračanin. Interactive visual analysis and exploration of injection systems simulations. *IEEE Symposium on Proc. Visualization*, pages 391–398, 2005.

[19] W. Klösgen and M. May. Spatial subgroup mining integrated in an object-relational spatial database. In *Principles of Data Mining and Knowledge Discovery (PKDD)*, pages 275–286. Springer-Verlag, 2002.

[20] W. Klösgen and J. M. Zytkow, editors. *Handbook of data mining and knowledge discovery, chapter 16.3: Subgroup discovery*. Oxford University Press, Inc., 2002.

[21] W. H. Ludwig Fahrmeir, Gerhard Tutz. *Multivariate Statistical Modelling Based on Generalized Linear Models (2nd edition)*. Springer, New York, NY, USA, 2001.

[22] J. Mcclave and T. Sincich. *Statistics (10th Edition)*. Prentice Hall. Inc., Upper Saddle River, New Jersey, USA, 2003.

[23] R. Panigrahy. An improved algorithm finding nearest neighbor using kd-trees. In *LATIN 2008*, pages 387–398, 2008.

[24] A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, Gatelli, D. Gatelli, M. Saisana, and S. Tarantola. *Global Sensitivity Analysis: The Primer*. John Wiley and Sons, New York, NY, USA, 2008.

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

[26] J. Yang, A. Patro, S. Huang, N. Mehta, M. O. Ward, and E. A. Rundensteiner. Value and relation display for interactive exploration of high dimensional datasets. In *IEEE Symposium on Proc. Information Visualization*, pages 73–80, 2004.