Nugget Browser: Visual Subgroup Mining and Statistical Significance Discovery in Multivariate Datasets

Zhenyu Guo, Matthew O. Ward, Elke A. Rundensteiner
Worcester Polytechnic Institute, Worcester Polytechnic Institute, Worcester Polytechnic Institute
zyguo@cs.wpi.edu, matt@cs.wpi.edu, rundenst@cs.wpi.edu

Abstract

Discovering interesting patterns in datasets is a very important data mining task. Subgroup patterns are local findings identifying the subgroups of a population with some unusual, unexpected, or deviating distribution of a target attribute. However, this pattern discovery task poses several compelling challenges. First, computational data mining techniques can generally only discover and extract pre-defined patterns. Second, since the extracted patterns are typically multi-dimensional arbitrary-shaped regions, it is very difficult to convey in an easily interpretable manner. Finally, in order to assist analysts in exploring their discoveries and understanding the relationships among patterns, as well as connections between patterns and the underlying data instances, an integrated visualization system is greatly needed. In this paper, we present a novel subgroup pattern extraction and visualization system, called the Nugget Browser, that takes advantage of both data mining methods and interactive visual exploration. The system accepts analysts' mining queries interactively, converts the query results into an understandable form, builds visual representations, and supports navigation and exploration for further analyses.

1 Introduction

Subgroup discovery [3] is a method to discover interesting subgroups of individuals, such as "the subgroup of students who study in small public high schools are significantly more likely to be accepted by the top 10 universities than students in the overall population". Subgroups are described by relations between independent (explaining) variables and a dependent (target) variable, as well as a certain interestingness measure. There are many application areas of subgroup discovery. For example, the extracted subgroups can be used for exploration and description, as well as understanding the relations between a target attribute and a set of independent attributes. Each subgroup or a set of subgroups is a pattern, i.e., a sub-region in the independent space. Detailed examination of such regions can be useful to improve understanding of the process that result in the pattern.

The subgroup discovery poses many challenges:

First, since the analysts may not know in advance what kind of interesting features the data contains, they may have to repeatedly re-submit queries and explore the results in multiple passes. For example, when the user submits a mining query, they need to specify the target attribute range of interest, such as the top 10 universities mentioned before. However, for different datasets and different application scenarios, the number of the top universities may be different, so they might have to try several times to find an appropriate range. This makes the mining process tedious and inefficient. Thus, we need an interactive mining process that allows analysts to submit queries dynamically and explore the results in an interactive manner.

Second, without visual support, users can only examine the mining results in text or tables. This makes it very hard to understand the relationships among different subgroups and how they are distributed in the feature space. Besides, when the user explores the mining results, the results are often in a descriptive or a abstracted form, such as summaries of the sub-regions. However, the examination of the instances in the region is also very important for understanding the data point distribution. Thus, without a visualization of the mining results, Users cannot build connections between the patterns and the instances.

Finally, adjacent subgroups should be aggregated and clustered when they are of the same interesting type. For example, given there are two subgroups of students, both of which have significantly higher acceptance rates than the population, and they are adjacent to each other in one independent attribute, such as the groups with medium and high income. Then the two subgroups should be aggregated, and reported or treated as a whole subgroup. One benefit is that this aggregate representation is more compact, which provides the users a smaller report list for easy examination. Another benefit is that the compact representation can be more efficiently stored in a file and loaded in computer memory. However, the clustered mining results generally tend to be multi-dimensional arbitrary-shaped regions, which are difficult to understand, report and visualize. Therefore, conveying the pattern in a compact, easily understandable, and visualizable form is desirable.

Focusing on these challenges, our main goal is to design a visual interface allowing users to interactively submit subgroup mining queries for discovering interesting patterns. Generally, the main users of our system are analysts who want to perform subgroup mining tasks but have difficulties in understanding the mining results. Without a visual representation of the results, analysts are have difficulties determining if the mining results are interesting, if there are any patterns in the results, and how to refine their queries. Another type of user for our system are analysts who have difficulties specifying queries. Like other types of mining queries and tasks, such as clustering and association rule mining, some parameters are needed to form the query, such as how to define subgroups and what is the target share range. Therefore, an exploratory process is strongly needed that supports analysts in examining mining results and refining queries. Specifically, our system can accept mining queries dynamically, extract a set of hyper-box shaped regions called *Nuggets* for easy understandability and visualization, and allow users to navigate in multiple views for exploring the query results. While navigating in the spaces, users can specify which level of abstraction they prefer to view. Meanwhile, the linkages between the entities in different levels and the corresponding data points in the data space are highlighted.

The primary contributions of this paper include:

- A novel subgroup mining system: we design a visual subgroup mining system where users can conduct a closed loop analysis involving both subgroup discovery and visual analysis into one process.
- An understandable knowledge representation: we propose a strategy for representing the mining results in an understandable form. In addition to storage benefits, this representation is easy for analysts to understand, and can be directly displayed using common multivariate visualization approaches.
- A 4-level structure model: we designed a layered model that allows users to explore the data space at different levels of abstraction: instances, cells, nuggets, and clusters.
- Visual representation for the nugget space: for each level, we design a view in which users are able to explore and select items to visualize. The connections between the adjacent layers are shown based on the user's cursor position.
- We implemented the above techniques in an integrated system called *Nugget Browser* in XmdvTool [21], a freeware multivariate data visualization tool.
- Case studies suggest that our visualization techniques are effective in discovering patterns in multivariate datasets.

2 Related work

Visual data mining techniques aim to combine information visualization with data mining [22, 17, 13]. A powerful data mining strategy should involve users in the visual analytics process. The users should be allowed to explore the discoveries and specify what they are looking for. The mining results should also be easily understandable. Recently, numerous visual analytics based systems have been presented to solve knowledge discovery tasks. Hao et al. [9] presented the Intelligent Visual Analytics Query (IV-Query) concept that combines visual interactions with automated analytical methods to support analysts in discovering the special properties and relations of the identified patterns. Yang et al. [23] presented the Nugget Management System (NMS) that allows users to extract patterns via interactive range queries and provided several mechanisms for users to manage their discoveries, such as filtering out similar nuggets and refining the discoveries. Guo et al. [8] presented a model space visualization system that assists users in discovering linear patterns in a dataset. The system can reveal multiple coexisting linear trends and provides users the flexibility to tune the discovered trends. uchs et al. [7] proposed a system that integrates interactive visual analysis and machine learning to support insight generation. Yu et al. [25] also proposed a closed loop between visual analysis of discoveries and data mining processes. They showed how this system can be effectively applied to multimedia datasets and continuous time series data. This paper follows these visual mining technique concepts allowing the analysts to interactively submit mining queries to discover interesting multi-dimensional patterns. In this paper, we focus on a specific data mining method, i.e., subgroup mining, to assist the users in discovering statistical significance in multivariate datasets.

Subgroup pattern mining is a very popular and simple form of knowledge extraction and representation [14]. In [15], an advanced subgroup mining system called “SubgroupMiner” is proposed, which allows the analysts discovering spatial subgroups of interest and visualize the mining results in a Geographic Information System (GIS). In [2], it is shown that the subgroup discovery methods benefit from the utilization of user background knowledge. In this paper, we not only allow the users to perform the subgroup mining in an interactive manner, but also visualize the mining results in different coordinated views, assisting the users in examining the patterns and understanding the multi-dimensional relationships among the patterns.

Since usually the extracted features in multivariate datasets are high-dimensional, a major problem is the difficulty in effectively visualizing such high-dimensional patterns and their relationships. There are several techniques that map high-dimensional patterns to a lower dimensional space. A linear mapping method takes the first two principal components obtained from Principal Component Analysis (PCA) [12] and maps the dataset to a 2 dimensional space. Multidimensional Scaling (MDS) [5] and Kohonens Self Organizing Maps (SOM) [16] are non-linear variants, requiring the minimization of a cost function of the distances. Somorjai et al. [18] proposed a relative distance plane method. $2N-3$ of the $N(N-1)/2$ interpattern distances are preserved in terms of two reference points. Radviz [10] is a radial visualization with dimensions assigned to points called dimensional anchors (DAs) placed on the circumference of a circle. We apply the layout strategies in our system for different views to reveal the relationships between multiple patterns and query results.

3 Visual Subgroup Mining and a Proposed 4-Level Layered Model

As mentioned in Sec. 1, a subgroup discovery problem can be defined in three main features: subgroup description, a target attribute, and a interestingness measure function.

A subgroup in a multivariate dataset is described as a sub-region in the independent attribute space, i.e., range selections on domains of independent attributes. For example, “male Ph.D. student in computer science department whose age is large (larger than 25)” is a subgroup with constraints in the 4 independent attribute space, i.e., *gender*, *degree program*, *department* and *age*. The sub-groups can be initialized by partitioning the independent attribute space. Given a multivariate dataset, pre-processing partitions the data space into small cells by binning each independent attribute into several adjacent subranges, such as low, median and high ranges. Each cell is a description of one subgroup element.

For the target attribute, based on the application and the cardinality, it can be continuous or discrete. The quality functions are different for these two target attribute types.

As a standard quality function, we uses the classical binomial test to verify if the target share is significantly different in a subgroup. The z-score is calculated as:

$$\frac{p-p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n}\sqrt{\frac{N}{N-n}}$$

This z-score quality function compares the target group share in the sub-group (p) with the share in its complementary subset. n and N are subgroup size and total population size. p_0 is the level of target share in the total population and $(p-p_0)$ means the difference of that target shares. For continuous target attributes and the deviating mean patterns, the quality function is similar, using mean and variance instead of share p and $p_0(1-p_0)$.

Users can submit queries on the target attribute to specify target range or a significant level to measure the interestingness of each group. The subgroups with high quality measures are query results, i.e., discovered patterns. Users can visually explore the extracted patterns and furthermore, can adjust the previous query and perform a new loop of query processing.

Intuitively, we use color to represent the mining result in the cell level. The cells (subgroups) are colored gray if their quality measure don’t satisfy the significance level (usually 0.05). If the z-score is larger than zero and the p-value is less than 0.05, the cells are colored red. This means that the target attribute share or the average target attribute value are significantly larger than the population. Similarly, for the cells whose z-score is less than zero and the p-value is less than 0.05, the cells are colored blue. This means that the target attribute share or the average target attribute value are significantly lower than the population. We say two subgroups are *of the same type* if they both satisfy the same query, i.e., both of them are significant and their z-scores are both larger than the positive critical value (1.96) or smaller than negative critical value (-1.96). we use different colors to represent different subgroup types.

A direct way to report the mining results is to return all the colored cells. Notice that the number of cells is exponential in the number of independent attributes. The query result can be very large, which makes it hard for the user to explore and understand. Specifically, a large set of unrelated cells may not be desired, because: 1. Users may only care about large homogeneous regions (subgroups of the same type) rather than a set of unrelated cells. 2. Users may want to know how many connected regions there are and what the sizes are. 3. The result should be in a compact manner for ease of understanding.

Towards these goals, we computationally extract two higher level of abstractions of the mining result, i.e., the nugget level and the cluster level.

In the cluster level, we aggregate neighbor cells of the same type to form a cluster i.e., a connected region (Fig. 1 (a)). The clustering results can be used to answer questions, such as how many connected regions there are and what the sizes (number of instances or cells) are. There are two benefits for the result in the cluster level besides to ease exploration. The first one is that the number of clusters can reveal the distribution of the mining result, such as a single continuous large cluster or a set of discontinuous small clusters scattered in the space. This can assist the users to better understand how the independent attributes influence the target share.

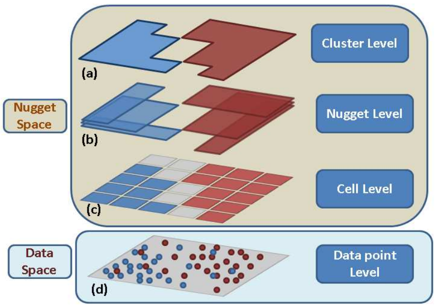


Figure 1: The proposed 4-level layered model. User can explore the data space in different levels in the nugget space.

Second, since the subgroups of the same type are generally treated as a whole set, the same treatment can be applied to all individuals in one cluster rather than each single cell. Since users might be only concerned with the large clusters, we can further filter out the small clusters, based on a user-specified threshold. This idea of clustering cells is similar to grid-based clustering and more benefits are discussed in[20, 1]. The difference is that we cluster the cells of the same type in terms of their interestingness based on the significance level for a target attribute, while most of the grid-based clustering techniques only consider the densities of each cell.

Although there are some benefits to representing the result as clusters, the largest problem is that the clusters are generally arbitrarily-shaped sub-regions in multi-dimensional space. This makes it very difficult for the users to understand the shape of a cluster and visually represent a cluster. To deal with these problems, we propose another level between the cell level and the cluster level, i.e., the nugget level. Specifically, we aggregate neighbor cells to form larger block-structured hyper-boxes for compact representation and easier perception. This aggregation of a set of adjacent cells is called a *nugget*. A nugget can be unambiguously specified and compactly stored by two cells, i.e., a starting cell and an ending cell, which are two corners of the corresponding hyper-box. A nugget has two important properties: *irreducibility* and *maximality*.

irreducibility: any sub-region of a nugget, also in the cell form, is still of the user's interest and meets the interestingness measure function requirement.

maximality: a nugget cannot be extended in any direction in any dimension to collect more cells to form a larger one.

The concepts of irreducibility and maximality were proposed by [4]. We extend this idea to a multi-dimensional space to generate a set of largest hyper-rectangular regions that satisfy the query.

The proposed 4-level layered model is shown Fig. 1. As shown in Fig. 1 (a), assume that the whole feature space is two dimensional (the gray plane) and the target dimension values (binary) are represented as the point color. In this example, assume the blue and red points are from two classes, e.g., USA cars and Japanese cars. Assume the user's query is requesting to find the subgroups where the target share (origin is USA) of the cars are significantly higher or lower than the population. To answer this, we first color the cells based on z-score: color the cell blue (red) if the percentage of cars from USA is significantly higher (lower) than the whole of the population. The partitioning and coloring results are shown in Fig. 1 (c). A gray cell means no significance is detected or are empty cells.

4 Nugget Extraction

In this section, we describe our proposed nugget representation and extraction method. Assume there are D dimensions in the feature space. As the discretization mentioned before, each dimension is partitioned into several bins. Assume there are B_k bins for dimension k. The cut points for dimension k are $C_{k,1}$ $(min) < C_{k,2} < \cdots < C_{k,B_k+1}$ (max). Here $C_{k,j}$ means the value of the j^{th} cut point in dimension k, assuming the first cut point is the minimum in this dimension.

For any cell x, we assign an index (entry) based on its value position in each dimension: [$I_{x,1}$, $I_{x,2}$, $\cdots$, $I_{x,D}$] (1 $\leq I_{x,k} \leq B_k$, for $1 \leq k \leq D$). For example, if the first dimension value lies between the minimum and the second cut point, i.e., $C_{1,1} \leq v < C_{1,2}$, the index value of the first dimension of this instance is 1.

Definitions and the nugget extraction algorithm are introduced below:

Sort all cells: we define a cell c_a as *ahead of* another cell c_b if for a dimension k, $I_{c_a,k} < I_{c_b,k}$, and for the previous indices, they are all the same, i.e., $I_{c_a,t} = I_{c_b,t}$ for $1 \leq t < k$. We sort all the cells according to this order. We call the sorted list *CellList*. Some positions could be missing if the cell with that index is empty.

Of the same type: two cells are *of the same type* if they both satisfy the same query. This means they have the same color.

Previous cell: c_a is the *previous cell* of cell c_b in dimension k if $I_{c_a,k} = I_{c_b,k}$ - 1, and for the other indexes, they are the same, i.e., $I_{c_a,k} = I_{c_b,k}$ for $1 \leq j \leq D$ and $j \neq k$. So usually one cell has D *previous cells* in terms of all the dimensions.

Between two cells: cell c_x is *between* c_a and c_b if for each dimension, the index of c_x is larger than or equal to c_a, and smaller than or equal to c_b, i.e., $I_{c_a,k} \leq I_{c_x,k} \leq I_{c_b,k}$, for $1 \leq k \leq D$. If cell c_x is between c_a and c_b, it means c_x is covered by the hyper-box taking c_a and c_b as two corners. Note that here '*between*' does not mean the location in *CellList*.

Reachable: cell c_b is *reachable* from c_a if a) c_a and c_b are of the same type, and b) all the cells *between* these two cells are of the same type as c_a and c_b. If c_b is *reachable* by c_a, then that means the hyper-box, taking c_a and c_b as corners, is colored uniformly.

Algorithm Description: To find all the nuggets, for each cell c_x, we fill a list of cells, called *reachList*. If cell c_y is in the *reachList* of c_x, that means c_y is reachable from c_x. We fill this list from an empty list for each cell in the order in *CellList*. This is because when filling the *reachList* for cell c_x, we have finished the lists of the D (maybe fewer) *previous cells* of c_x. Due to the property of *irreducibility*, we only examine the cells in the list of *previous cells* for filling the list for the current cell. After getting the union of all the *reachList*s of all the *previous* cells, we check each cell in the unioned list and delete unreachable cells. For this purging process, again only the previous cells' *reachList* require access. To fulfill *maximality*, those surviving cells, which can reach the current cell, have to be removed from the *reachlists* of the previous cells. The area between cell c_x and c_y (a cell in the *reachlists* of c_x) is a nugget.

5 Nugget Browser System

In this section, we introduce the system components, views, and the interactions. The overall mining and exploring procedure is as follows. Users start from a data space view and submit mining queries in this view interactively, such as changing the subgroup definition (the cutting point positions) and target share range. The mining results will be shown in real-time in both the data space view (Section 5.1) and the nugget space view (Section 5.2). Users can explore the visually represented mining results in different coordinated views, and then adjust their queries until an interesting pattern is found. Therefore, a closed loop is formed to guide users in fining interesting subgroups by refining their queries.

5.1 Data Space

We employ Parallel Coordinates (PC), a common visualization method for multivariate datasets [11], to visualize the data points and nuggets. In parallel coordinates, each data point is drawn as a poly-line and each nugget is drawn as a colored translucent band (Fig. 6), whose boundaries indicate the values of the lower range (starting cell) and upper range (ending cell) for each dimension. The color blue and red indicate the sign of the z-score and darker color means higher significance is discovered for the subgroup. We provide interactions in the nugget navigation space view so that users can select which data points to view in the cell, nugget and cluster level. The last dimension (axis) is the target attribute that guides the user in submitting queries and changing the target share ranges. The query ranges are shown during adjustment (vertical colored bars on the last axis). To assist users filtering out uninteresting nuggets, a brush interaction is provided. Users can submit a certain query range in the independent attribute space and all the nuggets that don't fully in the query range will be hidden in the nugget view. An example of a query is to select all the subgroups within a certain age range.

5.2 Nugget Space

In the nugget space view, three coordinated views, i.e., cluster view, nugget view, and cell view are shown in different 2D planes (Fig. 7). The linkages show the connections between adjacent views [6].

Cluster View. In the cluster view (Fig. 7 left), we employ a small "thumbnail" of a parallel coordinate view to represent each cluster. The size of each thumbnail is proportional to the number of instances each cluster contains, so that large clusters attract the user's attention. When the user moves the cursor onto a cluster, the parallel coordinate icon is enlarged and the connections are shown from this cluster to all the nuggets in the nugget view that comprise this cluster. Meanwhile, the corresponding instances are shown in the data space view.

Since the clusters consist of the data points in a high-dimensional space, to preserve the high-dimensional distances among the clusters we employ an MDS layout [5] to reveal latent patterns. The question is how to measure the similarity of two clusters. A commonly used and relatively accurate method for measuring the distance between two groups of instances is to average all the Euclidean distances of each instance pair from different groups. The problem is that for large clusters, the computational cost is high. We therefore calculate the distance in a upper level of the proposed 4-level model, i.e., using the average Euclidean distances between all cell pairs. As a result, the cost reduces as it depends on the number of cells, which is much smaller. The cell distance is calculated as the Euclidean distance between two cell centroids.

Nugget View. As mentioned before, each nugget is a hyper-rectangular shape. A single star glyph with a band, as proposed in [24], can thus be used to represent a nugget (Fig. 7 middle). The star glyph lines show the center of the nugget, and the band fades from the center to the boundaries. Similar to the cluster view, connections between the nugget view and the cell view are displayed according to the user's cursor position. The corresponding data points are also highlighted.

We again use an MDS layout for the nugget view, but the distance metrics are calculated differently from the cluster view. This is because any two nuggets could overlap in space, thus an instance could be covered by multiple nuggets. To reveal the distance between two nuggets, we designed two different distance measurements: one for overlapping nuggets and one for non-overlapping nuggets.

When the two nuggets have common cells, the distance metric indicates how much they overlap:

$$Dis(Nugget_A, Nugget_B) = \frac{|A| + |B| - 2|A \cap B|}{|A| + |B|}$$

Here $|A|$ means the number of cells that cluster A includes. When the two cells have a very small overlapping area, i.e., almost non-overlap, the distance is near 1. When the two cells almost fully overlap on each other, the distance is near 0.

When the two nuggets do not have any common cells, we use the Manhattan distance as the measurement. For each dimension, the distance is measured by using a grid as a single unit, called *grid distance*. For example, the grid distance for dimension k is 0 if on that dimension the two nuggets' boundaries meet without any gaps, or the two nuggets have overlapping bins (note that the two nuggets do not overlap in space, but may overlap in certain dimensions). The grid distance of dimension k is 1 if there is a one-bin gap between the two nuggets on that dimension. The distance in any dimension is the cell distance $+1$ indicating how many steps they are away from each other:

$$Dis(Nugget_A, Nugget_B) = \sum_{k=1}^{D} (GridDistance_k(A, B) + 1)$$

Note that the minimal distance is 1 for two non-overlapping nuggets, which is also the maximal distance for two overlapping nuggets. Hence in the MDS layout view, the nuggets in a cluster will tend to stay together to help reveal patterns.

Cell View. In the cell view (7 right), each cell is represented as a square. The cell colors are consistent with the colors in other views. The cell is highlighted when the user is hovering the cursor on it. Meanwhile, all the data points in this cell are shown in the data space view. The curves indicating connections between the cell level and the nugget level are also shown for the cells the cursor points to. Instead of a single curve, multiple ones are shown as a cell could be included in multiple nuggets.

6 Case Studies

In this section, we discuss a case study showing the effectiveness of our system. The dataset was obtained from the UCI Machine Learning Repository called "Mammographic Mass Dataset" [19]. Mammography is the most effective method for breast cancer screening. The dataset size is 961 (830 after removing instances with missing values). 5 independent attributes, such as the *age* of the patient and the *density* of the mass, are extracted and the target attribute is *Severity* (benign or malignant).

There are two main goals for analyzing the dataset. The first one is to understand how the independent attributes influence the target. This can help the doctors find the important attributes impacting the diagnosis results. The second goal is to discover the subgroups that the benign (malignant) rate is significantly higher or lower than the population. For a future diagnosis, if a patient is discovered in those groups, more attention should be paid or some conclusion about the diagnosis result could be drawn.

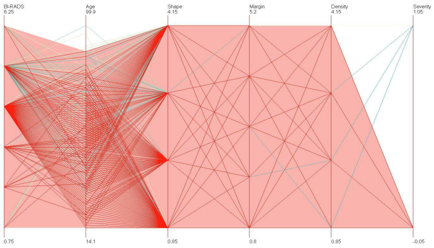


Figure 2: This is the data space view (parallel coordinate). The red poly-lines are brushed benign instances. The pink region is the brushed area.

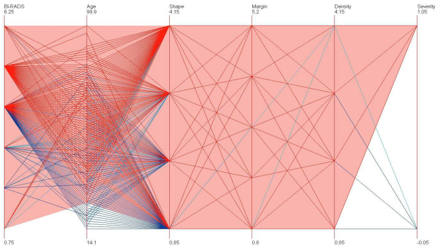
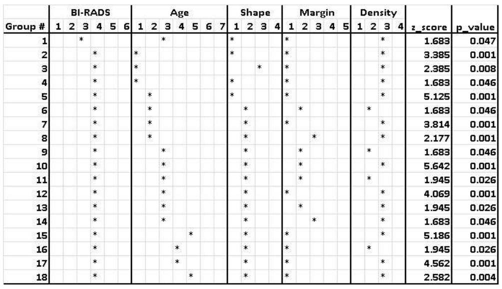


Figure 3: Similar to Fig. 2: the red poly-lines are brushed malignant instances.

To show the difficulty of finding how the independent attributes influence the target attribute using common multivariate data visualization techniques and interactions, we first display the dataset using Parallel Coordinates in XmdvTool. As shown in Fig. 2 and 3, the highlighted instances are selected using the brush technique (range query) on the target attribute. Fig. 2 shows the query result on all the benign instances (red color poly-lines) and Fig. 3 shows the query result on all the malignant instances. The pink area shows the bounding box of all the instances in the query. It can be observed that for each query, the instances cover almost the whole attribute ranges and all different values in different dimensions. This shows the common visualization technique, even with interactive range queries, can hardly reveal the relationship between the independent attributes and the target attribute.

We then show the insufficiency of the traditional subgroup mining technique without visualization in providing a compact and easily understandable mining results. We performed the mining as follows. The target share value is benign in the target attribute. This query examines the subgroups with significantly higher benign rate and significantly lower benign rate. Note that significantly lower benign rate does not necessarily mean significantly higher malignant rate, which can be examined by specifying another mining query that takes share value as malignant in the target attribute.



	BI-RADS						Age							Shape				Margin					Density					
Group #	1	2	3	4	5	6	1	2	3	4	5	6	7	1	2	3	4	1	2	3	4	5	1	2	3	4	z_score	p_value
1			*						*					*				*							*		1.683	0.047
2				*			*							*				*							*		3.385	0.001
3				*			*									*		*							*		2.385	0.008
4				*			*							*				*							*		1.683	0.046
5				*				*						*				*							*		5.125	0.001
6				*				*							*				*					*			1.683	0.046
7				*				*							*			*							*		3.814	0.001
8				*				*							*					*					*		2.177	0.001
9				*					*						*				*					*			1.683	0.046
10				*					*						*				*						*		5.642	0.001
11				*					*						*				*					*			1.945	0.026
12				*					*						*			*							*		4.069	0.001
13				*					*						*				*						*		1.945	0.026
14				*					*						*					*					*		1.683	0.046
15				*							*				*			*							*		5.186	0.001
16				*						*					*			*						*			1.945	0.026
17				*						*					*			*							*		4.562	0.001
18				*							*				*			*							*		2.582	0.004

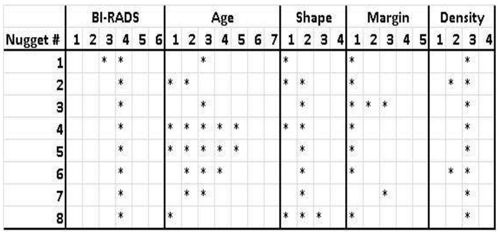
Figure 4: The mining results are represented in a table before aggregating neighbor subgroups. Each row is one subgroup and each subgroup is described using stars.

The independent attribute space is portioned by binning each attribute. Specifically, for the attribute whose cardinality is smaller than 7, the bin number is the same as the cardinality, such as *density*. For numerical attribute (*age*), the bin number is set to 7. We chose 7 because for lower values, the patterns are very similar, but less clear. While higher number of bins results in a lower number of instances in each group, which reduces the reliability of significance due to the small sample size. After the binning, the whole dataset is partitioned into a set of subgroups. Each subgroup consists of a group of individuals whose attribute values are similar or the same in all dimensions. Each subgroup is examined using the p-value and z-score of the statistical test as the interestingness measure.

Parts of the mining results are shown in Fig. 4 as a table. The star means the description of each subgroup in each dimension. 18 subgroups have the benign rate significantly larger than the population. It is clear that without the visualization, analysts cannot understand how the subgroups are distributed in the space and the relationships between the subgroups. Also, for some subgroups, such as number 12, 13, and 14, they are adjacent to each other and can be reported as a single group for a compact representation.

From the previous discussions, we can observe several difficulties: 1. it is hard to understand how the independent attributes influence the target using common visualization techniques, and 2. it is hard to understand the distribution of the subgroups, and 3. the mining results are not reported as a compact knowledge representation form. Next we will show how to use the Nugget Browser system to better solve the subgroup mining problem. Fig. 5 shows the higher level, i.e., the nugget level representation of the mining result in a table form. 8 nuggets are reported in a more compact manner, compared to the result of traditional subgroup mining, i.e., a list of subgroups. Fig. 6 shows all the nuggets (translucent bands) extracted in the data space view. Color blue means a significantly higher benign rate and color red means a significantly lower benign rate.

It is very clear that subgroups with high benign rates can be differentiated from the low benign rate subgroups in most of the dimensions, which indicates that the independent attributes have a strong impact on the target. However, this influence can hardly be discovered in traditional multivariate data visualization techniques, even with range queries. Specifically, the high benign rate subgroups have lower values for attributes *BI-RADS*, *Age*, *Shape* and *Margin*, compared to the low benign rate subgroups. Most of the subgroups with significance discovered have *Density* value 3 (means low). More details of how the independent attributes influence the target will be discussed later.



	BI-RADS						Age							Shape				Margin					Density			
Nugget #	1	2	3	4	5	6	1	2	3	4	5	6	7	1	2	3	4	1	2	3	4	5	1	2	3	4
1			*	*					*					*				*							*	
2				*			*	*						*	*			*						*	*	
3				*					*						*			*	*	*					*	
4				*			*	*	*	*	*			*	*			*							*	
5				*			*	*	*	*	*				*			*							*	
6				*				*	*	*					*			*						*	*	
7				*				*	*						*					*					*	
8				*			*							*	*	*		*							*	

Figure 5: The mining results are represented in a table after aggregating neighbour subgroups. This representation is more compact.

Although the nugget representation, shown in Fig. 5, is more compact than the cell representation, without the visual representation, users still have difficulties understanding the distribution of the nuggets and build connections between the pattern and the instances. To better understand the mining results and further explore them, analysts can open the nugget space view (Fig. 7). Based on the distribution in the nugget view and the cluster view, the high benign rate cluster and the low benign rate cluster are separated from each other in the attribute space, indicating that the target is influenced by the independent attributes. We can also discover that a large red cluster and a large blue cluster are extracted. It is shown that the higher benign rate regions and low benign rate regions are continuous in the independent attribute space. More discoveries found during the exploration in the nugget space are as follows:

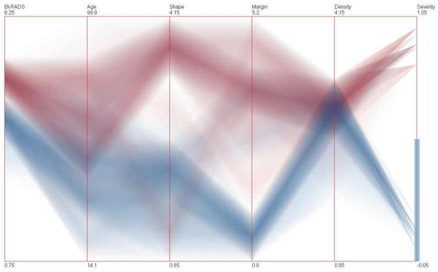


Figure 6: The data space view shows all the nuggets as the translucent bands. The rightmost dimension is the target attribute. The blue vertical region on the target dimension indicates the target range of the subgroup mining query.

1. For the low benign rate subgroups, there are two outliers outside the main cluster. By hovering the cursor and selecting on the two outliers, we can discover what causes the two outliers to differ from the main cluster: the *Shape* values of the main cluster (red) are 3 and 4, while the two outliers have *Shape* value 1. When showing these two outlier subgroup instances in the data space view, we can observe that no instances are benign and the group sizes are small. Thus, the doctors can consider that they are not typical and ignore these two outlier subgroups during analysis.

2. The shape value 4 is more important for the low benign rate. This can be discovered when displaying all the instances in the red cluster: the shape values are either 3 (means lobular) or 4 (means irregular), while for the value 4 , higher significance is found, which can be recognized by a darker color.

3. For lower age patients, higher benign rate tend to be discovered. This can be verified by the distribution of the interesting subgroups: no higher benign rate groups are in age bin 6 and 7; no lower benign rage groups are in age bin 1 and 2.

4. Attribute *BI-RADS* has a negative effect for higher benign rate, i.e., lower *BI-RADS* values tend to have higher benign rate. This can be discovered according to the distribution of subgroups with significance on this attribute. For the higher benign rate subgroups most of them have *BI-RADS* value 4. For low benign rate subgroup: most of them have *BI-RADS* value 5. The analysts can understand this trend better if the know the meaning of this attribute: each instance has an associated BI-RADS assessment. The lowest value means definitely benign and highest value means highly suggestive of malignancy.

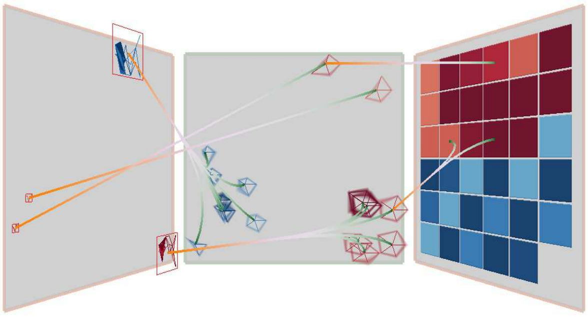


Figure 7: The nugget space view shows the mining result in 3 level of abstractions. The connecting curves indicate the connection between adjacent levels.

Conclusions

In this paper, we describe a novel visual subgroup mining system, called *Nugget Browser*, to support users in discovering patterns in multivariate datasets. We proposed a 4-level layered model that allows users to explore the mining result in different levels of abstraction. The nugget level mining results are represented as regular hyper-box shaped regions, which can be easily understood, visualized, and compactly stored. The layout strategies help users understand the relationships among extracted patterns. Interactions are supported in multiple related nugget space views to help users navigate and explore. The case studies show how our system can be used to reveal patterns and solve real life application problems.

In the future, we plan to extend our system to support more types of mining queries and pattern extraction methods. Furthermore, more complex mechanisms for managing the user's discoveries will be supported, such as adjusting nugget boundaries with domain knowledge and removing highly overlapping nuggets without reducing much accuracy. In addition, building an evidence pool that allows users to create a structured pattern graph with extracted nuggets is also one of our future goals. To better evaluate this system, we plan to conduct a formal user study to confirm how easy it is to learn this system, what types of users can benefit from using this system, what types of visual representations of the nugget space are better, pseudo-3D or purely 2D, and what interesting patterns analysts can have difficulty finding without a visual exploration.

Acknowledgements

This work is supported under NSF grant IIS-0812027.

References

[1] Rakesh Agrawal, Johannes Gehrke, Dimitrios Gunopulos, and Prabhakar Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. In *SIGMOD '98*, pages 94–105. ACM, 1998.

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

[3] Martin Atzmueller. Subgroup discovery. In *Künstliche Intelligenz*, volume 4, pages 52–53, 2005.

[4] Yonatan Aumann and Yehuda Lindell. A statistical theory for quantitative association rules. *J. Intell. Inf. Syst.*, 20(3):255–283, 2003.

[5] Ingwer Borg and Patrick Groenen. *Modern Multidimensional Scaling: Theory and Applications*. Springer, 1996.

[6] Christopher Collins and Sheelagh Carpendale. Vislink: revealing relationships amongst visualizations. *IEEE Trans Vis Comput Graph*, pages 1192–1199, 2007.

[7] Raphael Fuchs, Jürgen Waser, and Meister Eduard Groller. Visual human+machine learning. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1327–1334, 2009.

[8] Zhenyu Guo, Matthew O. Ward, and Elke A. Rundensteiner. Model space visualization for multivariate linear trend discovery. In *VAST '09*, pages 75–82. IEEE Computer Society, 2009.

[9] Ming C. Hao, Umeshwar Dayal, Daniel A. Keim, Dominik Morent, and Joern Schneidewind. Intelligent visual analytics queries. In *VAST '07*, pages 91–98. IEEE Computer Society, 2007.

[10] Patrick Hoffman, Georges Grinstein, Kenneth Marx, Ivo Grosse, and Eugene Stanley. Dna visual and analytic data mining. In *VIS '97*, pages 437–441. IEEE Computer Society Press, 1997.

[11] Alfred Inselberg and Bernard Dimsdale. Parallel coordinates: a tool for visualizing multi-dimensional geometry. In *VIS '90*, pages 361–378. IEEE Computer Society Press, 1990.

[12] J. Edward Jackson. *A User's Guide to Principal Components*. Wiley-Interscience, 2003.

[13] Daniel A. Keim. Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):1–8, 2002.

[14] Willi Klösgen and Jan M. Zytkow, editors. *Handbook of data mining and knowledge discovery, chapter 16.3: Subgroup discovery*. Oxford University Press, Inc., New York, NY, USA, 2002.

[15] Willi Klsgen and Michael May. Spatial subgroup mining integrated in an object-relational spatial database. In *PKDD, T. Elomaa, H. Mannila and H. Toivonen, eds, 6th European Conference*, pages 275–286, 2002.

[16] Teuvo Kohonen. *Self-Organizing Maps*. Springer, 2000.

[17] Ben Shneiderman. Inventing discovery tools: combining information visualization with data mining. *Information Visualization*, 1:5–12, 2002.

[18] R. L. Somorjai, B. Dolenko, A. Demko, M. Mandelzweig, A. E. Nikulin, R. Baumgartner, and N. J. Pizzi. Mapping high-dimensional data onto a relative distance plane: an exact method for visualizing and characterizing high-dimensional patterns. *J. of Biomedical Informatics*, 37(5):366–379, 2004.

[19] UC Irvine Machine Learning Repository. Mammographic Mass Data Set. http://archive.ics.uci.edu/ml/datasets/Mammographic +Mass.

[20] Wei Wang, Jiong Yang, and Richard R. Muntz. Sting: A statistical information grid approach to spatial data mining. In *VLDB '97*, pages 186–195. Morgan Kaufmann Publishers Inc., 1997.

[21] M. Ward. Xmdvtool: Integrating multiple methods for visualizing multivariate data. *Proc. IEEE Visualization*, pages 326–333, 1994.

[22] Pak Chung Wong. Guest editor's introduction: Visual data mining. *IEEE Computer Graphics and Applications*, 19(5):20–21, 1999.

[23] Di Yang, Elke A. Rundensteiner, and Matthew O. Ward. Analysis guided visual exploration of multivariate data. In *VAST '07*, pages 83–90, 2007.

[24] J. Yang, M. Ward, and E. Rundensteiner. Hierarchical exploration of large multivariate data sets. *Data Visualization: The State of the Art 2003*, pages 201–212, 2003.

[25] Chen Yu, Yiwen Zhong, Thomas Smith, Ikhyun Park, and Weixia Huang. Visual data mining of multimedia data for social and behavioral studies. *Information Visualization*, 8(1):56–70, 2009.