IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received September 15, 2017, accepted November 3, 2017, date of publication November 8, 2017, date of current version December 5, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2771237

FRIOD: A Deeply Integrated Feature-Rich Interactive System for Effective and Efficient Outlier Detection

XIAODONG ZHU[1], JI ZHANG[2], (Senior Member, IEEE), HONGZHOU LI[3], PHILIPPE FOURNIER-VIGER[4], JERRY CHUN-WEI LIN[4], AND LIANG CHANG[3]

[1]School of Economics and Management, Nanjing University of Information Science and Technology, Nanjing 210044, China
[2]Faculty of Health, Engineering and Sciences, University of Southern Queensland, Toowoomba, QLD 4350, Australia
[3]Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology, Guilin 541004, China
[4]Shenzhen Gradate School, Harbin Institute of Technology, Shenzhen 518055, China

Corresponding author: Ji Zhang (ji.zhang@usq.edu.au)

This work was supported in part by the Priority Academic Program Development of Jiangsu Higher Education Institutions, in part by the Top-Notch Academic Programs Project of Jiangsu Higher Education Institutions under Grant PPZY2015A072, in part by the National Natural Science Foundation of China under Grant 61503092, in part by the Guangxi Key Laboratory of Trusted Software under Grant kx201615, and in part by the Capacity Building Project for Young University Staff in Guangxi Province, Department of Education, Guangxi Province, under Grant ky2016YB149.

ABSTRACT In this paper, we propose an novel interactive outlier detection system called feature-rich interactive outlier detection (FRIOD), which features a deep integration of human interaction to improve detection performance and greatly streamline the detection process. A user-friendly interactive mechanism is developed to allow easy and intuitive user interaction in all the major stages of the underlying outlier detection algorithm which includes dense cell selection, location-aware distance thresholding, and final top outlier validation. By doing so, we can mitigate the major difficulty of the competitive outlier detection methods in specifying the key parameter values, such as the density and distance thresholds. An innovative optimization approach is also proposed to optimize the grid-based space partitioning, which is a critical step of FRIOD. Such optimization fully considers the high-quality outliers it detects with the aid of human interaction. The experimental evaluation demonstrates that FRIOD can improve the quality of the detected outliers and make the detection process more intuitive, effective, and efficient.

INDEX TERMS Outlier detection, space partitioning, human interaction, visualization.

I. INTRODUCTION

Outlier detection has been an important research problem in data management, particularly in the the areas of data mining and knowledge discovery. It aims to detect those data or objects from the given data source which, when compared with the major population of the data source, exhibit significantly abnormal, inconsistent or suspicious patterns. Outliers are usually important, or even critical, objects for the applications involved which requires special attentions or actions from users. Given its inherent importance, the problem of outlier detection has been intensively studied for the past decades and has enjoined a wide range of important applications such as network intrusion detection, environmental monitoring, financial and telecommunication fraud detection, to name a few.

The problem of outlier detection first originated from the area of statistics where outliers are primarily detected using statistical approaches. A certain data distribution needs to be assumed for the normal data in the data source and outliers are defined as those data which clearly do not fit in the assumed data assumption. This type of methods work reasonably well for simple, small datasets, but quickly suffers a performance degradation when the data scale increases. As the volume and dimensionality of data increases, the traditional statistical approaches quickly become insufficient to deal with this problem efficiently and effectively. Consequently, most of the recent research in outlier detection focuses on investigating various detection mechanisms from, for example, the distance and/or density perspectives, to model and detect outliers more effectively and efficiently.

In the era of big data when the volume and complexity of data are increased at an unprecedented rate, user-friendly human interaction together with the use of data visualization have been very useful to understand and interpret the data in question. They are capable of offering effective and efficient support to various data analytical tasks including outlier detection. Despite the intensive research work on outlier detection we have witnessed in literature during the past several decades, there is much less attention being paid on integrating human interaction as an effective means to assist and improve outlier detection.

In this paper, we propose FRIOD, an innovative outlier detection system. The technical contributions of FRIOD are summarized as follows.

- FRIOD integrates a rich set of interaction features in all the major stages of the underlying outlier detection algorithm it uses, contributing to its promising effectiveness and efficiency;
- Human interaction effectively helps FRIOD mitigate the long-standing difficulty of the existing outlier detection methods, especially in specifying appropriate values of the key parameters such as the density and distance thresholds;
- A novel approach is also proposed in FRIOD to optimize the grid-based space partitioning which fully leverages the good outliers detected by itself thanks to the human interaction integrated;
- The last but not least, the experimental evaluation results demonstrate that, through a deep integration of human interaction in FRIOD, the outlier detection process can be greatly streamlined and the detection performance can be improved noticeably when compared with the existing interactive and non-interactive outlier detection methods.

The remainder of this paper is organized as follows. In Section 2, we discuss the related work on outlier detection, covering both the traditional outlier detection methods as well as those integrating features of visualization and/or human interaction. Section 3 presents the basic outer detection algorithm that FRIOD uses. The rich set of human interaction features integrated into all the major stages of the basic outlier detection algorithm of FRIOD are elaborated in Section 4. The experimental results are reported in Section 5. The final section concludes this paper and highlights some future research directions.

II. RELATED WORK

There has been a rich body of research work conducted in the area of outlier detection. Depending on the mechanisms used for modeling data abnormality, the existing research work can be broadly categorized into distribution-based methods, distance-based methods, density-based methods and clustering-based methods. Distribution-based methods detect outliers by assuming a pre-determined distribution or probability model to fit the given dataset [3], [11]. Outliers are those data that significantly deviate the underlying model of the data. To improve the scalability of distribution-based methods for handling large datasets, distance-based methods use distance-based metrics to quantify the proximity between each data point and its neighborhood such as the nearest neighbors or dense regions/clusters [6], [7], [37], [39]. Those data points that are far from their respective neighbors are considered as outliers. Density-based methods use more complex mechanisms to model the outlier-ness of data points than distance-based methods [16], [19], [31], [35], [36], [40]. They usually involve investigating not only the local density of the data being studied but also the local densities of its nearest neighbors. Because of the close relationships between data clusters and outliers, clustering analysis can also be performed to assist the detection of outliers by defining outliers as data that do not lie in or located far apart from any clusters [2], [12], [21], [29], [30], [41].

Detecting outliers from increasingly large datasets is a very computationally expensive process. To improve the efficiency performance of outlier detection, a grid structure can be created through a space partitioning that discretizes each continuous attribute to a few intervals. Using the grid structure can considerably reduce the computational overhead as the major operation of detection is now performed on the grid cells which is typically of a much smaller number compared to the total number of data instances in the dataset. This makes the detection process much more scalable to datasets with a large number of instances. In addition, the grid structure can greatly facilitate the calculation of data synopsis to capture data distribution and characteristics for the purpose of outlier detection. Some related grid-based outlier detection and clustering methods (which can assist outlier detection) including DISTROD [31], the sparse cube search method [1], SPOT [32], Grid-k-Means [9] and Grid-DB [4]. Nevertheless, these methods are not equipped with interactive mechanism in any stage of their detection process, which limit their efficiency and effectiveness for outlier detection that may be achieved otherwise.

Although have being relatively sophisticated in the mechanism and procedure for detecting outliers, the existing outlier detection techniques mostly lack the facilities to support human interaction to effectively assist the detection process. This is somehow to our surprise since there has been little research in the area of interactive outlier detection, compared to the depth and width of the research in outlier detection that we have witnessed for decades. Furthermore, most of the existing interactive outlier detection methods provide very limited support for human interaction in the outlier detection process. Typically, they merely incorporate visualization and minimum human interaction on the final outliers detected. In [27], the final detected outliers are classified as *explainable* and *unexplainable* outliers. Those expendable outliers are removed immediately whereas the unexplainable ones will be further examined by human users. 2D and 3D visualization tools are developed to visualize the detected subspace outliers which are embedded in the low dimensional subspaces with two or three dimensions [28]. A few different types of view on the final outliers are presented by *VSOutlier* [23] which shows the outliers based on a query, displays a visual comparison of the qualified outliers of different queries and monitors the key performance metrics of the outlier detection algorithms. An outlier detection method incorporated with user feedback was proposed in [33], which allows users to decide the suspicious objects which are not directly classified as outliers by the system but each features a relatively high outlier-ness score. Visualization and interaction are also provided for feature selection using the evolutionary algorithm for subspace outlier detection [25], [26]. An exploration and visualization method for outlier detection was also proposed for dealing with log data [24] but it lacks the generality that cannot be directly applied to other types of data.

SODIT [34] is a recently proposed interactive outlier detection system which features mechanisms of human interaction inside the detection process, rather than only on the final detection result as in the above-referenced work. Interactive interfaces were developed to support the selection of dense regions of the dataset and distance calculation. It also mitigates the problem of using a single universal distance parameter for the whole dataset (as the case in many other methods) and introduced the concept of localized thresholding. Yet, it only provides some preliminary features of human interaction and suffers the following several major limitations: 1) The detailed data of the whole dataset are used to visualize the dense regions selected by users when they are scrolling through the bar-like control on the interface to select the optimal density threshold. This may be very slow and could seriously affect the experience and efficiency of human interaction in selecting the dense regions; 2) The coefficient of the distance threshold used in SODIT is not location aware. As the result, an appropriate value of the coefficient for one region of the dataset may not be appropriate for other regions. This may adversely affect the accuracy of outlier detection; 3) SODIT doesn't provide the advanced features such as the final outlier visual validation and the optimization of the space partitioning.

In summary, a deep integration of human interaction in supporting efficient and effective outlier detection hasn't yet been adequately addressed in the current literature and FRIOD is developed aiming to fill this research gap. FRIOD, the interactive outlier detection system proposed in this paper, can effectively solve the limitation of the system proposed by [34]. It improves the efficiency and accuracy of the selection of dense regions formed by the dataset and enhances the effectiveness of the location-aware distance thresholding. Besides that, our system offers additional important features such as interactive validation of the final detected outliers with advanced learning capacity and the optimization of the grid-based space partitioning.

III. THE BASIC OUTLIER DETECTION ALGORITHM

We first introduce the basic outlier detection method used in FRIOD, which serves as an ideal algorithmic framework for an deep integration of human interaction.

As the pre-processing step in FRIOD, the data space of the given dataset is undergone grid-based space partitioning which involves superimposing a grid structure into the data space under study. This partitioning results in a number of cells being created in the grid structure and each data in the dataset is mapped into one and only one cell. In FRIOD, we choose to utilize the grid-based equal-width space partitioning that partitions each dimension into intervals with an equal width. Compared to the alternative equal-depth partitioning method that partitions each dimension into a number of intervals such that each contains an equal number of data points, equal-width space partitioning is more advantageous in that it offers a more spatially balanced partitioning of the data space involved and is much easier and more efficient to implement than the equal-depth space partitioning and, therefore, contributes to the better efficiency for outlier detection.

In FRIOD, each dimension is partitioned proportionally into intervals with an equal width, meaning that the number of intervals generated for the dimension is in the right proportion to its range. Take the case of a two-dimensional data space for example which is represented by X and Y axis, we have

$$\frac{g(X)}{g(Y)} = \frac{Range(X)}{Range(Y)} \tag{1}$$

where $Range(X)$ and $Range(Y)$ can be calculated based on the minimum and maximum values in the dataset as

$$Range(X) = Max(X) - Min(X) \tag{2}$$

and

$$Range(Y) = Max(Y) - Min(Y) \tag{3}$$

For the ease of presentation, we only specify the granularity for X axis as $X = g$ in the rest of this paper. The granularity for Y axis can be obtained proportionally.

Once data partitioning is completed, the basic outlier detection algorithm is ready to perform, which will take the following several steps:

1) Each data in the input dataset is mapped into one and only one cell in the grid. Cell density is calculated and maintained for all the populated cells. That is, when a data point is assigned to a cell, then the density of this cell will be incremented by 1;
2) The populated cells will be ranked based on their density in a descending order and a specific number of most dense cells selected whose total number of data points have exceeded $r\%$ (for example 80%) of the number of data in the whole dataset. These cells are called *dense cells*. The centroids of these dense cells are called *representative data*. The data points that do not fall into the dense cells are called *outlier candidates* that require further evaluation in order to detect true outliers from them;
3) Using the extracted representative data, the outlier score of each outlier candidate will be calculated which is defined as the distance between each data and its nearest representative data;
4) The top n outlier candidates which have the highest outlier score are returned to users as the final result. The value of n is specified by end users which reflects

their requirement as to how many top outliers they are seeking.

Please note that, in the grid-based algorithm, the total number of cells will grow exponentially with regard to the partitioning granularity (i.e., the numbers of intervals for each dimension). Nevertheless, the actual number of populated cells only grows modestly with regard to the partitioning granularity. This ensures the efficiency of the outlier detection.

Thanks to the grid-based space partitioning which significantly reduces the computational complexity, the above basic outlier detection algorithm is highly efficient. It also serves as a very good platform for integrating deep human interaction to assist outlier detection in every stage of the algorithm, which will be detailed in the next section.

IV. INTEGRATION OF HUMAN INTERACTION IN FRIOD

In this section, we first present an overview of the architecture of FRIOD, followed by detailed discussions on the rich set of interactive features of FRIOD in all its major stages of outlier detection. The space partitioning optimization is also presented which considers the good set of outliers detected interactively by FRIOD.

Thanks to its interactive nature, FRIOD, like the existing interactive outlier detection systems, is very effective for outlier detection from datasets with two or three dimensions such as the spatial databases. For the datasets with more than three dimensions, there can be two scenarios where FRIOD can be applied for outlier detection. First, FRIOD can be used to carry out the detection of so-called *subspace outliers* from the given dataset. Due to the curse of dimensionality, outliers can only be detected in those low dimensional data spaces, many of which have two or three dimensions. Second, we can perform dimensionality reduction or feature selection to reduce its dimension to two or three in order to use FRIOD for interactive detection. Based on the identification information of data in the dataset, the detected outliers can be re-mapped to its original dimensionality if necessary at the end of the detection process. These two strategies ensure the general applicability of FRIOD in handling datasets with varying dimensions.

A. AN OVERVIEW OF FRIOD

FRIOD is an innovative interactive outlier detection system with deeply integrated human interaction modules to provide a rich set of interactive features for outlier detection. An overview of the system architecture of FRIOD is presented in Figure 1 where the interactive functional modules are particularly highlighted in the orange color. Those modules include dense cell selection, location-aware distance thresholding, final top outlier validation and grid-based space partitioning optimization.

In FRIOD, users are able to get instant feedback through visualization at the end of each stage regarding how well this stage has been performed towards producing a good outlier detection result in the end. This allows users to timely adjust the values of the key parameters of the algorithm as early as possible in the detection process to enable FRIOD to detect outliers in a more efficient and effective manner.

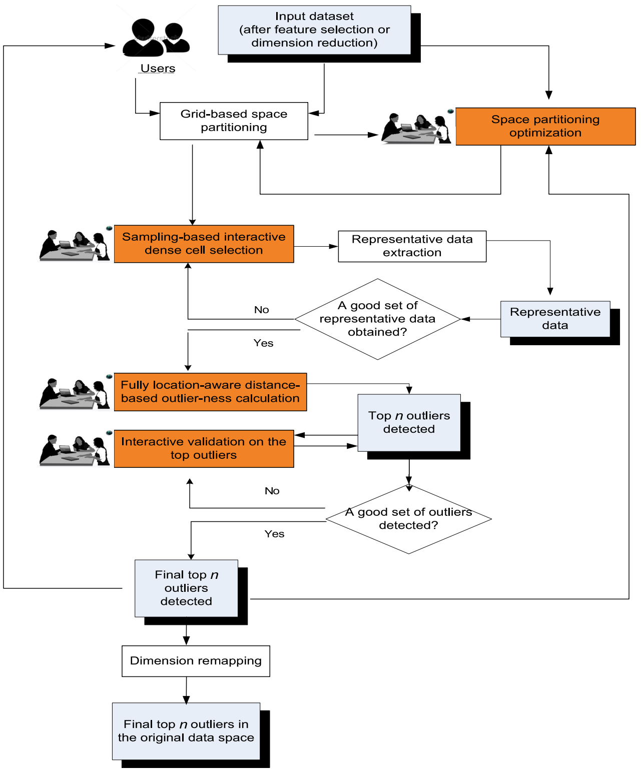


FIGURE 1. System architecture of FRIOD.

In real-life situations, users can choose to engage in all or some of the three interactive stages in FRIOD depending on their interests and availability of time. This can make FRIOD more flexible and adaptive to different situations and users. Based on our experience, if users can be involved in the earlier stages, the workload in the later stages, such as final outlier validation, can be somehow reduced due to the higher quality of outliers detected earlier in the process.

B. DENSE CELL SELECTION

Identifying dense cells is an important early step in FRIOD. Those dense cells are used to generate the representative data and calculate the outlier-ness score for each data point. Nevertheless, the term "dense" is a relative, subjective term which needs to be precisely defined, but this is not easy. The basic outlier detection algorithm defines the dense cells as those high-density cells whose total density reaches a certain fixed percentage (e.g., 80%). However, this definition may not be universally accurate for different datasets in question.

Human interaction is very helpful to provide an effective visual aid to determine the dense cells in the grid. In order to facilitate this process, we designed on the user interface a bar-like controller where users can scroll through to specify an appropriate density threshold. To enhance the user experience, whenever he releases the click of the mouse on the scrolling bar, FRIOD will immediately apply the corresponding selected value as the density threshold without the need to click another bottom to start this process. This simple design is quite effective in giving users a real-time feeling of the system and making the density threshold selection more efficient and streamlined.

SODIT, an existing interactive outlier detection system, also leverages human interaction to specify the dense cells. Nevertheless, for each selected density threshold, the whole dataset has to be loaded for visualization in SODIT. This renders this step very slow in practice when we are dealing with large datasets and compromises the user experience and efficiency in selecting dense cells. It is possible to build an index of the dataset for a speedup. However, such an index (which maintains the information regarding the cell which each data in the dataset belongs to) will incur a high computational and space overhead given the possibly very large size of the dataset. In addition, the workload of color coding the data points in the selected dense regions is almost equivalent to dealing with the whole dataset.

To solve this problem, we develop in FRIOD a more efficient mechanism for improving the real-timeliness of the human interaction in selecting dense cells. Instead of using the detailed data for visualization under each specified density threshold, we choose to display the dense cells of a sample generated from the original dataset for human inspection when users are scrolling through the bar control to tuning the value of the density threshold. A theorem has been derived by Guha *et al.* [5] to determine the minimum sample size required to ensure that a fraction of the cluster is always included in the sample with probability δ, which is ideal for the generation of dense cells and the extraction of representative data in FRIOD. Specifically, for a cluster u, if the sample size s satisfies

$$s \geq fN + \frac{N}{|u|} log(\frac{1}{\delta}) + \frac{N}{|u|}\sqrt{(log(\frac{1}{\delta}))^2 + 2f|u|log(\frac{1}{\delta})} \quad (4)$$

then the probability that the sample contains fewer than $f|u|$ points belonging to cluster u is less than σ, where N is the size of the dataset, $|u|$ is the size of the cluster u, $0 \leq f \leq 1$, $0 \leq \delta \leq 1$. FRIOD uses this theorem to determine the sample size and performs uniform sampling on large datasets to obtain a smaller sample. This theorem gives an insight on the minimum size of the randomly generated sample in order to ensure the representativeness of the sample. In FRIOD, we treat the data in each grid cell as forming a micro cluster in order to apply the above sampling theorem. This is important in our work to correctly visualize the dense regions formed by the dataset involved.

This sampling approach effectively waivers the need to carry out a possibly expensive clustering operation on the dataset. In the case that the sample generated is still be too large to fit entirely into the main memory, FRIOD can divide the sample into several smaller partitions, each of which can be loaded into the main memory sequentially for processing. This makes FRIOD flexible and yet effective in handling samples of all sizes.

This sampling approach is also very efficient as the density information of all the populated cells in the grid has already been obtained prior to the selection of dense cells. The sample size can be much smaller than that of the original dataset. The complexity of this interaction under each specified density threshold value is $O(|C_p| + s)$, where $|C_p|$ and s are the total number of the populated cells in the grid and the size of the sample, respectively, which are both considerably smaller than the number of data points in the dataset. Furthermore, this interactive process doesn't require any indexing to be built.

After a good set of dense cells have been selected using the data sample, users can optionally proceed to conduct the final validation by loading the whole dataset. This only requires at most one scan of the whole dataset. The data points in the dataset can be read into the main memory for processing sequentially like a data stream, thus this step can be completed with ease under system platforms with varying memory constraints.

To further assist the selection of dense regions on the cell level and improve its accuracy, we also use the *density plot*, presented in Figure 2, to display the density of all the populated cells in the grid and highlight the cells and their density in the plot in a real-time manner when they are selected by users. This can effectively provide a visual assistance to users to better understand the density transition from the densest cells to the sparsest ones in the selection process. This plot only includes the density information of the populated cells and does not include any empty cells as they are not involved in the dense cell selection process. The cells in the plot are sorted in a descending order based on their density, with the high-density cells being positioned on the left of the plot. Please note that the density of the last four cells (with the cell IDs of 13, 8, 15 and 6) in Figure 2 are not visible from the plot as their density are very low ranging from 0.01-0.06%. The cells in the density plot are synchronized in a real-time fashion with the selected dense cells and are highlighted using a darker color in the plot. Users can crosscheck the density of the cells that have been selected, together with their total density, through this plot to make sure that no dense cells have

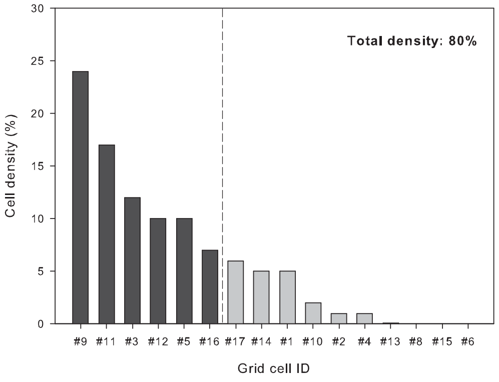


FIGURE 2. Density plot of populated cells.

been left out or excessively sparse cells have been included in the process. This can greatly facilitate the selection of good set of dense cells. Based on our experience, the selection of dense cells using a sample of the original dataset, coupled with an optional validation based on the whole dataset and the use of the density plot, is well adequate for accurate dense cell selection.

C. LOCATION-AWARE DISTANCE THRESHOLDING

Zhang *et. al* introduced the concept of localized distance thresholding in SODIT to consider the possibly different data characteristics in different regions formed by the dataset in question [34]. It quantifies the standard deviation for the data in populated cells and a data is labeled as an outlier if the distance from itself to its nearest representative data is q times as the standard deviation of the data in the cell where the nearest representative data is located. The coefficient q typically takes a value in the range of (1.5, 3). The major problem of this method is that whenever correction is made on the detection result through human interaction, the value of q might be changed as well. Unfortunately, the scope of this change is *global* in SODIT which will not only affect the current local region where the corrections made but also all the other regions of the dataset. A possible unwanted consequence of this global adjustment of the value of q is that it may adversely affect the correct detection results in other regions.

Our idea in FRIOD to solve this problem is to not only utilize the statistical information of different regions but also render q itself location aware. Under this principle, the adjustment of q is restricted only to the local region where the correction is performed and, therefore, will not adversely impact outlier detection results in other regions.

For a data point p whose labeling needs to be corrected in this step, the locally affected data points which require a re-evaluation in FRIOD include both those in the same cell of p and those in other cells which share the same nearest representative data as p. Mathematically, the set of all the affected data points can be presented as follows:

$$CorrectionSet = \{p_i | p_i \in cell(p)\} \cup \{p_j | nearestRep(p_j) = nearestRep(p)\} \quad (5)$$

Depending on how the labeling of p is corrected, the value of q will be updated in one of the following two ways:

1) p should be an outlier but incorrectly labeled as a normal data point by FRIOD. In this case, the value of q needs to be decreased as

$$q = \frac{dist(p, nearestRep(p))}{SD(nearestCell(p))} - \epsilon \quad (6)$$

where ϵ it is a small constant in the range of (0, 1) (such as $\epsilon = 0.1$) used for a minor value adjustment for q and $SD()$ calculates the standard deviations of the data in a cell;

2) p should be a normal data point but incorrectly labeled as an outlier by FRIOD. The value of q has to be increased in this situation as

$$q = \frac{dist(p, nearestRep(p))}{SD(nearestCell(p))} + \epsilon \quad (7)$$

After q is updated, the labeling of the set of affected data will be re-evaluated automatically and updated if necessary using the new value of q. This process is very efficient as typically only a (very) small number of data points need to be re-evaluated. By making both the coefficient and the standard deviation location aware, we are able to effectively solve the limitation of SODIT and make the human interaction process more effective.

D. FINAL TOP OUTLIER VALIDATION WITH LEARNING CAPACITY

Validating the correctness of outliers through visualization is fairly straightforward, intuitive and accurate. Human perception excels in identifying the dense data regions as well as outliers. Given the relatively small value of n (the number of top outliers sought) in most scenarios, validating all the top outliers produced by FRIOD is well manageable. In case that a relatively large number of outliers are requested by human users, FRIOD can request human attentions to verify only those weaker outliers in the top n list (i.e., the margin outliers), which feature relatively smaller outlier-ness scores. The stronger outliers in the top n list can be usually detected by FRIOD very accurately, thus they are not the top priorities for human validation. Users can also choose the percentage of the weak outliers to validate through our system easily based on his or her availability. In this validation process, if users think one particular outlier should not be in the top list, they can exclude it from the list and the $(n+1)^{th}$ strongest outlier will be added into the list. When outliers are validated, users only focus on visually evaluating whether they are outliers or not. Therefore, outlier validation, in theory, can be carried out simultaneously by multiple users if necessary to make this process more efficient.

We also developed a learning module to train FRIOD on the corrected top outliers in the human validation process. Several major characteristic features are captured for each corrected outlier, if any, which are archived in FRIOD. For a corrected outlier o_c, the characteristics that are captured include the normalized density of the cell where o_c is located, the normalized density of the cell where the nearest representative data of o_c is located, the distance between o_c and its nearest representative data, and finally the average distance of the other data points in the same cell of o_c, if any, with respect to their own nearest representative data. These features captured describe the density and distance characteristics of each corrected outlier and its neighborhood. This training process can be performed off-line, which will not interfere with the human interaction process in FRIOD. The same set of features of the future top outliers can be compared with those of the previously corrected outliers. If a future outlier features has a high similarity with one or more of the corrected outliers archived, then it will be assigned a higher priority for human validation.

E. GRID PARTITIONING OPTIMIZATION

It has been well known that an improperly specified granularity for space partitioning may lead to a significant degradation of the detection performance for various grid-based data analytical methods (including FRIOD). On one hand, if the granularity is too small, the density of the cells where the normal data are located get increasingly close to that of the cells which contain outliers. On the other hand, if the granularity is too big, the outliers will be assigned to some high-density cells containing a large number of normal data, making them undetectable from the normal data. Specifying the appropriate granularity for space partitioning has become a long-standing problem for the existing grid-based data analytic methods. They lack this important feature and rely entirely on users to specify a predetermined, fixed granularity value for space partitioning.

Based on our observations, the granularity value in the middle ground of its reasonable range can generally produce better detection result than the extreme values on both ends of the spectrum. Thus, the aim of such optimization is to achieve a good granularity for space partitioning which is closer to the *middle ground* in the range while, at the same time, achieving the same or very similar set of the top n outliers obtained in the previous detection round with the aid of human interaction. In other words, the optimization is carried out by leveraging the correctly identified outliers achieved by the use of human interaction. The optimized space partitioning can be applied for future datasets if they are believed to have the same or similar distribution with the current dataset based on which the optimization is conducted.

Based on the top n outliers produced under a granularity value, the goodness of other possible granularity values has to be quantified. To this end, we use the metric of accuracy which is defined as the percentage of accurately detected outliers among the top n outliers returned by the system under a given granularity value. Mathematically, we have

$$Accuracy(g') = \frac{|topSet(g) \cap topSet(g')|}{n} \times 100\% \qquad (8)$$

where $topSet(g)$ and $topSet(g')$ are the top n outliers returned by FRIOD under the granularity of $x = g$ (g is the granularity value used in the previous iteration) and $x = g'$ (g' is another granularity value under evaluation), respectively. $||$ returns the cardinality (i.e., the number of elements) of a set. Among those top (e.g., 10%) granularity values which have the highest accuracy, we choose their *median* as the optimized granularity value. This allows users to choose the granularity of partitioning which is closest to the middle ground of the spectrum of granularity.

The above optimization routine can be performed continuously until the improvement of detection performance becomes negligible for the newly optimized granularity value, i.e., a convergence is achieved. A good feature of our space partitioning optimization method is that the process can be converged when multiple, if not one, optimization iterations are performed. This ensure that even under some poorly chosen granularity to start with, we still can, after several iterations of optimization, obtain the optimal or close to optimal granularity for space partitioning. This desirable phenomenon has been demonstrated by our convergence experiment. The details of the result is presented later in Section 5.

As mentioned earlier, the optimized space partitioning can continually be used for other datasets believed to have the same or very similar characteristic with the current one. At any point of time, users can also choose to opt out the optimized granularity if the subsequent dataset to be processed is believed to have significantly different characteristics. This can be achieved based on domain knowledge or with the help of the existing methods for detecting concept drifts in datasets.

Please note that the optimization process is performed entirely automatically without any involvement of human interaction. Therefore, it can be executed in an off-line manner. Depending on the number of different granularity configurations to be evaluated and the size of the datasets to be processed, this optimization process sometimes can be time-consuming. To mitigate this issue, we develop a method that uses progressive sampling in the optimization. We start with a small sample for all the granularity values at the beginning and then gradually increase the sample size for those granularity values with a good performance. This is able to help achieve a balance between speed and effectiveness for the optimization. Users can decide how the sample size is increased as the optimization progresses. In FRIOD, the initial sample size is 10% of the original dataset and the subsequent sample size is increased by 10% whenever the number of the remaining granularity values for further evaluation is decreased by 10%. Given this design, the relationship between the number of granularities for further evaluation and the sample size at different stages can be described (both in percentage) as $GranularityLeft + SampleSize = 110\%$.

F. PSEUDOCODE OF FRIOD

After introducing the basic outlier detection algorithm and all the interactive features that FRIOD use, we now present the pseudocode of FRIOD (presented in Figure 3) to give a big picture of the system. The three functions appearing in the pseudocode, *denseCell*(), *locationAwareDist*() and *validate*(), represent the three interactive strategies integrated in FRIOD.

V. EXPERIMENTAL RESULTS

We carried out extensive experiments to evaluate the performance of FRIOD, with a focus on the contribution of various interactive strategies on its performance improvement in outlier detection. The results are reported in this section.

To cover datasets with possibly varying characteristics and distributions, both synthetic and real-life datasets were used for our experimental evaluation. To facilitate more efficient evaluation, two-dimensional synthetic datasets were generated by our synthetic data generator. The advantage of our

```
Algorithm FRIOD $(D, n)$
Input: Dataset $D$ and number of top outliers returned $n$;
Output: Top $n$ outliers detected from $D$;
1. $RepSet = \emptyset$; /* $Repset$ denotes the set of representative data*/
2. $topOutliers = \emptyset$;
3. Feature selection on $D$ if necessary;
4. Grid space partitioning optimization if necessary;
5. Superimpose a grid structure to data space;
6. $RepSet \leftarrow denseCell(D)$;
7. IF $RepSet$ is satisfactory
8.   THEN Goto Step 10;
9.   OTHERWISE Goto Step 6;
10. $topOutliers \leftarrow locationAwareDist(D, k)$;
11. $validate(topOutliers)$;
12. IF $topOutliers$ is satisfactory
13.   THEN Goto Step 15;
14.   OTHERWISE Goto Step 11;
15. Dimension remapping if necessary;
16. Return($TopOutliers$);
```

FIGURE 3. Pseudocode of FRIOD.

synthetic data generator is that we can easily control the percentage of normal data and outliers in the resulting dataset. Figure 4 presents the five synthetic datasets used. The colored regions refer to various clusters of data in the dataset while the black dots are outliers which are located in low-density areas and far from the data clusters. In addition, two multi-dimensional datasets from UCL machine learning repository, i.e., *Letter Image* and *Musk*, were also used in the evaluation. For each real-life dataset, we performed dimension reduction by randomly selecting three different attribute pairs which generated three two-dimensional datasets. In this way, we produced a total of 11 two-dimensional datasets (considering both synthetic and real-life datasets) for the evaluation. To facilitate the evaluation, the true top n outliers were first obtained as the ground truth for all the datasets by runing FRIOD by several experienced users to generate the detection results which were agreed unanimously by all of them.

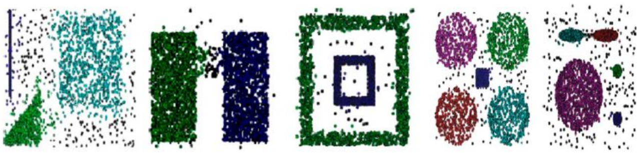


FIGURE 4. Synthetic datasets.

We recruited 20 postgraduate students to participate in the evaluation as human interaction plays a vital role in FRIOD. The necessary mechanism has been implemented in our study to ensure a good selection of students being selected to participate in the study. Prior to using FRIOD, they did not have any knowledge about the datasets that are involved in the evaluation or the optimal parameter values that should be applied. They haven't had used FRIOD before either. Information sessions were organized before the commencement of the study to get the students familiarized with the basic knowledge of outlier detection, along with the ideas of all the outlier detection methods to be evaluated in the study. Afterwards, all the students have gone through an evaluation process which involves evaluating their knowledge about outlier detection through a written quiz. A post-study survey was also conducted which asked the students about their experience in this study including whether they have tried their best to produce the best possible outlier detection results within the shortest possible time to ensure their evaluation is fair and accurate. Please note that we do not let the participating students use all the outlier detection systems involved in the study before the evaluation began in order to simulate the scenario that the systems are used by new users who haven't have any experience with them beforehand.

The tasks that each participating user is required to complete involve detecting the top n outliers from all the given datasets using a number of different outlier detection methods which include: FRIOD (our proposed method with a full suite of human interaction features), the basic detection algorithm used by FRIOD (presented in Section 3 of this paper which doesn't have any support for human interaction), SODIT (the outlier detection system with some preliminary interaction features. Two variants of SODIT were evaluated, i.e., SODIT equipped with and without the optimized space partitioning technique proposed in our work), DB-Outlier [37], kNN-Outlier [39], LOF [38] (three arguably the most popular outlier detection methods) and finally Grid-Outlier [40] and k-Means-Outlier [41] (two recently proposed outlier detection methods). The aforementioned outlier detection methods selected in our study for evaluation represent a good mixture of methods. Both the interactive and non-interactive methods as well as the widely used methods versus the recently proposed ones have been studied. In our evaluation, the comparison amongst FRIOD, the basic algorithm used by FRIOD and SODIT are considered as the *internal evaluation* because the three methods use a similar set of parameters, while the comparison between FRIOD and other non-interactive methods are considered as the *external evaluation* as they are quite different in terms of their outlier detection mechanisms and the parameters used.

A total of three values of n are considered in the evaluation, i.e., $n = 5$, $n = 15$ and $n = 30$. For the non-interactive outlier detection methods, only the visualization of the final detected outliers is available for assisting users to potentially improve the parameter values. The participating users can execute multiple rounds of those non-interactive methods until the satisfactory detection results are achieved. For SODIT and FRIOD, the users were required to provide their interaction as soon as they possibly can in order to produce an accurate measurement of the time involved in using the systems.

We carried out two types of evaluations in our study. We first carried out an objective, quantitative evaluation to evaluate the efficiency and effectiveness of FRIOD. We also conducted a subjective, qualitative evaluation on FRIOD through a survey on the participating users.

All the methods involved in the study are implemented using C/C++ and Microsoft Visual Studio on desktop computers configured with Intel I7 processor with 8G of RAM.

A. QUANTITATIVE EVALUATION

In the quantitative evaluation through simulation experiments, we compared FRIOD with all the other competitive methods in terms of the accuracy of the detected outliers and the elapsed time the detection process takes.

1) DETECTION ACCURACY COMPARISON

The quality of outliers is measured by detection accuracy that is based on the ground truth produced by experts. It is defined as the percentage of the outliers accurately detected in the top n outliers by an outlier detection method when compared with the ground truth result. Its mathematical definition has been introduced in Section 4.5. It is worthwhile pointing out that, since we only evaluate the final detection result based on a specific number of the top outliers detected, thus the metric of accuracy we define here is identical to precision and recall, two commonly used metrics in the information retrieval domain for performance evaluation. To minimize the bias, the accuracy performance of each method is averaged across all the users on all the datasets.

The internal experiment compares the accuracy performance of FRIOD, the basic detection algorithm and SODIT under different numbers of runs the algorithms are executed. Figure 5 present the accuracy comparison of the three methods. Only one run of the outlier detection algorithm is performed in FRIOD and SODIT because both can produce a relatively good set of outliers in a single run thanks to the human interaction incorporated. Due to this reason, only a single horizontal line is used to present the accuracy performance for each of them in Figure 5. With the aid of human interaction in FRIOD and SODIT, the accuracy of the detected outliers is noticeably higher than the basic algorithm without any human interaction. The quality of detected outliers in the basic algorithm is gradually improved as more runs of the algorithm are performed, but the accuracy is still inferior to that of the interactive counterparts. Furthermore, among the two interactive methods, FRIOD outperforms SODIT in terms of accuracy even when it is equipped with the grid space partitioning optimization mechanism developed in this work. By making the distance coefficient q location aware and incorporating the interaction for the final outlier validation, FRIOD is generally more accurate than SODIT. Without the grid space partitioning optimization mechanism, the accuracy performance of SODIT is much worse than that of FRIOD.

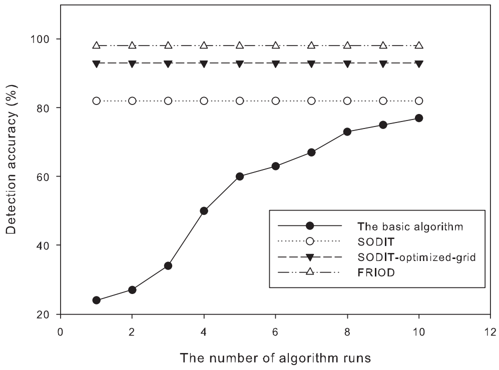


FIGURE 5. **Internal detection accuracy comparison.**

We also evaluated the accuracy of FRIOD under different combinations of the interactive strategies. In this experiment, we use 1, 2 and 3 to represent the three interactive strategies, namely dense cell selection, location-aware distance thresholding and final top outlier validation, and ? is used here as a placeholder symbol to indicate the corresponding interactive strategy is not used in the outlier detection method, which can be used in the first, second or third location in the string. For instance, 12? means that only the first two strategies are used while the third one is left out. ??? represents the scenario where none of the interactive strategies are utilized in FRIOD, which effectively reduces FRIOD to the basic detection algorithm. We can see from the results, presented from Figure 6, that the interactive strategy employed in the dense cell selection is the most important one in improving the accuracy of FRIOD, followed by the location-aware distance thresholding, while the final top outlier validation contributes the lowest portion in accuracy enhancement.

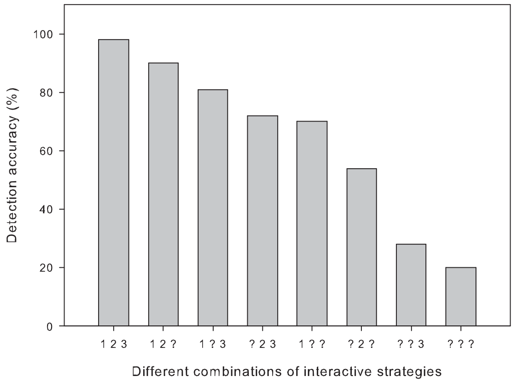


FIGURE 6. **Detection accuracy of FRIOD when using different combinations of interactive strategies.**

The external experiment was also conducted to compare the accuracy of FRIOD with other non-interactive methods. The result is presented in Figure 7. It shows that the pro-

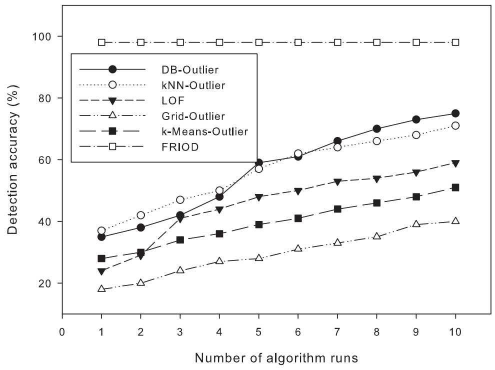


FIGURE 7. **External detection accuracy comparison.**

gression of accuracy shares the same pattern with that of the basic detection algorithm observed in the internal evaluation. This is simply because that none of them have any interactive features incorporated in any stages of the detection algorithm. Users need to execute multiple rounds of the algorithms of non-interactive methods to gradually pick up better values of parameters used in the method, resulting in a gradually improving accuracy performance. However, their accuracy performance is still inferior to that of FRIOD. Looking at the accuracy performance of the non-interactive methods themselves, we can further observe that they differ from each other in terms of how the detection accuracy is improving when more rounds of the algorithms are executed. Among them, DB-Outlier and kNN-Outlier enjoy a faster improvement of accuracy compared to other methods. A possible explanation of this phenomenon is due to the fact that those two methods are simple and they use parameters which can be better tuned by users in the process.

2) ELAPSED TIME COMPARISON

We evaluated the elapsed time of FRIOD and the other competitive methods under different values of n. For a fair comparison, we do not consider the time taken in space partitioning optimization for FRIOD.

In the internal evaluation, we evaluated the three methods under the same granularity value ($G_x = 15$), averaged across all users on all the datasets used. The elapsed time comparison is presented in Figure 8. We can see from the result that FRIOD is more efficient than SODIT even when it is equipped with the grid space partitioning optimization mechanism developed in this work. This is mainly due to the following two reasons. First and most importantly, FRIOD visualizes the selected dense cells through a smaller data sample, rather than the whole dataset, under each selected density threshold. Second, the tuning of the distance coefficient q through human interaction in FRIOD is more convenient and efficient than SODIT. Both FRIOD and SODIT feature a considerably shorter elapsed time than the basic detection algorithm. This is very interesting and surprising to us at first as we know that the elapsed time will surely be increased dramatically by considering the time involved in human interaction. A further investigation reveals that this is because the users typically need to run many rounds of the basic algorithm when human interaction is not allowed in order to gradually pick up the optimal values of the algorithm parameters in this process before the final satisfactory detection result is produced. In comparison, users only need to go through one run of the whole algorithm in both FRIOD and SODIT in order to obtain the satisfactory result. Another interesting finding we obtained from Figure 8 is that when the value of n increases, the elapsed time for the basic detection algorithm also increased quickly. This is because that it becomes more difficult for users to correctly adjust the value of parameters when they are dealing with a larger value of n. In comparison, FRIOD and SODIT are much more insensitive to n because of the integrated human interaction.

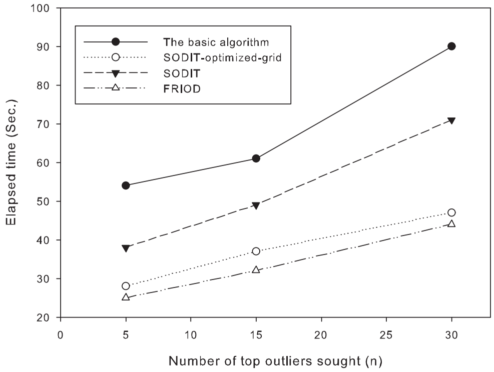


FIGURE 8. Internal elapsed time comparison.

We also evaluated the contribution of individual interactive strategy employed in the three stages of FRIOD to its efficiency enhancement. Figure 9 shows the breakdown of the elapsed execution time of FRIOD in its three interactive stages under different values of n. It shows that the elapsed time of the dense cell selection and location-aware distance thresholding is independent of the value of n and, therefore, should be identical for FRIOD under different values of n. The variations shown in the figure is due to the average time of different users which may slightly differ from each other. In contrast, the execution time of the final top outlier validation is increased when the values of n goes up. This is because a higher workload is incurred for FRIOD to validate a high number of the final top outliers detected. We also present the execution time of the three stages in percentage in Figure 10. It shows that those three stages are balanced in terms of their execution time which suggests that there is no salient performance bottleneck in FRIOD when the value of n is small. Yet, the percentage of the execution time of the final stage, i.e., the final top outlier validation, is increased when the values of n increases, while that of the first two stages are reduced as a result.

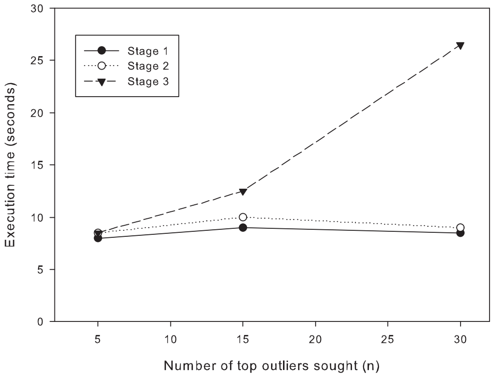


FIGURE 9. Elapsed time comparison for the three stages of FRIOD.

In the external experiment, the elapsed time between FRIOD and other non-interactive method were investigated. The comparison result is presented in Figure 11. It shows that FRIOD is considerably more efficient than all the

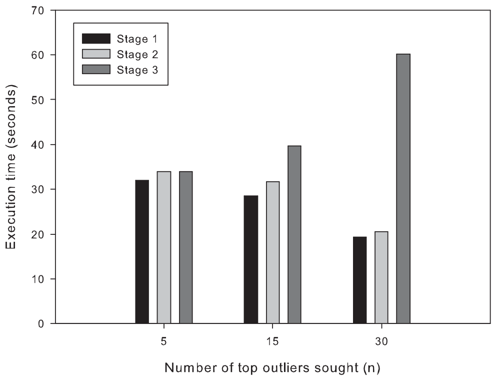


FIGURE 10. Elapsed time comparison for the three stages of FRIOD (in percentage).

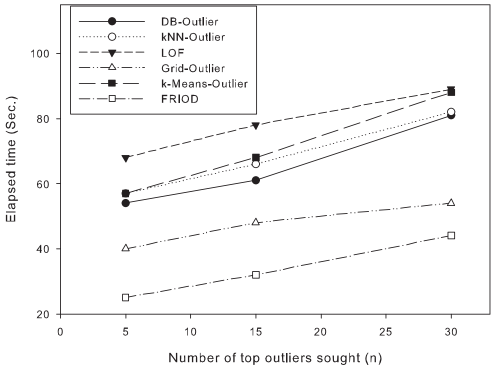


FIGURE 11. External elapsed time comparison.

non-interactive methods evaluated in the study. Users are able to spend a much shorter time by using FRIOD to achieve a satisfactory detection result than the non-interactive methods. Grid-Outlier is the most efficient non-interactive methods in the evaluation due to its use of grid structure to accelerate the detection process.

3) p-VALUE ANALYSIS

Besides evaluating the performance improvement achieved by FRIOD thanks to its extensive interactive features incorporated, we also study the statistical significance of such improvement using *p*-value analysis. In the context of our evaluation, *p*-value represents the probability that the performance improvement using FRIOD occurs by pure chance against another competitive method. In this experiment, *p*-values are computed for FRIOD against all the other competitive methods by counting the percentage of experimental runs where FRIOD does not perform better than another method. Thus, the lower the *p*-value is, the better performance achieved by FRIOD will be in a statistically significant manner. Our *p*-value analysis was conducted under two different experimental measures, i.e., accuracy and elapsed time. A total of 100 experimental runs are completed for each *p*-value analysis. Table 1 and 2 shows the *p*-value results for the measures of accuracy and elapsed time, respectively. The results from the tables show that, under multiple experimental runs, FRIOD is 95% of chance more accurate than the interactive method SODIT (that uses space partitioning optimization) and above 98% of chance more accurate than those non-interactive methods. It is also above 90% of chance faster than other competitive methods in detecting outliers.

4) CONVERGENCE STUDY ON THE SPACE PARTITIONING OPTIMIZATION

In this experiment, we investigated the convergence of FRIOD in space partitioning optimization. Comparison with the other competitive methods is not applicable for this experiment as none of them supports space partitioning optimization. The range of granularity values evaluated in this experiment is between 1 and 30 and we choose three groups of initial granularity values to start the optimization process, namely the values that are at the two ends (either very small or very big) and in the middle ground of the granularity spectrum. We tested a total of nine granularity values, with three being in each group. The values of granularity is in the range of [3-5], [14-16] and [28-30] respectively for the three groups. We investigated the optimized granularity value obtained as the number of optimization iterations increases for different granularity groups. The result, presented in Figure 12, shows that 1) irrespective of the initial granularity value, FRIOD exhibits a good convergence behavior after several iterations optimization. By leveraging such an optimization process, we can (gradually) obtain a fairly good granularity for space partitioning, which can be used effectively on future similar datasets; 2) for the initial granularity values in the middle ground of the granularity spectrum, the optimization process requires a smaller number of iterations than the other two groups, indicating a faster convergence speed. This experiment suggests that, as a rule of thumb, it is a good choice for users to start with an initial granularity that is generally in the

TABLE 1. p-value analysis based on accuracy.

p-value	Basic	SODIT	SODIT-optimized-grid	DB-Outlier	kNN-Outlier	LOF	Grid-Outlier	kMeans-Outlier
FRIOD	0	0.01	0.05	0.01	0.02	0.02	0	0

TABLE 2. p-value analysis based on elapsed time.

p-value	Basic	SODIT	SODIT-optimized-grid	DB-Outlier	kNN-Outlier	LOF	Grid-Outlier	kMeans-Outlier
FRIOD	0	0	0.1	0	0	0	0.05	0

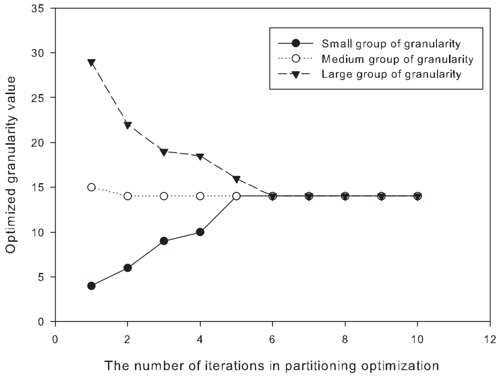


FIGURE 12. Convergence of space partitioning optimization in FRIOD.

middle ground when they don't have much knowledge about the dataset.

We also compared the accuracy performance of FRIOD when using the optimized granularity for space partitioning against the cases when the poorly selected granularity (either too large or too small) are used. The result is presented in Figure 13. It shows that the average performance of FRIOD using the optimized granularity for space partitioning (the value in the middle ground of the granularity spectrum) is significantly higher than other two groups of granularity, with the use of the exceedingly small granularity leading to the worst effectiveness performance.

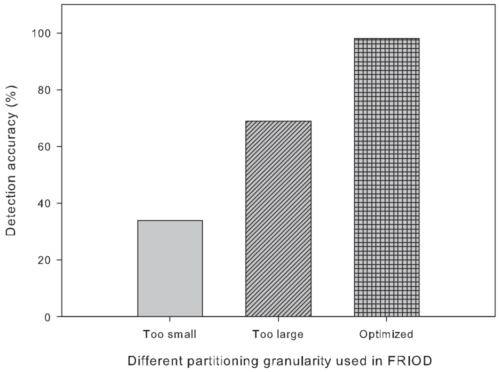


FIGURE 13. Detection accuracy of FRIOD under different granularity for space partitioning.

B. QUALITATIVE EVALUATION

As a qualitative evaluation, we have conducted a survey asking the participating students involved in this study for their experience and feedback after using FRIOD. Their responses have been very positive. The feedback from 90% of the users shows that FRIOD is more interactive and user-friendly than the SODIT. This is attributed to the fact that FRIOD is more responsive in human interaction and values of the thresholds used in FRIOD can be better controlled by users than SODIT. 100% of the users responded that FRIOD is also more efficient and much more easy to use than the non-interactive methods which human interaction is absent. The primary reason is that, in most cases, FRIOD only needs to be run once to acquire the satisfactory detection result while the non-interactive methods needs to be executed multiple rounds to achieve this, which is more cumbersome and time consuming.

C. DISCUSSIONS

Our extensive performance evaluation of FRIOD, as well as the comparative study between FRIOD and a number of major existing outlier detection methods, demonstrates that FRIOD achieves a very good performance in terms of both effectiveness and efficiency.

The evaluation reaffirms that human interaction which incorporates valuable human perception is very effective in achieving highly accurate outlier detection results. The evaluation results demonstrate that the detection accuracy of interactive detection methods, such as FRIOD and SODIT, is significantly higher than that of non-interactive alternatives. It is also very interesting to find that FRIOD is more efficient overall than the non-interactive alternatives. Non-interactive methods are generally faster than FRIOD in a single run of the algorithm. However, a single run of the algorithms in most cases is inadequate to produce satisfactory detection result. Users typically have to run multiple rounds of the algorithms for non-interactive methods. Consequently, this leads to a (much) longer elapsed time than FRIOD. Compared with the latest interactive outlier detection method SODIT, FRIOD is also advantageous by integrating more interactive features in various stages of its algorithm, helping FRIOD achieve a better overall performance. FRIOD has been proven to be a very effective means to overcome the long-standing difficulty in specifying the optimal values for various key parameters used in outlier detection methods, which contributes to the ease-of-use, accuracy and efficiency for the method as a whole. The better performance of FRIOD is also statistically significant, as evidenced by our p-value analysis.

The most important reason why FRIOD is able to achieve a better performance compared to the existing outlier detection methods, particular those non-interactive ones, is due to the fact that interactive features are incorporated in all the important stages of the detection process in FRIOD. Therefore, users are able to have a good assurance that each stage is completed in a high quality and can produce a good input into the subsequent stage. This contributes to a good detection performance overall for the whole method. In comparison, the existing methods, due to their limited or lack of interactive features, cannot guarantee a good execution of its internal stages. This often makes their detection results difficult to control in terms of quality of the output, which adversely affects their overall detection performance.

The study involving actual human users (students) shows that FRIOD, by virtue of its human-friendly and intuitive interactive interfaces and underlying mechanisms, can be easily used by users who have only fundamental knowledge about outlier detection and can produce very satisfactory detection results. This is conducive to the potentially wide adoption of FRIOD by users with varying levels of knowledge on outlier detection.

VI. CONCLUSIONS AND FUTURE DIRECTIONS

In this paper, we propose FRIOD, a novel interactive system to integrate human interaction for effective and efficient outlier detection. In FRIOD, the long-standing difficulty in specifying threshold values can be effectively mitigated and the space partitioning can be optimized to generate the optimal (or close optimal optimal) setup. We are impressed by the improvement of user-friendliness and performance of outlier detection in FRIOD by incorporating human interaction.

In the future, we are interested in investigating how human interactions can be integrated with other existing outlier detection methods to establish a more general approach for outlier detection with human interaction. We are also interested in developing a query language for outlier detection, which can not only deliver the function of outlier detection, but also makes the best use of interactive outlier detection mechanisms that we have developed.

REFERENCES

[1] C. C. Aggarwal and P. S. Yu, "Outlier detection for high dimensional data," in *Proc. SIGMOD*, 2001, pp. 37–46.

[2] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan, "Automatic subspace clustering of high dimensional data for data mining applications," in *Proc. ACM SIGMOD Int. Conf. Manage. Data (SIGMOD)*, 1998, pp. 94–105.

[3] V. Barnett and T. Lewis, *Outliers in Statistical Data*, 3rd ed. Hoboken, NJ, USA: Wiley, 1994.

[4] M. Elahi, X. Lv, M. W. Nisar, and H. Wang, "Distance based outlier for data streams using grid structure," *Inf. Technol. J.*, vol. 8, no. 2, pp. 128–137, 2009.

[5] S. Guha, R. Rastogi, and K. Shim, "CURE: An efficient clustering algorithm for large databases," in *Proc. ACM SIGMOD Int. Conf. Manage. Data (SIGMOD)*, Seattle, WA, USA, 1998, pp. 73–84.

[6] E. M. Knorr and R. T. Ng, "Algorithms for mining distancebased outliers in large datasets," in *Proc. 24th Int. Conf. Very Large Data Bases (VLDB)*, New York, NY, USA, 1998, pp. 392–403.

[7] E. M. Knorr and R. T. Ng, "Finding intensional knowledge of distance-based outliers," in *Proc. VLDB*, Edinburgh, U.K., 1999, pp. 211–222.

[8] J. L. Y. Koh, M.-L. Lee, W. Hsu, and W. T. Ang, "Correlation-based attribute outlier detection in XML," in *Proc. ICDE*, Apr. 2008, pp. 1522–1524.

[9] L. Ma, L, Gu, B. Li, L. Zhou, and J. Wang, "An improved grid-based *k*-means clustering algorithm," *Adv. Sci. Technol. Lett.*, vol. 73, pp. 1–6, Dec. 2014.

[10] M. Gupta, J. Gao, C. C. Aggarwal, and J. Han, "Outlier detection for temporal data: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 9, pp. 2250–2267, Sep. 2014.

[11] D. M. Hawkins, *Identification of Outliers*. London, U.K.: Chapman & Hall, 1980.

[12] A. Hinneburg and D. A. Keim, "An efficient approach to clustering in large multimedia databases with noise," in *Proc. KDD*, 1998, pp. 58–65.

[13] E. Schubert, A. Zimek, and H.-P. Kriegel, "Generalized outlier detection with flexible kernel density estimates," in *Proc. SDM*, 2014, pp. 542–550.

[14] S. Vijayarani and P. Jothi, "An efficient clustering algorithm for outlier detection in data streams," *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 2, no. 9, pp. 3657–3665, 2013.

[15] L. Su, W. Han, S. Yang, P. Zou, and Y. Jia, "Continuous adaptive outlier detection on distributed data streams," in *Proc. HPCC*, Houston, TX, USA, 2007, pp. 74–85.

[16] J. Tang, Z. Chen, A. Fu, and D. Cheung, "Enhancing effectiveness of outlier detections for low density patterns," in *Proc. PAKDD*, Taipei, Taiwan, 2002, pp. 535–548.

[17] B. Sheng, Q. Li, W. Mao, and W. Jin, "Outlier detection in sensor networks," in *Proc. MobiHoc*, Montreal, QC, Canada, 2007, pp. 219–228.

[18] M. E. Otey, A. Ghoting, S. Parthasarathy, "Fast distributed outlier detection in mixed-attribute data sets," *Data Mining Knowl. Discovery*, vol. 12, nos. 2–3, pp. 203–228, 2006.

[19] W. Jin, A. K. H. Tung, and J. Han, "Mining top-n local outliers in large databases," in *Proc. SIGKDD*, San Francisco, CA, USA, 2001, pp. 293–298.

[20] H. Dutta, C. Giannella, K. Borne, and H. Kargupta, "Distributed top-k outlier detection from astronomy catalogs using the DEMAC system," in *Proc. SDM*, Minneapolis, MN, USA, 2007, pp. 473–478.

[21] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. SIGKDD*, Portland, OR, USA, 1996, pp. 226–231.

[22] P. Chhabra, C. Scott, E. D. Kolaczyk, and M. Crovella, "Distributed spatial anomaly detection," in *Proc. INFOCOM*, Phoenix, AZ, USA, Apr. 2008, pp. 1705–1713.

[23] L. Cao, Q. Wang, and E. A. Rundensteiner, "Interactive outlier exploration in big data streams," *Proc. VLDB Endowment*, vol. 7, no. 13, pp. 1621–1624, 2014.

[24] I. Louhi, L. Boudjeloud-Assala, and T. Tamisier, "Exploration and visualization approach for outlier detection on log files," in *New Trends in Intelligent Information and Database Systems*. Geneva, Switzerland: Inderscience Publishers, 2015, pp. 3–12.

[25] L. Boudjeloud-Assala, "Visual interactive evolutionary algorithm for high dimensional outlier detection and data clustering problems," *Int. J. Bio-Inspired Comput.* vol. 4, no. 1, pp. 6–13, 2012.

[26] L. Boudjeloud and F. Poulet, "Visual interactive evolutionary algorithm for high dimensional data clustering and outlier detection," in *Proc. PAKDD*, Hanoi, Vietnam, 2005, pp. 426–431.

[27] R. M. Konijn and W. Kowalczyk, "An interactive approach to outlier detection," in *Rough Set and Knowledge Technology*. Beijing, China: Springer, 2010, pp. 379–385.

[28] D. Liu, Q. Gao, H. Wang, and J. Zhang, "A web-based interactive data visualization system for outlier subspace analysis," in *Proc. SEDE*, 2010, pp. 275–280.

[29] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: An efficient data clustering method for very large databases," in *Proc. ACM Int. Conf. Manage. Data (SIGMOD)*, Montreal, QC, Canada, 1996, pp. 103–114.

[30] J. Zhang, W. Hsu, and M. L. Lee, "Clustering in dynamic spatial databases," *J. Intell. Inf. Syst.*, vol. 24, no. 1, pp. 5–27, 2005.

[31] J. Zhang, X. Tao, and H. Wang, "Outlier detection from large distributed databases," *World Wide Web J.*, vol. 17, no. 4, pp. 539–568, 2014.

[32] J. Zhang, Q. Gao, H. H. Wang, Q. Liu, and K. Xu, "Detecting projected outliers in high-dimensional data streams," in *Proc. DEXA*, 2009, pp. 629–644.

[33] C. Zhu, H. Kitagawa, S. Papadimitriou, and C. Faloutsos, "Example-based outlier detection with relevance feedback," *DBSJ Lett.*, vol. 3, no. 2, pp. 1–4, 2005.

[34] J. Zhang, H. Wang, X. Tao, and L. Sun, "SODIT: An innovative system for outlier detection using multiple localized thresholding and interactive feedback," in *Proc. ICDE*, Brisbane, QLD, Australia, 2013, pp. 1364–1367.

[35] J. Zhang, M. Lou, T. W. Ling, and H. Wang, "Hos-Miner: A system for detecting outlying subspaces of high-dimensional data," in *Proc. 30th Int. Conf. Very Large Data Bases (VLDB)*, Toronto, ON, Canada, 2004, pp. 1265–1268.

[36] J. Zhang and H. Wang, "Detecting outlying subspaces for high-dimensional data: The new task, algorithms, and performance," *Knowl. Inf. Syst.*, vol. 10, no. 3, pp. 333–355, 2006.

[37] E. M. Knorr, R. T. Ng, and V. Tucakov, "Distance-based outliers: Algorithms and applications," *VLDB J.*, vol. 8, nos. 3–4, pp. 237–253, 2000.

[38] M. Breuning, H.-P. Kriegel, R. Ng, and J. Sander, "LOF: Identifying density-based local outliers," in *Proc. ACM SIGMOD Int. Conf. Manage. Data (SIGMOD)*, Dallas, TX, USA, 2000, pp. 93–104.

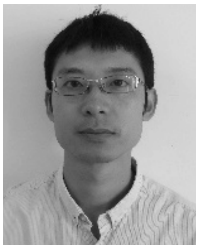
[39] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms for mining outliers from large data sets," in *Proc. ACM SIGMOD Int. Conf. Manage. Data (SIGMOD)*, Dallas, TX, USA, 2000, pp. 427–438.

[40] J. Lee and N.-W. Cho, "Fast outlier detection using a grid-based algorithm," *PLoS ONE*, vol. 11, no. 11, p. e0165972, 2016.

[41] J. Manoharan, S. H. Ganesh, and J. G. R. Sathiaseelan, "Outlier detection using enhanced k-means clustering algorithm and weight-based center approach," *Int. J. Comput. Sci. Mobile Comput.*, vol. 5, no. 4, pp. 453–464, 2016.

XIAODONG ZHU is currently a Lecturer with the Department of Information Systems, School of Economics and Management, Nanjing University of Information Science and Technology, China. His research focuses on management information systems, data mining, and operations management.

JI ZHANG (SM'17) received the Ph.D. degree in computer science from Dalhousie University, Canada, in 2008. He served the Principal Advisor for Research in the Division of ICT Services, University of Southern Queensland, Australia, from 2010 to 2013, where he is currently an Associate Professor (Reader) in computing. He has published over 100 papers, some appearing in top-tier international journals, including the IEEE Transactions on Dependable and Secure Computing, *Information Sciences*, *WWW Journal*, *Bioinformatics*, *Knowledge and Information Systems*, *Soft Computing*, the *Journal of Database Management*, and the *Journal of Intelligent Information Systems*, and international conferences, such as VLDB, ACM CIKM, ACM SIGKDD, IEEE ICDE, IEEE ICDM, WWW, DASFAA, DEXA, and DaWak. His research interests are knowledge discovery and data mining, health informatics, and information privacy and security. He is an Australian Endeavour Fellow, a Queensland Fellow, and an Izaak Walton Killam Fellow (Canada).



HONGZHOU LI was an Academic Visitor to the University of Southern Queensland in 2013. He is currently an Associate Professor with the School of Life and Environmental Science, Guilin University of Electronic Technology, China. His research interests are data mining and sensor networks.

PHILIPPE FOURNIER-VIGER received the Ph.D. degree in computer science from the University of Quebec, Montreal, in 2010. He is currently a Professor with the Shenzhen Graduate School, Harbin Institute of Technology, China. He is also the Founder of the popular SPMF open-source data mining library, which has been cited in over 430 research papers since 2010. He has published over 140 research papers in refereed international conferences and journals, which have received over 1,300 citations. His research interests include data mining, pattern mining, sequence analysis and prediction, text mining, e-learning, and social network mining. He has received the title of Youth 1000 Talent from the National Science Foundation of China. He is the Editor-in-Chief of the *Data Mining and Pattern Recognition Journal*.

JERRY CHUN-WEI LIN received the Ph.D. degree in computer science and information engineering from National Cheng Kung University, Tainan, Taiwan, in 2010. He is currently an Associate Professor with the Harbin Institute of Technology, Shenzhen, China. He is also the Co-Leader of the popular SPMF open-source data mining library. He has published over 200 research papers in refereed international conferences and journals, which have received over 1000 citations. His research interests include data mining, privacy-preserving and security, big data analytics, and social network. He is the Editor-in-Chief of the *Data Mining and Pattern Recognition Journal*.

LIANG CHANG received the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, in 2008. He is currently a Professor with the School of Computer Science and Engineering, Guilin University of Electronic Technology, China. His research interests include knowledge representation and reasoning, formal methods, and intelligent planning.