2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC)
October 6-10, 2024. Sarawak, Malaysia

AutoEDA: Iterative Data Focusing and Exploratory Analysis Based on Attribute Frequency

Tong Wu[1], Song Wang[1]∗ Xin Peng[1]

***Abstract*— The paper proposes an automated data exploration and analysis method based on Attribute Frequency Statistical Feature Ratio (AFSFR). It integrates AutoVis and Data Preprocessing Methods to design and develop AutoEDA-Segment. Addressing the Concentrate on Field Sequences (CFS) problem in data exploration and analysis, this study employs various classification models and combines AFSFR with field type and the Elbow Inflection Point (EIP) of index features to design a field type identification and field value assessment method. For evaluating the effectiveness of focus analysis, the approach provides clustering visualization effects and an analysis scheme based on Field Type Search Tree (FST) and cluster comparison profiles, using a custom CFS approach. Additionally, to enhance the value of focused subset analysis data, the approach introduces a Parallel Coordinates-Based Data Filter (PCF), forming an EDA feedback loop to achieve Iterative Exploratory Data Analysis for User-inferred Cognition (IEDA-UC). Finally, we engaged graduate students with varying levels of experience in visualization research for collaboration and discussion, validating the effectiveness and feasibility of the approach using structured data from Kaggle.**

I. INTRODUCTION

With the rapid development of information technology, big data has become an important asset in modern society. However, the fields and analysis targets of big data are often complex and diverse, and the data volume is extremely large, which presents numerous challenges in data processing and analysis. Firstly, the analysis targets of big data often involve multiple levels and dimensions, leading to frequent changes and adjustments in the analysis process. Secondly, the vast amount of data makes traditional data processing tools and technologies face bottlenecks, necessitating more efficient methods for data management, filtering, and organization. Moreover, the low efficiency of data visualization makes it difficult to extract valuable information from massive data, affecting the timeliness and accuracy of decision-making. Lastly, there are often discrepancies in the quality of analysis results, which may be due to data noise, improper or limited data processing methods. Facing these challenges, how to improve the efficiency and quality of big data analysis has become an important topic in current data science research and practice.

A. Related Work

In automated data exploration and analysis, statistical feature analysis of field attributes is a crucial aspect. Epperson et al. [7] proposed the AutoProfiler tool, which implements continuous data summarization and allows analysts to view interactive data summaries in real-time. Zhang et al. [19] introduced the AdaVis system, which models relationships among data features, datasets, and visualization choices through an embedded knowledge graph. Qian et al. [15], [14] developed an end-to-end visualization recommendation system based on machine learning, capable of automatically handling new datasets and generating appropriate visualizations. Hu et al. [9] proposed VizML, a machine learning-based visualization recommendation method that predicts design choices by learning from large-scale datasets and visualization pairs. Iizuka et al. [11] proposed an automatic target attribute selection method based on decision trees or correlation coefficient matrices as early as 1998. Hull et al. [10] developed VISGRADER, which can automatically evaluate D3 data visualizations, including aspects like data binding and visualization encoding. These methods primarily focus on visualization recommendation or evaluation, while our AFSFR method emphasizes frequency statistical features of field attributes, providing a more comprehensive field perception capability.

In the realm of data visualization and clustering analysis, Harris et al. [8] proposed the SpotLight system, which helps users quickly understand data by recommending relevant groups of insights. Law et al. [12] reviewed tools for automated data insight recommendations, presenting twelve types of automated insights. Li et al. [13] introduced a dynamic data exploration method based on pattern classifiers, capable of automatically identifying new pattern categories. Zeng et al. [18] proposed an evaluation framework for comparing various visualization recommendation algorithms. Anand and Talbot [1] proposed an automated method for creating effective small multiples displays. Shen et al. [16] provided a comprehensive review of visualization-guided natural language interfaces (V-NLI), offering new ideas to enhance user experience. Wongsuphasawat et al. [17] explored how analysis goals and context influence exploratory data analysis (EDA) through interviews with data analysts. These works lay a significant foundation for automated data visualization and clustering analysis, but most consider only partial or single static fields. Our approach, driven by CFS, enables more targeted automated visualization and clustering profiling. In iterative data exploration analysis, Badam et al. [2] explored interface design guidelines for progressive visualization analysis (PVA) to support early decision-making based on progressively updated results. Bertini and Santucci [3] proposed a feature preservation method that models

[1]Tong Wu, Song Wang and Xin Peng all of the Department of Computer Science, School of Computer Science and Technology, Southwest University of Science and Technology, Mianyang, China
Song Wang is the corresponding author, wangsong@swust.edu.cn

2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC) | 978-1-6654-1020-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/SMC54092.2024.10832030

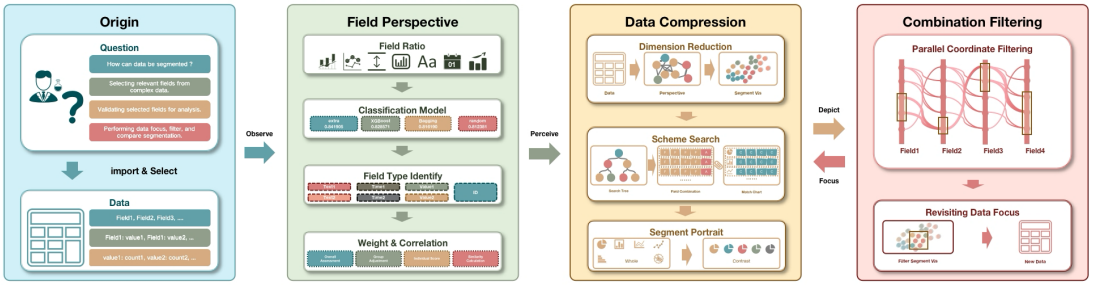

Fig. 1. Overview of Methods and Exploratory Processes in AutoEDA-Segment

visual features of large-scale data through virtual space. Ellis and Dix [5] studied clutter measurement in parallel coordinate plots. Elmqvist and Fekete [6] proposed a multi-scale information visualization model to reduce visual clutter through hierarchical aggregation. Ding et al. [4] developed the QuickInsights technique, which can quickly and automatically discover interesting patterns in multidimensional data. These works provide important insights into iterative data exploration. Integrating clustering profiling exploration with parallel coordinates-based progressive analysis schemes could enhance the personalization of the progressive exploration process. Our IEDA-UC method incorporates these elements, achieving a more efficient iterative data exploration and analysis process.

B. Contribute

The uncertainties in data analysis goals, data noise, and limitations of processing methods often constrain analysts' creative analysis and CFS selection. This paper addresses these issues, summarizing the key needs and contributions as follows.

- **Data Field Perception Based on AFSFR.** By combining AFSFR's field categorization and value assessment with network visualization, this method provides users with a means to quickly perceive field characteristics, assisting in the initial clarification of analysis objectives amidst extensive data fields.
- **FST-Driven CFS Portrait.** By integrating field categorization analysis with search trees, this approach generates and compares cluster feature sequences for cluster portraits. Visualizing CFS provides users with decision-making references for field sequence selection.
- **Iterative EDA for User-Inferred Cognition.** This method offers PCF-style data filtering and selection capabilities. By combining cluster scatter portraits and analysis scheme cluster portraits, it iteratively refines data and CFS based on user experience to better infer and filter data noise, offering analytical insights and clarifying analysis goals.

II. METHOD

When facing a new structured dataset, users often have specific analysis needs, such as how to slice the data, select fields, and compare focused data with the original dataset.

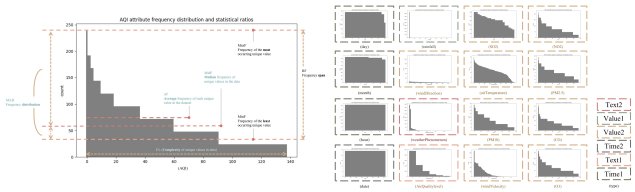

Fig. 2. Bar Charts of Attribute Frequency for Each Field in Structured Data and AFSFR Distribution Diagram

TABLE I
DEFINITION AND DESCRIPTION OF FIELD TYPE

field type	describe
Text1	Text with sparse attribute categories
Text2	Text with dense attribute categories
Time1	Discrete time like year and month
Time2	Continuous time distribution, such as timestamps
Value1	Some sparsely distributed discrete numerical values
Value2	Some continuously distributed numerical values
ID	Unique identifier in the data, akin to a primary key

The data requirements in this paper stem from these user queries, with the core data extracted through AFSF of the data frame. The specific method is illustrated in Figure 1. After extracting frequency data, a classification model is first used to identify field categories, assess value, and calculate relevance based on defined frequency statistical features. Following this data perception, users cluster field sequences in the field perception network graph for data dimensionality reduction and then assess the validity of field sequence selection based on the CFS portrait generated from clustered scatter plots. Finally, data distribution is analyzed using CFS-Driven PCF. If more detailed features and analytical conclusions are needed or data filtering is required, the process returns to the dimensionality reduction step, thereby forming an iterative data exploration analysis process guided by user experience. The following sections will provide a detailed description of the three methods and exploration process.

A. Data Field Perception Based on AFSFR

This paper categorizes fields into seven types based on distribution characteristics and value categories, as shown in Table I, according to common field categories found in structured data.

The methods and system functionalities of this study begin with AFSFR, which reflects the suspected categories and value of data fields from perspectives such as attribute distribution patterns and attribute richness.

As illustrated in Figure 2, this figure shows the AFSFR distribution of fields in meteorological pollution data, with the positions of the seven statistical feature attributes used in this paper depicted in Figure 2 (AQI). The remaining charts display the frequency distribution characteristics of other fields. As indicated by the legend in the lower-left corner of Figure 2, there is a certain correlation between field categories and distribution characteristics. The specific calculation methods for these metrics are detailed in Table II, where N represents the total number of data fields, n denotes the number of unique values, and f_i indicates the frequency of the i-th unique value. To mitigate the impact of data length on these indices, some indices use N as the denominator to dilute the effect of data length on the indices. Additionally, $\mathbf{f}$ is defined as a vector that contains all unique frequency values f_i, while $\mathbf{1}$ represents a vector of ones with the same length as $\mathbf{f}$, facilitating the computation of metrics involving sums and averages.

TABLE II

FIELD FREQUENCY STATISTICAL CHARACTERISTIC RATIO

AFSFR	formula
UL	$\frac{1}{N} \times \text{length}(\mathbf{f})$
AF	$\frac{1}{N \times n} \times \mathbf{1}^T\mathbf{f}$
MinF	$\frac{1}{N} \times \min(\mathbf{f})$
MaxF	$\frac{1}{N} \times \max(\mathbf{f})$
MidF	$\frac{1}{N} \times \text{median}(\mathbf{f})$
MAD	$\frac{\max(\mathbf{f}) - \text{median}(\mathbf{f})}{\max(1, \text{median}(\mathbf{f}) - \min(\mathbf{f}))}$
RF	$\frac{1}{N} \times \sum_{i=1}^{n} \left\lvert f_i - \frac{1}{n} \times \mathbf{1}^T\mathbf{f} \right\rvert$

The primary focus of this section is to explore how classification models can learn these relationships. As shown in Figure 3, thirteen classification models were evaluated using structured data from public sources such as Kaggle for training and testing. The models were analyzed for recognition accuracy, average accuracy, and total training time over 5, 10, and 15 iterations, with the training and test sets split at a 7:3 ratio.

From the scores in different iteration counts displayed in Figure 3 (b1), (b2), (b3), it is evident that the four selected classification models (highlighted with red bars) performed well across different iterations, with their performance improving as the number of iterations increased. Additionally, the training scores shown in Figure 3 (a1), (a2), (a3) indicate that these four models (with bold lines) exhibited better stability compared to others. Finally, comparing training durations, Figure 3 (c1), (c2), (c3) reveals that the four chosen models (indicated with red bars) offer moderate response times relative to most models. Considering time efficiency, accuracy, and prediction stability, the study selects ExtraTrees, XGBoost, Bagging and RandomForest as the preferred classification models.

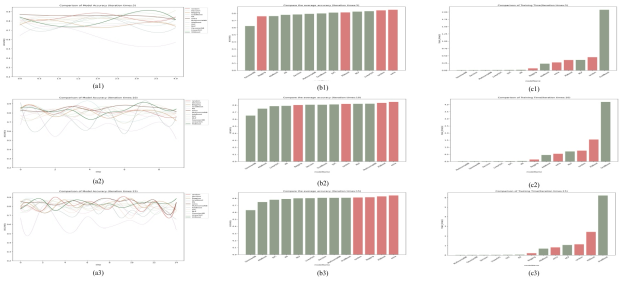

Fig. 3. Comparison of Accuracy, Average Accuracy Ranking, and Total Time Consumption of 13 Classification Models under 5, 10, and 15 Iterations

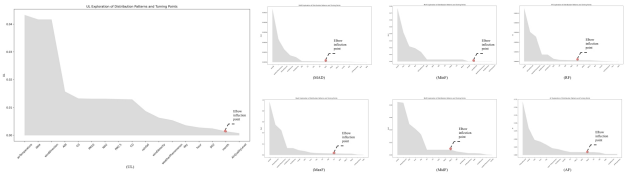

Fig. 4. Elbow Method for Identifying Inflection Points in AFSFR: An Example Using Urban Meteorological Pollution Data

After classifying and identifying data by fields, a value assessment is performed based on AFSFR. The value assessment utilizes a method combining the Index Feature EIP, with AFSFR EIP locations in the example meteorological dataset shown in Figure 4. First, the AFSFR sets for each field in the structured data are combined and sorted: $\mathrm{S}(r_i)$ (the AFSFR of the i-th field). The second-order derivative of the set is then computed: $\mathrm{S}''(r_i) = \left(\frac{\partial^2 \mathrm{S}(r_i)}{\partial r_1^2}, \frac{\partial^2 \mathrm{S}(r_i)}{\partial r_1 \partial r_2}, \ldots, \frac{\partial^2 \mathrm{S}(r_i)}{\partial r_n^2}\right)$. The distance from each inverted point to the minimum second-order derivative point is calculated as the field score: $field_{score} = \frac{1}{1+|r_i - \min(\mathrm{S}''(r_i))|}$. Additionally, the correlation between fields is derived by computing the Euclidean distance of AFSFR. Finally, the field perception image is displayed as a network graph, as shown in Figure 6 (A2).

B. FST-Driven CFS Portrait

Using the aforementioned methods for field classification and value assessment, the next step is to visualize CFS, including the CFS images. The most crucial part of the visualization is the CFS image. Figure 5 illustrates the FST combinations used in this study: Figure 5 (a) Text1, Figure 5 (b) Time1&Time2, Figure 5 (c) Value1&Value2, and Figure 5 (d) Text2. Each FST generated by a different field type is divided into four layers, corresponding to the field type and aggregation scheme used in the CFS image analysis charts. From top to bottom, the five layers are: field1 (encoding1, typically representing the x-axis and pie chart categories), field2 (encoding2, usually representing color categories, stacked groups, etc.), field3 (encoding3, generally representing color categories), field4 (encoding4, usually representing numeric values or the y-axis), and aggType (including sum, mean, and count aggregation methods).

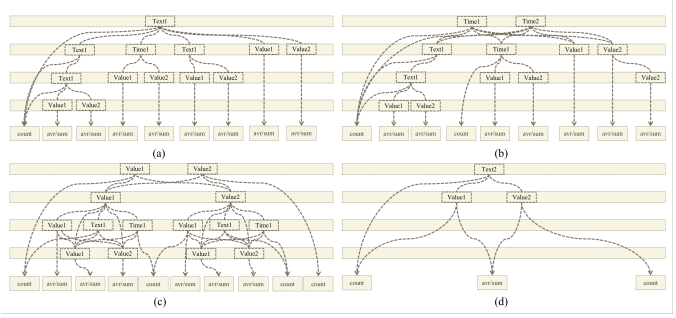

Fig. 5. Field Type Analysis Scheme Combination Search Tree

Based on the combination schemes searched by FST, a value assessment is required for ranking. Initially, individual scores are calculated and grouped according to field values, with competitive relationships among fields considered. Let $\mathbf{f} = [f_1, f_2, \ldots, f_n]$ represent the vector of field scores, where each f_i corresponds to a specific field score.

The adjusted score vector $\boldsymbol{\delta f}$ is computed as:

$$\boldsymbol{\delta f} = \mathbf{D}^{-1}\mathbf{f} - \frac{\mu \mathbf{f}}{\max(\mathbf{f})} + \Delta \mathbf{1},$$

where $\mathbf{D} = \Omega(\text{field}_{\text{type}}) + \lambda \mathbf{I}$ is a diagonal matrix, μ is a precision control parameter, Δ is the lower bound scalar (the minimum score), and $\mathbf{1}$ is a vector of ones.

Next, to evaluate the value of the combination scheme, the scores of the four field encodings are combined into a scalar ω, given by:

$$\omega(\text{combination}) = \frac{\|\boldsymbol{\delta f}\|_1}{\Omega(\mathbf{f})} + \Delta,$$

where $\|\cdot\|_1$ denotes the L_1 norm, and $\Omega(\mathbf{f})$ represents the number of fields with actual scores.

Finally, considering the recommended field type matrix $\mathbf{F}$ for each encoding attribute in the table and chart, and the chart attribute length vector $\mathbf{C}$, the similarity between the scheme and the chart is calculated as:

$$\theta(\text{combination, chart}) = 1 - \|\mathbf{F} - \mathbf{C}\|,$$

where $\mathbf{F} = [F_1, F_2, F_3, F_4]$ is a vector representing the recommended lengths of the independent attributes for each field, and $\mathbf{C} = [C_1, C_2, C_3, C_4]$ is the corresponding vector of actual chart attribute lengths.

C. Iterative EDA for User-inferred Cognition

Based on the aforementioned data perception and image generation methods, we have developed and designed an interactive visualization system based on the B/S architecture. The functional modules include data import (Figure 6(A1)), data perception (Figure 6(A2)), CFS Clustering Portrait (Figure 6(B1)), PCF (Figure 6(B2)), CFS Portrait (Figure 6(C1)), and CFS-Driven Cluster Comparison Portrait (Figure 6(C2)). These portraits collectively bridge the gap from data to iterative EDA.

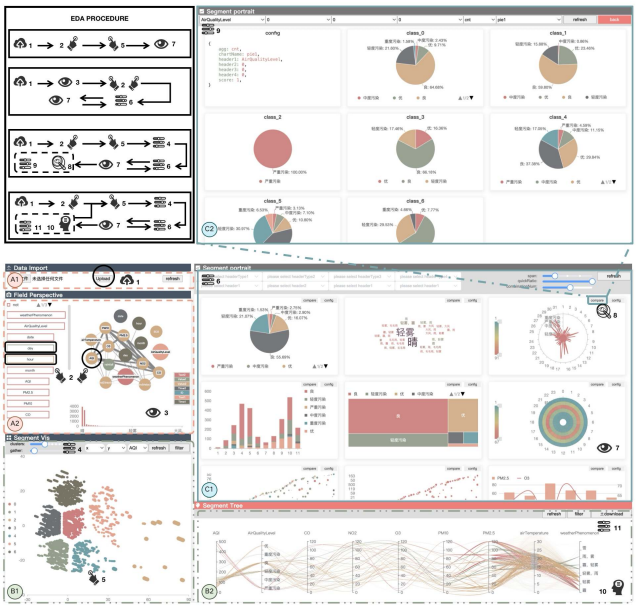

Fig. 6. Overview and Exploratory Analysis Workflow of Interactive Systems

TABLE III

CHART ATTRIBUTE LENGTH AND RECOMMENDED FIELD TYPE

Chart	C_1	C_2	C_3	C_4	F_1	F_2	F_3	F_4
Words1	3	4	4	4	$Te_{1,2}$	N	N	$V_{1,2,N}$
Pie1	1	4	4	4	Te_1, Ti_1	N	N	$V_{1,2,N}$
Rose2	1	1	4	4	Te_1, V_1	Te_1, Ti_1	N	$V_{1,2,N}$
Polar2	2	1	4	4	Te_1, Ti_1	Te_1, V_1	N	$V_{1,2,N}$
Area1	3	4	4	4	$Te_1, Ti_{1,2}$	N	N	$V_{1,2,N}$
Column1	2	4	4	4	Te_1, Ti_1, V_1	N	N	$V_{1,2,N}$
Column2	2	1	4	4	Te_1, Ti_1, V_1	Te_1	N	$V_{1,2,N}$
Heat2	2	1	4	4	Te_1	Te_1	N	$V_{1,2,N}$
Heat3	2	1	1	4	Te_1	Te_1	Te_1, Ti_1, V_1	$V_{1,2,N}$
Scatter2	3	3	4	4	V_2	V_2	N	$V_{1,2,N}$
Scatter3	3	3	1	4	V_2	V_2	T_1	$V_{1,2,N}$
Sankey2	1	1	4	4	Te_1, Ti_1	Te_1	N	$V_{1,2,N}$
Sankey3	1	1	1	4	Te_1, Ti_1	Te_1, Ti_1	T_1	$V_{1,2,N}$
DualAxes3	2	3	3	4	$Te_1, Ti_{1,2}$	$V_{1,2,N}$	T_1	$V_{1,2,N}$
Doublebar3	2	2	2	4	Te_1, Ti_1	$V_{1,2,N}$	T_1	$V_{1,2,N}$
Tree2	1	1	4	4	Te_1, Ti_1	Te_1, Ti_1	N	$V_{1,2,N}$
Tree3	1	1	1	4	Te_1, Ti_1	Te_1, Ti_1	Te_1, Ti_1	$V_{1,2,N}$

This paper outlines the data exploration and segmentation processes as illustrated in Figure 6 using the EDA PROCEDURE framework: 1) EDA Driven by CFS Chart Generation Schemes, 2) EDA Combined with CFS-Based Data Filtering Schemes, 3) EDA of CFS-Based Clustering with Cluster Comparison Portrait, 4) PCF-Driven Iterative EDA. This approach enables a focused exploration from data fields to analysis schemes, and further from cluster division to precise data segmentation.

III. RESULT

This section proceeds with case studies on urban meteorological pollution data, medical examination data, and medical hospitalization reimbursement information to validate the efficacy of AutoEDA-Segment.

A. Case1: Case Study on Urban Meteorological and Air Pollution Data Analysis

The dataset on urban meteorology and air pollution primarily consists of time, meteorological attributes, and pol-

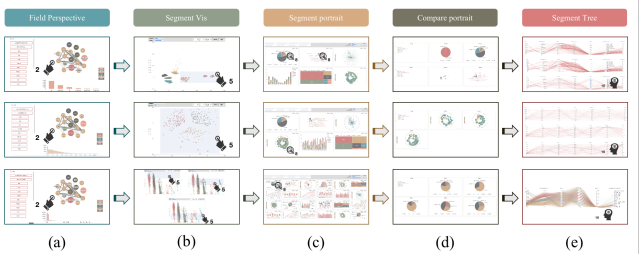

Fig. 7. Case Study on Urban Meteorological and Air Pollution Data Analysis under Different CFS and Exploratory Analysis Workflows

lution indicators. As shown in Figure 7(a), data analysts selected three CFS combinations: AQI with temperature, wind speed, and precipitation data; AQI with six pollutants including PM2.5; and AQI with time and other fields. These selections led to the generation of three EDA PROCEDIRE frameworks. Through screening of the CFS Clustering Portrait and exploration of the CFS Portrait, as depicted in Figures 7(b) and (c), a strong correlation was observed between AQI and PM2.5, month, hour, precipitation, and wind speed. Further analysis using PCF revealed that when AQI is high ($>= 300$), wind speed is low (< 1), precipitation is almost zero, PM2.5 and PM10 levels are elevated ($>= 100$), and pollution exhibits a seasonal and diurnal distribution (higher AQI values are more prevalent from March to May, and pollutants such as PM2.5 are more abundant from 10:00 to 20:00). Additionally, the predominant weather phenomena during these periods are sunny, haze, and cloudy conditions. Through these EDA PROCEDIRE frameworks, researchers gained clarity on the objectives of analyzing and exploring this dataset.

B. Case2: Segment Analysis and Cluster Classification of Medical Examination Data

The medical examination dataset primarily consists of case summary tables, including fields such as age, gender, diagnosis, surgery, number of hospitalizations, and affected organs. As illustrated in Figures 8(a1) through (a4), researchers employed an EDA PROCEDIRE framework similar to Case 1 for data exploration, focusing on two main exploration records. By determining the number of clusters and selecting relevant CFS, the researchers chose diagnosis, affected organs, age, and gender as clustering features, with the number of clusters set to six. They developed a health profile exploration system based on patient clusters, as shown in Figure 8(b). Additionally, based on the analysis schemes in Figures 8(a3) and (a4), they designed heatmaps for organ-to-organ, condition-to-condition, and condition-to-organ analyses, as well as organ portraits and association analysis images. The data and research were ultimately focused on pulmonary and cardiovascular diseases.

C. Case3: Comparative analysis of cluster features of medical insurance data and iterative EDA case study

For the case analysis of the medical hospitalization reimbursement data, which is similar in scope to Case 2,

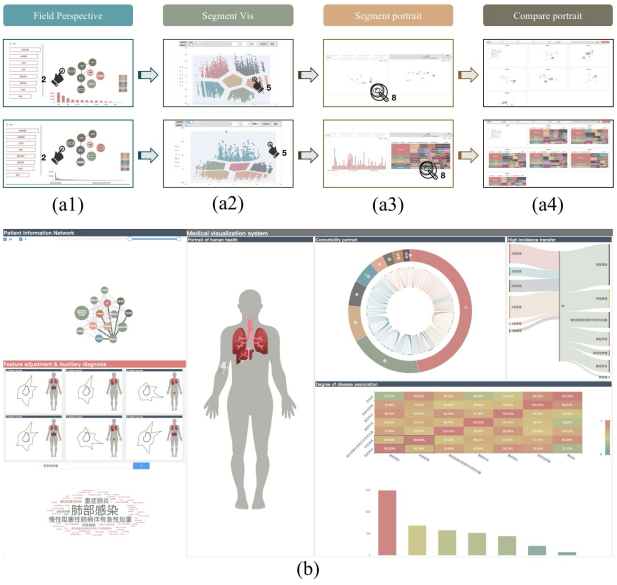

Fig. 8. Case Study on Medical Examination Data Analysis and Exploratory Analysis Workflow under Different CFS

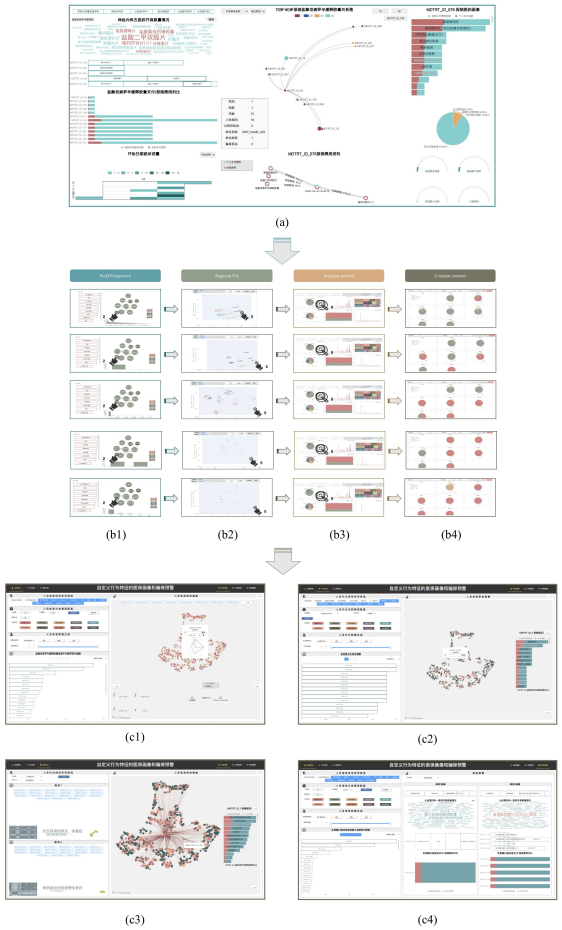

Fig. 9. Case Study on Health Insurance Reimbursement Data Analysis and Exploratory Analysis Workflow under Different CFS

researchers applied the EDA PROCEDIRE framework used in Case 2, retaining five key exploration records as shown in Figures 9(b1) through (b4). Initially, as depicted in Figure 9(a), the team had a healthcare insurance visualization system that explored reimbursement behavior based on classification of individuals. After incorporating AutoEDA-Segment, they discovered that cluster features could be further refined and selected. Consequently, they initially explored the original features, then added additional features such as unit and account types for clustering. Ultimately, they found that removing user features such as age, ethnicity, and gender improved clustering results. Therefore, they redefined the cluster features and, inspired by the EDA PROCEDIRE framework, developed a custom feature-based healthcare insurance exploration and analysis platform, as illustrated in Figures 9(c1) through (c4), to explore the feasibility of additional CFS and segmentation schemes.

D. Case Study Survey Questionnaire

Following discussions with the data analysis researchers, the results are summarized in Table IV. The findings indicate that, despite some variability in scores across different functionalities and researchers, the evaluations consistently remain at a high and satisfactory level when addressing various objectives and tasks.

TABLE IV
SURVEY QUESTIONNAIRE FOR DIFFERENT ANALYSIS TASKS AND DATASETS

Case ID	Researcher	Target	CFS Clustering	CFS Portrait	PCF
Case 1	1	Vis	5	4	4
Case 1	1	Vis&Segment	4	4	5
Case 1	3	Segment	3	5	5
Case 2	4	Vis&Cluster	3	4	4
Case 2	5	Cluster&Segment	3	4	5
Case 3	6	Cluster	4	3	4
Case 3	7	Cluster&Segment	5	3	4
Case 3	8	Cluster&Vis	5	4	3

IV. CONCLUSION

In this study, attribute frequency statistics combined with clustering portraits from visualizations and iterative EDA using parallel coordinate filters effectively assist users in clearly analyzing objectives and focusing on attribute and field targets. However, we recognize the critical role of visualization in this process. Future work will explore expanding the range of chart types and incorporating additional interactive algorithms to enhance the fluidity of iterative EDA.

ACKNOWLEDGMENT

This work was supported by Natural Science Foundation of Sichuan Province (Grant No. 2022NSFSC0961), the Special Research Foundation of China (Mianyang) Science and Technology City Network Emergency Management Research Center (Grant No. WLYJGL2023ZD04).

REFERENCES

[1] Anushka Anand and Justin Talbot. Automatic selection of partitioning variables for small multiple displays. *IEEE transactions on visualization and computer graphics*, 22(1):669–677, 2015.

[2] Sriram Karthik Badam, Niklas Elmqvist, and Jean-Daniel Fekete. Steering the craft: Ui elements and visualizations for supporting progressive visual analytics. In *Computer Graphics Forum*, volume 36, pages 491–502. Wiley Online Library, 2017.

[3] Enrico Bertini and Giuseppe Santucci. Quality metrics for 2d scatterplot graphics: automatically reducing visual clutter. In *International Symposium on Smart Graphics*, pages 77–89. Springer, 2004.

[4] Rui Ding, Shi Han, Yong Xu, Haidong Zhang, and Dongmei Zhang. Quickinsights: Quick and automatic discovery of insights from multi-dimensional data. In *Proceedings of the 2019 international conference on management of data*, pages 317–332, 2019.

[5] Geoffrey Ellis and Alan Dix. The plot, the clutter, the sampling and its lens: occlusion measures for automatic clutter reduction. In *Proceedings of the working conference on Advanced visual interfaces*, pages 266–269, 2006.

[6] Niklas Elmqvist and Jean-Daniel Fekete. Hierarchical aggregation for information visualization: Overview, techniques, and design guidelines. *IEEE transactions on visualization and computer graphics*, 16(3):439–454, 2009.

[7] Will Epperson, Vaishnavi Gorantla, Dominik Moritz, and Adam Perer. Dead or alive: Continuous data profiling for interactive data science. *IEEE Transactions on Visualization and Computer Graphics*, 2023.

[8] Camille Harris, Ryan A Rossi, Sana Malik, Jane Hoffswell, Fan Du, Tak Yeon Lee, Eunyee Koh, and Handong Zhao. Insight-centric visualization recommendation. *arXiv preprint arXiv:2103.11297*, 2021.

[9] Kevin Hu, Michiel A Bakker, Stephen Li, Tim Kraska, and César Hidalgo. Vizml: A machine learning approach to visualization recommendation. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–12, 2019.

[10] Matthew Hull, Vivian Pednekar, Hannah Murray, Nimisha Roy, Emmanuel Tung, Susanta Routray, Connor Guerin, Justin Chen, Zijie J Wang, Seongmin Lee, et al. Visgrader: Automatic grading of d3 visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 2023.

[11] Yuichi Iizuka, Hisako Shiohara, Tetsuya Iizuka, and Seiji Isobe. Automatic visualization method for visual data mining. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 174–185. Springer, 1998.

[12] Po-Ming Law, Alex Endert, and John Stasko. Characterizing automated data insights. In *2020 IEEE Visualization Conference (VIS)*, pages 171–175. IEEE, 2020.

[13] Jie Li, Huailian Tan, and Wentao Huang. Active pattern classification for automatic visual exploration of multi-dimensional data. *Applied Sciences*, 12(22):11386, 2022.

[14] Xin Qian, Ryan A Rossi, Fan Du, Sungchul Kim, Eunyee Koh, Sana Malik, Tak Yeon Lee, and Nesreen K Ahmed. Personalized visualization recommendation. *ACM Transactions on the Web (TWEB)*, 16(3):1–47, 2022.

[15] Xin Qian, Ryan A Rossi, Fan Du, Sungchul Kim, Eunyee Koh, Sana Malik, Tak Yeon Lee, and Joel Chan. Learning to recommend visualizations from data. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 1359–1369, 2021.

[16] Leixian Shen, Enya Shen, Yuyu Luo, Xiaocong Yang, Xuming Hu, Xiongshuai Zhang, Zhiwei Tai, and Jianmin Wang. Towards natural language interfaces for data visualization: A survey. *IEEE transactions on visualization and computer graphics*, 29(6):3121–3144, 2022.

[17] Kanit Wongsuphasawat, Yang Liu, and Jeffrey Heer. Goals, process, and challenges of exploratory data analysis: An interview study. *arXiv preprint arXiv:1911.00568*, 2019.

[18] Zehua Zeng, Phoebe Moh, Fan Du, Jane Hoffswell, Tak Yeon Lee, Sana Malik, Eunyee Koh, and Leilani Battle. An evaluation-focused framework for visualization recommendation algorithms. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):346–356, 2021.

[19] Songheng Zhang, Yong Wang, Haotian Li, and Huamin Qu. Adavis: Adaptive and explainable visualization recommendation for tabular data. *IEEE Transactions on Visualization and Computer Graphics*, 2023.