111-2 DBMS Final Project: ECSQL

Access databases with one button

LIN, BO-YONG  
 NTU BEBI  
 First Grade

KUO, TING-YI  
 NTU BEBI  
 First Grade

TSENG, YU-HSUAN  
 NTU BEBI  
 First Grade

ZHANG, YU-JIE  
 NTU BEBI  
 First Grade

ABSTRACT

In this sample-structured document, neither the cross-linking of float elements and bibliography nor metadata/copyright information is available. The sample document is provided in “Draft” mode and to view it in the final layout format, applying the required template is essential with some standard steps.

These steps, which should require generation of the final output from the styled paper, are mentioned here in this paragraph. First, users have to run “Reference Numbering” from the “Reference Elements” menu; this is the first step to start the bibliography marking (it should be clicked while keeping the cursor at the beginning of the reference list). After the marking is complete, the reference element runs all the options under the “Cross Linking” menu.

For accuracy check of the structured paper, user can run the option **Manuscript Validation**. It informs the user of the wrong or missing values in the paper. The user must correct the paper as per validation messages and rerun **Manuscript Validation**.

Now, to generate the required layout of the paper, the user should select one of the template styles under the Define Template Style option (choose the required layout design, i.e. choose between Journals and Proceedings).

∗Article Title Footnote needs to be captured as Title Note

†Author Footnote to be captured as Author Note

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

*WOODSTOCK’18, June, 2018, El Paso, Texas USA*

© 2018 Copyright held by the owner/author(s). 978-1-4503-0000-0/18/06...$15.00

https://doi.org/10.1145/1234567890

Some specific values are required to create a standard layout by choosing a template for the journals or proceedings. So once the user chooses one of the template layout styles, the respective Journal or Conference details dialog box (i.e. **journal/conference acronym, DOI, ISBN, copyright, year, etc.**) will appear as a prompt during the Define Template Style functionality. The user should fill these values, after which the template creates the desired layout of the paper. The user can now create a PDF of his/her manuscript using the “*Save as PDF*” option.

If the user is adding any new data, they should make sure to style it as per the instructions provided in previous sections. Carry out the steps for Cross-linking, Fundref data, adding Document History (specific to journal submission), and finally, Manuscript validation and placing the respective metadata (Bibstrip/copyright text)[[1]](#footnote-2) while applying the required template.

CCS CONCEPTS

•Information systems~Data management systems~Data structures~Data access methods

<https://blog.csdn.net/weixin_41971366/article/details/107722049>

KEYWORDS

Insert keyword text, Insert keyword text, Insert keyword text, Insert keyword text

ACM Reference format:

FirstName Surname, FirstName Surname and FirstName Surname. 2018. Insert Your Title Here: Insert Subtitle Here. In *Proceedings of ACM Woodstock conference (WOODSTOCK’18). ACM, New York, NY, USA, 2 pages.* https://doi.org/10.1145/1234567890

2  PRIOR WORK

2.1 EXCEL

Managing large datasets in Excel has long been a significant challenge for users, prompting the exploration of various approaches to improve efficiency. One common strategy is to split the data into multiple worksheets or workbooks, allowing for better organization and easier manipulation. Another technique involves utilizing pivot tables and filtering options, enabling users to analyze and summarize large datasets more effectively. Additionally, the use of extensions and add-ins has become popular, providing users with additional functionalities and tools to enhance data management and analysis capabilities.

While these prior works have made significant contributions in addressing the challenges of working with large datasets in Excel, there is ongoing research in this field to further enhance data management capabilities, improve performance, and provide more seamless integration with databases.

2.2 PCA

While PCA has been widely used and well-established, the prior work also recognized certain limitations. For instance, the interpretability of the resulting principal components and their relationship to the original features has been a subject of interest. Researchers have sought to develop techniques to better understand and interpret the meaning and contribution of individual features in the reduced-dimensional space.

Additionally, the state of the art acknowledged the need for exploring alternative dimensionality reduction techniques beyond PCA. Researchers have investigated methods such as t-SNE, LLE, and ISOMAP, among others, aiming to address specific challenges and provide complementary approaches to dimensionality reduction.

3  SOLUTION

3.1 EXCEL

We propose a user-friendly solution to facilitate non-programmers in accessing data from databases through a user interface designed in Excel. The solution involves the following functions:

1. Database Connection: Users input the relevant information of the server and establish a connection to the desired database.
2. Download and Upload Data: Users can easily download data from or upload data to a specific table within the chosen database. The downloaded data is automatically saved in a worksheet named after the table.
3. Create New Tables: To prevent users from uploading data that does not exist in the database, we have implemented a feature that allows users to input the information of the table they want to create. This includes specifying the table name, attributes, and attribute types.

By providing an intuitive user interface within Excel, our solution empowers non-programmers to interact with databases efficiently. Users can seamlessly retrieve and manipulate data, as well as create new tables with the assurance of data integrity. This approach bridges the gap between non-programmers and databases, enabling easier access to valuable data resources.

3.2 PCA

In this study, we employed Principal Component Analysis (PCA), a statistical method widely utilized for dimensionality reduction in data analysis, to address the challenges associated with high-dimensional data. High-dimensional data often poses various difficulties, including feature correlation, computational costs, and overfitting problems.

Feature correlation is a common issue in high-dimensional data, where certain features exhibit strong correlations with each other, leading to problems such as multiple solution ambiguity and redundancy. To mitigate this challenge, PCA examines the data comprising multiple variables, identifies correlations between these variables, and determines the optimal combination of values that effectively captures the differences in the results. By employing these combined feature values, PCA facilitates the construction of a more concise feature space.

Furthermore, computational costs can be a significant concern when dealing with large sample sizes and a high number of features. Excessive features demand increased memory requirements, thereby reducing operational efficiency. PCA addresses this challenge by reducing the dimensionality of the data, allowing for more efficient storage and processing.

Additionally, high-dimensional data can make training models more susceptible to overfitting issues. With numerous complex features, models tend to capture excessive noise and detail, hindering their ability to generalize to new data. PCA aids in overcoming this problem by extracting the most informative components, ensuring that the retained principal components represent the essential underlying structure of the original data.

In our project, we applied the PCA function to our dataset, which initially consisted of 13 dimensions. Through PCA, we were able to retain three principal components that captured the maximum variability in the original data. The concept of explained variance ratio was utilized to measure the contribution of each principal component to the overall variability of the original data.

By employing PCA as a dimensionality reduction technique, we effectively addressed the challenges posed by high-dimensional data in our analysis. The retained principal components provided a compact and informative representation of the original data, enabling us to better understand its underlying structure and facilitate subsequent analysis and interpretation.

4  RESULT

4.1 EXCEL

In our study, we developed a user interface that facilitates easy data input and management. The user interface, depicted in Figure 1 and 2, features a left column where users can input relevant information of the server. A dropdown menu displays all available databases and tables, allowing users to select a specific database and table for data upload. Once a database and table are chosen, the corresponding attributes are shown under the button. Users can then record data based on the displayed attributes.

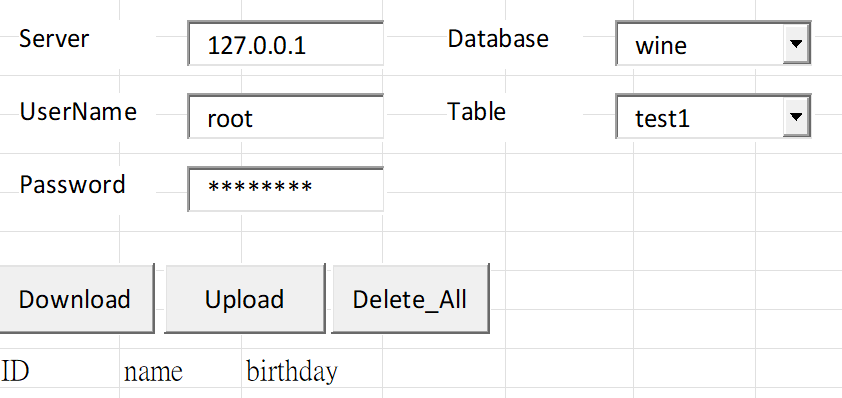


Figure 1: Database connection / Download and upload data

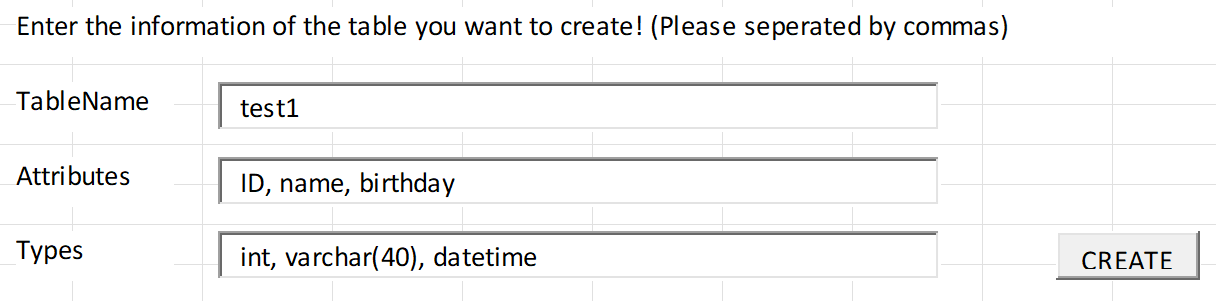


Figure 2: Create new tables

Upon pressing the Upload button, the data undergoes a two-step processing procedure involving PCA and the kNN algorithm. PCA is applied to reduce the dimensionality of the data and extract meaningful features. Subsequently, the reduced-dimensional data is classified using the kNN method. The kNN algorithm assigns each data point to a specific class based on its proximity to neighboring data points in the feature space.

Following the classification step, the data is stored in the database. We leverage the classification results to organize the data into different tables based on their assigned classes. This approach enables faster retrieval of relevant data by querying specific tables based on the desired classification. By employing PCA and kNN in the data processing pipeline, we enhance the efficiency and accuracy of data storage and retrieval within the database.

4.2 PCA

The analysis of Table 1 reveals that a significant portion of the data variability (99.81%) can be effectively captured by the first principal component alone. The second and third principal components, although less easily interpretable, exhibit variations that are specific to different datasets, making them valuable and reserved for future utilization.

Moreover, scatterplot matrices serve as intuitive visualization tools, enabling the observation of correlations between variables, scatter patterns, and potential trends. The scatterplot matrix analysis further supports the effectiveness of the principal components. Specifically, combinations that incorporate the first principal component tend to provide superior discrimination between the three distinct data labels when represented in a two-dimensional plot. Conversely, combinations that do not include the first principal component fail to achieve a clear resolution of the different data labels.

These findings emphasize the importance of the first principal component in capturing the most significant variability within the data. Leveraging this principal component in combination with others, when necessary, enhances the distinction between data labels and facilitates a comprehensive understanding of the dataset's underlying structure. The scatterplot matrix analysis serves as an additional validation of the effectiveness and utility of the selected principal components in enhancing data visualization and interpretation.

Table 1. The explained variance ratio is the percentage of variance that is attributed by each of the selected components

|  |  |  |  |
| --- | --- | --- | --- |
| Principal Components | PC1 | PC2 | PC3 |
| Explained Variance Ratio (%) | 99.81 | 0.17 | 0.01 |

5  CRITIQUE

5.1 EXCEL

One notable strength of our approach is the seamless integration of Excel with a database through our user interface. By connecting Excel to a database, we provide users with the familiar Excel environment and its extensive range of functionalities, while also enabling them to interact with and manipulate database data. This integration eliminates the need for users to switch between different software applications or learn new tools, making it convenient and efficient for users who are already proficient in Excel.

However, it is important to acknowledge that our approach has certain limitations. One of the main limitations is the restricted scope of database manipulation offered by our user interface. Although it provides basic functionality for data input, retrieval, and storage, it may not fulfill the requirements of users who need to perform more advanced operations commonly found in dedicated database management systems. Complex queries, custom data structures, and advanced data processing algorithms are beyond the capabilities of our interface, which could limit the flexibility and sophistication of data manipulation.

5.2 PCA

We provided a clear and concise explanation of PCA and its relevance in reducing the dimensionality of data while retaining maximum variability. We effectively highlighted the specific challenges of feature correlation, computational costs, and overfitting problems in high-dimensional datasets, establishing a strong foundation for the necessity of employing PCA as a solution.

Furthermore, we successfully communicated our methodology and its implementation in the project. Our decision to retain three principal components from the original 13 dimensions of data was appropriately justified. We utilized the concept of explained variance ratio to measure each principal component's contribution to the overall variability of the data. By incorporating these details, we demonstrated a robust approach to dimensionality reduction and the selection of informative components.

6  POSSIBLE EXTENSION

Some potential extensions that can be explored based on the findings and methodology presented in this project are:

1. In order to provide users with more advanced functionality and enhance their experience with the database, possible extensions to our user interface include integrating additional features such as advanced querying capabilities, data visualization tools, and data mining algorithms. By incorporating these extensions, users will be able to perform complex data analysis, generate insightful visualizations, and uncover valuable patterns and trends within their data.
2. Feature Importance and Interpretability: While PCA effectively reduces the dimensionality of the data, the resulting principal components may lack direct interpretability. To address this, future research could focus on exploring techniques to assess the importance of individual features within each principal component and their contribution to the overall variability. Methods such as feature loading analysis or correlation analysis between the original features and principal components could be employed to gain insights into the significance and interpretability of the features in the reduced-dimensional space. This extension would enhance the understanding of the relationship between the original features and the principal components obtained through PCA.
3. Application to Different Datasets: Extending the analysis to different datasets and domains would provide a broader perspective on the applicability and effectiveness of PCA. Investigating how PCA performs on diverse datasets with varying characteristics and complexities would highlight its strengths and limitations in different contexts. This extension could involve applying PCA to datasets from various domains, such as healthcare, finance, or image analysis, and evaluating its performance in terms of dimensionality reduction, information retention, and subsequent analysis. By examining the behavior of PCA across different datasets, a more comprehensive understanding of its generalizability and utility can be gained.

Exploring these potential extensions would contribute to the advancement of knowledge in the field of dimensionality reduction and provide valuable insights into the interpretability of principal components as well as the applicability of PCA across diverse datasets and domains.

ACKNOWLEDGMENTS

Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here.

REFERENCES

[1] Zoomer Analytics, L. (2023). API Reference - xlwings Documentation. https://docs.xlwings.org/zh\_TW/latest/api/index.html

[2]

UPDATED PRESENTATION VIDEO

WORK ASSIGNMENT TABLEConference Name:ACM Woodstock conference

|  |  |
| --- | --- |
| Member | Work |
| LIN, BO-YONG | Organize meetings and project information, implement PCA method, map workflows |
| KUO, TING-YI | Establish connection between Excel and database, design an intuitive user interface, introduction |
| TSENG, YU-HSUAN |  |
| ZHANG, YU-JIE |  |

Conference Name:ACM Woodstock conference

Conference Short Name:WOODSTOCK’18

Conference Location:El Paso, Texas USA

ISBN:978-1-4503-0000-0/18/06

Year:2018

Date:June

Copyright Year:2018

Copyright Statement:rightsretained

DOI:10.1145/1234567890

RRH: F. Surname et al.

Price:$15.00

1. The existing Bibstrip data, copyright text and permission block in the sample file are dummy values, so the user needs to provide the correct values required for the submission in the metadata dialog box. [↑](#footnote-ref-2)