ISBN 978-981-14-4787-7
Proceedings of 2020 the 10th International Workshop on Computer Science and Engineering (WCSE 2020)
Shanghai, 19-21 June, 2020, pp. 6-12
doi: 10.18178/wcse.2020.06.002

A Visualization Recommendation Approach based on Machine Learning

Shichao Wei[1,2,3,4], Xin Li[1,2,3] , Peiyin Song[2,3,4,5] , Xiaofeng Zhou[1,2,3+], Yichi Zhang[1,2,3], Shuai Li[1,2,3,4]

[1]Shenyang Institute of Automation,Chinese Academy of Sciences,Shenyang,110016,China

[2]Key Laboratory of Network Control System, Chinese Academy of Sciences, Shenyang, 110016, China

[3]Institutes for Robotics and Intelligent Manufacturing, Chinese Academy of Sciences, Shenyang, 110016, China

[4]University of Chinese Academy of Sciences, Beijing, 100049, China

[5]Shenyang University of Technology, China

Abstract. As an important means of data analysis, data visualization is used by more and more people. For most people who don't have visualization technology expertise, data visualization has some Visualization recommendation aims to lower the barrier to exploring basic visualizations by automatically generating results for analysts to search and select.This paper proposes a visual recommendation method based on machine learning, which can learn the most meaningful visualization results from many visualization practice datasets and mark them.Firstly, 22 data features and corresponding meaningful visualization types are extracted from 30 real visualization datasets. Then, binary classifiers are used to train the classification model, from which we can learn meaningful visualization and use crowdsourced testsets to test the accuracy.Finally, the results of multiple classifiers are fused to vote for multiple meaningful charts in the datasets.Experiments show that this method can effectively learn the meaningful visualization types in datasets, mark and recommend them to users.

Keywords: visualization recommendation, machine Learning, classification model

1. Introduction

Knowledge workers—from business to science research—increasingly use data visualization to generate images,communicate,and make decisions[1][2][3][4].Knowledge workers can use basic data processing software (such as Microsoft Excel, Google spreadsheets, etc.) and professional BI analysis software (such as Tableau, Qlik, etc.) to generate simple visualization quickly[5]. But these software have limited visualization forms and low flexibility.In order to make visualization more expressive and have more control over visualization, command API (such as OpenGL, HTML canvas) can be used to generate customized visualization[6]. However, the command API requires professional programming skills and a large number of working hours. For most people, the learning curve is steep and the operation is difficult.

In order to balance the direct relationship between speed and presentation, researchers developed declarative canonical grammar, such as ggplot2[7],D3[8], Vega[9]and Vega-Lite[10], etc. Similarly, these grammars are accompanied by a steep learning curve, and few people are proficient in using them.For example, the application of D3 is difficult for people who are not familiar with JavaScript language, because it is designed based on JavaScript and combines the professional knowledge of computer graphics. Even for users who are familiar with JavaScript, it is difficult to apply, ggplot2 and Vega are no exception. In order to solve these challenges, researchers developed some tools[11][12][13][14][15] to automatically design

+ Corresponding author. Tel.: +15040373027;
E-mail address: zhouxf@sia.cn.

effective visualization and guide users in the visualization exploration, and the visualization recommendation system came into being.

Most visual recommendation systems are based on the set of "If then" rules[16], and generate visualization automatically through explicit enumeration and heuristic methods. For example, APT[17], BOZ[11] and SAGE[15] generate and rank visualizations using rules informed by perceptual principles. Explicit enumeration and heuristic rules limit the extensibility of these applications. Recent systems such as Voyager2[18], Show Me[19] have extended these approaches by supporting for column selection. But these systems needs more human interventions and operations, and face the problem of cold-start. Consequently researchers turned their attentions to the data characteristics of visual analysis and data-driven visual recommendation.SeeDB[20] defines a visual result that deviates from the user's given reference as "interesting" and serves as a recommendation candidate by finding the deviation and exception between data, but it takes a lot of calculation time. By contrast, the visual recommendation system based on machine learning (ML) directly learns the relationship between data and visualization through training model, which provides a new idea for visual recommendation. Data2vis[21] trains an end-to-end neural translation model, which represents visual generation as a language conversion problem, learns the vocabulary and syntax of visual specifications from a given data set, and directly maps the data specifications to the visual specifications in the declarative language (Vega-lite) .VizML pays more attention to design choices, which collects a large number of training datasets from the online visualization platform to predicts their design options.Although these methods have achieved high accuracy, they rely on a large number of datasets for training, and have poor extensibility, which is not conducive to be integrated with application.

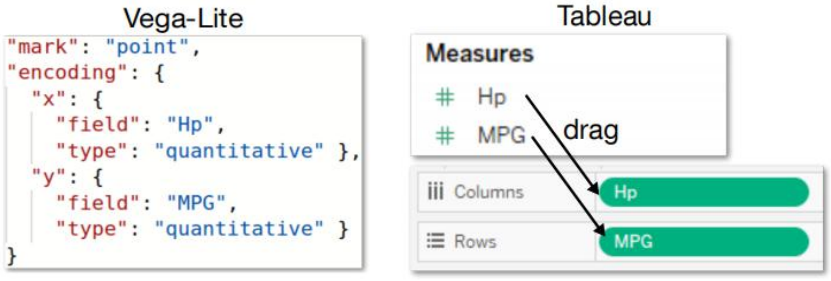


Fig. 1: Vega-lite language and Tableau software.

In this paper, a visualization method named CL-viz is proposed, This method focuses on whether the visualization results are meaningful.We collect datasets and corresponding visualization results from a large number of real visualization cases, mark the visualization results that meet people's understanding and perception as "meaningful", and mark the messy visualization results as meaningless.Then 22 data features (data type, maximum value, minimum value, distribution law, etc.) of each data column are extracted for training classification model.Finally, a crowdsourcing test set is used to test the accuracy of the model, and the method of integrated learning is used to improve accuracy.Experiments show that the method can effectively get the meaningful visualization results of the corresponding data features and mark them out.

2. Related Work

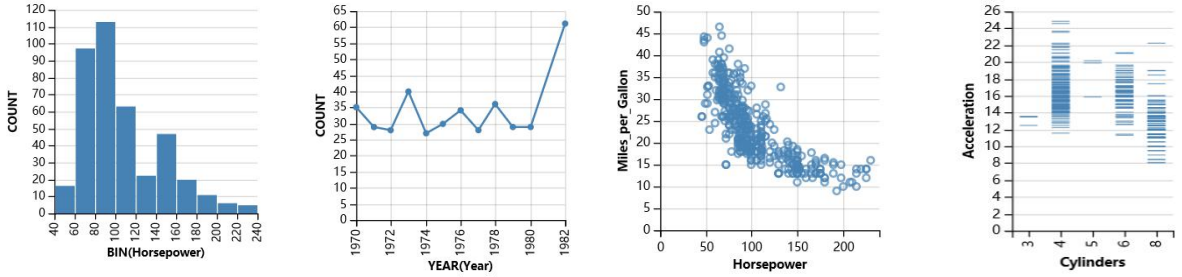
2.1. Meaningful Visualization

First of all, a real visualization application[22] was analyzed. Figure 2 is a real case from Voyager2 about the performance analysis of a variety of vehicle models. The data includes a time attribute, three category attributes and five digital attributes.

- Figure 2 (a) is a bar chart, where the x-axis represents the bin of vehicle horsepower, and the y-axis represents the number, among which the number of horsepower between 80-100 is the most, it clearly reflects the distribution characteristics of vehicle horsepower in the data set, which is called a meaningful visualization result.
- Figure2(b) is a line chart, where the x-axis represents the year, and the y-axis represents the number of vehicle samples in the data, among which the number of vehicles produced in 1982 is the largest,

reflecting the relationship between the number of vehicles and the year of production, which is called a meaningful visualization result.

- Figure2(c) is a scatter chart without any transformation operation. The x-axis represents the value of horsepower, and the y-axis represents the value of miles per gallon. It can be clearly seen that there is an obvious inverse relationship between horsepower and miles per gallon, which is called a meaningful visualization result.
- Figure2(d) is a stacked chart, where the x-axis represents Cylinders and the y-axis represents acceleration. It seems that the number of cars with 4 cylinders is the most and the number of cars with 3 cylinders is the least. However, this information is more suitable for bar chart representation than the visualization results shown in the figure. It is also difficult to see the relationship between Cylinders and acceleration from the chart.The relationship between Cylinders and acceleration is not suitable to be represented by stack graph, which is called meaningless visualization results.



(a) (b) (c) (d)

Fig. 2: Sample visualizations for the car from voyager2

We selected 15 workers with visualization experience to evaluate the visualization results in the above examples. They agreed that the visualization in Figure2 (d) could not clearly show the relationship between Cylinders and acceleration, which could be considered as meaningless chart.

Generally, finding a good visualization result is a process of exploration and mining. For those who have professional visualization knowledge, it is also difficult. Therefore, learning visual design patterns automatically from real visual practice and filtering out meaningless visual results is particularly important to retain meaningful results.

For the label of training data set, it usually needs human calibration. At present, there are some online data analysis and visualization platforms, including well-designed visualization results. In the process of data analysis based on Bi software, we have accumulated a large number of visualization results of real datasets, which are considered as meaningful visualization. In order to further improve the data quality, we enumerate all possible visualization types (including bar, pie, line, and scatter charts) in the experimental data set and send them to 15 workers with visualization experience for manual marking. The results show that the professional experience of worker members is roughly the same as that of existing online software for marking meaningful charts.

3. Method

3.1. Data

Qlik is a BI analysis software for data visualization. It allows users to upload datasets, which are related to each other in the way of star link or snowflake link, and create more than 20 types of interactive charts manually. The creation of visualization is manually specified, including selecting chart types and specifying dimensions and measures.

With the help of BI analysis tools for data visualization, we have accumulated a large number of data visualization results. In the past visualization, the number of columns and rows of datasets is very different. Although some datasets contain hundreds of column attributes, the vast majority of datasets are less than 25 columns(Fig.3). The datasets with the best visualization practice usually contain about 10 attribute columns. Although there are many chart types to display datasets, more than 85% of visualization results can be represented by bar, pie, line, or scatter charts(Fig.4). Therefore, we select a typical data set with 30 attributes

listed between 4-25 for training model, which includes time series attributes, classification attributes, and numerical attributes. A part of training sets information is shown in table 1.

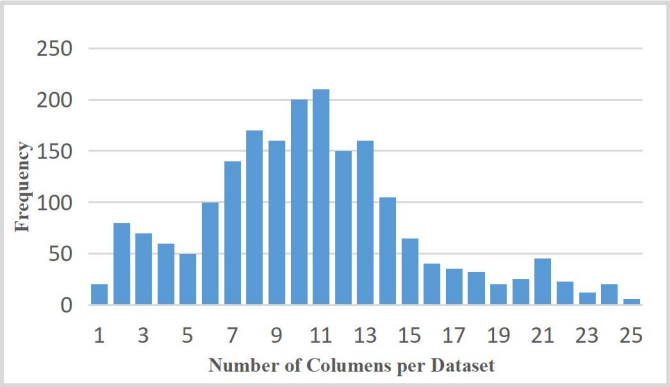


Fig. 3: Distribution of columns per dataset after removing more than 25 columns

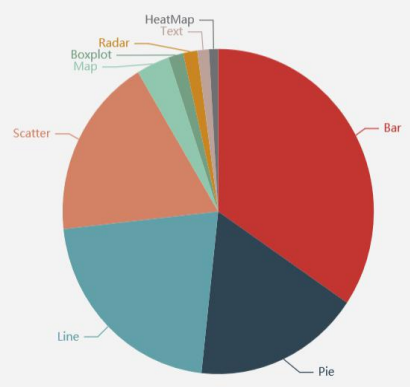


Fig. 4: Frequency of chart types

Table 1: Training datasets information

Dataset	Columns	Categorical	Quantitative	Temporal
Foreign Visitor	4	Y	Y	N
Clothing sales	6	Y	Y	Y
Water treatment	24	N	Y	Y
Forest fires	12	Y	Y	N
Train maintenance	12	Y	Y	Y

3.2. Features

For training datasets, the characteristic descriptions of column attributes are listed in table 2. Features are divided into six parts(length, type, distinct values, statistical, distribution and Pairwise-column features). Types features capture whether a column is categorical, temporal, or quantitative. For categorical, ratio, entropy and gini is calculated. For quantitative, maximum, minimum, mean, median, mode, variance, standard, median absolute deviation and distribution is calculated. For pairwise column features, depend on the individual column types determined through single-column feature extraction. For instance, the Pearson correlation coeffificient requires two numeric columns, and the χ^2 feature requires two categorical columns.

Table 2: Dataset features

Feature	description
Length	The length of values
Type	Categorical (C), quantitative (Q), temporal (T)
Distinct values(C)	The number of distinct values in column,The ratio of distinct values,Entropy,Gini
Statistical(Q)	Maximum,minimum,mean,median,mode,variance,standard,median absolute deviation
Distribution(Q)	Normality,skewness,kurtosis
Pairwise-column features	Correlation(Q-Q), χ^2 (C-C)

In daily visualization practice, more than 85% of visualization results can be represented by bar, pie, line, or scatter charts. In this paper, we only consider the recommendation of four kinds of visualization graphics, and mark bar, pie, line and scatter charts with 0-3 respectively. Four types of visualization charts from one column and two columns are enumerated.For a data set with m columns, there are 4m cases for each column and 2m(m-1) possible visualizations for two columns.According to the characteristics listed in table 2, 21453 feature vectors are extracted from 30 datasets.

3.3. Experiment

30 real datasets from different fields are used for training , which include automobile, chemical industry, transportation, sales and other fields. Some of the column attributes contained in the data are listed in table 1. According to the visualization practice in BI analysis software, we label each data set and enumerate all possible visualization results. Combined with Bi software and 15 researchers with visualization experience, we label all 21453 visualization results respectively. Finally, 680 meaningful visualization results are obtained. By using decision tree (DT), support vector machine (SVM) and Bayes classifier, 21453 pieces of data collected are used for model training, and six test datasets are selected on UCI for testing, which is shown in table 3.

Table 3: Testing datasets information

Dataset	Columns	Categorical	Quantitative	Temporal
Adult	14	Y	Y	N
Most Profitable Stories	8	Y	Y	Y
Credit Approval	15	Y	Y	N
Flag design	10	Y	Y	N
Summer Olympic	9	Y	Y	Y
Baby names	6	Y	Y	N

Table 4 shows the accuracy of three classifiers on 6 testing datasets. Decision tree has the highest accuracy in the six test data sets, with an average accuracy of 0.8609. Bayes has the lowest accuracy, with an average accuracy of 0.7223. The output of the three classifiers is the judgment of all possible visualization results. 1 is meaningful visualization results, 0 is meaningless visualization results, and marks its index in the data set. Through the index, the meaningful visualization performance type can be located.

Table 4: The accuracy of DT, SVM and Bayes

Dataset	DT	SVM	Bayes
Adult	0.8667	0.8412	0.7269
Most Profitable Stories	0.8633	0.8278	0.7178
Credit Approval	0.8912	0.8224	0.7333
Flag design	0.8467	0.8019	0.7065
Summer Olympic	0.8542	0.8533	0.7367
Baby names	0.8435	0.8166	0.7129
AVE	0.8609	0.8272	0.7223

In order to improve the accuracy, we were inspired by ensemble learning methods. A model combining three simple classifiers is trained, and the class with the most votes is marked as the output result by the relative majority voting method. If multiple class marks get the highest votes, one of the class labels is randomly selected as the output.

Fig.5 shows the ensemble learning algorithm flowchart. Table 5 shows the accuracy of ensemble learning model. It can be seen from the table that the accuracy of ensemble learning is improved slightly.

Table 5: The accuracy of DT+SVM+Bayes

Dataset	DT+SVM+Bayes
Adult	0.8776
Most Profitable Stories	0.8833
Credit Approval	0.9125
Flag design	0.8662
Summer Olympic	0.8937
Baby names	0.8647
AVE	0.8830

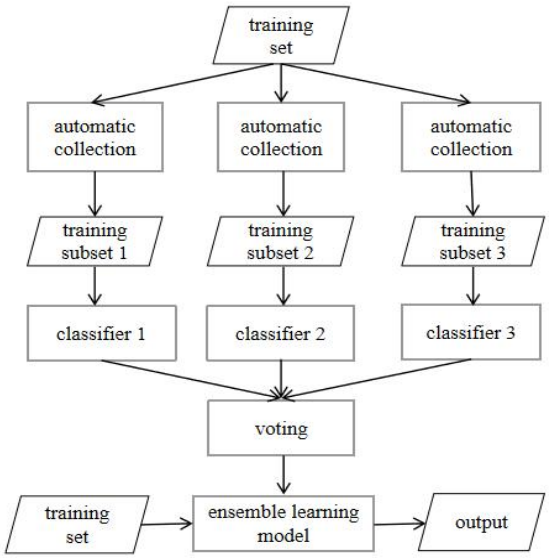


Fig. 5: The ensemble learning algorithm flowchart

But beyond that, the recommended chart types are recorded to find patterns between data types and chart types. Table 6 shows the recommended chart types by six test datasets.It can be seen from the table that bar is the most popular type. Almost all datasets can be represented by bar charts. Line chart prefers to represent some data with temporal attribute, because it is more about the trend of data over time. Two datasets with temporal attribute are both recommended line chart.

Table 6: The recommended chart types

Dataset	Bar	Pie	Scatter	Line
Adult	Y	Y	N	N
Most Profitable Stories	Y	Y	Y	Y
Credit Approval	Y	Y	N	N
Flag design	Y	Y	Y	N
Summer Olympic	Y	N	Y	Y
Baby names	Y	Y	Y	N

4. Conclusion

In this paper, a visual recommendation method based on machine learning is proposed. This method can learn the most meaningful visualization results from many visualization practice datasets and mark them. It can eliminate the redundancy of visualization results caused by the large number of visualization types and enumeration search space. For each visualized dataset, the features of different column attributes are extracted, which is of great significance to model learning. The binary classifier used in this paper and it can effectively learn the meaningful visualized results. The experimental results show that the approach achieves high accuracy. Meanwhile, we are inspired by ensemble learning methods.Multiple binary classifiers are integrated and the results show that the accuracy has been slightly improved, which provides a new idea for the next research.

5. Future Work

Although the method proposed in this paper has certain accuracy in the judgment of meaningful and meaningless visual results, there are still some areas to be improved.For example, our method is to learn meaningful visualization results from many visualization practice data sets and build a model. The output is the mark of visualization results, which does not involve the presentation of the final chart. That requires additional visualization language coding, which is a heavy task.In addition, this training set comes from long-term visualization practice, and the quality and quantity of training set affect the accuracy of the

model.In this paper, only 22 features of the datasets are extracted, and more features may improve the accuracy of the experiment. In the next research, we will analyze the impact of each feature on the results, and try to find the most effective feature combination.

6. Acknowledgements

This work was supported by LiaoNing Revitalization Talents Program under Grant XLYC1808009.

7. References

[1] K. Z. Hu, M. A. Bakker, Stephen Li, etc. VizML: A Machine Learning Approach to Visualization Recommendation. *CHI Conference on Human Factors in Computing Systems (CHI),2019.*

[2] S. Kandel, A. Paepcke, J. M. Hellerstein, and J. Heer. Enterprise Data Analysis and Visualization: An Interview Study. *IEEE Transactions on Visualization and Computer Graphics,* 2012, 18(12):2917-2926.

[3] E. Segel and J. Heer. Narrative Visualization: Telling Stories with Data. *IEEE Transactions on Visualization and Computer Graphics,* 2010, 16(6):1139-1148.

[4] E. Brynjolfsson and K. McElheran. The Rapid Adoption of Data-Driven Decision-Making. *American Economic Review,* 2016, 106(5):133-39.

[5] O.Donovan, Mark. *Qlik Sense for Beginners*, 2014.

[6] Randi J. Rost. *OpenGL shading language*. Addison Wesley, 2004.

[7] H. Wickham. A layered grammar of graphics. *Journal of Computational and Graphical Statistics*, 2010, 19(1): 3-28.

[8] M. Bostock, V. Ogievetsky, and J. Heer. D3: Data-driven documents. *IEEE TVCG (Proc. InfoVis), 2011.*

[9] A. Satyanarayan, R. Russell, J. Hoffswell, and J. Heer. Reactive vega: A streaming dataflow architecture for declarative interactive visualization. *IEEE TVCG (Proc. InfoVis), 2016.*

[10] A. Satyanarayan, D. Moritz, K. Wongsuphasawat, and J. Heer. Vegalite: A grammar of interactive graphics. *IEEE TVCG (Proc. InfoVis),* 2017.

[11] S. M. Casner. Task-analytic Approach to the Automated Design of Graphic Presentations. *ACM Trans. Graph,* 1991, 10(2):111–151.

[12] C.Demiralp, C. Scheidegger, G. Kindlmann, D. Laidlaw, and J. Heer. Visual embedding: A model for visualization. *IEEE CG&A,* 2014.

[13] G. Kindlmann and C. Scheidegger. An algebraic process for visualization design. *IEEE TVCG*, 2014, 20:2181–2190.

[14] T. Siddiqui, A. Kim, J. Lee, K. Karahalios, and A. Parameswaran. Effortless data exploration with Zenvisage. *PVLDB*, 2016, 10(4):457–468.

[15] S. F. Roth, J. Kolojechick, J. Mattis, and J. Goldstein. Interactive graphic design using automatic presentation knowledge. *In ACM Human Factors in Computing Systems (CHI),* 1994.

[16] F. Hayes-Roth. Rule-based Systems. Commun. ACM, 1985, 28(9):921-932.

[17] J. D. Mackinlay. Automating the Design of Graphical Presentations of Relational Information. *ACM Trans. Graphics,* 1986, 5(2): 110-141.

[18] K. Wongsuphasawat, Z. Qu, D. Moritz, R. Chang, F. Ouk, A. Anand, J. Mackinlay, B. Howe, and J. Heer. Voyager 2: Augmenting visual analysis with partial view specifications. *In ACM CHI, 2017.*

[19] J. Mackinlay, P. Hanrahan, and C. Stolte. Show me: Automatic presentation for visual analysis. *IEEE TVCG,* 2007, 13(6):1137-1144.

[20] M. Vartak, S. Rahman, S. Madden, A. Parameswaran, and N. Polyzotis. SeeDB: Efficient data-driven visualization recommendations to support visual analytics. *PVLDB,* 2015, 8(13):2182-2193.

[21] Victor Dibia, Cagatay Demiralp. Data2Vis: Automatic Generation of Data Visualizations Using Sequence to Sequence Recurrent Neural Networks. *IEEE Computer Graphics and Applications,* 2018

[22] Voyager2. https://vega.github.io/voyager2/