Visualization of High-Dimensional Machine Learning Data

Jessica Pauli de C. Bonson, Dalhousie University

Abstract

Visualization can be used to better understand the data for categorization tasks and effectively apply machine learning techniques. This work presents a visualization tool for the analysis of high-dimensional classification data, that integrates scatter matrix, histograms and star glyph representations. The system has the features of selection, filtering, and zooming, including options for automatic filtering using feature selection algorithms. The system was tested extensively to avoid bugs, and provide smoothly integrated behaviors.

Keywords

information visualization, machine learning, high-dimensional.

I. Introduction

VISUALIZATION techniques graphically presents the information in order to allow a better understanding and analysis of the data by the users. It is a powerful method to improve the decision-making process based on data for fields such as biological sciences, engineerings, and computer science. In the case of the machine learning domain, the visualization of the dataset is essential to gather insights, find patterns and define the most effective techniques to process and analyze the data. Scatterplots [1] are one of the most popular approaches to visualize multidimensional data with simplicity, expressiveness and clarity. Scatter matrices [2] expands scatter plots, allowing the visualization of various dimensions at the same time. Star glyphs [3], or star plots, are also a recognized visualization technique for multi-dimensional data, where a glyph can be used to represent a single data point, all the values for a single attribute, or a summary of the data point values across all attributes. Considering the classification domain, this project represents each class by a glyph, and each dimension by an axis. Where the axis range is the minimum and maximum values for that attribute, and the points composing the glyph polygon are the average of the values for that attribute for a class. Since that the projection of all the dimensions may be too complex and hide the data main characteristics, feature selection and feature extraction techniques [4] can be used to improve the dataset visualization.

This work uses Python, Javascript and D3 to implement a system for the visualization and analysis of datasets for classification tasks in machine learning problems. It is composed of an integrated visualization of the dataset using scatter matrix, histograms and star gryphs. The users are able to select data points, that are shown across scatter plots and are updated on the star plot. Another feature is the filtering of attributes, that also modifies the visualization on both charts. The last feature is the zooming of charts on the matrix, so that the data can be seen in more details. The next sections are organized as follows. Section 2 will briefly compare this work against previous works. In the Section 3 the system implementation and details of its features will be explained. Section 4 presents the conclusion and the future work.

II. Related Work

This section will give an overview of similar works that focus on visualization of datasets using scatter matrices, scatter plots, or star plots. In [5] the authors work on reduction of information clutter, in order to improve the understanding of the visualization of a dataset without modifying the data itself. The authors applied their technique to parallel coordinates, scatter plot matrices, and star plots. They produced better visualization, with clearer shapes for the star plots and a more organized distribution of the data for the scatter matrix. In a future work, this project could be improved with information clutter reduction to further enhance the data analysis. The work in [6] investigate new interactions to explore multidimensional data using scatter plots and scatter matrices. Some of the new features presented are interactive queries, dimension reordering, and 3D navigation, all focused on comparing and correlating dimensions across the dataset. [7] also focus on exploring the dimensions of the dataset. Their technique consists of creating a hierarchy of scatter plots using decision trees and subplots, instead of analyzing the dimensions using a scatter matrix. Their approach allows the user to decide how she wants to interactively visualize and relate the different dataset dimensions using an exploratory method.

[8] developed a technique for multidimensional data, Star Coordinates. The method is similar to a star plot, but uses points instead of a glyph to represent the data, and each point represents a data point, instead of a summary of the values. Each axis have different sizes, depending on the range of values in the dataset. The authors goal was to enhance the user ability to obtain insights through the data visualization. In [9] work, star glyphs and parallel coordinates were merged as a way to improve the clarity of parallel coordinates for high-dimensional datasets. [10] present StarClass, an interactive visual classification method

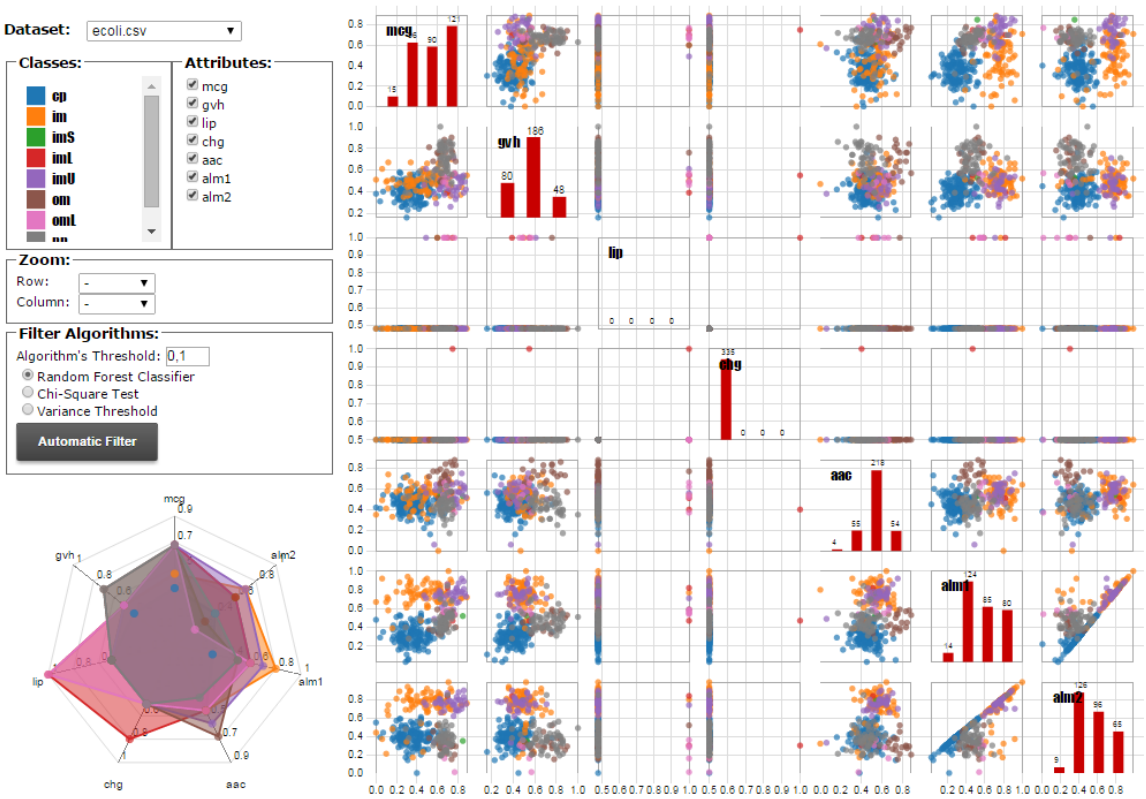

Fig. 1: System interface for the Ecoli dataset, composed of menu (upper left), star plot (lower left), and scatter matrix (right).

for multi-dimensional data that uses star coordinates to display the data points of the dataset. Besides the location on the star coordinates, each data point is represented as a set of centered lines, where each line size means the value of the data point for an attribute.

The main difference of this project regarding the previous work is that it uses the integrated visualization of scatter matrix, histograms and star glyphs, so that all visualization are integrated for all the possible interactions (selection, filtering, and zooming). Among the related work, this project uses star glyphs differently, by presenting a summary of information, for each class and each attribute. Each glyph represents the minimum, maximum, and average values of all the data points, or only for the selected data points. The goal of this work is to produce a tool that could be used for real-world machine learning problems, that is able to analyze any classification dataset in the CSV format. The related work is interesting as a way to further improve the current project.

III. Visualization Tool

A. Main System

This work implements a visualization system for the analysis of high-dimensional classification data for machine learning problems. The server side of the code was implemented in Python, using the Web framework Flask [11] and the machine learning library scikit-learn [12], that was used to implement the feature selection algorithms. Javascript was used on the client side, along with D3 [13] and JQuery. Figure 1 shows the tool interface. The main parts of the system are the scatter matrix, the star plot, and the interactive menu. The input is a CSV file with the attributes, classes, and data points. It is possible to work with any data in the CSV format. The current code has the files for the Iris [14] and Ecoli [15] datasets. In Figure 1 the system is loaded with the Ecoli dataset, and on Figure 2 the system is loaded with the Iris dataset, that have a very different quantity of classes, attributes, and data points. The tool is able to generalize for datasets with various quantities of classes, attributes and data size successfully. In the next subsection details of the system parts will be further explained.

1) Menu: The first part of the menu, 'Classes', has all the classes in the dataset and its colors representations. The second column presents the attributes of the dataset, where the checkboxes can be used to filter them. Below, the 'Zoom' options enable the user to choose the indexes of the chart that she wants to zoom. It isn't a very practical approach, so in the future it may be

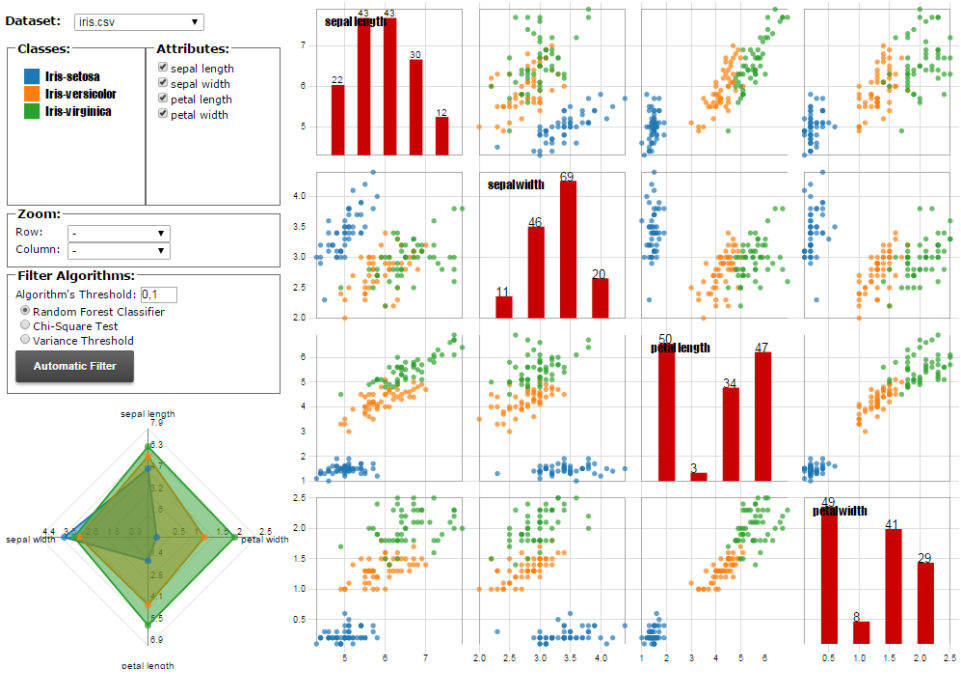

Fig. 2: Visualization of the Iris dataset.

improved to work with the mouse wheel or with a smaller sample version of the scatter matrix. The last option on the menu is 'Filter Algorithms', that provides three machine learning approaches for feature selection. All the three approaches rank the attributes importance and use a threshold to choose the best ones. The user is informed if the threshold choice filtered out all or none of the attributes, and in both cases the data visualization isn't modified.

2) Scatter Matrix: The scatter matrix crosses all attributes data and shows scatter plots for each combination of attributes. For the cells were row and column represents the same attribute, a histogram is presented showing the distribution of the attributes across range of values. The name of each attribute is located at the histograms. The scatter matrix axis has the horizontal and vertical axis for each scatter plot, organized in ticks of five values.

3) Star Plot: The star plot has an axis for each attribute, where the range of the axis is proportional to the range of the attribute values. Each class is represented as a colored glyph. The glyphs show the average of the values for each class, for all selected data points. If no data point is selected, it shows the average for all the dataset. The maximum and minimum values in the axis is the maximum and minimum possible values for that attribute.

B. Interactions

1) Selection: The user can use brushing to select data points in one of the scatter plots. All the points representing the same data in the other scatter plots will also be selected. The star plot is updated to represent the average of the selected data points for each class. Figure 3 shows the integrated visualization of the selected points in the scatter matrix and the star plot. Besides the integrated selection, the histograms and the star plot also have unique selections, as shown on Figure 4. By passing the mouse over the bar of a histogram, it is highlighted and shows the range of values that are contained in the bar. For the star plot, the user can pass the mouse over a point or a colored area of a class. The polygon representing that class will be more opaque while the other classes will become more transparent, so the class's glyph can be clearly seen. If the user pass over a point, the value of that point will be shown as a tooltip.

2) Zoom: The zoom function enables the user to expand a chart in the matrix, either a scatter plot or a histogram. The zoomed chart occupy the whole matrix space, and the data is shown in more details. Figure 5 presents a histogram and a scatter plot with zoom. The horizontal and vertical axis use twice ticks to represent the data range, and the histogram also uses two times the number of bars to represent the data distribution.

3) Manual and Automatic Filtering: Users can filter the dataset by selecting the attributes checkboxes, or by applying automatic filter methods. The scatter matrix and the star plot are modified to the new quantity of attributes. There are three algorithms available: Random Forest Classifier [16], Chi-Square Test [17] and Variance Threshold [18]. The Random Forest method scores the features by their importance using the construction of multiple random decision trees. Chi-Square Test calculates dependence

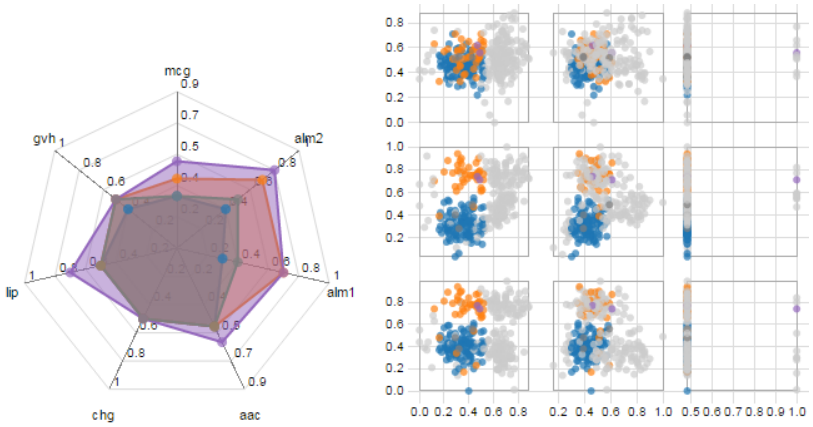

Fig. 3: Brushing the scatter matrix changes the star plot to show the average values of the selected points.

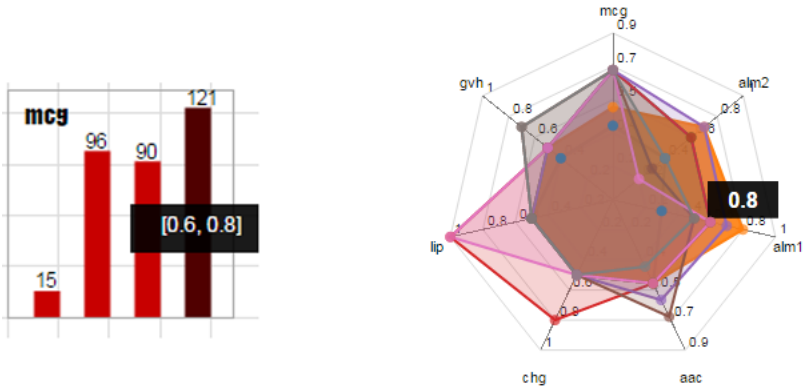

Fig. 4: Selection of column in histogram (left) and selection of class and its attribute value in the star plot (right).

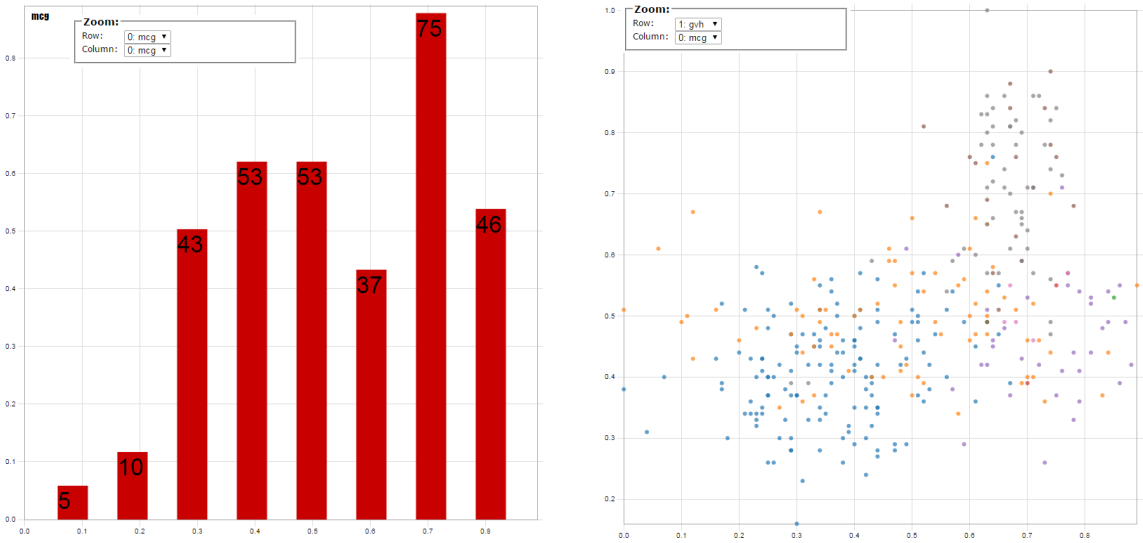

Fig. 5: Zoom of histogram (left), and of scatter plot (right).

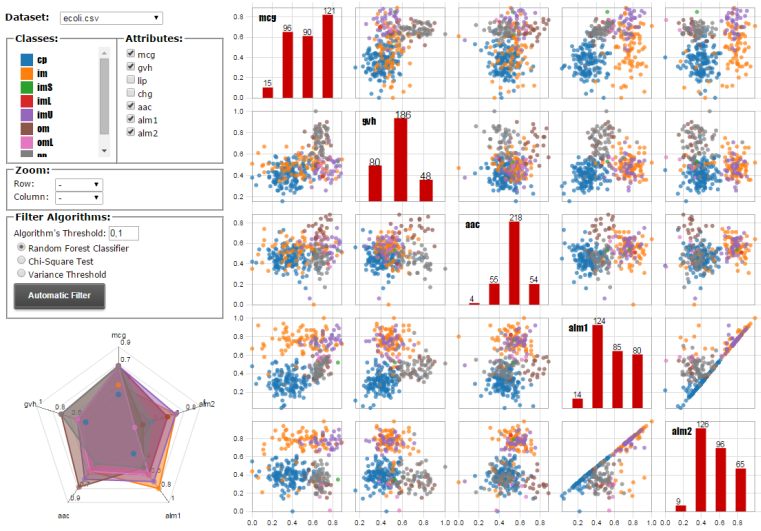

Fig. 6: Result of filtering using Random Forest with threshold 0.1. The attributes lp and chg were filtered, modifying the scatter matrix, the star plot and the interactive interface.

between stochastic variables, to remove features that are likely to be independent of class, i.e. that are irrelevant for the classification task. The Variance Threshold method keeps all features with non-zero variance, or to a variance higher than the threshold, to remove features that have similar values for all data samples. As a matter of curiosity, among the three methods the one who provided the best results was the Random Forest technique.

IV. Conclusion and Future Work

This paper present a visual analysis tool for datasets, focused on classification problems in the machine learning domain. The main components of the system are the integrated visualization of scatter matrix, histograms and star plot. The final user can interact with the data by using selection, filtering, and zooming, that work integrated across visualizations, and also has specific interaction with each component. The star plot summarizes the dataset by providing, for each attribute and each class, the minimum, maximum, and average values for the selected data points. The system helps the understanding of the dataset by providing ways to visualize it using various dimensions, such as the average of values per class, how the attributes interact with each other, and how they are distributed across the possible values. The tool also provides automatic feature selection, so the user can focus in the main parts of the dataset. It is possible to obtain insights of the overall dataset, of specific selection of attributes, and of a group of data points. Some features should be improved in a future work. The zooming feature could be more practical and be used by scrolling the mouse wheel or by clicking in a smaller version of the scatter matrix. Another filtering option could be added, to allow the user to filter the dataset by classes. Also, a relevant modification would be to show the histograms as stacked bars, organized by classes colors, so that the user would be able the analyze how the attributes values are distributed across range values and classes. A last improvement would be to apply the technique described by [5] and [6] to reduce the information clutter and allow the reordering of the charts in the matrix, enhancing the user ability to understand the data.

References

[1] J. Utts, *Seeing through statistics*. Cengage Learning, 2014.

[2] D. B. Carr, R. J. Littlefield, W. Nicholson, and J. Littlefield, "Scatterplot matrix techniques for large n," *Journal of the American Statistical Association*, vol. 82, no. 398, pp. 424–436, 1987.

[3] M. O. Ward, "Multivariate data glyphs: Principles and practice," in *Handbook of Data Visualization*. Springer, 2008, pp. 179–198.

[4] D. D. Lewis, "Feature selection and feature extraction for text categorization," in *Proceedings of the workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992, pp. 212–217.

[5] W. Peng, M. O. Ward, and E. A. Rundensteiner, "Clutter reduction in multi-dimensional data visualization using dimension reordering," in *Information Visualization, 2004. INFOVIS 2004. IEEE Symposium on*. IEEE, 2004, pp. 89–96.

[6] N. Elmqvist, P. Dragicevic, and J.-D. Fekete, "Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 14, no. 6, pp. 1539–1148, 2008.

[7] M. Eisemann, G. Albuquerque, and M. Magnor, "A nested hierarchy of localized scatterplots," in *Proceedings of the 2014 27th SIBGRAPI Conference on Graphics, Patterns and Images*. IEEE Computer Society, 2014, pp. 80–86.

[8] E. Kandogan, "Star coordinates: A multi-dimensional visualization technique with uniform treatment of dimensions," in *Proceedings of the IEEE Information Visualization Symposium*, vol. 650, 2000, p. 22.

[9] E. Fanea, S. Carpendale, and T. Isenberg, "An interactive 3d integration of parallel coordinates and star glyphs," in *Information Visualization, 2005. INFOVIS 2005. IEEE Symposium on*. IEEE, 2005, pp. 149–156.

[10] S. T. Teoh and K.-L. Ma, "Starclass: Interactive visual classification using star coordinates." in *SDM*. SIAM, 2003, pp. 178–185.

[11] Flask, "Flask," http://flask.pocoo.org/, 2010, [Online; accessed 09-April-2015].

[12] scikit learn, "Scikit-Learn," http://scikit-learn.org/stable/, 2010, [Online; accessed 09-April-2015].

[13] D3, "D3," http://d3js.org/, 2013, [Online; accessed 09-April-2015].

[14] UCI, "Iris Data Set," http://archive.ics.uci.edu/ml/datasets/Iris, 1988, [Online; accessed 09-April-2015].

[15] ——, "Ecoli Data Set," http://archive.ics.uci.edu/ml/datasets/Ecoli, 1996, [Online; accessed 09-April-2015].

[16] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[17] D. S. Moore, "Chi-square tests." DTIC Document, Tech. Rep., 1976.

[18] H. Wang, T. M. Khoshgoftaar, and J. Van Hulse, "A comparative study of threshold-based feature selection techniques," in *Granular Computing (GrC), 2010 IEEE International Conference on*. IEEE, 2010, pp. 499–504.