Enhancing EDA and Misclassification Analysis through Tree Structure Visualization

Maor Hornstein

Tabular Data Science (89547) - Final Project

Abstract

This project aims to investigate how the use of tree-based visualization can address the limitations of scatter plot graphs. The first section of the study demonstrates how the tree structure can simulate the process carried out by data scientists when using Scatter Plots in the Exploratory Data Analysis (EDA) stage, and how this approach overcomes their shortcomings. The proposed method was tested on classification problems and was found preferred.

The second part of the study attempted to extend the application of tree-based visualization to another stage in the data-science pipeline - Misclassification Analysis. This attempt yielded limited success: it was discovered that tree-based visualization is particularly useful for local analysis of misclassification, yet for such cases there are already alternative tools available that were found to be more convenient to use.

1 Problem Description

The focus of the solution is the Exploratory Data Analysis (EDA) phase, particularly the use of Scatter Plots in this phase.

Scatter Plots are a useful tool in data analysis and visualization due to their simplicity and ability to convey the relationship between two or more groups. They are easy to comprehend and interpret, making them accessible for people with diverse backgrounds.

Despite their usefulness, Scatter Plots have limitations. when data points are tightly packed or overlap (often referred to as *Overplotting*), it can be challenging to accurately assess the actual density of the data. Another significant challenge is distinguishing between groups when they overlap or when there is a small sample size. Last, Scatter plot displays the distribution of the data only in two dimensions and cannot give insights into density or distribution in additional dimensions.

2 Solution overview - The ICC plot

The proposed solution is a tree-based representation offered as an alternative to Scatter Plots to address their limitations.

The tree structure mimics the process a researcher goes through when analyzing a scatter plot, visually dividing the plane into subplanes and examining the density and distribution of samples in each subplane, as illustrated in figure 1.

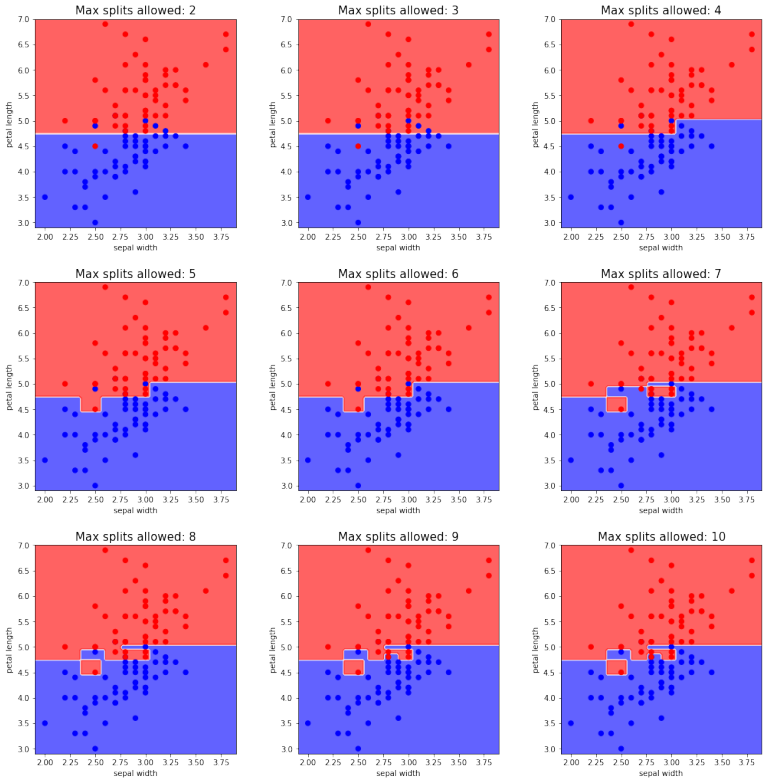

Figure 1: Illustrating how a researcher visually divides the Scatter plot had he or she used only horizontal and vertical lines. The simulation is based on a subset of the Iris dataset containing only the Versicolor and Virginica classes, and the Sepal width and Petal length attributes.

The selection of a tree structure representation was inspired by the widespread use of tree-representation in daily decision making. Also, unlike the scatter plots that only feature points, a tree enables to display additional information at its vertices and arcs. Last, the branching structure of the tree enables to facilitate the analysis of multiple dimensions of the data. Visual features such as colors and gradients aid in emphasizing the data's distribution, while confusion matrices at the tree nodes depict the precise

data's density.

I named the representation ICC plot, as an acronym of its 3 components: **I**nduction tree, **C**onfusion matrices and **C**olors.

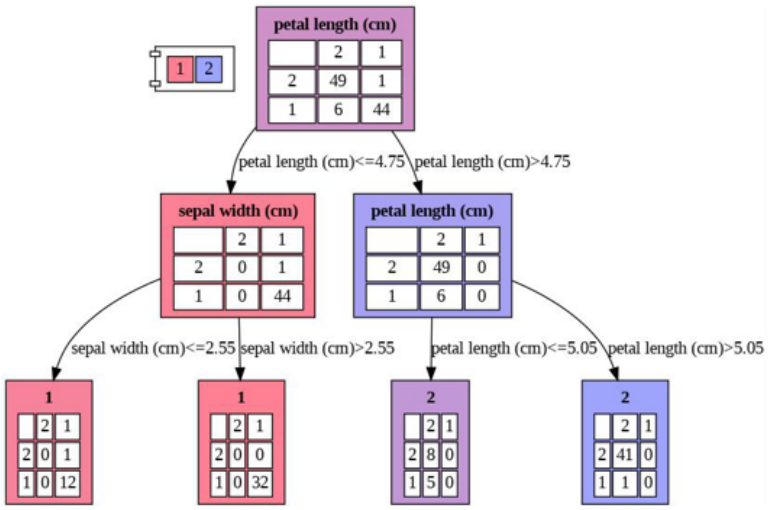

Figure 2: ICC with 3 nodes simulate a plain with 3 splits. The data in this figure is the same as in figure 1.

The ICC offers additional functionality to the researcher, such as adjusting the tree depth to increase or decrease the number of simulated splits of the plain, hiding the confusion matrices (to focus on the data spread instead of specific quantities), and adjusting colors, as illustrated in figure 4.

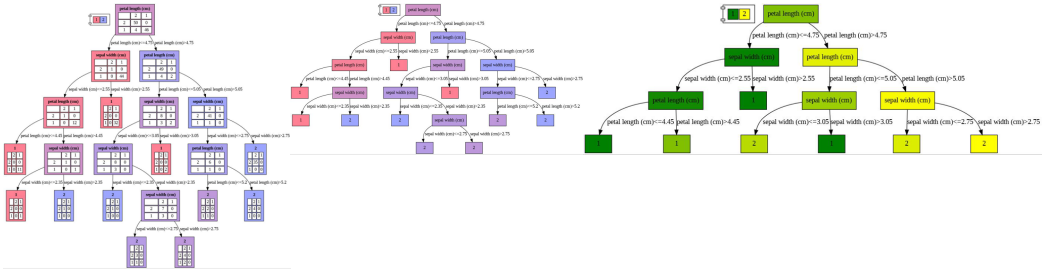

Figure 3: Different variations of ICC.

3 Experimental evaluation

I evaluated the performance of the ICC plot by comparing it to Scatter and Jitter plots, as a baseline. To do this, I presented four CS students majoring in Data Science two graphs representing a given dataset: one Scatter (or Jitter) plot, and the other - ICC plot.

In the first stage, I asked the students to write down as many observations as possible from each graph. In the second stage, I asked them to determine the relevance of their observations to the classification problem the dataset is used for. Finally, I asked them to share the advantages and disadvantages they came across during this evaluation.

I repeated this process with three well-known datasets: Iris (to classify iris species), Titanic (to classify survivors and fatalities in the Titanic disaster), and Wisconsin breast cancer data[1] (to classify a cell sample as malignant or benign).

The results are in table 1.

[1]Only ten out of the thirty possible attributes in the Wisconsin breast cancer dataset were used in order to simplify the evaluation process.

		Iris data	Titanic data	Wisconsin breast cancer data
Scatter\ Jitter plot	Mean number of observations according to the graph	4.25	2	8.5
	Percentage of observations relevant to the classification process	27.9%	50%	23.6%
ICC	Mean number of observations according to the graph	2	2	3.5
	Percentage of observations relevant to the classification process	79.1%	79.1%	75.4%

Table 1: Quantity and percentage relevance of insights provided by respondents for the classification problem discussed.

The results show that Scatter and Jitter Plots are easier to draw insights from. On average, the scatter plots enabled to generate twice as many insights about the data compared to the ICC plot. However, the percentage of relevant insights for classification is low: for the ICC plot, almost 4/5 of

the insights were relevant to the classification process compared to only one third for the Scatter and Jitter plots.

Based on the feedback from the subjects, it was revealed that the ICC graph was better suited for classification problems analysis but had two major drawbacks compared to the Scatter plot: firstly, it was less intuitive and user-friendly and therefore required learning before using it. Secondly, the scatter plot made it easier to identify correlations, which was not possible in the ICC graph.

4 Taking it one step forwards - The MAGIC tool

As the ICC plot provides a convenient method to visualize data, it could potentially be utilized to analyze cases where a problem inherent in the data itself (rather than in the training process) results in misclassification. To test this, I propose algorithm 1 and provide an API that implements it[2].

Algorithm 1 MAGIC algorithm

1: Build an ICC graph based on the training data.

2: Trace the path of the test data on the ICC graph:

2.1: Update the confusion matrices accordingly at each node.

2.2: Mark nodes no test-data has passed through at all with white (These represent parts of the space that were present in the training data but are missing in the test data).

2.3: Highlight nodes where samples of a different class than the one that appeared in the training set have reached with a double frame.

[2]The API not only generates the resulting tree but also enables extracting predicates representing a given test-sample according to it.

I named the API the MAGIC (Misclassification Analysis by Graphic Induction-Tree Classifier) tool.

A sample of the MAGIC algorithm output is presented in figure 4.

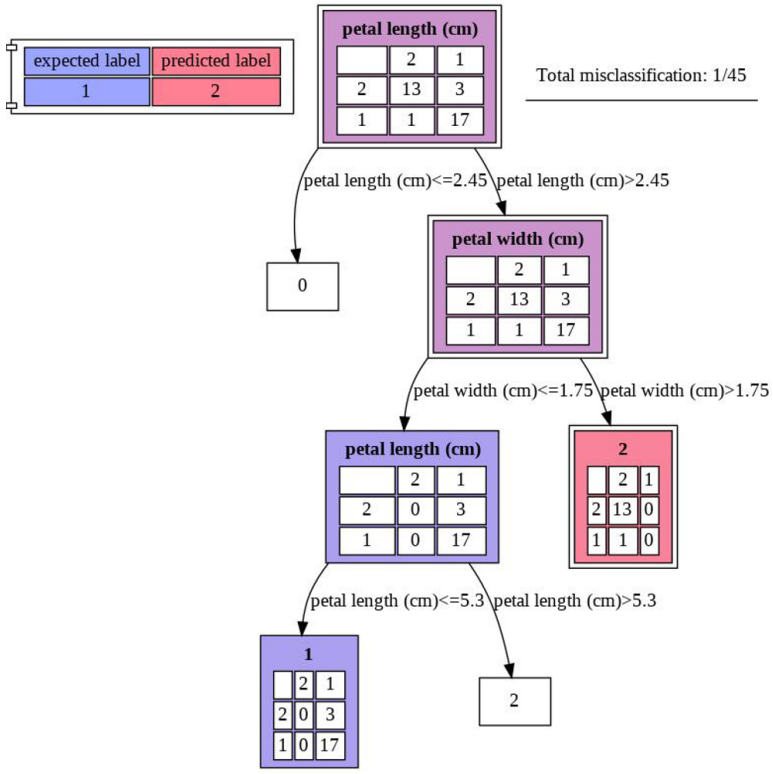

Figure 4: Examining test-data of the Iris dataset: The white vertices represent samples that are not present in the test data (for example, irises with a petal length of 2.45 or less). The double-framed vertices reveal that one sample in the test data was misclassified as type 2 when it is actually type 1. This error is understandable as 13 type 2 samples were correctly classified for similar reasons.

4.1 MAGIC Evaluation Process

I conducted the evaluation process of the MAGIC tool in two stages.
The first stage involved a qualitative comparison of the MAGIC tool's performance with the Scatter Plot on the three datasets used in the ICC evaluation, as well as on a forth synthetic dataset demonstrating Hidden Stratification.
In the second stage of the evaluation, I used a rather humorous questionnaire for a quantitative measurement. Three out of the four subjects participated and were presented with the data distribution of each of the four datasets using both the MAGIC tool and the Scatter plot. The questionnaire was used to determine the cases where they were able to successfully identify the reason for misclassification.

4.2 MAGIC Evaluation Results

From the qualitative research, it was found that the scatter plot graph is more effective in characterizing global phenomena, such as overlap in the characteristics of samples belonging to different classes, while the MAGIC tool is more useful for analyzing local cases, such as samples with identical or almost identical characteristics belonging to different classes. Furthermore, the MAGIC tool was significantly useful for datasets where most of the characteristics were categorical because the Scatter Plot suffered from overplotting. The MAGIC tool was also found to be more useful in cases where misclassification was due to Hidden Stratification. The questionnaire results are presented on table 2.

Reason for misclassificaion	Number of subjects who correctly identified the phenomenon using a Scatter plot	Number of subjects who correctly identified the phenomenon using the MAGIC tool
samples with identical characteristics	0/3	2/3
Samples with overlapping characteristics	3/3	2/3
Hidden Stratification	1/3	2/3

Table 2: Quantity and percentage relevance of insights provided by respondents for the classification problem discussed.

The subjects reported that the MAGIC tool was reasonably useful for the presented tasks, yet its main disadvantage is its visual complexity, which requires a significant learning curve.

5 Related work

5.1 Scatter plots limitation

Although having quite a few limitations, Scatter plots have been widely used in the field of statistics and data analysis. Over time, researchers have proposed different improvements to overcome these limitations. I will briefly review a selection of works and solutions.

Scatter plots can get cluttered with large datasets and overlapping points,

preventing the researcher from correctly assessing the density of the data. This is especially crucial when comparing between different groups. The oldest and most famous improvement for that is the Jitter Scatter plot, that randomly adjusts point positions to show data distribution better. Woodruff et al. (1998) proposed a system named VIDA that plots dots for objects in dense regions and polygonal outlines for objects in sparser regions. Dang et al. (2010) suggested addressing the above by arranging points in a third dimension.

Another problem that the Overplotting poses is the difficulty of separating overlapping groups in the data. Lee et al. (2012) offer to use hierarchical multi-class sampling method to create a simplified, feature-preserving scatter plot visualization. Mayorga and Gleicher (2013) propose to group dense data points and reveal subgroup relations through the use of smooth contour lines.

Oftentimes, two or three variables may not provide enough information to fully understand the relationships in the data. As plotting in four or more dimensions is impossible, Scatterplot Matrix (SPLOM) is commonly used. While solving the above, it imposes new challenges as it doesn't scale with the features quantity[4][13].

Additional improvements might include employing Scagnostics methods[18], Spatial distortion[10] and focusing on local patterns[17].

5.2 Visualizing human decision making processes by trees

Trees are broadly used across different disciplines to analyze complex decision making processes. For example, the Ethno-graphic decision tree modeling[7] is a research method designed to identify the factors that groups

of people use in their decision making. In game theory, the decision-making process is formalized and visualized using a tree, known as a *game tree*[6]. Trees are available to professionals, such as psychological counselors, to aid in the counseling process[3] and for service providers to assess service provision[8]. In political science, decision trees serve as a tool to analyze election results[2][9].

5.3 Misclassification Analysis

Misclassification analysis is a crucial step in the Data Science pipeline, and as such, development environments designed for data scientists provide a variety of dedicated tools for it. For instance, Gestalt-20[15] enables researchers to compare different samples side by side, making it possible to visually analyze differences between images in entity extraction tasks, or between texts in sentiment analysis tasks. ModelTracker[1] is another tool that enables local examination of classification errors.
An example of a system in the text domain is EluciDebug[11], which reports to the user why it classified a certain email in the way it did.
Finally, it is important to mention the well-known LIME algorithm[16], which suggests using simpler and more interpretable models to examine the nearby and local environment of the misclassified sample.

6 Conclusion

In this project I proposed a novel approach to data visualization using tree-structures. This approach was found to be particularly beneficial in the exploratory data analysis (EDA) process, overcoming the limitations of scatter plots. Although its usefulness in Misclassification Analysis was limited, it

proved to be useful in specific cases, such as identifying Hidden Stratification.

Throughout this project, I first and foremost learned about the research process and its essential steps. One of the most important takeaways was the significance of thoroughly defining the research question before initiating the coding and investigation process.

In particular, I came across the LIME algorithm at an advanced stage in the work on the project, before it was discussed in class. I was somewhat disheartened to discover that there was a tool that provides some of the functionality of the tools I offer (mostly the MAGIC tool) in a simpler and more convenient manner. This also reinforced the feedback from the subjects, which taught me that the ease of use and user-friendliness of the tool is a crucial factor. On the other hand, I appreciate the elegance of the LIME algorithm. It was also reassuring to see that it relies on simpler and more interpretable models, similar to the tree structure I chose for my own work. In addition, it was encouraging to see that the tools I suggested, could help in discovering cases like Hidden Stratification more effectively. In any case, if I had to name one work that inspired me, it would be the LIME algorithm. Its elegant, mathematically-supported, and user-friendly approach aligns perfectly with the main goal of visualization tools - to provide clear, precise and easy-to-use solutions.

Overall, exploring diverse data visualization tools and approaches was a fascinating experience. It provided me with a profound comprehension of the objectives and procedures that researchers undergo when encountering various visualizations. The knowledge and insights gained are valuable to me as a researcher in general and as a data scientist in particular.

References

[1] Saleema Amershi, Max Chickering, Steven M Drucker, Bongshin Lee, Patrice Simard, and Jina Suh. Modeltracker: Redesigning performance analysis tools for machine learning. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 337–346, 2015.

[2] Eva Armengol and Àngel García-Cerdaña. Decision trees as a tool for data analysis. elections in barcelona: A case study. In *Modeling Decisions for Artificial Intelligence: 17th International Conference, MDAI 2020, Sant Cugat, Spain, September 2–4, 2020, Proceedings 17*, pages 261–272. Springer, 2020.

[3] Kirk A Beck. Ethnographic decision tree modeling: A research method for counseling psychology. *Journal of Counseling Psychology*, 52(2):243, 2005.

[4] Daniel B Carr, Richard J Littlefield, WL Nicholson, and JS Littlefield. Scatterplot matrix techniques for large n. *Journal of the American Statistical Association*, 82(398):424–436, 1987.

[5] Tuan Nhon Dang, Leland Wilkinson, and Anushka Anand. Stacking graphic elements to avoid over-plotting. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1044–1052, 2010.

[6] Robert Gibbons. An introduction to applicable game theory. *Journal of Economic Perspectives*, 11(1):127–149, 1997.

[7] Christina H Gladwin. *Ethnographic decision tree modeling*, volume 19. Sage, 1989.

[8] Paula A Harrison, Rob Dunford, David N Barton, Eszter Kelemen, Berta Martín-López, Lisa Norton, Mette Termansen, Heli Saarikoski, Kees Hendriks, Erik Gómez-Baggethun, et al. Selecting methods for ecosystem service assessment: A decision tree approach. *Ecosystem services*, 29:481–498, 2018.

[9] NJ Joyner and ML Joyner. Simulating electoral college results using ranked choice voting if a strong third party candidate were in the election race.

[10] Daniel A Keim and Annemarie Herrmann. *The gridfit algorithm: An efficient and effective approach to visualizing large amounts of spatial data*. IEEE, 1998.

[11] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces*, pages 126–137, 2015.

[12] S Lee, M Sips, and H Seidel. Ieee transactions on visualization and computer graphics. 2012.

[13] Dirk J Lehmann, Georgia Albuquerque, Martin Eisemann, Marcus Magnor, and Holger Theisel. Selecting coherent and relevant plots in large scatterplot matrices. In *Computer Graphics Forum*, volume 31, pages 1895–1908. Wiley Online Library, 2012.

[14] Adrian Mayorga and Michael Gleicher. Splatterplots: Overcoming overdraw in scatter plots. *IEEE transactions on visualization and computer graphics*, 19(9):1526–1538, 2013.

[15] Kayur Patel, Naomi Bancroft, Steven M Drucker, James Fogarty, Amy J Ko, and James Landay. Gestalt: integrated support for implementation and analysis in machine learning. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 37–46, 2010.

[16] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[17] Lin Shao, Timo Schleicher, Michael Behrisch, Tobias Schreck, Ivan Sipiran, and Daniel A Keim. Guiding the exploration of scatter plot data using motif-based interest measures. *Journal of Visual Languages & Computing*, 36:1–12, 2016.

[18] Harshita Sharma, Alexander Alekseychuk, Peter Leskovsky, Olaf Hellwich, Radhey Shyam Anand, Norman Zerbe, and Peter Hufnagl. Determining similarity in histological images using graph-theoretic description and matching methods for content-based image retrieval in medical diagnostics. *Diagnostic pathology*, 7(1):1–20, 2012.

[19] Allison Woodruff, James Landay, and Michael Stonebraker. Constant density visualizations of non-uniform distributions of data. In *Proceedings of the 11th annual ACM symposium on User interface software and technology*, pages 19–28, 1998.