Enhancing EDA and Misclassification Analysis through Tree Structure Visualization

Maor Hornstein

Tabular Data Science (89547) - Final Project

Abstract

This project aims to explore how tree-based visualization can overcome the limitations of scatter plot graphs. First, I'll shows how the tree structure simulates the process followed by data scientists when using scatter plots during the EDA stage. Then, I attempt to extend its application to the misclassification analysis stage. The presented methods were tested on four classification problems and were found most useful for the EDA stage.

1 Problem Description

Scatter plots are a popular tool for data analysis and visualization due to their ease of interpretation. However, they also have limitations: overplotting occurs when data points are tightly packed, making it difficult to assess the true density of the data. Another challenge is distinguishing between different groups when they overlap. Additionally, Scatter plots only display data distribution in two dimensions, and cannot provide insights in four or more dimensions. This limitations should be taken into consideration during the EDA and misclassification analysis stages, in which scatter plots are used.

2 Solution overview

2.1 The ICC plot

The proposed plot is a tree-based representation offered as an alternative to scatter plots to address their limitations during the EDA phase. 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.

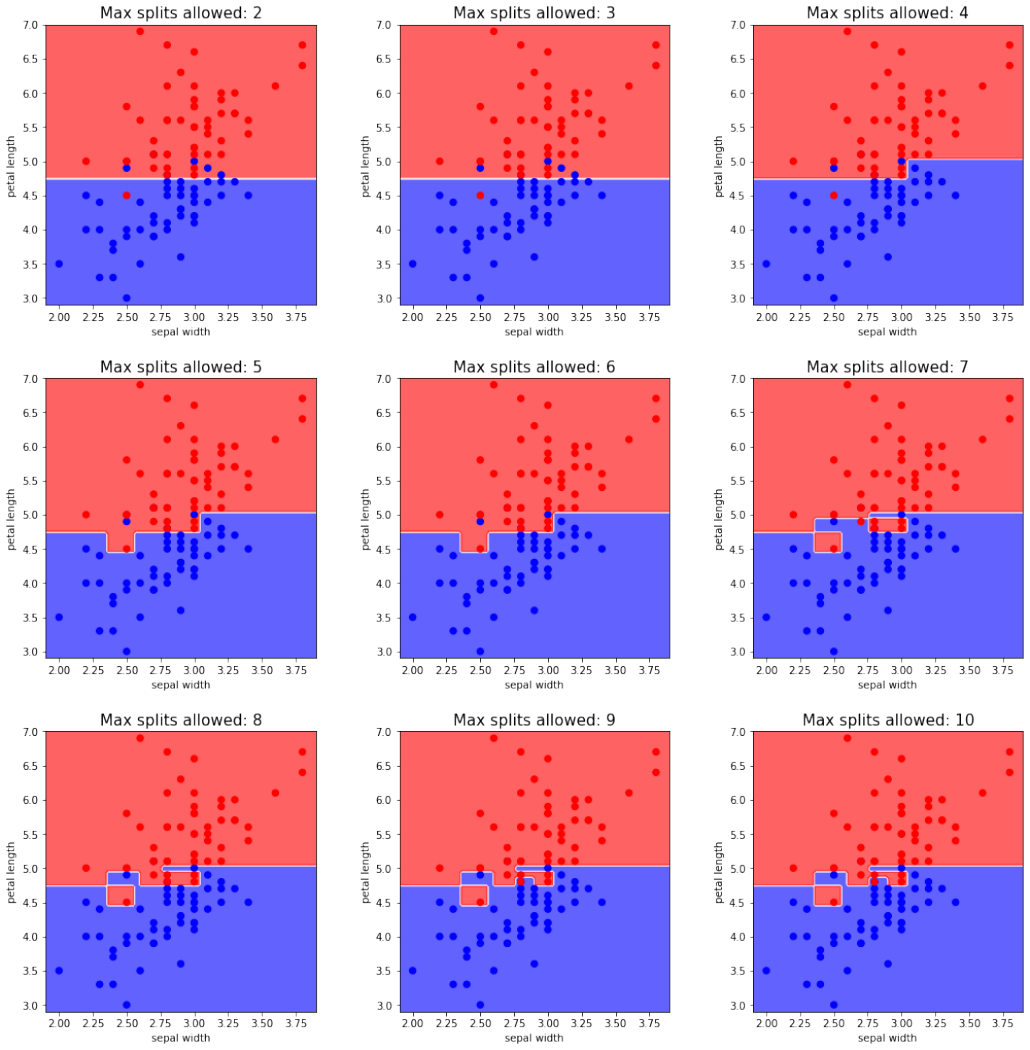

Figure 1: Illustrating how a researcher visually divides the scatter plot had he or she used only horizontal and vertical lines. This simulation uses a subset of the Iris dataset containing only the Versicolor and Virginica classes, and the Sepal width and Petal length attributes.

The choice of a tree structure representation was influenced by its common use in decision making and its ability to display additional information in its vertices and edges, unlike scatter plots that feature only points. The branching structure of the tree facilitates 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 this representation ICC plot, as an acronym of its 3 components: **I**nduction tree, **C**onfusion matrices and **C**olors.
The ICC plot provides researchers with additional functionality, such as the ability to adjust tree depth to increase or decrease the number of simulated splits of the plane, hide the confusion matrices to focus on data spread, and modify colors. These options are illustrated in figure 3.

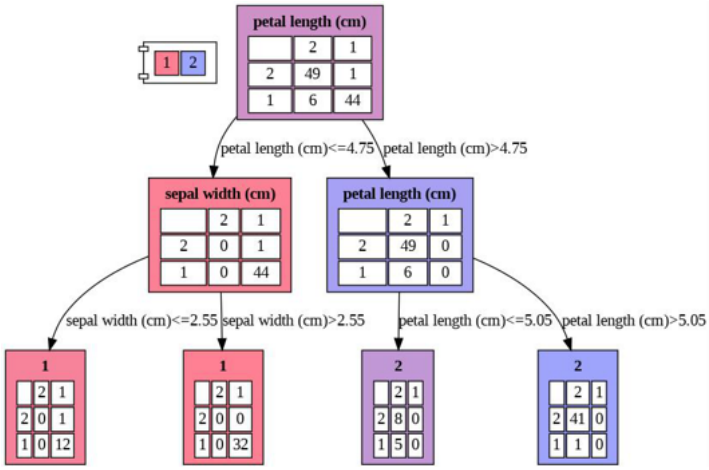

Figure 2: ICC with 3 nodes simulate a plain with 3 splits. The data is the same as in figure 1.

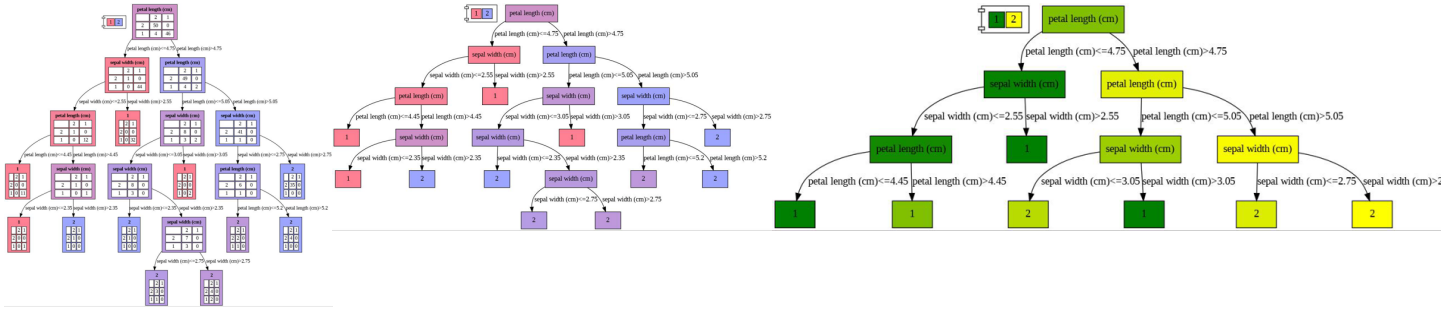

Figure 3: Different variations of ICC.

2.2 The MAGIC tool

The MAGIC tool is an extension of the ICC plot designed for the misclassification analysis stage. 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. For that end, I propose algorithm 1 and provide an API that implements it[1]. I named the API the MAGIC (**M**isclassification **A**nalysis by **G**raphic **I**nduction-Tree **C**lassifier) tool. A sample of the MAGIC tool's output is presented in figure 4.

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

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.

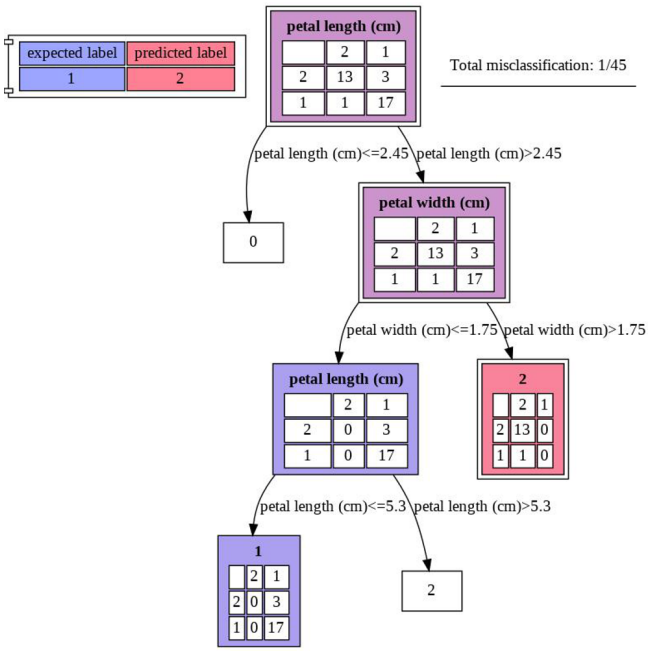

Figure 4: Examining test-data of the Iris dataset: The white vertices represent samples that are not present in the test data (for example, irises with a petal length of 2.45 or less). The double-framed vertices reveal that one sample in the test data was misclassified as type 2 when it is actually type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

3 Experimental evaluation

3.1 The ICC plot

I evaluated the performance of the ICC plot by comparing it to scatter and jitter plots, as a baseline. To do this, I presented six CS students majoring in Data Science two graphs representing a given dataset: one scatter (or jitter) plot, and the other - ICC plot. First, I asked the students to write down as many observations as possible from each graph. then, I asked them to determine the relevance of each observation to the classification problem the dataset is used for. I repeated this process with

three datasets: Iris, Titanic, and Wisconsin breast cancer[2]. The results are in table 1.

		Iris data	Titanic data	Wisconsin breast cancer data
Scatter\ Jitter plot	Mean number of observations according to the graph	6.1	2.1	9.1
	Percentage of observations relevant to the classification process	30.5%	47.6%	23.8%
ICC	Mean number of observations according to the graph	2.3	2.5	3.5
	Percentage of observations relevant to the classification process	86.9%	73.3%	66.6%

Table 1: Quantity and percentage relevance of insights for each classification problem.

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: almost 3/4 of the insights were relevant to the classification process for the ICC plot compared to only one third for the scatter and jitter plots.

The feedback from the students indicated that the ICC plot was better suited for analyzing classification problems than scatter plots, but had two major drawbacks: it was not very intuitive and required them to learn how to use it properly first. Additionally, scatter plots enable identification of correlations, which is not possible with the ICC plot.

3.2 MAGIC Evaluation Process

The evaluation of the MAGIC tool was conducted in two stages. The first stage involved a **qualitative comparison** of the MAGIC tool's performance with the scatter plot on the three datasets, as well as on a forth additional synthetic dataset demonstrating hidden stratification[3]. The second stage of the evaluation was **quantitative**: the six students were presented with the data distribution of the four datasets using both the MAGIC tool and scatter plot. A humorous questionnaire was used to

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

[3]The detailed qualitative comparison can be found in the Jupyter notebooks of the MAGIC tool, which are included in the project deliverables.

ask them to identify the reason for misclassification based on each of the visualization options.

The qualitative research 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 phenomena, such as samples with identical or almost identical characteristics belonging to different classes. The MAGIC tool was significantly useful for datasets where most of the characteristics were categorical because the scatter plot suffered from overplotting, as well as in cases where misclassification was due to hidden stratification. The questionnaire results are presented on table 2.

The subjects reported that the MAGIC tool was reasonably useful for the presented tasks, yet its main disadvantage is the significant learning required to understand its visual complexity.

Reason for misclassificaion	% of subjects correctly identified the phenomenon using scatter plot	% of subjects correctly identified the phenomenon using the MAGIC tool
Samples with identical characteristics	16.6%	83.3%
Samples with overlapping characteristics	83.3%	66.6%
Hidden stratification	33.3%	66.6%

Table 2: Comparison of correct identification of the cause of misclassification.

4 Related work

4.1 Scatter plots limitation

Although having quite a few limitations, Scatter plots are widely used. Researchers 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. To address it, **?** proposed the VIDA system that plots dots in dense regions and polygonal outlines in sparser regions. **?** suggested to arrange points in the third dimension instead. Another problem is the difficulty of separating overlapping groups in the data. **?** offer to use hierarchical multi-class sampling method to create a simplified, feature-preserving scatter plot visualization. **?** propose to group dense data points and reveal subgroup relations through the use of contour lines. As plotting in four or more dimensions is impossible, Scatterplot Matrix (SPLOM) is commonly used,

yet it imposes new challenges as it doesn't scale with the features quantity[**?**][**?**].
Additional improvements might include employing Scagnostics methods[**?**], Spatial distortion[**?**] and focusing on local patterns[**?**].

4.2 Visualizing 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[**?**] 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 using a tree, known as a *game tree*[**?**]. Trees are available to professionals, such as psychological counselors, to aid in the counseling process[**?**] and for service providers to assess service provision[**?**]. In political science, decision trees can serve as a tool to analyze election results[**?**][**?**].

4.3 Misclassification Analysis

Misclassification analysis is a crucial step in the Data Science pipeline, and as such, different tools were developed to assist with it. Gestalt-20[**?**] enables researchers to compare different samples side by side, making it possible to visually analyze differences between images in entity extraction tasks, or texts in sentiment analysis tasks. ModelTracker[**?**] is another tool that enables local examination of classification errors. Another example of a system in the text domain is EluciDebug[**?**], 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[**?**], which uses simple and interpretable models to investigate the local environment of a misclassified sample.

5 Conclusion

In this project I proposed a novel approach to data visualization using tree-structures, which was effective for EDA and overcame the limitations of scatter plots. While it had limited usefulness in misclassification analysis, it was found to be beneficial in specific cases such as identifying hidden stratification.
Throughout this project, I first and foremost learned about the essential steps of the research process. One important takeaways is the significance of thoroughly defining and investigating the research question. In particular, I came across the LIME algorithm at an advanced stage of the project, before it was discussed in class. I was somewhat disheartened to discover a tool that provides some of the functionality of the tools I offer (the MAGIC tool) in a simpler and more convenient manner. On the other hand, it was reassuring to see that the LIME algorithm relies on interpretable models, similar to the tree structure I chose. It was also encouraging to see that the tools I suggested 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 as it is an elegant, mathematically-supported, and user-friendly tool.

Overall, exploring different 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.