Automated Measures for Interpretable Dimensionality Reduction for Visual Classification: A User Study

Ilknur Icke[1] *
[1]The Graduate Center
The City University of New York
USA

Andrew Rosenberg[1,2] †
[2]Queens College
The City University of New York
USA

ABSTRACT

This paper studies the interpretability of transformations of labeled higher dimensional data into a 2D representation (scatterplots) for visual classification.[1] In this context, the term interpretability has two components: the interpretability of the visualization (the image itself) and the interpretability of the visualization axes (the data transformation functions). We define a data transformation function as any linear or non-linear function of the original variables mapping the data into 1D. Even for a small dataset, the space of possible data transformations is beyond the limit of manual exploration, therefore it is important to develop automated techniques that capture both aspects of interpretability so that they can be used to guide the search process without human intervention. The goal of the search process is to find a smaller number of interpretable data transformations for the users to explore. We briefly discuss how we used such automated measures in an evolutionary computing based data dimensionality reduction application for visual analytics.

In this paper, we present a two-part user study in which we separately investigated how humans rated the visualizations of labeled data and comprehensibility of mathematical expressions that could be used as data transformation functions. In the first part, we compared human perception with a number of automated measures from the machine learning and visual analytics literature. In the second part, we studied how various structural properties of an expression related to its interpretability.

Keywords: data transformation, visualization, user study

Index Terms: H.2.8 [Database Management]: Database applications—Data Mining

1 INTRODUCTION

Visual exploration of high dimensional labeled data requires mapping the data from its original variable space on to the 2D space so that the class structures are clearly visible (visual interpretability) and the relationships between the original variables and the visualization axes (meaning–semantics- of the visualization) are easy to interpret. Therefore, it is desirable to have access to the explicit mapping functions that reveal the meaning of the visualization. However, in a standard dimensionality reduction scheme, most techniques either ignore the label information or do not create explicit transformation functions for the users to inspect.

In [3], we present a genetic programming based dimensionality reduction tool that explores the space of data transformations searching for 2D views of the data showing clear class structures created by easily interpretable (simple) data transformation functions. The algorithm automatically constructs new features by optimizing the two aspects of interpretability of the data transformations: *visual interpretability* (of the image itself) and *semantic interpretability of the view* (interpretability of the visualization axes). We do not constrain the transformation function to be strictly linear, they can be in any arbitrary form. The search process aims to optimize both aspects of interpretability simultaneously. The goal of this paper is to present our investigation on how much the automated measures of interpretability of data transformations (the view and the transformation functions separately) match human perception. The motivation for studying such measures is clear: we would like to incorporate them into our automated search process in order to increase interpretability of the results on both aspects: the view and the visualization axes.

*e-mail: iicke@gc.cuny.edu
†e-mail: andrew@cs.qc.cuny.edu

[1]In this paper, the term visual classification refers to data classification on lower dimensional representations of the dataset, rather than specifically to classification of visual (image) data

2 RELATED WORK

Recently a number of user studies were reported in the visual analytics field concentrating on the visual interpretability aspect. The Vizrank algorithm uses the K-nearest Neighbor classifier as a quality measure for 2D scatterplots of labeled data [4]. A total of four visual quality measures have been proposed for labeled data in [7] and [8]. All authors report user studies showing correlation between human judgement and their proposed automated techniques.

Studies on interpretability of data transformation functions focus on the linear projections. Morton defines the interpretability of the linear projection functions in terms of parsimony (simplicity) and explores the use of rotation and entropy based methods to simplify the coefficients while preserving the quality of the view [6]. However, there is no guarantee that a linear function is always the best model for the problem. Any arbitrary function of the original variables can be used as a data transformation function in order to create a visualization provided that we can devise a measure of interpretability of the function.

3 USER STUDY ON VISUAL INTERPRETABILITY OF 2D SCATTERPLOTS

We have developed a Java language based application that administered the user study without investigator intervention. Twenty graduate students from such fields as computer science and physics participated in the study. Thirteen of the students stated previous exposure to statistics or data mining. We selected a total of 40 2D scatterplots from four datasets (ten views each) and a total of ten automated measures (*classifier accuracy*: K-nearest Neighbors (KNN), Naive Bayes (NB), Support Vector Machine with Polynomial Kernel (SVM), C4.5 Decision Tree (DT), *cluster validity indices*: Dunn Index (Dunn), Davies-Bouldin Index (DB) and C Index (C), *visual quality measures*: Class Consistency Measure (CCM) [7], Linear Discriminant Analysis (LDA) Index [5] and 2D-Histogram Density Measure (HDM-2D) [8]). The datasets were: Wisconsin Diagnostic Breast Cancer [2], Wine [2], Segment [2] and Italian Olive Oils [1]. For each participant, the application displayed the visualizations one at a time and in random order asking for them to rate how good the view was on a continuous scale from very bad (0.0)

to very good (1.0). The participants did not receive any information on what a good view meant. Unknown to the participants, the first five views were calibration images which were not used in the analysis of the results. Since we studied visual interpretability independent from the projection axes, we did not reveal any background information other than specifying that different colors represented different groups. Our experimental protocol is different than those reported in [7, 8]: we display the visualizations one by one, while those authors ask the users to sort a screen (or sheet) full of images from good to bad.

Our results indicated that other than the LDA Index and the C Index, correlations between the human perception and automated measures were statistically significant at the 95% confidence level when all scatterplots are considered.

measure	R^2	p-value
SVM	0.5506	< 0.05
Naive Bayes	0.5406	< 0.05
CCM	0.5246	< 0.05
Dunn Index	0.5047	< 0.05
K-Nearest Neighbors	0.4914	< 0.05
Decision Tree	0.4404	< 0.05
Davies-Bouldin Index	0.4197	< 0.05
HDM-2D	0.4050	< 0.05
LDA Index	0.0664	0.11
C Index	0.0295	0.29

Table 1: Summary of linear relationships between the automated measures and human perception on all scatterplots (R^2: correlation coefficient, p-value: computed by the two-tailed t-test where H_0: Pearson's correlation is not 0, degree of freedom=38, $\alpha = 0.05$, p-value < 0.05 means significant correlation)

On individual datasets, we found that all correlations were statistically significant on the Wine dataset (where the class structures are closer to round shapes). The Dunn Index did not correlate with human perception on the Olive Oils dataset and the LDA, Davies-Bouldin and C Indices did not correlate with human perception on the Wisconsin Breast Cancer and Segment datasets.

measure	R^2	p-value
HDM-2D	0.7592	< 0.05
SVM	0.7495	< 0.05
K-Nearest Neighbors	0.6786	< 0.05
Naive Bayes	0.6759	< 0.05
Decision Tree	0.6667	< 0.05
Dunn-Index	0.6057	< 0.05
CCM	0.4430	< 0.05
Davies-Bouldin Index	0.3516	0.07
LDA Index	0.2616	0.13
C Index	0.2337	0.16

Table 2: WDBC (2 classes)

measure	R^2	p-value
Dunn Index	0.8061	< 0.05
SVM	0.7363	< 0.05
CCM	0.6764	< 0.05
Decision Tree	0.6535	< 0.05
C Index	0.6266	< 0.05
Naive Bayes	0.5761	< 0.05
LDA Index	0.5741	< 0.05
Davies-Bouldin Index	0.5519	< 0.05
K-Nearest Neighbors	0.5357	< 0.05
HDM-2D	0.4531	< 0.05

Table 3: Wine (3 classes)

measure	R^2	p-value
HDM-2D	0.7140	< 0.05
Dunn Index	0.6152	< 0.05
Naive Bayes	0.6041	< 0.05
SVM	0.6035	< 0.05
K-Nearest Neighbors	0.5582	< 0.05
Decision Tree	0.5222	< 0.05
CCM	0.4352	< 0.05
Davies-Bouldin Index	0.2927	0.11
LDA Index	0.0847	0.41
C Index	0.0002	0.98

Table 4: Segment (7 classes)

measure	R^2	p-value
Naive Bayes	0.7420	< 0.05
Davies-Bouldin Index	0.7405	< 0.05
K-Nearest Neighbors	0.7275	< 0.05
CCM	0.7100	< 0.05
C Index	0.6702	< 0.05
HDM-2D	0.6496	< 0.05
Decision Tree	0.6445	< 0.05
SVM	0.5513	< 0.05
LDA Index	0.4950	< 0.05
Dunn Index	0.1252	0.32

Table 5: Olive Oils (9 classes)

Similar to the results of [8], we found the Class Consistency and the 2D-Histogram Measures to be in agreement with human perception on all four datasets. Furthermore, we trained a linear regression model of the *predicted human response (PHR)*:

$PHR = (-0.7772 * DT) + (0.8155 * SVM) - (0.4305 * C) - (0.4588 * DB) + (0.6586 * CCM) + (0.3285 * HDM - 2D) + 0.3606$

Dataset	R^2	p-value
All	0.8582	< 0.05
WDBC	0.8629	< 0.05
Wine	0.6331	< 0.05
Segment	0.9139	< 0.05
Olive oils	0.5702	< 0.05

Table 6: Match between the combined model (PHR) and human perception

None of the automated measures (including the best linear combination of the measures, despite being significantly correlated to human ratings on all four datasets) consistently outperformed all measures across different datasets hinting for *no free lunch* in finding an automated measure that closely matches human perception across all data. The participants also commented on their ratings. The comments reveal that views with overlapping groups were not favored. Similarly, groupings that were clumped into small areas were not preferred since it was hard to distinguish the points.

4 User Study on Generic Data Transformation Functions

The goal of this study was to investigate how humans judge the interpretability of arbitrary data transformation functions independent of the visualization. We chose five generic variables t, u, x, y, z, numerical coefficients and arithmetic operators ($+, -, *, /$, logarithm, square-root, exponential and power) in order to create 30 mathematical expressions. No actual visualization was generated or displayed. The shortest expression was of length 2 (one operator and one operand) and longest expression was of length 19.

We adopted a syntax-tree representation and identified a number of structural characteristics: number of operators and operands, tree depth, number of blocks (compact sub-trees), average length of blocks and total size. The tree depth relates to the nestedness and the number and size of the blocks indicate the distinguishable components of the expression. The participants were shown each expression one by one and in random order. They were given 10 seconds to study the expression and then asked to write down the expression and provide a rating for its difficulty. We developed a linear model which is highly predictive of the human ratings with respect to structural characteristics (Pearson's R=0.9598, Df=23, $p < 0.05$): $difficulty = 0.0854 * \textit{Tree Depth} - 0.2568 * \textit{Number of Blocks} - 0.1014 * \textit{Avg. Block Size} + 0.0899 * \textit{Total Size} + 0.2151$. We infer that longer and nested expressions are difficult to interpret while the existence of small number of compact blocks improve interpretability.

5 Conclusion

We argued that interpretability measures were crucial for automated search for interpretable data transformations. In this paper, we studied the interpretability of the visualizations and the data transformation functions separately. Presenting the visualization and the transformation functions together might introduce a bias due to the user's trade off between the quality of the view and complexity of data transformation functions. Such a study is left for future work.

References

[1] Italian olive oils dataset. http://www.ggobi.org/book/data/olive.csv.

[2] Uci machine learning data repository. http://www.ics.uci.edu/ mlearn.

[3] I. Icke and A. Rosenberg. Multi-objective genetic programming for visual analytics. In S. Silva, J. A. Foster, M. Nicolau, M. Giacobini, and P. Machado, editors, *Proceedings of the 14th European Conference on Genetic Programming, EuroGP 2011*, volume 6621 of *LNCS*, pages 323–334, Turin, Italy, 27-29 Apr. 2011. Springer Verlag.

[4] G. Leban, B. Zupan, G. Vidmar, and I. Bratko. Vizrank: Data visualization guided by machine learning. *Data Mining and Knowledge Discovery*, 13:119–136, 2006. 10.1007/s10618-005-0031-5.

[5] E.-K. Lee, D. Cook, S. Klinke, and T. Lumley. Projection pursuit for exploratory supervised classification. *Journal of Computational and Graphical Statistics*, 14(4):831–846, December 2005.

[6] S. C. Morton. *Interpretable Projection Pursuit*. PhD thesis, Stanford University, 1989.

[7] M. Sips, B. Neubert, J. P. Lewis, and P. Hanrahan. Selecting good views of high-dimensional data using class consistency. *Computer Graphics Forum*, 28(3):831–838, 2009.

[8] A. Tatu, P. Bak, E. Bertini, D. Keim, and J. Schneidewind. Visual quality metrics and human perception: an initial study on 2d projections of large multidimensional data. In *Proceedings of the International Conference on Advanced Visual Interfaces*, AVI '10, pages 49–56, New York, NY, USA, 2010. ACM.