A SURVEY ON VISUALIZATION FOR EXPLAINABLE CLASSIFIERS

by

YAO MING

Department of Computer Science and Engineering
The Hong Kong University of Science and Technology
Supervised by Prof. Huamin Qu

October 2017, Hong Kong

TABLE OF CONTENTS

A SURVEY ON VISUALIZATION FOR EXPLAINABLE CLASSIFIERS

by

YAO MING

Department of Computer Science and Engineering

The Hong Kong University of Science and Technology

ABSTRACT

Classification is a fundamental problem in machine learning, data mining and computer vision. In practice, interpretability is a desired property of classification models (classifiers) in critical areas like security, medicine and finance. For instance, a quantitative trader may prefer a more interpretable model with less expected return due to its predictability. Unfortunately, most best-performing classifiers in many applications (e.g., deep neural networks) are complex machines whose predictions are difficult to explain. Thus, there is a growing interest in using visualization to understand, diagnose and explain machine learning systems in both academia and industry. Many challenges need to be addressed in the formalization of explainability and design principles and evaluation of explainable intelligent systems.

The survey starts with an introduction on the concept and background of explainable classifiers. Existing work in both visualization and machine learning communities is categorized in terms of data types and purposes of explanation. Then the survey ends with a discussion on the challenges and future research opportunities of explainable classifiers.

CHAPTER 1

INTRODUCTION

1.1 Motivation

Classification is the problem of identifying if an observation or object belongs to a set or not, or which of several sets. It is a fundamental problem in machine learning, data mining and computer vision. With the support of the increasing capacity of computation resources and growing volume of available data, the last decades have witnessed an explosion of breakthroughs in these fields. Nowadays, classification models (classifiers) are widely adopted to solve real world tasks, including face recognition [], handwritten recognition [], sentiment analysis [] and spam filtering []. Take image classification for instance, a well-designed convolutional neural network can achieve human-level performance in a number of benchmark datasets [,].

Despite their promising capability, an often-overlooked aspect is the important role of humans [4]. When humans are to understand and collaborate with these autonomous systems, it is desirable if we have explanations of their outputs. For instance, a doctor using a machine classifier to assist identifying early signs of lung cancers would need to know why the classifier "thinks" there might be a cancer so that he/she can make a more confident diagnosis. The research for explainable intelligent systems can be traced backed to 1980s, when expert systems are created and proliferated [1, 3, 5]. These early works focused on reducing the difficulty of maintaining the complicated if-then rules by designing more explainable representations. Recently, DARPA launches the Explainable Artificial Intelligence (XAI) project [2], which aims to develop new techniques to make the new generations of AI systems explainable.

However, the best-performing classifiers (e.g., neural networks) are becoming increasingly complex, in terms of the number of parameters and operations they employed, which makes them difficult to be explained.

1.2 Challenges and Research Issues

1.3 Overview

CHAPTER 2

EXPLAINABLE CLASSIFIERS

2.1 Classification

2.1.1 Definition

2.1.2 Classifiers

2.2 Explanation of Classifiers

2.2.1 Definition

2.2.2 Instance-level Explanation

2.2.3 Model-level Explanation

2.3 Evaluation

CHAPTER 3

VISUALIZATION FOR EXPLAIANBLE CLASSIFIERS

3.1 Visualization for Model Development

3.1.1 Model Understanding

3.1.2 Model Diagnosing

3.2 Visualization for Trust Establishment

3.3 Evaluation

CHAPTER 4

CONCLUSION

Placeholder for Conclusion.

BIBLIOGRAPHY

[1] W. Clancey, "The epistemology of a rule-based expert system: A framework for explanation," Stanford, CA, USA, Tech. Rep., 1981.

[2] D. Gunning, "Explainable artificial intelligence (XAI)," *Defense Advanced Research Projects Agency (DARPA)*, 2017.

[3] R. Neches, W. Swartout, and J. Moore, "Enhanced maintenance and explanation of expert systems through explicit models of their development," *IEEE Transactions on Software Engineering*, vol. SE-11, no. 11, pp. 1337–1351, Nov 1985.

[4] M. T. Ribeiro, S. Singh, and C. 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*. New York, NY, USA: ACM, 2016, pp. 1135–1144.

[5] W. Swartout, C. Paris, and J. Moore, "Explanations in knowledge systems: design for explainable expert systems," *IEEE Expert*, vol. 6, no. 3, pp. 58–64, June 1991.