Intrinsically Interpretable Machine Learning In Computer Aided Diagnosis

S. Ghosh, Sreejita

DOI:
10.33612/diss.175627883

Chapter 1

Introduction

With the rise in efficient and cost-effective data collection technologies increasing number of anthropocentric applications are using machine learning (ML) or artificial intelligence (AI) to *serve us better*. In recent years thankfully many humans have thought to question the exact decision-making process of these tools. This has consequently led to the demand for transparency in ML classifiers and regressors, or what is currently termed as explainable AI (XAI). It is not just performance that we should focus on, but also whether it is biased, robust, prone to noisy data, trustworthy, explainable, and which of our data it is using to arrive at a decision (Arrieta et al. 2020). Being denied EMI scheme on a telephoto lens is not that big a deal. But imagine not getting the correct treatment for yourself or your loved ones at the right time because the unquestionable and inscrutable algorithm decided so. Imagine being rejected insurance money or a much needed loan for your startup business or your first house, because the algorithm computed so. If a human or group of humans were to make these decisions would we not have asked them to explain their decision to us? Then why not question the algorithms as well?

This PhD thesis focuses on development of four broad variants of the Learning Vector Quantization classifier, which is a family of nifty intrinsically interpretable classifiers, i.e., classifiers whose working logic one can understand and explain. The LVQ family of classifiers follow Nearest Prototype Classification (NPC) scheme of learning. We have applied our developed classifiers on a real-life medical dataset of rare and inborn disorders of steroidogenesis. The clinicians wanted not just to have a crisp label of patients, but also why a diagnosis was made, how sure the classifier is of a diagnosis, and also to extract knowledge about these rare conditions. Being prototype based, the LVQ classifiers are able to provide one or more representative example(s) of each labelled class it learns from, thus throwing light on each of the investigated groups or conditions. Thus clinicians could verify and/or learn what a typical patient profile of a certain disorder looks like. However, being a real-life medical dataset it had the challenges of missing data, imbalanced classes, and heterogeneous measurements. The technological contributions in this thesis address these issues. For the interested readers we provide a set of publicly available synthetic datasets which are loosely modelled after the real-life medical dataset barring

in dimensionality. Using these synthetic datasets we systematically investigated the effect of the type and amount of missing data and sample size of training data, on two broad variants of LVQ classifiers, those which use Euclidean distance, and those using angle-based dissimilarity.

The major contributions in the thesis are the angle-dissimilarity based variants of the popular (1) Generalized Matrix Learning Vector Quantization (GMLVQ) (Schneider et al. 2007),(Bunte et al. 2012), (Bunte et al. 2010), (2) Localized GMLVQ (LGMLVQ) (Bunte et al. 2012), (Bunte et al. 2010), (3) Localized Limited Rank Matrix LVQ (LLiRaMLVQ) (Bunte et al. 2012), (Bunte et al. 2010), and (4) probabilistic variants of GMLVQ and newly introduced angle GMLVQ which use Kullback-Liebler (KL) divergence as cost function. We also introduce a model agnostic technique of dealing with imbalanced classes and a geodesic averaging technique which combines the power of multiple learners with the explainability aspect inherent in the shallow LVQ models. We compare these new developments with some of the state-of-the-art shallow ML techniques such as Random Forests (RF), $k-$Nearest Neighbours, and Linear Discriminant Analysis (LDA). Since our medical collaborators are interested in extraction of biological knowledge from the applied ML classifiers we compare the type of knowledge which can be straightforwardly extracted from each of the newly introduced LVQ classifiers to that extracted from RFs. It should however be mentioned that the kind of knowledge extracted is the same as those offered in the Euclidean predecessors of the newly introduced angle-dissimilarity based LVQ variants. The *novel* knowledge would be the classifier's confidence obtained from the probabilistic variants of LVQ. (Villmann et al. 2018) generalized Robust Soft LVQ (RSLVQ) by using information theoretic principles, using Cross-Entropy as cost-function. However the probabilistic LVQ we introduce approach the problem from a different perspective.

1.1 Scope of this thesis

The thesis is presented in two broad parts. The first part introduce the readers to the urgent need for transparency in ML classifiers, gives a brief review of the terms associated with XAI and interpretability. It also explains the challenges in a typical biomedical dataset and to which extent the available model-agnostic and model-specific ML techniques can handle them and when they cannot. The latter part provided the technical motivation for the *novel* contributions of this thesis. This part contains a new model-agnostic approach of handling imbalanced classes, which was built on an oversampling strategy introduced in (Chawla et al. 2002). We developed a variant of this technique such that oversampling could be done on a hypersphere instead of a hyperplane. The motivation for this was the angle-based classifier we developed. The second part contains the technical contributions which were angle-

based successors of GMLVQ, LGMLVQ, LLiRaMLVQ, the probabilistic versions of GMLVQ and angle GMLVQ, and introduces the technique of geodesic averaging of LVQ models, both global and local versions. This part of the thesis also compared how similarly or differently the Euclidean distance and angle-based dissimilarity, within the scope of LVQ classifiers, were affected by size and quality of datasets. It is not just the performance of the developed classifiers which are reported but also how interpretable they are based on the knowledge extracted from them.

1.2 Outline

Chapter 2 introduces the concepts of interpretability, explainability, transparency, trustworthiness, understandability, and comprehensibility in the context of ML. It discusses the nuances between these terminologies in the mentioned context, and the reason for rising demand for these traits in ML algorithms being deployed in anthropocentric applications. Following this it explains the types of interpretability and levels of model transparency. Next it describes some of the available interpretable ML techniques and discusses how interpretable and transparent they are. Finally it discusses the trade-off between a model's interpretability and its perfromance.

Chapter 3 discusses the main challenges of biomedical datasets which prevent the straightforward application of some of the available shallow ML classifiers which are intrinsically explainable. It discusses the available state-of-the-art techniques which can deal with some of these challenges and point out their limitations. This chapter also introduces a model-agnostic imbalance handling technique which performs oversampling on a hypersphere.

Chapter 4 describes the two main datasets on which our developed LVQ variants were trained and tested. One is a set of synthetic datasets differing in the type and amounts of missingness, but with equal proportion of samples from all three classes. The other is a real-life medical dataset of rare and inborn disorders of steroidogenesis. This dataset is not publicly available during the time of submission of the thesis.

Chapter 5 introduces the angle-based variant of GMLVQ. It provides a methodical investigation of how different dissimilarity measures in GMLVQ could be affected by sample size of training set, and amount and type of missing values. It also compares the performance of three other popular shallow ML classifiers, capable of multi-class classification. This chapter also introduces geodesic averaging technique of LVQ models which the interpretability aspect of LVQ with the power of multiple learners.

Chapter 6 contains the angle based variants of LGMLVQ and LLiRaMLVQ. In addition to applying these classifiers on the aforementioned datasets we also tested

them on a synthetic dataset with non-linear decision boundary, and a UCI heart disease dataset with non-linear decision boundaries when treated as a multi-class problem. Since the thesis introduces the mentioned four new variants of LVQ, for easier comparison of the pros and cons of these classifiers we performed some investigations, such as application of the geodesic averaging technique, on only the GCMS dataset. Knowledge extraction was performed on both the UCI heart disease dataset and the GCMS dataset for understanding of complexity of each of the datasets.

Chapter 7 discusses the four possible types of probabilistic LVQ classifiers, (1) using Euclidean distance metric and inclusive KL divergence, (2) using Euclidean distance and exclusive KL divergence, (3) using angle-based dissimilarity and inclusive KL divergence, and (4) using angle-based dissimilarity and exclusive KL divergence. We investigate which of these four proposed probabilistic variants are feasible and robust, and also how the feasible-in-practice ones are affected by the limitations of sample size of training data, and type and amounts of missing data. Similar to the previous three chapters this chapter also illustrates and discusses the interpretability offered by these variants of LVQ.

Lastly, Chapter 8 provides a summary of the thesis, discusses its main findings, and proposes possible future research ideas which could be built up on this thesis.

Part I

Interpretability, Missingness, and Datasets explored