Intrinsically Interpretable Machine Learning In Computer Aided Diagnosis

S. Ghosh, Sreejita

DOI:
10.33612/diss.175627883

Summary

Due to relative ease of availability of all kinds of data and the technological advancements in these recent times, we might be trusting *the AI algorithm* just like our ancestors or some of our family and/or friends who we consider blind 'believers', trust(ed) the so-called divine machinations and the high priests. Why should we then consider ourselves any less blink and gullible when we too are trusting *that which is inscrutable*? Sure some complex mathematics is involved which has been validated in the widely accepted top tier journals and conferences. But what is the surety that the black-box model learned what it should really have learned?

An alarmingly harmful effect of trusting such black-box models is when such models are applied in healthcare, as discussed in (Obermeyer et al. 2019). This study investigated how a black-box model, which was trusted and therefore implemented in the healthcare system in a particular country, had actually learned medically irrelevant features to make critical medical decisions. The study reported how the model was in fact biased and had led to the loss of much need medical care of the people belonging to a certain ethnicity. To prevent such bias we need AI models or algorithms whose decisions are explainable.

This thesis presents a set of intrinsically interpretable machine learning models which were applied on real-world medical datasets, a synthetic dataset, and a publicly available dataset from the UCI repository, which posed the challenges of heterogeneous measurements, imbalanced classes, and systematic missingness. The interpretability of the presented set of classifiers are in terms of (1) the classifier's confidence in assigning a class label to a presented sample (instead of just crisp labels), (2) straightforward visualization of the decision boundaries of a presented problem as learned by the classifier, (3) implicit feature relevance computation, and (4) extraction of typical profile(s) of each of the learned classes (prototypes) by the classifier. These newly introduced set of classifiers are nearest prototype based clas-

sifiers (NPCs) which belong to the family of Learning Vector Quantization (LVQ). During training, a LVQ model learns one or more (as defined by the user) representative sample of each class present in the training set; these are known as prototypes. When a new sample is presented to the model the distance or dissimilarity of that sample to each of the prototypes is computed and the sample is assigned the class label of the prototype to which it is most similar, following the winner takes all (WTA) strategy. Later versions of LVQ, have used an adaptive dissimilarity instead of absolute dissimilarity, since not all the features of a training set are not equally relevant to a classification problem. Learning the feature relevances during training leads to an embedded feature selection.

The most popular choice of metric for computation of adaptive dissimilarity is the Euclidean distance. However we investigated and found out that in the presence of missingness and low number of training samples angle based (cosine) dissimilarity is more robust than Euclidean distance. This thesis first presents the angle-dissimilarity based variants of Generalized Relevance LVQ (GRLVQ), Generalized Matrix Relevance LVQ (GMLVQ), Local metric tensor LVQ (LGMLVQ) and Localized Limited Rank Metric LVQ (LLiRAM LVQ). Next, probabilistic variants of the GMLVQ and angle GMLVQ are presented. These models provide the classifier's confidence in assigning different class to a sample. To handle class imbalance this thesis also introduced two strategies: (1) a geodesic variant of Chawla et al's technique of synthetic oversampling of the minority class samples, and (2) user-defined penalization of misclassification of a sample from minority class to an incorrect class.

These newly developed models not just have comparable performance to that of Random Forests, they also help in medical knowledge-extraction from the dataset they are trained on. Random Forest is an ensemble of Decision Trees (DTs) which renders it powerful while compromising with the interpretability of DTs. In this thesis we introduced a geodesic averaging technique which combined the power of ensembling while maintaining the interpretability aspect of the LVQ models.

The performances of these newly developed and presented intrinsically interpretable models showed promise for wider applications in industries of different sectors. It is hoped that industries working on anthropocentric applications using AI will particularly benefit from these models, especially following the recent regulations by the European Commission which states that high risk AI systems, such as those having direct impact on the lives of people, should be as transparent as possible, so that the output from them are easily interpretable. The thesis also discussed the different levels of interpretability and tried to motivate its readers to inspect what a model has learned rather than being satisfied solely by the model's performance metrics.