Journal of Biomedical Informatics 46 (2013) 1125–1135

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Learning classification models from multiple experts

Hamed Valizadegan, Quang Nguyen, Milos Hauskrecht *

Department of Computer Science, University of Pittsburgh, United States

ARTICLE INFO

Article history:
Received 13 April 2013
Accepted 17 August 2013
Available online 13 September 2013

Keywords:
Classification learning with multiple experts
Consensus models

ABSTRACT

Building classification models from clinical data using machine learning methods often relies on labeling of patient examples by human experts. Standard machine learning framework assumes the labels are assigned by a homogeneous process. However, in reality the labels may come from multiple experts and it may be difficult to obtain a set of class labels everybody agrees on; it is not uncommon that different experts have different subjective opinions on how a specific patient example should be classified. In this work we propose and study a new multi-expert learning framework that assumes the class labels are provided by multiple experts and that these experts may differ in their class label assessments. The framework explicitly models different sources of disagreements and lets us naturally combine labels from different human experts to obtain: (1) a consensus classification model representing the model the group of experts converge to, as well as, and (2) individual expert models. We test the proposed framework by building a model for the problem of detection of the Heparin Induced Thrombocytopenia (HIT) where examples are labeled by three experts. We show that our framework is superior to multiple baselines (including standard machine learning framework in which expert differences are ignored) and that our framework leads to both improved consensus and individual expert models.

1. Introduction

The availability of patient data in Electronic Health Records (EHRs) gives us a unique opportunity to study different aspects of patient care, and obtain better insights into different diseases, their dynamics and treatments. The knowledge and models obtained from such studies have a great potential in health care quality improvement and health care cost reduction. Machine learning and data mining methods and algorithms play an important role in this process.

The main focus of this paper is on the problem of building (learning) classification models from clinical data and expert defined class labels. Briefly, the goal is to learn a classification model f: $x \rightarrow y$ that helps us to map a patient instance x to a binary class label y, representing, for example, the presence or absence of an adverse condition, or the diagnosis of a specific disease. Such models, once they are learned can be used in patient monitoring, or disease and adverse event detection.

The standard machine learning framework assumes the class labels are assigned to instances by a uniform labeling process. However, in the majority of practical settings the labels come from multiple experts. Briefly, the class labels are either acquired (1) during the patient management process and represent the decision of the human expert that is recorded in the EHR (say diagnosis) or (2) retrospectively during a separate annotation process based on past patient data. In the first case, there may be different physicians that manage different patients, hence the class labels naturally originate from multiple experts. Whilst in the second (retrospective) case, the class label can in principle be provided by one expert, the constraints on how much time a physician can spend on patient annotation process often requires to distribute the load among multiple experts.

Accepting the fact that labels are provided by multiple experts, the complication is that different experts may have different subjective opinion about the same patient case. The differences may be due to experts' knowledge, subjective preferences and utilities, and expertise level. This may lead to disagreements in their labels, and variation in the patient case labeling due to these disagreements. However, we would like to note that while we do not expect all experts to agree on all labels, we also do not expect the expert's label assessment to be random; the labels provided by different experts are closely related by the condition (diagnosis, an adverse event) they represent.

Given that the labels are provided by multiple experts, two interesting research questions arise. The first question is whether there is a model that would represent well the labels the group of experts would assign to each patient case. We refer to such a group model as to the (group) consensus model. The second question is whether it is possible to learn such a consensus model purely from label assessments of individual experts, that is,

* Corresponding author. Tel.: +1 412 624 8845.
E-mail addresses: hamed@cs.pitt.edu (H. Valizadegan), quang@cs.pitt.edu (Q. Nguyen), milos@cs.pitt.edu (M. Hauskrecht).

1532-0464/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2013.08.007

without access to any consensus/meta labels, and this as efficiently as possible.

To address the above issues, we propose a new multi-expert learning framework that starts from data labeled by multiple experts and builds: (1) a *consensus model* representing the classification model the experts collectively converge to, and (2) *individual expert models* representing the class label decisions exhibited by individual experts. Fig. 1 shows the relations between these two components: the experts' specific models and the consensus model. We would like to emphasize again that our framework builds the consensus model without access to any consensus/meta labels.

To represent relations among the consensus and expert models, our framework considers different sources of disagreement that may arise when multiple experts label a case and explicitly represents them in the combined multi-expert model. In particular our framework assumes the following sources for expert disagreements:

- *Differences in the risks annotators associate with each class label assignment*: diagnosing a patient as not having a disease when the patient has disease, carries a cost due to, for example, a missed opportunity to treat the patient, or longer patient discomfort and suffering. A similar, but different cost is caused by incorrectly diagnosing a patient. The differences in the expert-specific utilities (or costs) may easily explain differences in their label assessments. Hence our goal is to develop a learning framework that seeks a model consensus, and that, at the same time, permits experts who have different utility biases.
- *Differences in the knowledge (or model) experts use to label examples*: while diagnoses provided by different experts may be often consistent, the knowledge they have and features they consider when making the disease decision may differ, potentially leading to differences in labeling. It is not rare when two expert physicians disagree on a complex patient case due to differences firmly embedded in their knowledge and understanding of the disease. These differences are best characterized as differences in their knowledge or model they used to diagnose the patient.
- *Differences in time annotators spend when labeling each case*: different experts may spend different amount of time and care to analyze the same case and its subtleties. This may lead to labeling inconsistency even within the expert's own model.

We experiment with and test our multi-expert framework on the Heparin Induced Thrombocytopenia (HIT) [23] problem where our goal is to build a predictive model that can, as accurately as possible, assess the risk of the patient developing the HIT condition and predict HIT alerts. We have obtained the HIT alert annotations from three different experts in clinical pharmacy. In addition we have also acquired a meta-annotation from the fourth (senior) expert who in addition to patient cases have seen the annotations and assessments given by other three experts. We show that our framework outperforms other machine learning frameworks (1) when it predicts a consensus label for future (test) patients and (2) when it predicts individual future expert labels.

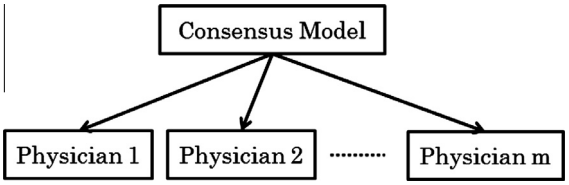

Fig. 1. The consensus model and its relation to individual expert models.

2. Background

The problem of learning accurate classification models from clinical data that are labeled by human experts with respect to some condition of interest is important for many applications such as diagnosis, adverse event detection, monitoring and alerting, and the design of recommender systems.

Standard classification learning framework assumes the training data set $D = \{(x_i, y_i)\}_{i=1}^{n}$ consists of n data examples, where x_i is a d-dimensional feature vector and y_i is a corresponding binary class label. The objective is to learn a classification function: f: $x \rightarrow y$ that generalizes well to future data.

The key assumption for learning the classification function f in the standard framework is that examples in the training data D are independent and generated by the same (identical) process, hence there are no differences in the label assignment process. However, in practice, especially in medicine, the labels are provided by different humans. Consequently, they may vary and are subject to various sources of subjective bias and variations. We develop and study a new *multi-expert classification learning framework* for which labels are provided by multiple experts, and that accounts for differences in subjective assessments of these experts when learning the classification function.

Briefly, we have m different experts who assign labels to examples. Let $D^k = \{(\mathbf{x}_i^k, y_i^k)\}_{i=1}^{n_k}$ denotes training data specific for the expert k, such that $\mathbf{x}_i^k$ is a d-dimensional input example and y_i^k is binary label assigned by expert k. Given the data from multiple experts, our main goal is to learn the classification mapping: f: $x \rightarrow y$ that would generalize well to future examples and would represent a good consensus model for all these experts. In addition, we can learn the expert specific classification functions g_k: $x \rightarrow y^k$ for all $k = 1, \ldots, m$ that predicts as accurately as possible the label assignment for that expert. The learning of f is a difficult problem because (1) the experts' knowledge and reliability could vary and (2) each expert can have different preferences (or utilities) for different labels, leading to different biases towards negative or positive class. Therefore, even if two experts have the same relative understanding of a patient case their assigned labels may be different. Under these conditions, we aim to combine the subjective labels from different experts to learn a good consensus model.

2.1. Related work

Methodologically our multi-expert framework builds upon models and results in two research areas: *multi-task learning* and *learning-from-crowds*, and combines them to achieve the above goals.

The multi-task learning framework [9,27] is applied when we want to learn models for multiple related (correlated) tasks. This framework is used when one wants to learn more efficiently the model by borrowing the data, or model components from a related task. More specifically, we can view each expert and his/her labels as defining a separate classification task. The multi-task learning framework then ties these separate but related tasks together, which lets us use examples labeled by all experts to learn better individual expert models. Our approach is motivated and builds upon the multi-task framework proposed by Evgeniou and Pontil [9] that ties individual task models using a shared task model. However, we go beyond this framework by considering and modeling the reliability and biases of the different experts.

The learning-from-crowds framework [17,18] is used to infer consensus on class labels from labels provided jointly by multiple annotators (experts). The existing methods developed for the problem range from the simple majority approach to more complex consensus models representing the reliability of different experts.