Women's Health Risk Assessment Using Active Learning Techniques

Ishita Dasgupta
Email: ishitadg@cs.umass.edu

Lopamudra Pal
Email: lpal@cs.umass.edu

Sruthi Chilamakuri
Email: schilamakuri@cs.umass.edu

I. Problem

There is an international challenge in Women's health risk assessment. Sex-based health disparities are evident throughout the world; especially in women belonging to developing countries due to widespread unawareness in these areas[1]. Some machine learning analysis has been carried out for predicting women's health risk by collecting their social and health data[2].However challenges in health risk assessement are aggravated by insufficient data, either due to sensitivity of the information involved or mishandling of data. This calls for application of different forms of Active Learning[8] on existing classification algorithms to help solve the problem of missing data while maintaining or improving prediction accuracy.

II. Methodology

- The main components of our approach : 1. Data Collection and Preprocessing 2. Model Selection and Training 3. Hyperparameter Tuning 4. Classification 5. Testing
- Models/Algorithms - Logistic Regression (as baseline), Support Vector Classifier, Neural Networks along with Active Learning
- Optimization Techniques - Optimizing the Fisher Information matrix and using Maximum likelihood estimation of joint probability distribution.
- Code Libraries - Numpy, TensorFlow/Pytorch, pandas, matplotlib
- Hardware Platforms - The hardware platforms we are targeting is general purpose computers, although we may incorporate some part of the code in a distributed environment.

III. Related Work

A lot of work has been carried out in the field of active learning for the purpose of medical diagnosis. [10] describes the approach of selecting a batch of unlabeled data for manual labeling as opposed to a single unlabeled example to reduce the number of times the model must be retrained. [11] incorporates clustering to handle unlabeled data. The algorithm first constructs a classifier on the set of the cluster representatives, and then propagates the classification decision to the other samples via a local noise model. [6] handles active learning on high dimensional image data by making use of Bayesian Convolutional Networks. [5] details geometric methods of choosing a set of images to label such that they cover the set of unlabeled images as closely as possible. We will attempt to implement a combination of these active learning techniques in order to classify data.

IV. DATASETS USED

This dataset has been collected from a competition that was organized to create a machine learning model which will automatically classify women into different health risk segments and sub groups based on the information collected from the participants by Cortana Intelligence Gallery[9]. The data was provided in accordance with Bill & Melinda Gates Foundation open data access policy. Around 9000 women aged between 15 and 30 from 9 under developed regions were surveyed, approximately 1000 from each region. Women participants were asked to answer few basic questions regarding their education level, sexual health, etc[3]. Only the training dataset of 5000 women was made available publicly which we will use for carrying out our experiments at this point. We have split the training data set into 80% -20% training-test data subsets. Further description of the dataset is available here[4]. We have requested the organization to provide us with the remaining 4000 test dataset if possible.

V. EXPERIMENTS

The classification model selected in [10] is Logistic Regression and they have applied Active Learning by optimizing the Fisher Information matrix. The paper has given a greedy solution to the problem of optimizing the objective function when the number of unlabeled data is high. We plan to extend the same problem with other classification algorithms like Support Vector Machines and Neural Networks in our current experiment. [11] suggests the use of clustering to perform Active Learning. The optimization Technique used is maximum likelihood estimation of the sum of labeled data and unlabeled data. According to this framework, the clustering information is explicitly incorporated by introducing the hidden cluster label. Once the cluster label is known we can find the joint probability distribution of x,y and the clustered label.
We plan to implement the above two variations along with the methodology explained on the top of classification strategies applied on the dataset as in [2]. We may also try Bayesian techniques for Active Learning if time permits. We shall perform apply hyperparameter tuning by performing n-fold cross-validation over the 80% of the data by iteratively changing the training and validation data. We shall evaluate our results by computing the classification accuracy and comparing with our baseline.

VI. Overlap Statement

Project does not overlap with past or current work at all for each of the team members involved.

VII. Collaboration specifics

The entire project will be carried out in the form of 6 steps, with each member of the team overlooking each process differently.

1) Preprocessing and understanding data will be done collaboratively
2) Feature selection will be done collaboratively
3) Once features are selected, we use different methods of active learning - Lopamudra will implement batch mode learning, Ishita will work on active learning with pre-clustering and Sruthi will execute Bayesian methods in active learning
4) Next we will apply these active learning methods on various classifiers. Lopamudra will implement logistic regression, Sruthi will perform SVC and Ishita will experiment with neural networks
5) Hyperparameter tuning will be done for each model individually
6) Performance analysis and report generation will be done collaboratively.
7) We will use different visualizations to convey our work. This will be done collaboratively.

References

[1] WHO, *"Womens health," in World Health Organization,*, 3rd ed. Harlow, England: Addison-Wesley, 1999.

[2] Sharathkumar Anbu, Bhaskarjit Sarmah *Machine Learning Approach for Predicting Womens Health Risk*, IEEE 2017

[3] *Dataset*, http://az754797.vo.msecnd.net/competition/whra/data/WomenHealth_Training.csv

[4] *Dataset Description*, https://az754797.vo.msecnd.net/competition/whra/docs/data-description.docx

[5] *A Geometric Approach to Active Learning for Convolutional Neural Networks*, https://arxiv.org/abs/1708.00489

[6] *Deep Bayesian Active Learning with Image Data*, https://arxiv.org/abs/1703.02910

[7] *Active Learning with a Neural Network*, https://www.cs.cmu.edu/afs/cs/project/jair/pub/volume4/cohn96a-html/node4.html

[8] *Active Learning*, https://en.wikipedia.org/wiki/Active_learning_(machine_learning)

[9] *Womens Health Risk Assessment Competition*, https://gallery.cortanaintelligence.com/Competition/Womens-Health-Risk-Assessment-1

[10] *Steven C. H. Hoi,Rong Jin,Jianke Zhu,Michael R. Lyu, (2006) "Batch mode active learning and its application to medical image classification", doi 10.1145/1143844.1143897*

[11] *Hieu T. Nguyen,Arnold Smeulders (2004) "Active learning using pre-clustering", doi10.1145/1015330.1015349*,