# Analysis of the Impact of Errors Made During Health Data Collection Using Mobile Phones: Exploring Error Modeling and Automatic Diagnosis

Sukhada Palkar
Language Technologies Institute, CMU
spalkar@cs.cmu.edu

Emma Brunskill
Computer Science Department, CMU
ebrunskill@cs.cmu.edu

## 1. INTRODUCTION

Mobile phones are near ubiquitous, and can be easily used to gather and store health data in remote or low resource settings. There exist many systems for supporting such data gathering, including Commcare, Frontline SMS, and OpenData Kit. Survey and health data is often collected by community health workers and frequently includes errors, due to mistakes, challenges with the input interface, systematic error or neglect [1,5]. Automatic detection of errors is important because of its potential impact on aggregate health statistics, and on individual patient treatment. In some important cases, such as tuberculosis diagnosis and monitoring, the space of possible medical diagnoses will generally be significantly smaller than the possible set of symptoms recorded. This suggests that it may be possible to build diagnostic systems whose recommendations are fairly robust to errors in the recorded patient symptoms.

In this abstract we describe preliminary work towards this goal. First, we construct a model of tuberculosis symptom data entry errors made by health workers as as observed by Patnaik et al. [1]. Second, we train machine learning classifiers to predict diagnosis decisions based on a simulated dataset generated from the error records: simulated data was used since the original dataset was very limited. Finally we analyze the sensitivity of our results to the particular learned error model parameters, by computing a Bayesian posterior over the parameters and sampling parameters from that distribution.

## 2. ERROR MODELING

Previously Patnaik et al. [1] conducted a lab study to investigate the accuracy of health care workers at entering tuberculosis symptoms using three different mobile phone interfaces: cue-card guided SMS entry, Java forms, and voice interaction with a human operator. Though the voice entry had a very low error rate (0.4%), both forms and SMS had an error rate of just above 4%. Since it was a lab study, the dataset included ground truth data. To model the pattern of errors made in the SMS and form interfaces, we created a hierarchical probabilistic error model. We explored several error models, and evaluated them by the likelihood of the dataset under the proposed model and learned features.

Of the models tried, the one with the highest likelihood is a 3-layer multinomial error model. The input to the model is a health record with the following features: temperature, weight, cough type and presence of nausea, coughing with blood, yellow eyes, chest pain, night sweats, loss of appetite and fatigue. All features are discrete or are discretized before further analysis. The model first samples the number of errors that will be present in the record from a multinomial over 0, 1 or 2 errors (no more than 2 errors were ever observed in the dataset). Next, given the number of errors (1 or 2), we sample the possible symptom errors from a joint multinomial over the set of features: this stage selects which features have errors. Cough is the only non-binary feature. If the cough feature is sampled as one of the error features, we then sample the cough error from another multinomial. We fit the model parameters using maximum likelihood estimation, but added a small amount of smoothing to account for unseen errors that might be present in a larger dataset.

We would like to build a classifier that can automatically predict a tuberculosis diagnosis based on the input features, and evaluate how sensitive it is to erroneous input data. The typical approach to doing this would be to separate a labeled dataset into a training and test set, and train the model on the training set, and evaluate its accuracy on the dataset. Unfortunately, Patnaik et al.'s dataset is too small (52 records) to be further subdivided into a training and test set. Though we wish to later work with a larger dataset, we initially circumvented this issue by generating a larger dataset. We did this by sampling (with replacement) the original Patnaik dataset to get 623 total data points. We labeled each data point as one of three tuberculosis classes, which we generated by looking at the literature ([2], [3]): no tuberculosis (TB), TB or TB along with an opportunistic disease. Note that another limitation of our use of the Patnaik dataset is that the class of symptoms designed for the authors' lab study is not necessarily reflective of typical tuberculosis symptom combinations.

## 3. TRAINING A DIAGNOSTIC CLASSIFIER

We trained a classifier to predict the diagnostic label given some (potentially erroneous) record. We trained two popular classification algorithms: Naive Bayes and Random Forests [6]. It is certainly possible that further gains in accuracy could be achieved by a more exhaustive exploration of classifiers, but as our interest was in the relative classification performance on observed vs true data, we chose to focus on two frequently used approaches. Both Naive Bayes and Random Forests are available on a number of software packages and we used the nltk implementation for Naive Bayes and the sklearn implemen- tation for Random Forests. Learning and testing a classifier on the true records provided us with an upper

*DEV '13,* January 11-12, 2013 Bangalore India
Copyright 2013 ACM 978-1-4503-1856-3/13/01 ...$15.00.

**Table 1: Classification accuracy.$^{*}$ indicates that a classifier was significantly better than a chance classifier $^{+}$ indicates that a result was significantly worse than training and testing on the original dataset using the same classifier.**

| | | Classifier Accuracy | | |
|---|---|---|---|---|
| Training | Test | Naive Bayes | Random Forests | Chance |
| Error | Error | .89 [.83-.93]* | .88 [.83-.93]*$^{+}$ | .41 [.24-.58] |
| Original | Error | .87 [.82-.91]* | .89 [.85-.91]*$^{+}$ | .41 [.24-.58] |
| Original | Original | .91 [.87-.93]* | .96 [.91-1.0]* | .41 [.24-.58] |

bound of the performance we could hope to achieve on the given dataset, and a rela- tive measure by which to compare a classifier learned and trained on the observed records which had additional error. We also constructed a simple baseline classifier, which simply labels all records by whichever class had the highest number of items in the training set, This provides a lower bound on the performance we would hope to achieve.

## 4. EXPERIMENTAL RESULTS

We performed 10-fold cross validation. For each fold, we computed the each classifier's accuracy. When testing on records with possible interface errors ("Error"), it is not immediately obvious if it is better to train on error-free records ("Original"), or on error records. Therefore we tried both options. The test set always consisted of records that could contain errors.

The learned classifiers were quite accurate at predicting the diagnostic class of the error records test set (see Table 1). When training on error records (Error & Error), both Naive Bayes and Random Forests achieved an average accuracy significantly better than the chance baseline classifier. There appeared to be no definitive advantage to training on error records versus original records, as the average accuracy fell slightly for Naive Bayes, and rose slightly for Random Forests (Original & Error). These slight differences were not statistically significant. This is encouraging, because it suggests that it is possible to train a good diagnostic classifier even if the only available records may have interface errors. It is aiso instructive to see how much accuracy is lost compared to a classifier trained on the original records (Original & Original, our upper bound). For both classifiers the actual difference in accuracy is fairly small. This indicates that the decrease in diagnostic accuracy due to interface error is fairly small.

## 5. ERROR MODEL SENSITIVITY ANALYSIS

One important issue is that our original error model parameters were set by computing the maximum likelihood estimates. Given the small set of available data, there is a fair amount of uncertainty over these underlying estimates. Therefore, assuming the same error model structure, the parameter values of how frequently errors occur, and the symptom errors themselves, may be significantly differ from the maximum likelihood estimates. We wished to estimate the impact of this uncertainty on the classification error rates. To evaluate this, we constructed two different error models, with the same hierarchical probabilistic structure as our original model, but using different parameter estimates.

Recall that our hierarchical probabilistic error mode first samples from a multinomial over the number of errors. We used the observed frequency counts of the number of errors made as the parameters to a Dirichlet distribution, which specifies a probability distribution over the parameters of a multinomial. We then sampled 100 multinomial parameter sets from this Dirichlet and ranked the samples according to the expected number of errors. We then selected the sets with the third lowest and ninety-eighth highest and tenth lowest and ninetieth highest expected error. We did this to approximate a 95th percentile and an 80th percentile confidence interval over the expected error rate parameters. We followed a similar approach for generating alternatef error model parameters for the probability of individual symptom errors.

**Table 2: Error Model Sensitivity**

| Error Model | Naive Bayes Classifier Accuracy |
|---|---|
| Maxmimum likelihood model | .91 [.85-.95] |
| Dirichlet 3% | .88 [.85-.93] |
| Dirichlet 10% | .88 [.82-.93] |
| Dirichlet 90% | .84 [.79-.87]$^{+}$ |
| Dirichlet 98% | .82 [.74-.90]$^{+}$ |

We then repeated all the prior analysis using these four additional error models and the Naive Bayes classifier. The classifier results for the different Dirichlet error models are detailed in Table 2. Note that Dirichlet 3% corresponds to a model with the lowest number of expected errors of the four alternate models, and Dirichlet 98% corresponds to the highest. These results suggest that the classifier accuracy is fairly robust to variability in the underlying error model parameters, though as might be expected, the classifier performance decays slightly as more errors are common.

## 6. CONCLUSION AND FUTURE WORK

This preliminary work leads to many interesting future directions. One is to create classifiers that explicitly leverage the known error model. We are also interested in using this process to build similar models for other diseases, and see if perhaps this process could be automated.

The key limitations of the present work is the error model is built on lab study data, and evaluated on simulated data with hand crafted labels. Such data is unlikely to reflect the distribution of data, and data entry errors, present in normal field operations. Therefore, we hope to collect or obtain access to real field data with expert provided diagnoses in the future. Ultimately we are very interested in helping make a clinical care decision support or reminder system more robust to input data errors. Despite the aforementioned limitation, our preliminary results are encouraging in that they suggest that it is likely to be possible to build classifiers for use in health systems that are robust to data entry errors.

## 7. REFERENCES

[1] S. Patnaik, E. Brunskill, and W. Thies. Evaluating the accuracy of data collection on mobile phones: a study of forms, sms, and voice. ICTD, 2009.

[2] Tuberculosis. Technical report, World Health Organization, Atlanta, GA, USA, April 2012.

[3] Tuberculosis (tb). Technical report, Centers for Disease Control and Prevention, 2012.

[4] K. Chen, H. Chen, N. Conway, J. Hellerstein, and T. Parikh. Usher: Improving data quality with dynamic forms. *IEEE Transactions in Knowledge and Data Engineering*, 2010.

[5] L.Breiman. Random forests. *Machine Learning*, 45:5–32, October 2001.