A Simplified Bayesian Approach for Context and Characteristic Identification Based on Demographic Variables

Saurabh Amin,
Department of Civil Eng.
Univ. of Texas at Austin,
Austin, TX-78712
samin@mail.utexas.edu

The present approach characterizes the contexts and the gender based on demographic variables only i.e. it completely ignores the sensor data. The premise behind this approach is that given the demographic variables, Bayes rule can be used to estimate the probability that a subject with these demographics is likely to pursue a certain kind of activity. For example persons belonging to a working age group are more likely to do work. Though the same logic does not extend to gender prediction, we still try to predict the gender based on demographics. We compare the Bayesian prior model for both context and gender identification to the most naive global prior model.

The Bayesian model is as follows:

*P(Context|Characteristic)=P(Characteristic|Context)*P(Context)/P(Characteristic)*
Where:
P(Context)= Prior probability of a context
P(Characteristic|Context)= Conditional likelihood or the probability of characteristic given the Context
P(Characterisitc)= Likelihood of the characteristic
P(Context|Characteristic)= Posterior or the probability of a context given the chactacteristic

The two demographic variables given in the data are CHARACTERISTIC1 and CHARACTERISTIC2. CHARACTERISTIC1 is an ordinal variable and CHARACTERISTIC2 is binary. Preliminary observation shows that CHARACTERISTIC2 is not a good candidate for characterizing ANNOTATION because most of the users with CHACTERISTIC2 value of 1 contribute very little to the annotated contexts. In addition, all the users with CHACTERISTIC2 value of 1 belong to GENDER type 1. On the other hand, CHRACTERISTIC1 can be used for better generalization because of its ordinal nature.

All the sessions which did not even have a single labeled context i.e. the user did not annotate the activity he/she was doing are removed from the training data. The remaining sessions are divided into 70% TRAINING SET and 30% TESTING SET using random stratified sampling. Further, all the sub-sessions with unlabled context were removed from the TRAINING SET to form the LABELED TRAINING SET. For the purpose of context identification, ANNOTATION is relabled to CLASSLABEL with nominal values 1 (CONTEXT 1), 2 (CONTEXT 2), 3 (ALL OTHERS). For all the three values of CLASSLABEL, counts of records of values of CHARACTERISTIC1 in range

[1,100] are obtained and a three parameter Weibull is fitted. Weibull distribution is frequently used in life data analysis and offers an advantage that it can assume different shapes based on the values of its parameters. The scale and location parameters of each Weibull are decided by trial and error by observing the plot fitted distribution and the observed counts. The third parameter is the shape parameter and it determines the spread of the distribution. The shape parameter was found to affect the results of the final model to a great extent. The shape parameters of the Weibulls are optimized by observing the performance of the model over a range of values and selecting the values that give best performance on the TESTING SET.

Once the interpolated counts of CHARACTERISTIC1 values for all the values of CLASSLABEL are obtained, the P(Context), P(Characterisitc1) can be computed by row and column summation and dividing by total counts. *P(Characteristic|Context)* can also be calculated by dividing by total counts. Finally, *P(Context|Characteristic)* can be computed by applying Bayes rule. The accuracy of P(Context|Characteristic) model is evaluated on the TESTING SET and the confusion matrix is calculated. The performance of the model is judged using the following criterion fixed by the organizers.

Score = 0.3(TP/(TP+FN))+ 0.7*(TN/(TN+FP))* {For Context Indentification}}

The Bayesian model can also be compared to the more naive global prior model which identifies the contexts based on their probability of occurrences in the LABELED TRAINING SET. While the Bayesian model gives a score of 0.689, the naive prior model gives 0.463 accuracy on the TESTING SET. The optimal shape parameters for the three CLASSLABEL values are 2.0, 2.0, and 6.0 respectively.

The procedure can be repeated for calculating the *P(Gender|Characteristic)*. Since all the records have GENDER labels, TRAINING SET is used directly instead of LABELED TRAINING SET. The criterion for GENDER prediction is:

Score = 0.5(Correct0/(Correct0+Incorrect1))+ 0.5*(Correct1/(Correct1+Incorrect0))*
{For Gender Indentification}

The Bayesian model and the naive prior model scored 0.51 and 0.49 on the TESTING SET. The optimal parameters in this case are 2.0 and 5.0 for the two GENDER labels.

Since the COMPETITION TEST SET is representative of the training data in terms of demographic variables and it is expected that the Bayesian model obtains similar performance on the COMPETITION TEST SET. Of course, since the only the one demographic feature is used to characterize the contexts and gender, the usefulness of the model is limited. However, we plan to extend the model for modeling sensor data as follows:

P(Context|Sensor Data)=P(Sensor Data|Context)*P(Context)/P(Sensor Data)

Each context can be characterized by a probability distribution and the parameters can be estimated from the TRAINING SET. Given the sensor data, the context class where there sensor data has a maximal probability of belonging to that class can be chosen.

Acknowledgements: The author gratefully acknowledges help provided by Mr. Gunjan Gupta in problem formulation and coding. Thanks are due to Dr. Peter Stone for his encouragement.