Review of The Paper

Ved Thakur

January 31, 2021

1 Pros of the Proposed Model

The paper clearly shows how its proposed model performs much better than primitive imputation techniques. Also the real life data in most cases is never complete. The data calculated in social science experiments is often in complete as several people refrain from giving the information; or a certain category of information would be removed due to error in data management. This all causes missing data.

In cases where we record sensor data, it might be possible that certain sensor is not working,in such cases the data recorded is of NULL value. It might be also possible that there is some error in the sensor; therefore a lot of outliers might have been recorded and so these outliers should be deleted for our model to have maximum accuracy.This also creates missing data.

In all such cases it is not possible to manually impute the missing value as the number of missing data is very large. Further standard imputation techniques like replacing by zeroes or replacing by mean would change the outcome of the model being trained. Hence, our proposed model would outperform in all such cases and also would lead to higher accuracy of the trained model.

Further in the paper it is also visible the VB-EM performs much better in cases when there are fewer training data to train the GMM. Also experiments show than VB-EM algorithm is superior in terms of density estimation when data is scarce. Another important question arises whether ignoring data with missing features is better than using our model or any standard imputation. Ignoring the data would eliminate the chances of incomplete data training issues but this would not solve issues with testing data since we cannot ignore a data point because of one or two missing features.

Therefore we would have to rely on some ad-hoc method(filling in zeroes or unconditional mean); it would be better to use the proposed model since it is principled does not rely on any ad-hoc method and also performs better than those methods. This proves that our proposed model works in most of the cases.

Also average of all 10 trials was represented in the graph which was fair enough to compare the different GMM models. Also it is visible that the experiments have been considered with data sets from different fields. We are testing our model on an IONOSPHERE data set and also on a BREAST CANCER data set. Both these data sets belong to different values and have different set of values which ensures that our model works with data sets of all fields.

All the concepts regarding logistic regression were properly implemented. Concepts regarding GMMs and EM were also well written. Also the experiments were considered appropriately; avoiding the chances of over fitting by randomly sorting the data.

2 Cons of the Proposed Model

Since we are using concepts of Gaussian Matrix Model (GMM) and Expectation Maximization, the maths and the calculation involved is much more complex than standard imputation techniques. This would lead to more time taken to train a model. Further this problem would get more severe when the data increases. Also we can observe from the graphs;[Here graphs refer to those in Figure 2 of paper] that if we increase the percent of training data our proposed model performs similar to that of imputation methods.

Also a major con of the paper which I feel is the size of data set used for experiment purposes. The data sets are smaller with 351 and 569 data points. This created doubt in the mind whether our model would work fine with data sets having more than 10,000 data points.

Another major problem which I feel is that our proposed model is compared with standard imputation techniques but never compared with method when we drop all the data points which have some missing features. It still creates a doubt in the mind of reader whether dropping those data points with missing features would lead to better accuracy or not.

3 Technical Errors and Suggestions

I feel that the numerical parts of the paper should have be written in a different method for beautification and to look different from normal text. Also the numerical part has not been explained in detail. It's clearly mentioned in the paper, "TO CONSERVE SPACE"; which conveys the idea that the author is not interesting in explaining the concept in detail. Adding to that there are several usages of Greek symbols which have not be explained in depth but only in brief. Further the flow of equations is not easily understandable; the author makes it complicated to understand for a normal amateur. There are not much mentions of citations wherever required.

Technical Errors related to Experimentation are mentioned in detail in the cons of the paper. The author has not used a data set with large data points; this might create error when we use our model with large and varied data sets.

It would be better if the above technical errors are solved. Providing citations wherever necessary. Modifying the numerical part so that a normal person understands it easily and also changing the font type of the numericals.

Despite all these the theory part is well written. It is neat and easy to understand. All the basic rules of grammar and punctuation have been followed. The Experiment figures are easy to understand since there is a usage of various colours.

4 Directions for Future Research

Future work would be examining the possibility of a classifier constructed from incomplete data that can outperform a classifier constructed from the complete data. This hypothesis may be true if features that confuse the class separation(outliers) are present.

Additional work would also be of investigating the use of Dirichlet's process to address choosing the number of components for the GMM.