Bayesian Multiple Imputation Approaches for One-Class Classification

Shehroz S. Khan, Jesse Hoey, and Daniel Lizotte

David R. Cheriton School of Computer Science
Univesity of Waterloo
{s255khan,jhoey,dlizotte}@uwaterloo.ca

Abstract. One-Class Classifiers build classification models in the absence of negative examples, which makes it harder to estimate the class boundary. The predictive accuracy of one-class classifiers can be exacerbated by the presence of missing data in the positive class. In this paper, we propose two approaches based on Bayesian Multiple Imputation (BMI) for imputing missing data in the one-class classification framework called Averaged BMI and Ensemble BMI. We test and compare our approaches against the common method of Mean imputation and Expectation Maximization on several datasets. Our preliminary experiments suggest that as the missingness in the data increases, our proposed imputation approaches can do better on some data sets.

1 Introduction

Missing or incomplete data imposes severe limitations on inference and estimation of machine learning algorithms. To handle missing data, analysts either employ ad-hoc approaches, e.g. ignore/delete instances with missing values or estimate them using some process and fill in the missing values (called *imputation*). Ignoring the missing values is not efficient as it involves loss of information and may either lead to underfitting or exhaustion of training instances. The predictive performance of classifiers deteriorates on incomplete data, therefore data imputation is generally employed before training a classifier as a pre-processing step [8]. A common data imputation technique is Mean Imputation (MEI), where in the missing values of an attribute are replaced by the mean or mode value of the attribute.

One-Class Classification (OCC) [9][7] is a machine learning paradigm that learns classifiers primarily in the absence of negative examples. Since either the negative examples are absent, statistically insignificant, or very few, this inflicts challenge in defining the class boundary across the positive data. The OCC problem can be compounded if there are missing values in the positive data itself, which imposes two-fold challenge in terms of defining the class boundary on the positive data and handling missing values in the dataset. Su et al. [8] show that for multi-class classification methods, using multiple imputation techniques such as Bayesian Multiple Imputation (BMI) as a pre-processing step on data with missing values improves the predictive performance of conventional

machine learning algorithms. In this paper, we extend the use of BMI for the one-class classification problems and propose two approaches called Averaged BMI (ABMI) and Ensemble BMI (EBMI) to learn one-class classifiers with missing data.

The rest of the paper is organized as follows. In Section 2, we briefly discuss the BMI method. Section 3 provides a short overview of one-class SVM classifiers. In Section 4, we present our proposed approach for integrating BMI with OCC on missing data, followed by experimental design and results in Section 5. Section 6 concludes our presentation with pointers to future work.

2 Bayesian Multiple Imputation

BMI follows a Bayesian framework by specifying a parametric model for the complete data and a prior distribution over unknown model parameters θ. Then it draws m independent trials from the conditional distribution of missing data given the observed data using Bayes' Theorem. MCMC can be used to simulate the entire joint posterior distribution of the missing data to obtain estimates of posterior parameters. We use a version of BMI that assumes a multivariate normal distribution when generating imputations for the missing values [6].

Let the complete data be denoted by $Y = (Y_{obs}, Y_{mis})$, where Y_{obs} and Y_{mis} are observed and missing part of Y. Also, let $P(Y|\theta)$ model the complete data, and the parameter θ is the mean and covariance matrix that parameterize a normal distribution, let $P(\theta)$ be the prior distribution over θ. The posterior predictive distribution for Y_{mis} is given by [6]

$$P(Y_{mis}|Y_{obs}) = \int P(Y_{mis}|Y_{obs}, \theta)P(\theta|Y_{obs})d\theta \tag{1}$$

where

$$P(\theta|Y_{obs}) \propto P(\theta) \int P(Y_{obs}, Y_{mis}|\theta)dY_{mis} \tag{2}$$

BMI imputes the missing data by repeating the following steps for $j = 1, 2, ..., m$

1. Imputation I-step: Generate missing values Y_{mis}^{j+1} from $P(Y_{mis}|Y_{obs}, \theta^j)$
2. Posterior P-Step: Draw parameter θ^{j+1} from $P(\theta|Y_{obs}, Y_{mis}^{j+1})$

Repeating the above I-step and P-step generates the following Markov Chain

$$(Y_{mis}^1, \theta^1), (Y_{mis}^2, \theta^2), \ldots, (Y_{mis}^j, \theta^j) \tag{3}$$

These two steps are iterated with a starting value θ^0, until the distribution $P(Y_{mis}, \theta|Y_{obs})$ is stabilized and provides reliable imputed datasets [6]. Once the m-imputed data sets are generated using the above process, they can be combined for analysis and inference.

3 One-Class Classification

Tax and Duin [9] propose a OCC method that seek to distinguish the positive class from all other data objects by constructing a hyper-sphere with the minimum radius around the positive class data that encompasses almost all points in the data set. This method is called the Support Vector Data Description (SVDD). Schölkopf et al. [7] present an alternative approach for solving this problem by constructing a hyper-plane instead of a hyper-sphere around the data, such that this hyper-plane is maximally distant from the origin and can separate the regions that contain no data. They propose to use a binary function that returns +1 in 'small' region containing the data and -1 elsewhere. The data is mapped into the feature space corresponding to the kernel and is separated from the origin with maximum margin. In practical implementations, this method and the SVDD method operate comparably and both perform best when the Gaussian kernel is used. Tax and Duin [9] mention that SVDD requires large number of training data to support a flexible classification boundary. Their method becomes inefficient when the dataset has high dimension. This method also doesn't work well when large density variation exists among the objects of data set.

4 Proposed Approaches

In the formulation of SVDD, Tax [10] did not consider the presence of missing values in the data as they may introduce extra complication while learning a one-class classifier. Cohen et al. [2] use the one-class SVM [7] for nosocomial infection detection. They replace the missing values in the data with the class-conditional mean for continuous variables and the class-conditional mode for nominal ones. In our literature review we did not find any reported research work that uses BMI to impute missing values and then train the one-class classifier. In the following subsections we present two proposed approaches that are based on BMI and can be used in the OCC framework to handle missing data.

4.1 Average BMI

In the Average BMI (ABMI) method, positive data with missing values is imputed m times using BMI, thereby generating m instances of original data that are different in their imputed values. The imputed data is then averaged across all m imputations for every attribute.

4.2 Ensemble BMI

Similar to ABMI, in EBMI the positive data with missing values is imputed m times using BMI, and therefore m instances of original data are generated. Each of these m imputed datasets are supplied to a separate one-class classifier that takes individual decision as to whether a given test instance belongs to *target* or *outlier*. A combined prediction from all these m classifiers can be arrived at by a majority voting scheme.

5 Experimental Analysis

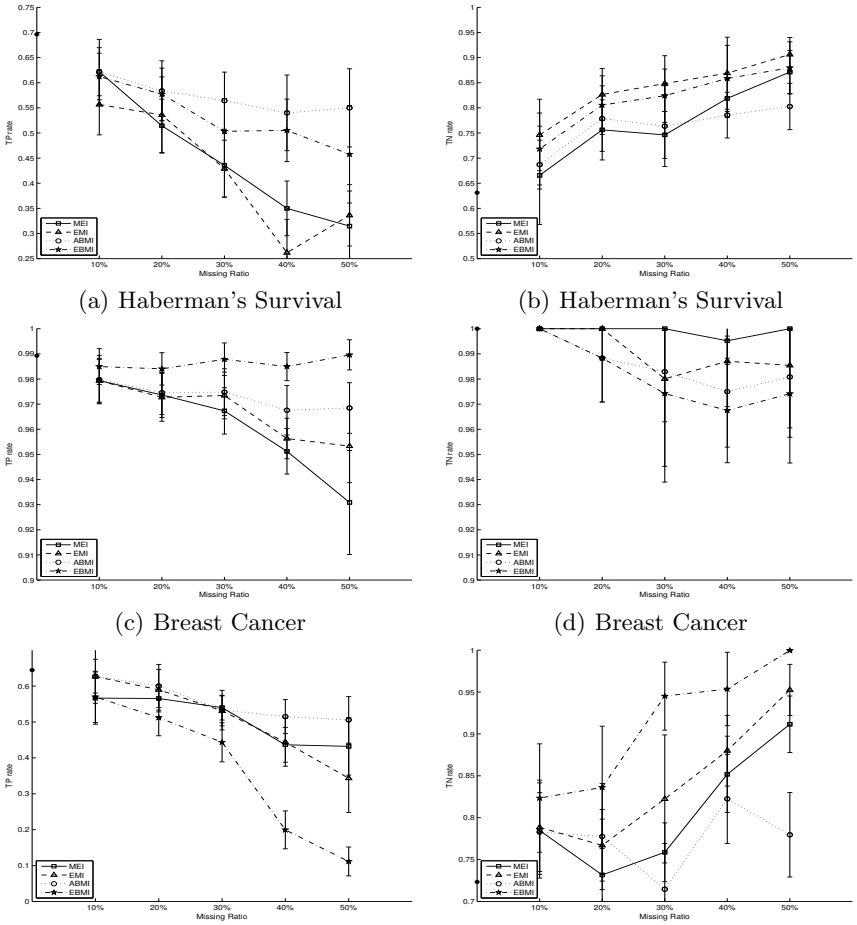
The datasets we took from UCI Machine Learning Repository [3] are: Breast Cancer, Indiana Diabetes and Haberman's Survival. In each of the datasets, the positive class corresponds to the instances from healthy individuals as they are easy to collect and negative samples require someone to actually develop those ailments. We use the One-Class SVM ($OSVM$) [1] with ν=0.05 and $\gamma = \frac{1}{\#attributes}$ (except for Indiana Diabetes data with $\gamma = 0.01$).

5.1 Methodology and Results

For evaluating the classification accuracy on the complete dataset, 10 fold cross-validation is performed 10 times. For every iteration, the data is randomized, the *target* and *outlier* classes are divided into 10 folds, and missingness is introduced in the *target* class. For training the classifiers, only the data objects from the *target* class are used and those from *outlier* class are ignored [5]. For testing, the remaining objects (non-imputed) from the *target* and *outlier* class are used. The average of mean True Positive rate (TPR) and True Negative rate (TNR) at different missingness across 10 folds over 10 times is used as performance metric.

From each dataset, incomplete datasets are generated by artificially deleting values from across the attributes 'completely at random', that is 'missing completely at random (MCAR)' in that the missingness does not depend on the observed data. For a dataset, the missingness in attribute values is varied in several steps from 10% to 50%. The proposed approach (see Section 4) of ABMI and EBMI is then employed to impute these missing values and learn the one-class classifiers. For comparison purposes, the common method of MEI and Expectation Maximization (EMI) [4] is also employed on the same incomplete data as presented to EBMI and ABMI and corresponding one-class classifiers are built. The value of multiple imputations for ABMI and EBMI is kept at 5. The y-axis does not start from origin and the symbol • (see Figure 1) represents the mean accuracy of the classifier on the complete data (or missingness=0)

Figure 1(a), 1(c) and 1(e) show that with the increase in missingness the TPR drops for all datasets using all the imputation methods except for Cancer data with EBMI. This indicates that as the missingness increases and imputation is performed, the $OSVM$ starts rejecting the positive samples. For missingness up to 40%, the TPR of proposed methods on Haberman data are better than MEI and EMI. For Cancer data, TPR of EBMI is higher than MEI but in Diabetes data it is lower and ABMI's acceptance rate is better at lower missingness (10% to 20%). With the proposed methods, TNR on Haberman and Diabetes data are higher than MEI (Figure 1(b) and 1(f)) but on Cancer data MEI's TNR is higher (Figure 1(d)) (with corresponding worst TPR). A joint TPR and TNR plot on Haberman and Diabetes data (see Figure 2) show that for moderate missingness, EBMI and ABMI can give good acceptance and rejection rate. We believe that tuning the proposed methods can do better in some situations. For missingness beyond 30% the TPR of all the imputation methods degrades and no imputation method can be statistically deemed better than other. The reasons for the



(a) Haberman's Survival (b) Haberman's Survival

(c) Breast Cancer (d) Breast Cancer

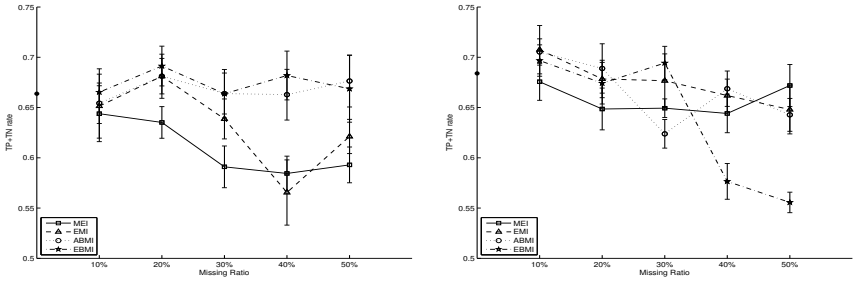
(e) Indiana Diabetes (f) Indiana Diabetes

Fig. 1. TPR and TNR at different degree of missingness (with standard error bars)

poor performance of *OSVM* at high missingness are small training datasets and dependability on parameters, also the increase in the ratio of imputed values in place of the actual values effect the flexibility of classification boundary.

6 Conclusions and Future Work

In this paper we propose two new approaches, ABMI and EBMI, to learn one-class classifiers in the presence of missing data and compare our result with common method of MEI and EMI. Our preliminary results on UCI datasets indicate that for moderate missingness the proposed methods of EBMI and ABMI give better TPR rates than MEI on all datasets and better than EMI on two datasets. The results are encouraging and show that EBMI and ABMI can help in building better one-class classifiers on data with missing values in comparison



(a) Haberman (b) Diabetes

Fig. 2. Combined TPR and TNR at different degree of missingness (with standard error bars)

to conventional methods. We observe that OSVM is very susceptible to choice of parameters and does not perform well on small training sets, therefore in future we would like to perform parameter tuning on large datasets to evaluate the efficiency of the proposed methods. We would also like to attempt other one-class classifiers such as nearest neighbours that can learn with less positive examples.

References

1. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 27:1–27:27 (2011)
2. Cohen, G., Hilario, M., Sax, H., Hugonnet, S., Pellegrini, C., Geissbuhler, A.: An application of one-class support vector machines to nosocomial infection detection. In: et al., M.F. (ed.) In Proc. of Medical Informatics. IOS Press (2004)
3. Frank, A., Asuncion, A.: UCI machine learning repository. http://archive.ics.uci.edu/ml (2010)
4. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: An update. In: in SIGKDD Explorations. 1, vol. 11 (2009)
5. Hempstalk, K., Frank, E., Witten, I.H.: One-class classification by combining density and class probability estimation. In: Proc European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. pp. 505–519. Antwerp, Belgium (2008)
6. Schafer, J.L.: Analysis of Incomplete Multivariate Data. Chapman and Hall (1997)
7. Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., Williamson, R.C.: Estimating the support of a high dimensional distribution. Tech. Rep. MSR-TR-99-87, Microsoft Research (1999)
8. Su, X., Khoshgoftaar, T.M., Greiner, R.: Using imputation techniques to help learn accurate classifiers. In: in 20th IEEE International Conference on Tools with Artificial Intelligence. pp. 437–444 (2008)
9. Tax, D.M.J., Duin, R.P.W.: Support vector data description. Machine Learning 54(1), 45–66 (2004)
10. Tax, D.: One Class Classification. Ph.D. thesis, Delft University of Technology (2001)