Special Issue Published in International Journal of Trend in Research and Development (IJTRD), ISSN: 2394-9333, www.ijtrd.com

Comparative Study of Naïve Bayes Classifier and K Nearest Neighbor in Imputation of Missing Values

[1]Priya.S and [2]Dr.Antony Selvadoss Thanamani,
[1]Research Scholar, Bharathiar University and Assistant Professor, Department of Computer Science, Government First Grade College, KGF, Karnataka, India
[2]Professor and Head, Research Department of Computer Science, NGM College, 90, Palghat Road, Pollachi, Coimbatore District, Tamilnadu, India

***Abstract*:** Predictive classification as a wide range of application in data mining. Most real data set have missing values which affects the accuracy of classifiers. This paper will investigate predictive performance of missing data using two classifier techniques naive Bayes classifier and Knn classifier. Among the two classifiers naive bayesian is least sensitive and provides a good accuracy to handle missing data but K nearest neighbour is the most sensitive to missing data. NB is one of the classifiers that handle missing data very well, it just excludes the attribute with missing data when computing posterior probability (i.e. probability of class given at a data point).

Keywords: *Classifiers, Naive Bayes Classifier And Knn Classifier, Predictive.*

I. INTRODUCTION

Data Mining (DM) is the process of discovering interesting knowledge from large amounts of data stored either in databases, data warehouse, or other information repositories [1]. There are a lot of functions of DM, such as description, association analysis, classification and prediction, clustering analysis etc. Among all, classification and prediction are widely used in many fields. However, in real-world datasets, there are many problems in data quality such as incompleteness, redundancy, inconsistency, noise data etc. All these serious data quality problems affect the performance of DM algorithms [2].

Missing data is a quite common problem of data quality in real-life datasets. Then, what is the reaction of classifier to missing data? For the large amount of missing data in real-word datasets, several methods have been proposed. For example, case deletion, mean imputation and model prediction. The basic idea of model prediction is using prediction model, built on known data, to predict and fill in the missing data. The performance of the missing treatment relies on the prediction model. Then, which classifier will be suitable for handling missing data?so we have chosen two predictive classifiers and compared to find which classifier is least sensitive and most sensitive to missing data.

II. CLASSIFICATION AND PREDICTION

A. Classifier

Classification means constructing a classifying function or model from the known data. Such function or model can also be called 'classifier', which can classify the records in database into given classes, thus can predict the unknown variables under some given conditions [3]. Classifiers differ greatly in prediction accuracy, training time and number of leaves (Decision Trees). There is not a classifier which performs best in all aspects [4]. Prediction accuracy of classifiers can be affected by the factors as follows [3].

B. Number of records in training subset.

Classifier needs to learn from training set. Therefore, larger training set makes the classifier more reliable. But the training time is also become longer.

C.Data quality

Problems such as noise data, missing data, data inconsistency etc. bring a lot of wrong information which will lead to wrong classify. It is impossible to build a convictive classifier with incompleteness or wrong data.

D.Attribute quality

Attributes provide information for classifying. The prediction accuracy can be improved by including more attributes. However, more attributes means calculating more attribute combinations and more training time. It is essential to choose attributes which are valuable for classification.

E.Characteristics of the records to be predicted

If Characteristics of the records to be predicted are different from records in training set, it may lead to high incorrect rate.

III. MISSING TREATMENT METHODS USING CLASSIFIER

In the popular methods for missing data handling, using classifier to predict and fill in missing data is a set of fast developing methods. Among the so many classifiers, which one is suitable and how to use is still in the research. Methods have been used are introduced as follow:

A.K-Nearest Neighbor Imputation (KNN)

This method uses k-nearest neighbor algorithms to estimate and replace missing data. The main advantages of this method are that:

- It can estimate both qualitative attributes (the most frequent value among the k nearest neighbors) and quantitative attributes (the mean of the k nearest neighbors).
- It is not necessary to build a predictive model for each attribute with missing data, even does not build visible models. Efficiency is the biggest trouble for this method. While the k-nearest neighbor algorithms look for the most similar instances, the whole dataset should be searched. However, the dataset is usually very huge for searching. On the other hand, how to select the value "k" and the measure of similar will impact the result greatly.

B.Bayesian Iteration Imputation (BII)

Naive Bayesian Classifier is a popular classifier, not only for its good performance, but also for its simple form. It is not sensitive to missing data and the efficiency of calculation is very high. Bayesian Iteration Imputation uses Naive Bayesian

Classifier to impute the missing data. It is consisted of two phases:

- Decide the order of the attribute to be treated according to some measurements such as information gain, missing rate, weighted index, etc.;
- Using the Naive Bayesian Classifier to estimate missing data. It is an iterative and repeating process. The algorithms replace missing data in the first attribute defined in phase one, and then turn to the next attribute on the base of those attributes which have be filled in. Generally, it is not necessary to replace all the missing data (usually 3~4 attributes) and the times for iterative can be reduced [7]

III. EXPERIMENT AND ANALYSIS

A.Sensitivity Analysis

Sensitivity analysis (SA) is to study the impacts of one or more input variables on the outputs of a model, that is, the sensitivity of the model to one parameter or a combination of parameters [8]. If a tiny change of an input leads to great changes of the output, the model is highly sensitive to that input. SA can help to identify the decisive input parameter of the model [8]. In our experiments, the proportion of missing data in the datasets is the parameter which affects the results of the classification models. The effect of missing data on the prediction accuracy will be investigated through the tiny changing of the missing rate

B.Design of Experiments

This paper has selected 2 classifiers to study the influence of missing data to classifiers, that is, Naive Bayesian classifier (NB), K-Nearest Neighbours classifier (KNN) datasets were collected from UCI repository and, are used in the experiments, as shown in Table 1.

Three indexes are used to evaluate the missing influence on classifier: Prediction accuracy (Pa), Prediction profit (Pp) [10] and Prediction losing (Pl), which are defined as follows.

$$Pa = \frac{number\ of\ correctly\ predicted\ record}{number\ of\ all\ records} \times 100\%$$

$$Pp = \frac{Pa - MD}{MD} \times 100\%$$

$$Pi = \frac{AC - Pa\ under\ certain\ missing\ rate}{AC} \times 100\%$$

AC is the prediction accuracy without missing data.

MD is the proportion of the class in the dataset which having the greatest number of records.

Without any prediction model, if all the records are classified into that class, MD will be the prediction accuracy of the dataset. Prediction profit is proposed by Peng Liu, Elia El-Darzi et al. (2004) to evaluate the performance of the classifier [9].

Then, a given percentage, 10%, 20%, … , 80% of missing data is artificially inserted into the training subsets at completely random. Finally, the two classification algorithms mentioned above are applied into the training subset to build up classifiers, and, these classifiers are used to classify instances in testing subset to investigate the classification accuracy. The average prediction accuracy of the 2 classifiers under different missing percentages is displayed from Figure 1 . Limited by the space, only part of the results is showed in this paper.

C. Results and Analysis

The figure interprets that with the increase in missing rate, the prediction accuracies of all the classifiers have an obvious trend of decrease. In general, when the proportion of missing data in the dataset is less than 10%, they have little adverse impact on the classifiers. If the missing rate is between 10% and 20%, the impact should not be neglected. The average Prediction losing rises to 4.63%. However, the adverse impact can be reduced significantly by some simple methods, such as replacing missing data by an approximation. If the missing rate exceeds 20%, there is an obvious decrease in the prediction accuracy and the missing data should be handled with high cautiousness

Table 1: Average Prediction Losing Of Classifier(%)

Missing data	NB	Knn	Average
0	0	0	0
10%	0.08	4.29	2.19
20%	0.59	13.22	6.91
30%	0.47	16.45	8.46
50%	1.24	22.45	11.85
70%	2.7	24.95	13.83
80%	9.73	29.03	19.38

Appropriate methods should be chosen to eliminate the adverse impact of the missing data and optimize the performance of classifiers. In the real world, there are great quantities of missing data in databases and, usually, the proportion of missing data exceeds 20%. However, if the proportion of missing data exceeds 50%, the average prediction losing rises to more than 10%.

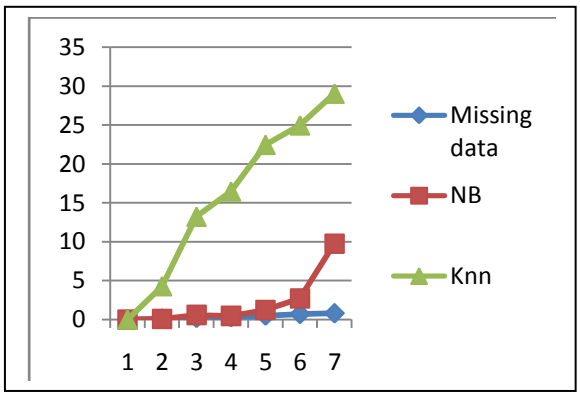

Figure 1: Average losing rate of classifiers

Obviously, the loss of prediction accuracy caused by missing data is quite huge.with the increase in the missing rate, the losing of prediction is rising with accelerated paces. That is to say, with the increase of quantity of the missing data, the little raising of the missing rate will result in a larger and larger decrease in the prediction accuracy.

The impact of missing data depends on classifiers. Among the classifiers, Naive Bayesian classifier is the least sensitive to missing data,. And K-Nearest Neighbor susceptible to missing data most,. With the increasing in the missing rate, the prediction accuracy of Naive Bayesian classifier is almost the same as that with no missing data. Only when more than 70% data are missing, the prediction accuracy drops obviously. the prediction losing of Naive classifier is always the lowest and that of the KNN is always the highest.

Further more, the pace of KNN is also very fast. When missing rate exceeds 10%, there is an obvious and sharp increase in prediction losing. The prediction profit of Naive Bayesian classifier drops with the lowest and smoothest pace. And the K-Nearest Neighbor has the sharpest trend. In summary, Naive Bayesian classifier is the least sensitive to missing data. Even if training subset has a lot of missing data, Naive Bayesian classifier can still make full use of the existing data and operate effectively.

Among all the classifiers, though the classification accuracy (with no data missing) of Naive Bayesian classifier isn't the highest, it is the most adaptive to missing data. The Naive Bayesian classifier is recommended for the datasets with great quantities of missing data. While, selecting Naive Bayesian classifier to deal with missing data can also get a better solution.

CONCLUSION

Missing data may reduce the accuracy of prediction models. This paper mainly studies the impact of missing data to classification algorithms. The sensitivity of classifiers to missing data is analyzed . The results showed that, with the increasing of the missing rate, the classification accuracies of all the classification algorithms have an obvious trend of decrease. If the proportion of missing data exceeds 20%, there is an obvious decrease in the accuracy of prediction. Methods for missing data treatment should be chosen cautiously to eliminate the negative impact on the classification accuracy and optimize the performance of classifiers. Among thetwo classifiers, the Naive Bayesian classifier is the least sensitive to missing data and K-Nearest Neighbour is the most sensitive. In conclusion, for datasets suffered with missing data, we prefer to use Naive Bayesian classifier to deal with missing.

Acknowledgment

I would like to thank my guide for his valuable suggestions and tips to write this paper.

References

[1] Han J., Kamber M. Data Mining Concepts and Technique. Morgan Kaufmann Publishers, 2000

[2] Cios K.J., Kurgan L. Trends in Data Mining and Knowledge Discovery. In N.R. Pal, L.C. Jain

[3] Tian Jinlan, Li Ben. Tools for Data Mining: Classifiers. Department of Computer Science, Tsinghua University. Computer World, 1999, 20th Periodical

[4] Tjen Sienlim, Wei Yinloh, Yu ShanShih. A Comparison of Prediction Accuracy, Complexity and Training Time of Thirty-three Old and New Classification Algorithms. Machine Learning, 40, 2000, 203~229, Kluwer Academic Publishers, Boston

[5] Marvin L., John F. Kros. Chapter VII-The Impact of Missing Data on Data Mining, Data Mining: Opportunities and Challenges, Idea Group Publishing, 2003

[6] Quinlan J. R., C4.5 Programs for Machine Learning. Morgan Kaufmann, CA, 1988.

[7] Liu P., Lei L. and Zhang X.F., A Comparison Study of Missing Value Processing Methods, Computer Science, 31(10):155-156 & 174, 2004.

[8] J. T. Yao. Sensitivity Analysis for Data Mining. Proceedings of The 22nd International Conference of NAFIPS, July 24-26, Chicago, U SA, 2003, 272~277

[9] Peng Liu, Elia El-Darzi et al. Comparative analysis of Data Mining Algorithms for Predicting Inpatient Length of Stay. Proceedings of the Eighth Pacific-Asia Conference on Information SystemsJuly 2004

[10] Peng Liu, Lei Lei, Naijun Wu, A Quantitative Study of the Effect of Missing Data in Classifiers

[11] S. Kanchana, Dr. Antony Selvadoss Thanamani, "Classification of Efficient Imputation Meth od for Analyzing Missing values",International Journal of Computer Trends and Technology, Volume-12 Part-I, P-ISSN: 2349-0829.

[12] S. Kanchana, Dr. Antony Selvadoss Thanamani, "Multiple Imputation of Missing Data Using Efficient Machine Learning Approach", International Journal of Applied Engineering Research, ISSN 0973-4562 Volume 10, Number 1 (2015) pp.1473-1482.