

# Discretization and Imputation Techniques for Quantitative Data Mining

Nuntawut Kaoungku\*, Phatcharawan Chinthaisong, Kittisak Kerdprasop, and Nittaya Kerdprasop

**Abstract**—Association rule mining from numerical datasets has been known inefficient because the number of discovered rules is superfluous and sometimes the induced rules are inapplicable. In this paper, we propose the discretization technique based on the Chi2 algorithm to categorize numeric values. We also handle missing values in the dataset with statistical methods. The discovered association rules are then evaluated with the four measurement metrics, that is, confidence, support, lift, and coverage. The dataset imputed with various missing value handling techniques has also been evaluated with the tree-based data classification method to assess predictive accuracy.

**Index Terms**—Association rule analysis, Data mining, Discretization, Missing value imputation

## I. INTRODUCTION

Current adoption of data mining technology can be seen in various fields such as economics, education, engineering, life science, medicine, and many more. The models automatically learned from data can facilitate future event prediction, as well as can explain current relations. Models built from datasets with some missing values can, however, cause error in the prediction. Efficient predictive model building, thus, requires the imputation of missing data.

In this research, we comparatively perform three schemes of missing value handling. These schemes are (1) removing record that show missing data, (2) imputation with average attribute value, and (3) imputation with the most correlated value. After data imputation, we investigate these missing value handling scheme through the decision tree induction technique. The decision tree induction is a data classification technique. This data mining task aims at inducing a model in a form of decision tree. This kind of model can be used to predict class of data that may occur in the future. We can call this kind of task as predictive data mining.

Manuscript received December 8, 2012; revised January 10, 2013. This work was supported in part by grant from Suranaree University of Technology through the funding of Data Engineering Research Unit.

N. Kaoungku is a master student with the School of Computer Engineering, Suranaree University of Technology, Nakhon Ratchasima 30000, Thailand (e-mail: b5111299@gmail.com).

P. Chinthaisong is a master student with the School of Computer Engineering, Suranaree University of Technology, Nakhon Ratchasima 30000, Thailand (e-mail: killuakaara@gmail.com).

K. Kerdprasop is an associate professor with the School of Computer Engineering, Suranaree University of Technology, Nakhon Ratchasima 30000, Thailand.

N. Kerdprasop is an associate professor with the School of Computer Engineering, Suranaree University of Technology, Nakhon Ratchasima 30000, Thailand.

Another kind of learning task is explanatory data mining. The purpose of such task is to explain existing relationships among various data attributes. Association rule mining is a type of data mining that will find the association among data objects and create a set of rules to model relationships. To perform association rule mining, data to be mined have to be categorical. Discretization of numerical values is thus an essential data preparation step for association rule mining.

In this paper, we propose a framework of missing value imputation and numerical data discretization as two major preparation steps for classification and association mining tasks. We also present evaluation results of classification and association rule mining using afferent benchmarks.

This research solves the problem by preparing dataset appropriately before association and classification of discretization methods for numeric in association rule and predicts of data missing that is closest to the most possible value.

## II. PRELIMINARIES AND RELATED WORK

Data can be in a variety of formats. For example, numeric data, nominal data, and a mixed type of numeric and nominal data. But data mining in some categories is not possible. For instance, to find the association rules from dataset with numeric attributes is impossible for some algorithms. Therefore, methods for managing numeric attribute data is essential. The common method to handle categorizing numeric values is discretization. Many current researches on how to divide the discretization in a variety of ways. For example in [3], Chi 2 algorithm was used as discretization method for handling numeric attributes. The discretization methods for numeric attributes in association rule analysis [7] had been applied with R language [3]. The algorithms used to discretization are, The Chi2 algorithm formed by  $\chi^2$  they are often used in statistics and discretization methods for numeric attributes.

The predictive value of the data missing is another important problem. We comparatively study the value of data missing technique, lost out in praise. The average value in that column if data missing is disrupted data or skew data. We are used median value. If the data in column aren't numeric. We used value that appears most often in the column. And how to use the value of the correlation between a column that has a data missing value with another column that is associated with the most. Other research also has using Rough Set theory [5] Include is used to determine the association between each column is set to create a rule that allows predicting. Datasets were used in this study was a series of patients that most data is dispersed across numeric data. With the numerical data will be grouped into ranges

(Discretized) are so easy to do the research to find the value of the data missing. And Jianhua's research [1] propose a technique to fill up the missing data by using Rough Sets theoretical and add a technique to compare the 3 methods: how to cut data rows that contain data values that are missing out, and data mining. How to select values that are come to missing data from data that contains values that appear most frequently in the column, and how to convert the entire datasets as a Discernibility matrix and create a rule for predicting the missing value. By using a series of six sets of data to compare efficiency, how to find the value of the data that is missing all three methods and data sets through the technique of value for the information that is missing, and the range, and then create a decision tree to test the data prepared for the test of efficiency technique to predict the best. The rules for an association with the four measurements the effectiveness and value of the gauge is decision of each of the algorithm for discretization methods for numeric attributes.

### III. METHODOLOGY

#### A. Framework

This research proposed discretization and imputation techniques for quantitative data mining. Figure 1 shows conceptual framework of the research. First, the missing value imputation has been applied. Second, the discretization has been performed on numeric attributes. Third, apply the association rule mining. Finally, the benchmarks on association rule mining result are to be evaluated.

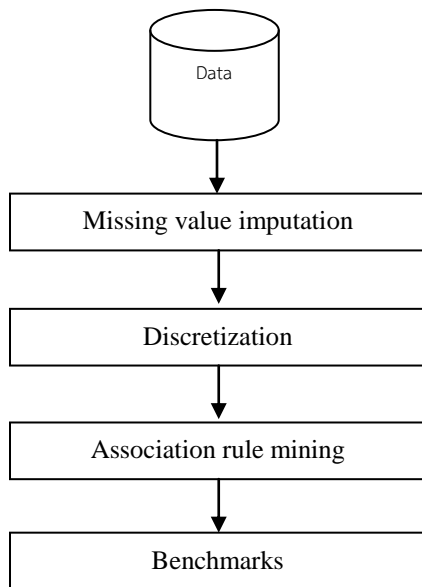


Fig. 1 Conceptual framework of the research

#### B. Predict the missing value

Techniques to handle missing values in our study are as follows:

- 1) Remove record that some values are missing.
- 2) Impute missing values with the average value of the attribute, if the data is normally distributed.
- 3) Use the correlation of column with missing values to another column, and impute with that column's value.

#### C. Algorithm Chi2

Chi2 algorithm that is based on the  $\chi^2$  statistics was used to perform discretization the numerical data [4]. The computation for  $\chi^2$  is as follows.

$$\chi^2 = \sum_{i=1}^2 \sum_{j=1}^k (A_{ij} - E_{ij})^2 / E_{ij} \quad (1)$$

where:

$k$  = number of classes,

$A_{ij}$  = number of patterns in the  $i$ th interval,  $j$ th class,

$E_{ij}$  = expected frequency of  $A_{ij} = R_i * C_j / N$

$R_i$  = number of patterns in the  $i$ th interval =  $\sum_{j=1}^k A_{ij}$

$C_j$  = number of patterns in the  $j$ th class =  $\sum_{i=1}^2 A_{ij}$

$N$  = total number of patterns =  $\sum_{i=1}^2 R_i$

The Chi2 algorithm is divided into two parts. The first part starts with a high level of significance, that is 0.5 (sigLevel = 0.5), for all numerical data. After that, it will sort all the numbers continuously.

Part 2 will be on the sideline of the first start of sigLevel0 as set forth in Part 1, then the consistency check after performing an individual attribute the inconsistency rate cannot exceed the assigned sigLevel [i] for inclusion attributes in the next round. This process stops when there is no value left in the attribute.

#### D. Benchmarks

The benchmarks in this study are the four measurements: support, confidence, lift, and coverage.

- 1) Support is the frequency of the event occurring, Compute support of equation (2).

$$Support(A \rightarrow B) = P(A \wedge B) \quad (2)$$

- 2) Confidence is the frequency of the incident with other events occurring together, Compute confidence of equation (3).

$$Confidence(A \rightarrow B) = Supp(A \rightarrow B) / Supp(A) \quad (3)$$

- 3) Lift is the influence of the association rule mining, Compute lift of equation (4).

$$Lift(A \rightarrow B) = Conf(A \rightarrow B) / Supp(A) \quad (4)$$

- 4) Coverage is considered the frequency of the association rules mining, Compute coverage of equation (5).

$$Coverage(A \rightarrow B) = Supp(A) = P(A) \quad (5)$$

IV. EXPERIMENTAL RESULTS

This research experimentation used Hepatitis dataset from the UCI Machine Learning Repository [7]. Hepatitis dataset has 20 attributes and 103 data instances.

For discretization and imputation techniques for quantitative data mining, we used classification and association mining for experimental result assessment. Table 1 and Fig.2 show comparative accuracy of classification both algorithm missing value and missing value + discretization of three models. Model 1 is removing records that contain missing values. Model 2 is missing value imputation with the attribute mean. Model 3 is missing value imputation with correlated value.

TABLE 1  
COMPARATIVE RESULTS OF CLASSIFICATION ACCURACY

Algorithm	Model 1	Model 2	Model 3
Missing value	65.95%	74.46%	80.85%
Missing value + Discretize	85.13%	89.36%	87.23%

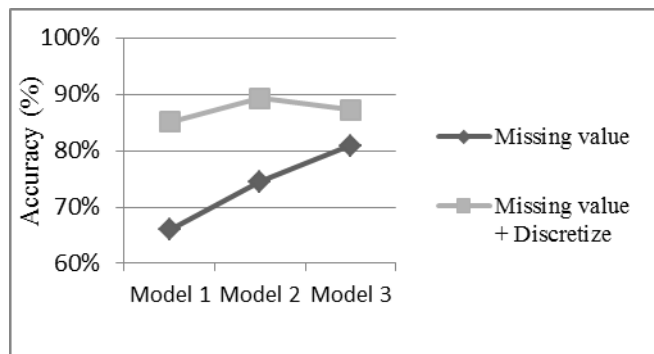


Fig. 2 Accuracy comparison for both algorithms: missing value and missing value + discretization

Table 2 show comparative results of association rule mining using the average of support, confidence, lift, and coverage values to measure performance.

TABLE 2  
COMPARATIVE RESULTS OF ASSOCIATION RULE MINING

Models	The average of support	The average of confidence	The average of lift	The average of coverage
Model 1	60.99%	97.66%	103.63%	62.65%
Model 2	62.56%	98.37%	102.94%	63.78%
Model 3	62.02%	98.33%	103.07%	63.27%

Fig. 3 compares the average of confidence and lift for three models. It can be seen from the result that model 3 is the highest compared to the other models.

Fig. 4 compares the average of support and coverage values for three models. It can be seen from the result that model 2 is the highest compared to the other models.

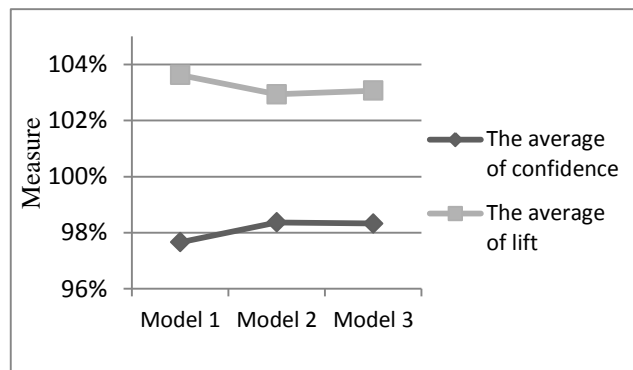


Fig. 3 Comparative the average of confidence and lift both three models

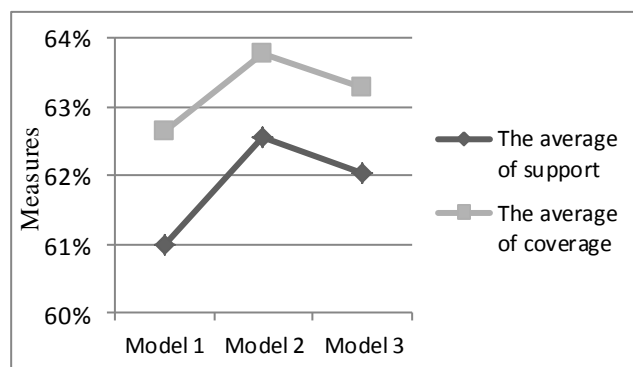


Fig. 4 Comparative the average of support and coverage both three models

V. CONCLUSION

This research aims to study discretization and imputation techniques for quantitative data mining. The results show that the best model of classification is model 2 that used missing value imputation with the average value if the data is normally distributed and used chi2 for discretization. The results also show that the best model of association rule mining is model 2. Therefore, it can be concluded that the model 2 that imputes missing values by attributes means gives the best result.

REFERENCES

- [1] Jianhua Dai, Qing Xu, and Wentao Wang (2011). "A Comparative Study on Strategies of Rule Induction for Incomplete Data Based on Rough Set Approach," *International Journal of Advancements in Computing Technology*, vol 3, no. 3, pp.176-183
- [2] Kittisak Kerdprasop (2012). "Data Mining Methodology and Development," Retrieved November 1, 2012, from <https://sites.google.com/site/kittisakthailand55/home/datamining2-55>
- [3] Huan Liu, Farhad Hussain, Chew Lim Tan, and Manoranjan Dash (2002). "Discretization: An Enabling Technique," *Data Mining and Knowledge Discovery*, vol.6, no.4, pp. 393-423
- [4] Huan Liu and Rudy Setiono (1995). "Chi2: Feature Selection and Discretization of Numeric Attributes," *Proceedings of the Seventh International Conference on Tools with Artificial Intelligence*, pp. 388-391
- [5] Fulufhelo V. Nelwamondo and Tshilidzi Marwala. (2007). "Rough Sets Computations to Impute Missing Data," CoRRabs/0704.3635
- [6] UC Irvine Machine Learning Repository, (1988) Hepatitis Data Set. Retrieved October 5, 2012 from <http://archive.ics.uci.edu/ml/datasets/Hepatitis>
- [7] Mohammed J. Zaki (2002). "Scalable Algorithms for Association Mining," *IEEE Transaction on Knowledge and Data Engineering*, vol.12, no.3, pp. 372-390

APPENDIX

Source code in R language to perform missing value imputation and discretization is presented as follows:

```
# Missing value imputation
hepatitis<-read.csv("hepatitis.csv", fill = TRUE)
hepatitis <- hepatitis [-c(62,199) , ]
predict1<- function(dataM){
  cutMissing <-na.omit(dataM)
  return(cutMissing)
}

predict2<- function(dataM,colM,more=F){
  if (more){
    dataM[is.na(dataM[[colM]]),colM]<-
    mean(dataM[[colM]],na.rm=T)
  }else{
    dataM[is.na(dataM[[colM]]),colM]<-
    median(dataM[[colM]],na.rm=T)
  }
  return(dataM)
}

lookCor<- function(crn){
  gg<-cor(crn,use='complete.obs')
  gp<-symnum(gg)
  return(gp)
}

creXY<- function(colM,dataM,NN){
  mM<-lm(colM,data=dataM)$coefficients[NN]
  mN<-mM[1][[1]]
  return(mN)
}

inputf<- function(oP){
  if ( is.na(oP) ) return(NA)
  else return ( (oP+(-mY))/mX )
}

cor.input<- function(colA,colB,dataM){
  dataM[ is.na ( dataM[[colA]] ),colA ] <-
  sapply ( dataM[ is.na (dataM[[colA]])],colB),inputf)
  return(dataM)
}

dataset1<-predict1(hepatitis)
dataset2<-predict2(hepatitis,"Chla",T)
dataset2<-predict2(dataset2,"Cl",T)
dataset2<-predict2(dataset2,"PO4",F)

mX<-creXY(oPO4~PO4,hepatitis,2)
mY<-creXY(oPO4~PO4,hepatitis,1)
dataset3<-predict3("PO4","oPO4",hepatitis)
dataset3<-predict3("Chla","oPO4",dataset3)

library(rpart)

rt.a1<-rpart(a1~.,data=dataset1[,1:12])
plot(rt.a1,uniform=T,branch=1, margin=0.1, cex=0.9)
text(rt.a1,cex=0.75)

rt.a2<-rpart(a1~.,data=dataset2[,1:12])
plot(rt.a2,uniform=T,branch=1, margin=0.1, cex=0.9)
text(rt.a2,cex=0.75)
```

```
rt.a3<-rpart(a1~.,data=dataset3[,1:12])
plot(rt.a3,uniform=T,branch=1, margin=0.1, cex=0.9)
text(rt.a3,cex=0.75)

testPred <- predict(rt.a1, newdata = test.hepatitis)
print(testPred)
table(testPred, test.hepatitis$a1)

#Discretization
hepatitisM<-read.csv("hepatitis.csv", fill = TRUE)
new.dataset<-chi2(hepatitisM,0.5,0.05)$Disc.data

#Association rules mining
rules <- apriori(new.dataset, parameter= list(supp=0.5,
conf=0.8))

# Benchmarks
quality(rules) <- cbind(quality(rules), coverage =
interestMeasure(rules, method = "coverage", tr))
WRITE(rules, file = "data_disc.csv", quote=TRUE, sep =
",", col.names = NA)
```