Data Preprocessing: Discretization and Imputation

A review of recent research output

Md Zahidul Islam, Md Geaur Rahman & Michael Furner

Charles Sturt University: School of Computing and Mathematics

Contact Information:
School of Computing and Mathematics
Charles Sturt University
Panorama Avenue, Bathurst, NSW

Phone: (02) 6338 4214
Email: zislam@csu.edu.au

Introduction

Organizations use Data Mining and Knowledge Discovery algorithms for making better decisions. A data mining algorithm extracts interesting patterns (such as logic rules and clusters) that could otherwise be extremely difficult for us to extract [15, 7]. Data preprocessing and cleansing play a vital role in data mining by ensuring good quality of data. Data-cleansing tasks include imputation of missing values, identification of outliers, and identification and correction of noisy data [5]. Another key preprocessing technique is discretization - the conversion of numerical attributes into categorical attributes [11]. This can be performed to allow the use of data mining techniques that require categorical attributes, increase the performance of data-mining techniques, and to convert a numerical attribute to categorical to be used in a classifier.

The data mining research group has published several papers exploring these concepts. This has lead to the development of a variety of new algorithms, including LFD [11] for discretization, and FIMUS [10], SiMI [9], FEMI [12] and MultiSiMI [2] for missing value imputation. This poster will discuss and provide a brief overview of these two topics, and provide a brief overview of each of the algorithms developed by the data mining research group.

Discretization

Many data mining algorithms, for example Naive Bayes, can only deal with categorical attributes and are unable to handle numerical attributes [18]. Some other data mining algorithms can handle numerical attributes. However, often the efficiency and effectiveness of a data mining algorithm increases when it makes use of a discretization algorithm [3, 18]. Discretization is also considered to be an important part of data preprocessing and cleansing that is likely to improve the quality of the results obtained by various data mining algorithms [5].

The LFD algorithm [11] has been developed for the express purpose of discretization. The FIMUS missing value imputation algorithm [10] includes a novel discretization technique, which will also be discussed here separately to its imputation context.

LFD

The Low Frequency Discretizer algorithm [11] is designed to target a fundamental flaw in many existing discretization techniques, namely that choosing an interval boundary that is in a region of the attribute space that has many occurances in the dataset causes a large amount of information about the similarity between ”close” values to be lost. In order to combat this, LFD automatically selects splitting points between categories that are in *Low Frequency* regions of the attribute space.

This process takes four steps:

1. Copy a full dataset D_F into D'_F.
2. Rank the numerical attributes of D'_F based on correlation ratio η.
3. Discretize all numerical attributes of D'_F.
4. Return the discretized dataset D'_F.

The numerical attributes are ranked since the discretization process happens one attribute at a time in order of ranking. Since the discretization process relies on the other attributes, it is crucial that the initial discretizations are of high quality. The attributes are discretized by finding high quality cut points automatically. The cut points need to be between attribute values with lower than average frequencies. The different possible cut points for an attribute are considered and the potential categories are voted on by considering the attribute-interdependency and uncertainty between the potential categories and the other categorical attributes in the dataset. The best splitting point is selected from the votes. This process is repeated until the vote no longer improves on a new iteration. This is then repeated for the next ranked numerical attribute.

LFD was shown to improve the quality of imputation techniques, classification techniques and noise detection techniques over the results obtained by other discretization algorithms including PD, FFD, EWD, EFD, and CAIM [18, 6, 17, 11].

FIMUS

In order to discretize an attribute, FIMUS splits the attribute into $N_c = \sqrt{s}$ categories where $s = max(A_i) - min(A_i) + 1$. This allows FIMUS to automatically determine the number of categories, a feature not found in some discretization techniques. The algorithm then divides the values into N_c categories, each containing values in intervals of the size $\frac{s}{N_c}$. Due to this process there may be empty categories, but these are ignored. Figure 1 shows discretization on a toy dataset performed by FIMUS' technique. This technique was shown to perform better than the PD [18] and FFD [18] discretization algorithms for the purposes of imputation using FIMUS [10].

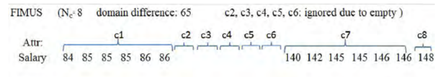


Figure 1: FIMUS discretization

Missing Value Imputation

Natural data sets often have missing values in them. The imputation of missing values as accurately as possible is an important data preprocessing task. Use of poor-quality data, having missing and incorrect values, can result in an inaccurate and non-sensible conclusion, making the whole process of data collection and analysis useless for the users [5, 8].

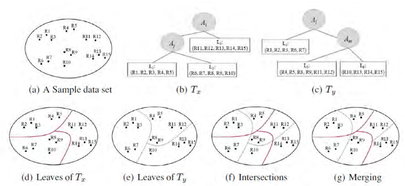
DMI [9], SiMI [9], MultiSiMI [2] are all imputation techniques that rely on EMI [14] and decision trees. FEMI [12] uses a fuzzy version of EMI and fuzzy c-means clustering in order to find an imputation. FIMUS [10] uses co-appearance, correlation and similarity analysis for imputations.

DMI, SiMI and MultiSiMI

A decision tree divides a data set into a number of leaves having sets of mutually exclusive records. A decision forest builds a number of decision trees. While these sets of records are usually used to classify records as part of a classification problem, they have been shown to contain sets of highly correlated and similar records [9]. The Expectation Maximisation Imputation algorithm (EMI) [14] relies on correlation and similarity, and is more effective when the records it is used on are highly correlated and similar.

DMI [9] was designed to make use of the sets created by decision trees to increase imputation accuracy. It does this by building a decision tree on the clean records of a dataset, and assigning the missing value records to a leaf. EMI is then performed on each of the subsets separately to impute numerical attributes, and categorical attributes use a mode imputation within the subset. This was shown to be a marked improvement over EMI [14], as well as several other imputation algorithms [9].

Figure 2: SiMI's tree intersections



SiMI [9] improved upon DMI's imputation accuracy even further, by replacing the role of a decision tree with a decision forest. In order to get very high correlation and similarity subsets which are mutually exclusive, SiMI finds the intersections between leaves from the different trees of a decision forest. This concept is illustrated in figure 2. SiMI [9] was expanded once again with the development of MultiSiMI [2], which instead of finding intersections to get mutually exclusive subsets of records simply performs EMI on every subset found by the decision trees in the forest. This gives n imputations for each record, where n is the number of trees in the forest. The results are then averaged - part of a concept known as multiple imputation [13].

FEMI

FEMI [12] is an advanced technique that uses a fuzzified version of EMI known as *FuzzyEM* in order to use the fuzzy c-means clustering algorithm to find subsets of records. The process consists of 6 steps:

1. Copy a full data set D_F into D_N and normalize all numerical attributes of D_N within a range between 0 and 1.
2. Divide the data set D_N into two sub data sets D_C (having only records without missing values) and D_I (having only records with missing values).
3. Find membership degrees of all records of D_C and D_I with all clusters.
4. Apply the *FuzzyEM* method to impute numerical missing values using all clusters.
5. Find the combined imputed value of a numerical attribute. Find the imputed value of a categorical attribute.
6. Combine records to form a completed data set (D'_F) without any missing values.

In a fuzzy clustering algorithm, every record has a degree of membership to every cluster. *FuzzyEM* [12] is used on each cluster, and makes use of the membership degrees of each record to impute using means and covariance specific to each cluster (using membership degrees as weights) for each cluster. This results in c imputations (where c is the number of clusters). These are combined, again using membership degrees, to find a single imputation result. Categorical attributes are also imputed using the membership degrees from the clustering process in order to affect where the imputation comes from.

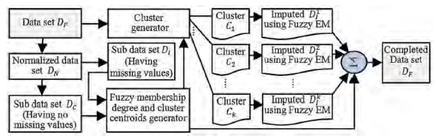


Figure 3: Block diagram of FEMI's process.

Figure 3 provides an overview of the technique. Like the other algorithms, FEMI has shown to significantly improve the results of older, more traditional methods of missing value imputation [12].

FIMUS

FIMUS [10] is an algorithm for imputing missing values that uses a combination of similarity, correlation, and co-appearance of values in records to determine a vote for the imputed value.

It does this as follows:

1. Initialize a missing matrix B from the input data set D^o.
2. Generalize all numerical attributes of D^o.
3. Generate co-appearance matrix (C), normalized similarity matrix S_j for attribute $A_j; \forall A_j \in A$, and correlation matrix K.
4. Impute missing values.
5. Repeat the imputation process (steps 2-4) until there is a change between two consecutive iterations.
6. Return a completed data set (D^o) without any missing values.

The similarity matrix describes the similarity between different values in the same attribute, allowing a measure of “closeness” to be used in the imputation calculations [4]. FIMUS works on one attribute at a time, imputing by considering two votes for each missing value for each non-missing attribute in the same record. These are

$$V_x^{N.p} = \frac{C_{xl}}{f_l}$$

$$V_x^{S.p} = \sum_{\forall a \in A_p} \frac{C_{xl}}{f_l} \times S_{la}$$

Where C_{xl} is the is the number of coappearances between the value l in non-missing attribute A_p and a candidate value x, f_l is the frequency of l in the whole dataset for A_p, and S_{la} is the similarity between values l and a in A_p. These are combined as $V_x^p = \{V_x^{N.p} \times \lambda + V_x^{S.p} \times (1 - \lambda)\} \times k_{jp}$. This is the vote in favour of the value x for the missing attribute considering attribute p. This is repeated for each attribute value in the missing attribute and each other available attribute in the record. Finally, the vote for a particular value for the missing attribute is expressed as $V_x^t = \sum_{\forall A_p \in A\ A_j} V_x^p$. The attribute value with the highest vote will be the result of the imputation.

Because the algorithm works on a discretized dataset, missing numerical attributes are initially imputed into a discrete category, and then FIMUS is repeated on a subset of the dataset found by selecting all records with the same category in the imputed attribute and replacing all of the generalised values with their original values in the imputed attribute. Each actual numerical value is then used as a category for a repeat of FIMUS which gives the numerical imputation. FIMUS was shown to provide significantly better results than DMI [12], SVR [16], EMI [14] and IBLLS [1].

References

[1] Kin-On Cheng, Ngai-Fong Law, and Wan-Chi Siu. Iterative bicluster-based least square framework for estimation of missing values in microarray gene expression data. *Pattern recognition*, 45(4):1281–1289, 2012.

[2] Michael Furner and Md Zahidul Islam. Multiple imputation on partitioned datasets. In *Proceedings of the 13th Australasian Data Mining Conference*. Australian Computer Society, Inc., 2015.

[3] Sergio Garcia, Julián Luengo, José Antonio Sáez, Victor Lopez, and Francisco Herrera. A survey of discretization techniques: taxonomy and empirical analysis in supervised learning. *Knowledge and Data Engineering, IEEE Transactions on*, 25(4):734–750, 2013.

[4] Helen Giggins and Ljiljana Brankovic. Vicus: a noise addition technique for categorical data. In *Proceedings of the Tenth Australasian Data Mining Conference-Volume 134*, pages 139–148. Australian Computer Society, Inc., 2012.

[5] Jiawei Han, Micheline Kamber, and Jian Pei. *Data mining: concepts and techniques: concepts and techniques*. Elsevier, 2011.

[6] Lukasz Kurgan, Krzysztof J Cios, et al. Caim discretization algorithm. *Knowledge and Data Engineering, IEEE Transactions on*, 16(2):145–153, 2004.

[7] Dorian Pyle. *Data preparation for data mining*, volume 1. Morgan Kaufmann, 1999.

[8] Md Geaur Rahman and Md Zahidul Islam. Data quality improvement by imputation of missing values. In *International conference on computer science and information technology (CSIT-2013), Yogyakarta, Indonesia*, pages 82–88, 2013.

[9] Md Geaur Rahman and Md Zahidul Islam. Missing value imputation using decision trees and decision forests by splitting and merging records: two novel techniques. *Knowledge-Based Systems*, 53:51–65, 2013.

[10] Md Geaur Rahman and Md Zahidul Islam. Fimus: A framework for imputing missing values using co-appearance, correlation and similarity analysis. *Knowledge-Based Systems*, 56:311–327, 2014.

[11] Md Geaur Rahman and Md Zahidul Islam. Discretization of continuous attributes through low frequency numerical values and attribute interdependency. *Expert Systems with Applications*, 2015.

[12] Md Geaur Rahman and Md Zahidul Islam. Missing value imputation using a fuzzy clustering-based em approach. *Knowledge and Information Systems*, pages 1–34, 2015.

[13] Joseph L Schafer and John W Graham. Missing data: our view of the state of the art. *Psychological methods*, 7(2):147, 2002.

[14] Tapio Schneider. Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. *Journal of Climate*, 14(5):853–871, 2001.

[15] Jason D Van Hulse, Taghi M Khoshgoftaar, and Haiying Huang. The pairwise attribute noise detection algorithm. *Knowledge and Information Systems*, 11(2):171–190, 2007.

[16] Xian Wang, Ao Li, Zhaohui Jiang, and Huanqing Feng. Missing value estimation for dna microarray gene expression data by support vector regression imputation and orthogonal coding scheme. *BMC bioinformatics*, 7(1):32, 2006.

[17] Andrew KC Wong and David KY Chiu. Synthesizing statistical knowledge from incomplete mixed-mode data. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6):796–805, 1987.

[18] Ying Yang and Geoffrey I Webb. Discretization for naive-bayes learning: managing discretization bias and variance. *Machine learning*, 74(1):39–74, 2009.