Chapter 14
Outliers, Missing Values, and Reliability:
An Integrated Framework for Pre-Processing of Coding Data

Swati Aggarwal
NSIT, India

Shambeel Azim
Vidyadaan Institute of Technology and Management, India

ABSTRACT

Reliability is a major concern in qualitative research. Most of the current research deals with finding the reliability of the data, but not much work is reported on how to improve the reliability of the unreliable data. This paper discusses three important aspects of the data pre-processing: how to detect the outliers, dealing with the missing values and finally increasing the reliability of the dataset. Here authors have suggested a framework for pre-processing of the inter-judged data which is incomplete and also contains erroneous values. The suggested framework integrates three approaches, Krippendorff's alpha for reliability computation, frequency based outlier detection method and a hybrid fuzzy c-means and multilayer perceptron based imputation technique. The proposed integrated approach results in an increase of reliability for the dataset which can be used to make strong conclusions.

INTRODUCTION

Data in the real world is dirty. They are noisy, incomplete and inconsistent. Data mining results are significantly affected by the quality of the data available. Low grade data can return unsatisfactory results. So to improve the quality of the data mining result, pre-processing of the raw data must be carried out. Reliability Analysis, Outlier detection and missing value imputations are the important part of knowledge discovery process. Reliability is a measure that tells how accurate the data is, to be used for a particular purpose. Generally reliability is classified into four classes, i.e. Inter-Coder or Inter-Rater Reliability,

DOI: 10.4018/978-1-5225-1008-6.ch014

Test-Retest Reliability, Parallel-Forms Reliability and Internal Consistency Reliability. Inter-coder reliability is an important element of Content Analysis. It is a measure of agreement to the extent different rater or observer agrees upon when rating the same item. High inter-coder reliability means that judges are in consensus with each other. A set of observation is needed to be judged by more than one judge to draw a conclusion that should be relatively strong. Since humans are associated with inconsistencies and mistakes, conclusion drawn out by a single observer cannot be taken as reliable one. Use of experts as a method of measurement is frequent in different fields like psychology, marketing, education, and others. For example, a candidate contesting an election can look out for an opinion before the poll, in articles published in different newspapers about him. For this he needs experts to draw conclusion from the different articles. On the basis of this conclusion he can determine his chances of winning or losing for which further strategies can be formulated. This is possible only if the inter-rater reliability is high.

Another point of concern is the outliers. Outlier is defined as an observed value that is aloof from the other observation values. They deviate from the other observations in the sample to the extent that they are noticeable. Another definition as given by Johnson (1992) is: An outlier as an observation in a data set which appears to be inconsistent with the remainder of that set of data. There may be a single or multiple outliers present in an observation. Occurrence of outliers may be due to the error, in calculating or recording the measurement value or due to the experimental error (Grubbs, 1969). For example, if four out of five judges assign 'yes' to a nominal question and one judge says 'no', then 'no' is said to be an outlier. The Outlier detection method has various applications like data cleansing, credit card fraud detection, intrusion detection system, medical diagnosis and several others.

Other important step involved in pre-processing of the data is missing value imputation. The missing value is an empty cell in a table that represent a dataset. Reasons for missing values are many, for example, human errors involved in the data collection tasks. Therefore, it is of great necessity to fill in these missing values or record to extract or find patterns from these datasets. Filling in the missing value is known as data imputation. Imputation is important because evaluation of complete data produces authentic results. Interpretations and inferences with complete data are more accurate (Abdella & Marwala, 2005).

RELATED WORK

Various works have been done by different researchers to estimate intercoder reliability in content analysis. Cohen's Kappa (Cohen's, 1960) is a reliability measure that works only on nominal data and takes into account occurrence of agreement by chance. Moreover the number of raters in this case is limited to two only. Fleiss' Kappa (Fliess, 1971) is an improvement over Cohen's Kappa as it works on more than two raters. It also has the same limitation as Cohen's Kappa i.e. it is limited to nominal data only. Cronbach's alpha is used most commonly as a reliability coefficient (Hogan et al., 2000). It is a measure of internal consistency of the test. The problem with Cronbach's alpha is that it is non robust, a single observation can greatly affect the coefficient value (Christmann & Van, 2006). An improvement to various other statistics for measuring the reliability of interrater data is Krippendorff's alpha (Hayes et al., 2007; Krippendorff, 2013, 2011). It is a flexible, reliable measure that can be used with any number of raters, with any metric or level of measurement and also with the missing data.

The Outlier detection method is of two types. One is univariate outlier detection method and the other is a multivariate outlier detection method. According to Pyle (1999), univariate outliers are those values