Data Pre-processing for knowledge discovery

Mortadha M. Hamad , Banaz A. Qader
College of Computer , University of Anbar

Abstract

Data pre-processing stage is also known as (data preparation) stage and it is a fundamental stage for data analysis and knowledge discovery. If there is much irrelevant and redundant information or noisy and unreliable data, then knowledge discovery during analysis and mining phase will be more difficult. Therefore we consider the pre-processing stage as an important step for knowledge discovery process and has a significant impact on predictive accuracy. Essentially, while each customer attribute may require special treatment for each algorithm, so the choices of data pre-processing (DPP) depend on the individual dataset or database used. In this paper we have chosen and explained two different pre-processing techniques which are (consistency, reduction) depending on our data warehouse of marketing which contains inconsistent attributes and also contains duplicated records. We have also proposed two new algorithms for reduction named (Removing Duplication Algorithm) and for consistency named (Resolving Inconsistency Algorithm) so that achieving the best performance for their data set. In this paper we applied and implemented our two new algorithms on our data warehouse using (C# programming language) and (Microsoft Access file), and gained cleaning data warehouse with consistent attributes and empty of duplicated records that is ready for preparing quality data as input to the algorithms of data mining process or any other analysis method which also influences of knowledge quality that is discovered during data mining process.

Keywords: data pre-processing, data mining, knowledge discovery, data cleaning.

1- Introduction

Recently, progress in computational and storage capacity has enabled the accumulation of ordinal, nominal, binary and unary demographic and psychographic customer centric data, inducing large, rich datasets of heterogeneous scales. Essentially, each customer attribute may require special treatment for each algorithm, such as discretization of numerical features, rescaling of ordinal features and encoding of categorical ones. The phase of data preprocessing (DPP) represents a complex prerequisite for data mining in the process of knowledge discovery in databases aiming to maximize the predictive accuracy of data mining [1]. In many computer science fields, such as pattern recognition, information retrieval, machine learning, data mining, and Web intelligence, we need to prepare quality data by pre-processing the raw data. Data preprocessing (preparation) is therefore a crucial research topic. However, much work in the field of data mining was built on the existence of quality data. The emergence of knowledge discovery in databases (KDD) as a new technology has been brought about with the fast development and broad application of information and database technologies. The process of KDD is defined as an iterative sequence of four steps which are (defining the problem, data pre-processing (data preparation) , data mining, and post data mining) [3]. So in this paper we study the pre-processing stage especially its some important techniques, where it is necessary to prepare the data for knowledge discovery and data mining. Because preprocessing stage consists of many analysis methods which transforms the raw data in to the cleaning and high quality data, this lead to get the accurate and high quality knowledge that is discovered during knowledge discovery phase by applying data mining.

Corporate data mining faces the challenge of systematic knowledge discovery in large data streams to support managerial decision making, because the application of each data mining algorithm requires the presence of data in a mathematically feasible format which achieved through DPP. Therefore, data pre-processing (DPP) represents a prerequisite phase for data mining in the process of knowledge discovery in databases (KDD). While research in operations research, direct marketing and machine learning focuses on the analysis and design of data mining algorithms, so there is need to interaction of data mining with the preceding phase of data pre-processing. Data pre-processing (DPP) tasks are distinguished in data reduction, aiming at decreasing the size of the dataset by means of instance selection and/or feature selection, and data projection, altering the representation of the data, e.g. mapping continuous variables to categories or encoding nominal attributes. While some of these are imperative for the valid application of a method, such as scaling for neural network, others appear to be more general to facilitate method performance in general [1].

2- Data Pre-Processing Definition

Data pre-processing is also known as (data preparation). It comprises those techniques concerned with analyzing raw data so as to yield quality data. It mainly includes data collection and integration, data transformation, data cleaning, data reduction, and data discretization [3].

Data pre-processing is an important step in the data mining process, because if there is much irrelevant and redundant information or noisy and unreliable data, then knowledge discovery during the analysis and training phase will be more difficult [7]. Data pre-processing is considered as a data mining technique that involves transforming raw data into an