Chapter 3

Data Discretization in Data Mining and Machine Learning

Dr. Balwinder Kaur
Department of Computer Science and Applications, Panjab University, Chandigarh
balwinder@pu.ac.in

Abstract

Data discretization is one of fundamental and essential preprocessing techniques used in Knowledge Discovery, Data Mining (DM) and Machine Learning (ML). The fundamental objective of discretization is to transform continuous attributes into discrete. This transformation allows quantitative data to be treated as qualitative. The transformation plays an important role in improving efficiency, interpretability, and compatibility of ML algorithms and specific models. This paper presents the basic concept of data discretization, its importance, discretization process, methods, framework, advantages, and associated limitations. The study covers theoretical concepts of discretization techniques such as equal-frequency binning, equal-width binning, clustering-based discretization, and entropy-based methods. Additionally, it also discusses the role of discretization in enhancing classification accuracy and reducing computational complexity.

Keywords: Discretization framework, Supervised and unsupervised discretization, Splitting, Merging

1. Introduction

With the exponential growth of data in various domains, effective data preprocessing is vital for successful data mining and machine learning applications. One essential step in data preprocessing process is data discretization as most of the real-world data is continuous in nature. Data discretization refers to transforming continuous data into discrete intervals or categories. Data in the datasets are usually a mix of formats like nominal, discrete, and/or continuous. Unlike nominal data, which lacks inherent ordering, both discrete and continuous variables are considered ordinal since their values follow a meaningful sequence. Discrete values are intervals in a domain of continuous values. While the number of continuous values for an attribute can be infinitely many whereas the number of discrete values is often few or finite[1], [2].

In the field of knowledge discovery, many learning methods –like Decision Tree, Bayesian Networks, Association Rules, etc. [3] can handle discrete values only. Therefore, before applying knowledge discovery process, it is important to encode continuous attribute into a discrete attribute constituted by a set of intervals. For example, the age of a person can be transformed into two discrete values: greater than equal to 18 (major) and less than 18 (minor). After discretization, the data can be handled as nominal data in both inductive and deductive data-mining processes.

Discretization simplifies data representation and is beneficial for algorithms that process categorical inputs, such as Decision Trees and Naïve Bayes classifiers. Discretization enhances model interpretability by making it easier to draw insights from voluminous and complex datasets. Discrete features are closer to knowledge-level representations than continuous ones,