19SE02IT058

Assignment 1

* Data mining functionalities

1. Classification

Classification is a data mining technique that categorizes items in a collection, based on some predefined properties. A training set containing items whose properties are known and is used to train the system to predict the category of items from an unknown collection of items. It uses methods like if-then, decision trees or neural networks to predict a class or essentially classify a collection of items.

2. Association Analysis

Also known as Market Basket Analysis for its wide use in retail sales, Association analysis aims to discover associations between items occurring together frequently. Association analysis is based on rules having 2 parts:

1. **antecedent (if)**
2. **consequent(then)**

An antecedent is an item in a collection, which, when found, also indicates a certain chance of finding a consequent in the collection. In other words, they are associated. From the data association inferences like,

If a customer buys potato chips, he is about 60% like to buy soft drinks along with it. So, over a big chunk of data, you see this association with a certain degree of confidence.

3. Cluster Analysis

Cluster Analysis, fundamentally similar to classification, where similar data are grouped together with the difference being that a class label is not known. Clustering algorithms group data based on similar features and

dissimilarities. Used in image processing, pattern recognition and bioinformatics, clustering is a popular functionality of data mining.

4. Class/Concept Description: Characterization and Discrimination

A class or concept implies there is a data set or set of features that define the class or a concept. A class can be a category of items on a shop floor, and concept could be the abstract idea on which data may be categorized like products to be put on clearance sale and non-sale products. There are two concepts here, one that helps with grouping and the other that helps in differentiating.

Data Characterization

Characterization involves summarization of general features of the data, resulting in specific rules that define a target class. A data analysis technique called Attribute-oriented Induction is employed on the data set for achieving characterization.

Data Discrimination

Discrimination is used to separate distinct sets of data, based on the disparity in attribute values. It is a comparison of features of a class with features of one or more contrasting classes.

5. Prediction

In data mining, there are primarily two types of predictions, numeric predictions and class predictions. Numeric predictions are made by creating a linear regression model that is based on historical data. Prediction of numeric values helps businesses ramp up for a future event that might impact business in a positive or a negative way. Class predictions are uses to fill in missing class information for products using a training data set where the class for products is known.

6. Outlier Analysis

Outlier analysis is important to understand the quality of data. If there are too many outliers, you cannot trust the data or draw patterns out of it. An outlier analysis of the data that cannot be grouped into any classes by the algorithms is pulled up. An outlier analysis tries to determine if there is something out of turn in the data and does it indicate a situation that a business needs to take account of and take measures to mitigate it. In most cases, outliers are data abnormalities. Still, it is important to take note of each so unusual activities or events that trigger a business impact can be detected well in advance.

7. Evolution & Deviation Analysis

Evolution Analysis pertains to study of data sets that undergo change over a time period. Evolution analysis models are designed to capture evolutionary trends in data helping in characterizing, classifying, clustering or discrimination of time-related data.

.