DATA MINING AND BIG DATA ANALYTICS (18MCA52C)

UNIT III: Classification and predictions

FACULTY

Dr. K. ARTHI MCA, M.Phil., Ph.D.,

Assistant Professor,

Postgraduate Department of Computer Applications,

Government Arts College (Autonomous),

Coimbatore-641018.

Unit III

Classification and predictions

- Classification is a form of data analysis that extracts models describing important data classes.
- Such models, called classifiers, predict categorical (discrete, unordered) class labels.
- For example, we can build a classification model to categorize bank loan applications as either safe or risky.
- Such analysis can help provide us with a better understanding of the data at large.
- Many classification methods have been proposed by researchers in machine learning, pattern recognition, and statistics.
- Most algorithms are memory resident, typically assuming a small data size.
- Recent datamining research has built on such work, developing scalable classification and prediction techniques capable of handling large amounts of disk-resident data.
- Classification has numerous applications, including fraud detection, target marketing, performance prediction, manufacturing, and medical diagnosis.

Basic Concepts

The general approach to classification as a two-step process.

In the first step, we build a classification model based on previous data.

In the second step, we determine if the model's accuracy is acceptable, and if so, we use the model to classify new data.

What Is Classification?

- A bank loans officer needs analysis of her data to learn which loan applicants are "safe"

  and which are "risky" for the bank.

- A marketing manager at AllElectronics needs data analysis to help guess whether a customer with a given profile will buy a new computer.

- A medical researcher wants to analyze breast cancer data to predict which one of three

  specific treatments a patient should receive.

- In each of these examples, the data analysis task is classification, where a model or classifier is constructed to predict class (categorical) labels, such as "safe" or "risky" for the loan application data; "yes" or "no" for the marketing data; or "treatment A," "treatment B," or "treatment C" for the medical data.

- These categories can be represented by discrete values, where the ordering among values has no meaning.
- For example, the values 1, 2, and 3may be used to represent treatments A, B, and C, where there is no ordering implied among this group of treatment regimes.
- Suppose that the marketing manager wants to predict how much a given customer will spend during a sale at AllElectronics.
- This data analysis task is an example of numeric prediction, where the model constructed predicts a continuous-valued function, or ordered value, as opposed to a class label. This model is a predictor.
- Regression analysis is a statistical methodology that is most often used for numeric prediction
- Classification and numeric prediction are the two major types of prediction problems.

General Approach to Classification

"How does classification work?"

- Data classification is a two-step process, consisting of a learning step (where a classification model is constructed) and
- a classification step (where the model is used to predict class labels for given data).
- The process is shown for the loan application data of Figure 8.1. (The data are simplified for illustrative purposes.
- In reality, we may expect many more attributes to be considered.
- In the first step, a classifier is built describing a predetermined set of data classes or concepts.
- This is the learning step (or training phase), where a classification algorithm builds the classifier by analyzing or "learning from" a training set made up of database tuples and their associated class labels.

- A tuple, X, is represented by an n-dimensional attribute vector X = (x1,x2,x3….xn)

  n database attributes, respectively, A1, A2, …. , An.

- Each tuple, X, is assumed to belong to a predefined class as determined by another database attribute called the class label attribute.
- The class label attribute is discrete-valued and unordered. It is categorical (or nominal) in that each value serves as a category or class.
- The individual tuples making up the training set are referred to as training tuples and are randomly sampled from the database under analysis.
- In the context of classification, data tuples can be referred to as samples, examples, instances, data points, or objects.
- Because the class label of each training tuple is provided, this step is also known as supervised learning (i.e., the learning of the classifier is "supervised" in that it is told to which class each training tuple belongs).
- It contrasts with unsupervised learning (or clustering), in which the class label of each training tuple is not known, and the number or set of classes to be learned may not be known in advance.
- For example, if we did not have the loan decision data available for the training set, we could use clustering to try to determine "groups of like tuples," which may correspond to risk groups within the loan application data.
- This first step of the classification process can also be viewed as the learning of amap-ping or function, y D f .X/, that can predict the associated class label y of a given tuple X.
- In this view, we wish to learn a mapping or function that separates the data classes.
- Typically, this mapping is represented in the form of classification rules, decision trees, or mathematical formulae.
- In our example, the mapping is represented as classification rules that identify loan applications as being either safe or risky (Figure 8.1a).
- The rules can be used to categorize future data tuples, as well as provide deeper insight into the data contents.
- They also provide a compressed data representation.

What about classification accuracy?"

- In the second step (Figure 8.1b), the model is used for classification. First, the predictive accuracy of the classifier is estimated.

- If we were to use the training set to measure the classifier's accuracy, this estimate would likely be optimistic, because the classifier tends to overfit the data (i.e., during learning it may incorporate some particular anomalies of the training data that are not present in the general data set overall).
- Therefore, a test set is used, made up of test tuples and their associated class labels.
- They are independent of the training tuples, meaning that they were not used to construct the classifier.
- The accuracy of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier.
- The associated class label of each test tuple is compared with the learned classifier's class prediction for that tuple.
- If the accuracy of the classifier is considered acceptable, the classifier can be used to classify future data tuples for which the class label is not known. (Such data are also referred to in the machine learning literature as "unknown" or "previously unseen" data.)
- For example, the classification rules learned in Figure 8.1(a) from the analysis of data from previous loan applications can be used to approve or reject new or future loan applicants.

Issues Regarding Classification and Prediction

Preparing the Data for Classification and Prediction

The following preprocessing steps may be applied to the data in order to help improve the accuracy, efficiency, and scalability of the classification or prediction process.

Data Cleaning:

- This refers to the preprocessing of data in order to remove or reduce noise (by applying smoothing techniques) and the treatment of missing values (e.g., by replacing a missing value with the most commonly occurring value for that attribute, or with the most probable value based on statistics.)
- Although most classification algorithms have some mechanisms for handling noisy or missing data, this step can help reduce confusion during learning.

Relevance Analysis:

- Many of the attributes in the data may be irrelevant to the classification or prediction task.
- For example, data recording the day of the week on which a bank loan application was filed is unlikely to be relevant to the success of the application.
- Furthermore, other attributes may be redundant. Hence, relevance analysis may be performed on the data with the aim of removing any irrelevant or redundant attributes from the learning process.
- In machine learning, this step is known as feature selection. Including such attributes may otherwise slow down, and possibly mislead, the learning step.
- Ideally, the time spent on relevance analysis, when added to the time spent on learning from the resulting "reduced" feature subset should be less than the time that would have been spent on learning from the original set of features.
- Hence, such analysis can help improve classification efficiency and scalability.

Data Transformation:

- The data can be generalized to higher – level concepts. Concept hierarchies may be used for this purpose.
- This is particularly useful for continuous – valued attributes. For example, numeric values for the attribute income may be generalized to discrete ranges such as low, medium, and high.
- Similarly, nominal – valued attributes like street, can be generalized to higher – level concepts, like city. S
- ince generalization compresses the original training data, fewer input / output operations may be involved during learning.
- The data may also be normalized, particularly when neural networks or methods involving distance measurements are used in the learning step.
- **Normalization** involves scaling all values for a given attribute so that they fall within a small specified range, such as – 1.0 to 1.0, or 0.0 to 1.0.
- In methods that use distance measurements, for example, this would prevent attributes with initially large ranges (like, say, income) from outweighing attributes with initially smaller ranges (such as binary attributes).

Comparing Classification Methods

Classification and prediction methods can be compared and evaluated according to the following criteria:

Predictive Accuracy: This refers to the ability of the model to correctly predict the class label of new or previously unseen data.

Speed: This refers to the computation costs involved in generating and using the model.

Robustness: This is the ability of the model to make correct predictions given noisy data or data with missing values.

Scalability: This refers to the ability to construct the model efficiently given large amount of data.

Interpretability: This refers to the level of understanding and insight that is provided by the model.

THANK YOU

This content is taken from the text books and reference books prescribed in the syllabus.