Chapter 4
Types of Data Manipulations

Study Exercise – Reviewing Data Reduction tools and concepts.

Use your knowledge, the preceding study material, or any other resources to answer the following questions.

1. What are the differences between **Numerosity Data Reduction** and **Dimensionality Data Reduction**?
2. When performing data reduction we must be careful to maintain a reasonable balance between two objectives, what are those two objectives?
3. What is the **Class Imbalance Problem** and how it is related to data reduction?
4. What is the difference between supervised and unsupervised dimension reduction? Mention three tools for performing both supervised and unsupervised dimension reduction.

Answers to questions

Q1: What are the differences between Numerosity Data Reduction and Dimensionality Data Reduction?

Numerosity Data Reduction reduces the data by reducing the data volume or the number of data objects (rows), and dimensionality data reduction reduces the data by reducing the data variety or the number of attributes (columns).

Q2: When performing data reduction we must be careful to maintain a reasonable balance between two objectives, what are those two objectives?

1. Keeping the integrity of the data and causing bias to the data as little as possible
2. Reducing the size of the data

Q3: What is the Class Imbalance Problem and how it is related to data reduction?

When the binary dependent attribute of a dataset has overwhelmingly higher repetition for one of the classes over the other. For instance, in a fraud detection dataset, most of the rows represent non-fraudulent records and that's the imbalance. This causes the algorithms to be skewed toward the more dominant class.

It is related to data reduction as one of the methods to deal with this problem is by resampling which is a form of data reduction.

Q4: What is the difference between supervised and unsupervised dimension reduction? Mention three tools for performing both supervised and unsupervised dimension reduction.

The Supervised dimension reduction is also called feature selection. It is the task of finding the subset of attributes that will lead to the best classification/prediction performance.

In contrast, unsupervised dimension reduction is about finding a new representation of the data that keeps the most important part of the data and cut the fluff out.

Three tools for supervised dimension reduction: Decision/Regression Tree, Random Forest, Brute-force dimension reduction

Three tools for unsupervised dimension reduction: Principal Component Analysis, Functional Data Reduction, Feature extraction