Numerosity reduction in data mining pdf

Numerosity reduction. Numerosity reduction example. Numerosity reduction in data mining. Dimensionality reduction and numerosity reduction.

The data reduction process reduces data size and makes it suitable and feasible for analysis. In the reduction process, the integrity of the data must be preserved, and data volume is reduced. Many techniques can be used for data reduction, and two primary methods of Data Reduction are Dimensionality Reduction and Numerosity Reduction. What is Numerosity Reduction? In the numerosity reduction, the data volume is decreased by selecting an alternative, smaller form of data representation. These techniques can be parametric or non-parametric. For parametric methods, a model can estimate the data so that only the data parameters need to be saved, instead of the actual data, for example, Log-linear models. Non-parametric methods are used to store a reduced representation of the data, including histograms, clustering, and sampling. Types of Numerosity Reduction This method uses an alternate, small forms of data representation, thus reducing data volume. There are two types of Numerosity reduction, such as: 1. Parametric This method assumes a model into which the data fits. Data model parameters are estimated, and only those parameters are stored, and the rest of the data is discarded. Regression and Log-Linear methods are used for creating such models. For example, a regression model can be used to achieve parametric reduction if the data fits the Linear Regression model. Regression: Regression can be a simple linear regression or multiple linear regression. When there is only a single independent attribute, such a regression model is called simple linear regression. If there are multiple independent attributes, then such regression models are called multiple linear regression. Linear Regression models a linear relationship between two attributes of the data set. We need to fit a linear regression model between two attributes, x, and y, where y is the dependent attribute, and x is the independent attribute or predictor attribute. The model can be represented by the equation y=wx+b, where w and b are regression coefficients. A multiple linear regression model lets us express the attribute y in terms of multiple predictor attributes. Log-Linear Model: Log-linear model can be used to estimate the probability of each data point in a multidimensional space for a set of discretized attributes based on a smaller subset of dimensional combinations. This allows a higher-dimensional data space to be constructed from lower-dimensional attributes. The Log-Linear model discovers the relationship between two or more discrete attributes. Assume we have a set of tuples in n-dimensional space; the log-linear model helps derive each tuple's probability in this n-dimensional space. NOTE: Regression and log-linear models can both be used on sparse data, although their application may be limited.

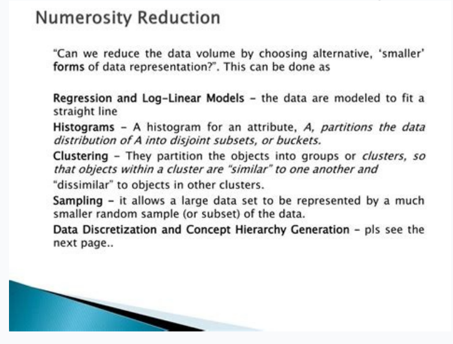
2. Non-Parametric A non-parametric numerosity reduction technique does not assume any model.

The non-Parametric technique results in a more uniform reduction, irrespective of data size, but it may not achieve a high volume of data reduction like the Parametric one.

There are at least four types of Non-Parametric data reduction techniques, Histogram, Clustering, Sampling, Data Cube Aggregation, Data Compression. Histograms: A histogram is the data representation in terms of frequency. It uses binning to approximate data distribution and is a popular form of data reduction. Suppose a histogram for an attribute A and divisions the data distribution of A into disjoint subsets or buckets. If each bucket defines only an individual attribute-value or frequency pair, the buckets are known as singleton buckets. Clustering: Clustering techniques consider data tuples as objects. They partition the objects into groups or clusters so that objects within a cluster are "similar" to one another and "dissimilar" to objects in other clusters. It is commonly defined in terms of how "close" the objects are in space, based on a distance function.

The quality of a cluster can be defined by its diameter, the maximum distance between any two objects in the cluster. Centroid distance is an alternative measure of cluster quality. It is represented as the average distance of each cluster object from the cluster centroid denoting the "average object," or average point in the area for the cluster. Sampling: Sampling can be used as a data reduction approach because it enables a huge data set to be defined by a much smaller random sample or a subset of the information. Sampling can reduce large data sets into smaller sample data sets to represent the original data set. There are four types of sampling data reduction methods. Simple Random Sample Without Replacement of sizes Simple Random Sample with Replacement of sizes Cluster Sample Stratified Sample Data Cube Aggregation: Data cube aggregation involves moving the data from a detailed level to fewer dimensions. The resulting data set is smaller in volume, without loss of information necessary for the analysis task. Data Cube Aggregation is a multidimensional aggregation that uses aggregation at various levels of a data cube to represent the original data set, thus achieving data reduction. Data Cube Aggregation, where the data cube is a much more efficient way of storing data, thus achieving data reduction, besides faster aggregation operations. Data Compression: It employs modification, encoding, or converting the structure of data in a way that consumes less space. Data compression involves building a compact representation of information by removing redundancy and representing data in binary form. Data that can be restored successfully from its compressed form is called Lossless compression. In contrast, the opposite, where it is not possible to restore the original form from the compressed form, is Lossy compression. Difference between Numerosity Reduction and Dimensionality Reduction There are two primary methods of Data Reduction, Dimensionality Reduction, and Numerosity Reduction. Data encoding or transformations are used to access a reduced or "compressed" depiction of the original data in dimensionality reduction. If the original data can be regenerated from the compressed data without any loss of data, the data reduction is known as lossless. If data reconstructed is only approximated by the original data, the data reduction is called lossy. Let's see the comparison between Dimensionality Reduction and Numerosity Reduction, such as: Numerosity Reduction Dimensionality Reduction In numerosity reduction, data volume is reduced by choosing alternating, smaller forms of data representation. In dimensionality reduction, data encoding or transformation are applied to obtain a reduced or compressed representation of original data. In numerosity reduction, regression and log-linear models can be used to approximate the given data. In linear regression, the data are modeled to fit a straight line. For example, a random variable, y (known as the response variable), can be modeled as a linear function of another random variable, x (known as a predictor variable), with the equation y = wx+b, where the variance of y is assumed to be constant. In dimensionality reduction, the discrete wavelet transform (DWT) is a linear signal processing technique that changes it to a numerically different vector, X', of wavelet coefficients when used to a data vector X. The two vectors are of the same length. When applying this technique to data reduction, it can consider each tuple as an n-dimensional data vector, that is, X=(x1,x2,...xn)depicting n measurements made on the tuple from n database attributes.

It is merely a representation technique of original data to a smaller form. It can be used for removing irrelevant and redundant attributes. There is no loss of data in this method, but the whole data is represented in a smaller form. In this technique, some data can be lost, which is inappropriate. Next TopicMarket Basket Analysis in Data Mining ReadDiscussCoursesPracticeImprove Article Save Article Like Article Prerequisite: Data preprocessing Why Data Reduction ? Data reduction process reduces the size of data and makes it suitable and feasible for analysis. In the reduction process, integrity of the data must be preserved and data volume is reduced. There are many techniques that can be used for data reduction. Numerosity reduction is one of them. Numerosity Reduction: Numerosity Reduction is a data reduction technique which replaces the original data by smaller form of data representation. There are two techniques for numerosity reduction- Parametric and Non-Parametric methods.INTRODUCTION:Numerosity reduction is a technique used in data mining to reduce the number of data points in a dataset while still preserving the most important information.



Numerosity Reduction

"Can we reduce the data volume by choosing alternative, 'smaller' **forms** of data representation?". This can be done as

Regression and Log-Linear Models – the data are modeled to fit a straight line

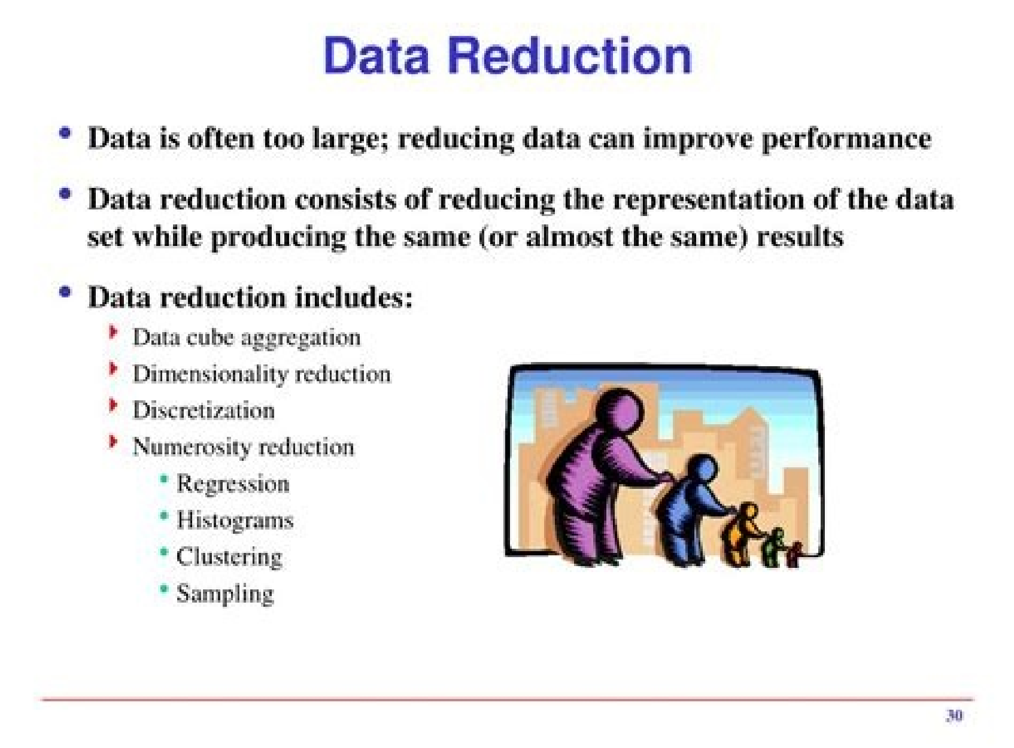
Histograms – A histogram for an attribute, *A, partitions the data distribution of A into disjoint subsets, or buckets.*

Clustering – They partition the objects into groups or *clusters, so that objects within a cluster are "similar" to one another and* "dissimilar" to objects in other clusters.

Sampling – it allows a large data set to be represented by a much smaller random sample (or subset) of the data.

Data Discretization and Concept Hierarchy Generation – pls see the next page..

This can be beneficial in situations where the dataset is too large to be processed efficiently, or where the dataset contains a large amount of irrelevant or redundant data points.Data Sampling: This technique involves selecting a subset of the data points to work with, rather than using the entire dataset. This can be useful for reducing the size of a dataset while still preserving the overall trends and patterns in the data.Clustering: This technique involves grouping similar data points together and then representing each group by a single representative data point.Data Aggregation: This technique involves combining multiple data points into a single data point by applying a summarization function.Data Generalization: This technique involves replacing a data point with a more general data point that still preserves the important information.Data Compression: This technique involves using techniques such as lossy or lossless compression to reduce the size of a dataset.It's important to note that numerosity reduction can have a trade-off between the accuracy and the size of the data. The more data points are reduced, the less accurate the model will be and the less generalizable it will be.In conclusion, numerosity reduction is an important step in data mining, as it can help to improve the efficiency and performance of machine learning algorithms by reducing the number of data points in a dataset. However, it is important to be aware of the trade-off between the size and accuracy of the data, and carefully assess the risks and benefits before implementing it.Parametric Methods –For parametric methods, data is represented using some model. The model is used to estimate the data, so that only parameters of data are required to be stored, instead of actual data. Regression and Log-Linear methods are used for creating such models. Regression: Regression can be a simple linear regression or multiple linear regression. When there is only single independent attribute, such regression model is called simple linear regression and if there are multiple independent attributes, then such regression models are called multiple linear regression. In linear regression, the data are modeled to a fit straight line. For example, a random variable y can be modeled as a linear function of another random variable x with the equation y = ax+b where a and b (regression coefficients) specifies the slope and y-intercept of the line, respectively. In multiple linear regression, y will be modeled as a linear function of two or more predictor(independent) variables.   Log-Linear Model: Log-linear model can be used to estimate the probability of each data point in a multidimensional space for a set of discretized attributes, based on a smaller subset of dimensional combinations.



Data Reduction

- **Data is often too large; reducing data can improve performance**
- **Data reduction consists of reducing the representation of the data set while producing the same (or almost the same) results**
- **Data reduction includes:**
  - Data cube aggregation
  - Dimensionality reduction
  - Discretization
  - Numerosity reduction
    - Regression
    - Histograms
    - Clustering
    - Sampling

30

This allows a higher-dimensional data space to be constructed from lower-dimensional attributes. Regression and log-linear model can both be used on sparse data, although their application may be limited.Non-Parametric Methods –These methods are used for storing reduced representations of the data include histograms, clustering, sampling and data cube aggregation. Histograms: Histogram is the data representation in terms of frequency. It uses binning to approximate data distribution and is a popular form of data reduction. Clustering: Clustering divides the data into groups/clusters. This technique partitions the whole data into different clusters. In data reduction, the cluster representation of the data are used to replace the actual data. It also helps to detect outliers in data. Sampling: Sampling can be used for data reduction because it allows a large data set to be represented by a much smaller random data sample (or subset). Data Cube Aggregation: Data cube aggregation involves moving the data from detailed level to a fewer number of dimensions. The resulting data set is smaller in volume, without loss of information necessary for the analysis task.Improved efficiency: Numerosity reduction can help to improve the efficiency of machine learning algorithms by reducing the number of data points in a dataset. This can make it faster and more practical to work with large datasets.Improved performance: Numerosity reduction can help to improve the performance of machine learning algorithms by removing irrelevant or redundant data points from the dataset. This can help to make the model more accurate and robust.Reduced storage costs: Numerosity reduction can help to reduce the storage costs associated with large datasets by reducing the number of data points.Improved interpretability: Numerosity reduction can help to improve the interpretability of the results by removing irrelevant or redundant data points from the dataset.Loss of information: Numerosity reduction can result in a loss of information if important data points are removed during the reduction process.Impact on accuracy: Numerosity reduction can impact the accuracy of a model, as reducing the number of data points can also remove important information that is needed for accurate predictions.Impact on interpretability: Numerosity reduction can make it harder to interpret the results, as removing irrelevant or redundant data points can also remove context that is needed to understand the results.Additional computational costs: Numerosity reduction can add additional computational costs to the data mining process, as it requires additional processing time to reduce the number of data points.In conclusion, numerosity reduction can have both advantages and disadvantages. It can improve the efficiency and performance of machine learning algorithms by reducing the number of data points in a dataset. However, it can also result in a loss of information and make it harder to interpret the results. It's important to weigh the pros and cons of numerosity reduction and carefully assess the risks and benefits before implementing it.Last Updated : 02 Feb, 2023Like Article Save Article