CHAPTER 6

Wavelet Transforms

6.1 Wavelet Transforms:

6.2 Principal Components Analysis

6.3 Numinosity Reduction

6.4 Regression and Log-Linear models:

6.5 Histograms:

6.6 Clustering:

6.7 Sampling:

6.8 Review Question

6.9 References

6.Wavelet Transforms

6.1 Wavelet Transforms

The discrete wavelet transform (DWT) is a linear signal processing technique. It transforms a vector into a numerically different vector (D to D') of wavelet coefficients. The two vectors are of the same length. However it is useful for compression in the sense that wavelet-transformed data can be truncated. A small compressed approximation of the data can be retained by storing only a small fraction of the strongest wavelet coefficient e.g., retain all wavelet coefficients larger than some particular threshold and the remaining coefficients are set to zero. The resulting data representation is sparse. Computations that can take advantage of sparsity are very fat if performed in wavelet space. Given a set of coefficients, an approximation of the original data con be got by applying the inverse DWT. The DWT is closely related to the discrete Fourier transform (DFT) a signal processing technique involving sine's and cosines. The general procedure for applying a discrete wavelet transform uses a hierarchical pyramid algorithm that halves the data in each iteration, resulting in fast computational speed. The method is as follows:

1. The length, L , of the input data vector must and integer power of 2.This condition can be met by padding the data vector with zeros as necessary.
2. Each transform involves applying two functions. The first applies some data smoothing, such as sum or weighted average .The second performs a weighted difference, which acts to bring out the detailed features of the data.
3. The two functions are applied to pairs of input data, resulting in two sets of data of length L/2. In general these represent a smoothed or low frequency version so he input data and the high frequency content of it.
4. The two functions are recursively applied to sets of data obtained in the previous loop, until the resulting data sets obtained are of length 2.
5. A selection of values from the data sets obtained in the above iterations are designated the wavelet coefficients of the transformed data.

Wavelet transforms can be applied to multidimensional data such as data cubes. Wavelet transforms have many real world applications, including the compression of fingerprint images, computer vision, and analysis of time-series data and data cleaning.

6.2 Principal Components Analysis

An intuitive introduction is provided for principal components analysis in this section. Let the data to be compressed consist of data vectors, from k dimensions. Principal Component Analysis or PCA searches for c k-dimensional orthogonal vectors that can be best used to represent the data, where c<=k. The original data are thus projected onto a much smaller space. PCA can be used to perform dimensionality reduction. The basic procedure is as follows:

1. The input data are normalized, so that each attribute falls within the same range. This step helps ensure that attributes with large domains will not dominate attributes with smaller domains.
2. PCA computes c orthonormal vectors the provide a basis for the normalized input data. These are unit vectors that each point in a direction perpendicular to the others .These vectors are referred to as the principal components. The input data are a linear combination of principal components.
3. Strength. The principal components essentially serve as a new set of axes for the data, by providing important information about various.
4. Since the components are sorted according to decreasing order of “ significance “ elimination the weaker components can reduce the size o f the data. Using the strongest principal components it should be possible to reconstruct the original data to a good approximation.

PCA is computationally inexpensive and it can be ordered or unordered.

6.3 Numinosity Reduction

Techniques of luminosity reduction can be applied for the purpose of representing data in “smaller”forms”.these techniques could be parametric or non-parametric. For parametric methods a model is used to estimate the data, so that typically only the data parameters need to be stored, instead of the actual data. Log-linear models, which estimate discrete multidimensional probability distributions, are an ex ample. Non parametric methods for storing data include histograms, clustering and sampling.

6.4 Regression and Log-Linear models:

Regression and log-linear models can be used to approximate the given data. In linear regression, the data are m modeled to fit a straight line. For example a random variable Y (called a response variablke3) can be modeled as a linear function of another variable, X (the predictor variable) with the equation.

$Y=\alpha+\beta X$

Where the variance of Y is assumed to be constant. The coefficients are known as regression coefficients; specify the Y intercept and the slope of the line respectively. The coefficients can be solved for by the method of least squares, which minimizes the error between the actual and the estimate. Multiple regressions is an extension of the linear regression.

Log-Linear models approximate discrete multidimensional probability distributions. The method can be used to estimate the probability of each cell in a base cuboids for a set of discretized attributes, based on smaller cuboids making up the data cube lattice.

Regression and log-linear models can both be used on sparse data although their application may be limited. These models will be further discussed later.

6.5 Histograms:

Histograms use binning too approximate data distributions and are a popular form of data reduction. A histogram for an attribute A partitions the data distribution of A into disjoint subsist, or buckets. The buckets are displayed on a horizontal axis, while the height (and area) typically reflects the average frequency of the values represented by the bucket. Often buckets represent continuous ranges for the given attribute.

6.6 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 another and “dissimilar” to objects in other clusters. Similarity is often defined on the basis of how close the objects are in space, based on a distance function. The “quality” of a cluster may be represented by its diameter, the maximum distance between any two objects in the cluster. In data reduction, cluster representations are used to replace the actual data. The effectiveness of this technique depends on the nature of the data.

6.7 Sampling::

Sampling can be used as a data reduction technique since it allows a larger data set to be represented by a much smaller random (or subset) of the data. Suppose a large data set D, contains N tuples some of the possible samples for D are:

- Simple random sample without replacement of size n: This created by drawing n of the N tuples from D (n<N), where the probability of drawing any tuple in D is I/N, that is all the tuples are equally likely.
- Simple random sample with replacement of size n: This is similar to the above except that each time a tuple is drawn from D, it is recorded and then replaced. That is after a tuple is drawn, it is placed back in D so that it could be drawn again.
- Cluster sample: If the tuples in D are grouped into M mutually disjount”clusters”then a simple random sample of m clusters can be obtained, where m<M. A reduced data set

representation can be obtained by applying say SRSWOR to the pages, resulting in a cluster sample of the tuples.

- Stratified sample: If D is divided into mutually disjoint parts called strata, a stratified random sample is obtained by simple random sample at each stratum.

The sampling techniques discussed above represent the most common forms of sampling for data reduction. When applied to data reduction, sampling is most commonly used to estimate the answer to and aggregate query.

6.8 Review Question

1.Expalin about Histograms, Clustering, Sampling

2 Explain about Wavelet Transforms:

3 Explain about Principal Components Analysis

4 Explain about Numinosity Reduction

5 Discuss Regression and Log-Linear models

6.9 References

[1]. Data Mining Techniques, Arun k pujari 1st Edition

[2] .Data warehousung,Data Mining and OLAP, Alex Berson ,smith.j. Stephen

[3].Data Mining Concepts and Techniques ,Jiawei Han and MichelineKamber

[4]Data Mining Introductory and Advanced topics, Margaret H Dunham PEA

[5] The Data Warehouse lifecycle toolkit , Ralph Kimball Wiley student Edition