CHAPTER-13

Mining Class Comparisons: Discrimination between DifferentClasses:

13.1 Introduction

13.2 Class Comparison Methods and Implementation

13.3 Presentation of Class Comparison Descriptions

13.4 Class Description: Presentation of Both Characterization and Comparison

13.5 Mining Descriptive Statistical Measures in Large Databases:

13.6 Measuring the Central Tendency

13.7 Quartiles, Outliers, and Boxplots

13.8 Graph Displays of Basic Statistical Class Descriptions

13.9 Review Questions

13.10 References

13 Mining Class Comparisons: Discrimination between DifferentClasses

13.1 Introduction

In many applications, users may not be interested in having a single class(or concept) described or characterized, but rather would prefer to mine a description that compares or distinguishes one class (or concept) from other comparable classes (or concepts). Class discrimination or comparison (here after referred to as class comparison) mines descriptions that distinguish a target class from its contrasting classes. Notice that the target and contrasting classes must be comparable in the sense that they share similar dimensions and attributes. For example, the three classes person, address, and item are not comparable. However, the sales in the last three years are comparable classes, and so are computer science students versus physics students.

Our discussions on class characterization in the previous sections handle Multilevel data summarization and characterization in a single class. The techniques developed can be extended to handle class comparison across several comparable classes. For example, the attribute generalization process described for class characterization can be modified so that the generalization is performed synchronously among all the classes compared. This allows the attributes in all of the classes to be generalized to the same levels of abstraction. Suppose, for instance, that we are given the ABCompany data for sales in 1998 and sales in 1999 and would like to compare

these two classes. Consider the dimension location with abstractions at the city, province_or_state, and country levels. Each class of data should be generalized to the same location level. That is/they are synchronously all generalized to either the city level, or the province_or_state level, or the country level. Ideally, this is more useful than comparing, say, the sales in Vancouver in 1998 with the sales in the India in 1999 (i.e. , where each set of sales data is generalized to a different level). The users, however, should have the option to overwrite such an automated, synchronous comparison with their own choices, when preferred.

13.2 Class Comparison Methods and Implementation

The general procedure for class comparison is as follows:

1. Data Collection: The set of relevant data in the database is collected by query Processing and is partitioned respectively into a target class and one or a set of contrasting class (es).
2. **Dimension relevance analysis**: If there are many dimensions and analytical comparisons is desired, then dimension relevance analysis should be performed on these and only the highly relevant dimensions are included in the further analysis.
3. **Synchronous generalization**: Generalization is performed on the target class to the level controlled by a user-or expert-specified dimension threshold, which

results in a prime target class relation/cuboid. The concepts in the contrasting class(es) are generalized to the same level as those in the prime target class relation/ cuboid, forming the prime contrasting class(es)relation/ cuboid.

4. **Presentation of the derived comparison**: The resulting class comparison description can be visualized in the form of tables, graphs, and rules. This presentation usually includes a "contrasting" measure (such as count)that reflects the comparisons between the target and contrasting classes. The user can adjust the comparison description by applying drill-down, roll-up, and other OLAP operations to the target and contrasting classes, as desired.

The above discussion outlines a general algorithm for mining analytical comparisons in databases. In comparison with analytical characterization, the above algorithm involves synchronous generalization of the target class with the contrasting classes so that classes are simultaneously compared at the same levels of abstraction.

"Can class comparison mining be implemented efficiently using data cube techniques?"A flag can be used to indicate whether or not a tuple represents a target or Since all of the other dimensions of the target and contrasting classes share the same portion of the cube, the synchronous generalization and specialization are realized automatically by rolling up and drilling down in the cube.

13.3Presentation of Class Comparison Descriptions

"How can class comparison descriptions be visualized?" As with class characterizations, class comparisons can be presented to the user in various forms, including generalized relations, crosstabs, bar charts, pie charts, curves, and rules. With exception of logic rules, these forms are used in the same way for characterization as for comparison. In this section, we discuss the visualization of class comparisons in the form of discriminant rules.

As is similar with characterization descriptions, the discriminative features of the target and contrasting classes of a comparison quantitatively by a quantitative discriminant rule, which associates a statistical interstingness measure, d-weight, with each generalized tuple in the description.

13.4 Class Description: Presentation of Both Characterization and Comparison

"Since class characterization and class comparison are two aspects forming a class description, can we present both in the same table or in the same rule ?" Actually, as long as we have a clear understanding of the meaning of the t-weight and d-weight measures and can interpret them correctly, there is no additional difficulty in presenting both aspects in the same table.

13.5 Mining Descriptive Statistical Measures in Large Databases:

Earlier in this chapter, we discussed class description in terms of popular Measures, such as count, sum, and average. Relational database systems provide five Built-in aggregate functions: count( ), sum( ), max( ), and min( ). These Functions can also be computed efficiently (in incremental and distributed manners) in data cubes. Thus, there is no problem in including these aggregate functions as basic measures in the descriptive mining of multidimensional data.

For many data mining tasks, however, users would like to learn more data characteristics regarding both central tendency and data dispersion . Measures of central tendency include mean, median, mode , and midrange, while measures of data dispersion include quartiles, outliers, and variance . These descriptive statistics are of great help in Understanding the distribution of the data. Such measures have been studied extensively In the statistical literature. From the data mining point of view, we need to examine how They can be computed efficiently in large multidimensional databases.

13.6 Measuring the Central Tendency

The most common and most effective numerical measure of the “center” of a set of Data is the (arithmetic)mean. Let[X], x-i, . - . , x, , be a set of n values or observations . The mean of this set of values is

$$\bar{X} = 1/n \sum_{i=1}^{n} X_i$$

This corresponds to the built-in aggregate function, average (avg ( ) in SQL), provided in relational database systems. In most data cubes, sum and count are saved in pre-computation. Thus, the derivation of average is straight forward, using the formula average=sum/ count.

Sometimes, each value a”, in a set may be associated with a weight Wj , for i= 1……n. The weights reflect the significance, importance, or occurrence frequency Attached to their respective values. In this case, we can compute

$$\bar{X} = \sum W_i X_i / \sum W_i$$

This is called the weighted arithmetic mean or the weighted average.

A measure was defined as algebraic if it can be computed from distributive aggregate measures. Since avg( ) can be computed by sum ( ) / count( ), where both sum( ) and count( ) are distributive aggregate measures in the sense that they can be Computed in a distributive manner, then avg( ) is an algebraic measure. One can verify That the eighted average is also an algebraic measure.

Although the mean is the single most useful quantity that we use to describe a set of data, it is not the only , or even always the best, way of measuring the center of a set of data. For skewed data, a better measure of the center of data is the median M is the middle value of the ordered set if the number of values n is an odd number otherwise (i.e. , if n is even), it is the average of the middle two values.

The median is neither a distributive measure nor an algebraic measure-it is a holistic measure in the sense that it cannot be computed by partitioning a set of value arbitrarily into smaller subsets, computing their medians independently, and merging the median values of each subset. On the contrary, count( ) ,sum( ),max( ),and min( ) can be Computed in this manner(being distributive measures) and are therefore easier to Compute than the median.

Although measure of central tendency is the mode. The mode for a set of data is the value that occurs most frequently in the set. It is possible for the greatest frequency to Correspond to several different values, which results in more than one mode. Data sets With one , two, or three modes are respectively called unimodal, bimodal, and trimodal. In general, a data set with two or more modes is multi modal. At the other extreme, if Each data value occurs only once , then there is no mode.

For unimodal frequency curves that are moderately skewed (asymmetrical), we have The following empirical relation:

Mean-mode=3 X (mean-median).

This implies that the mode for unimodal frequency curves that are moderately skewed can easily be computed if the mean and median values are known.

The midrange, that is, the average of the largest and smallest values in a dataset, can be used to measure the central tendency of the set of data. It is trivial to compute the midrange using the SQL aggregate functions, max( ) and min( ).

Measuring the Dispersion of Data The degree to which numeric data tend to spread is called the dispersion, or variance of the data. The most common measures of data dispersion are the five-number summary (based on quartiles), the interquartile range, and the standard deviation. The plotting of boxplots(which show outlier values) als0 serves as a useful graphical method.

13.7 Quartiles, Outliers, and Boxplots

The Kth percentile of a set of data in numerical order is the value x having the property that K percent of the data entries lie at or below X. Values at or below the median M (discussed in the previous subsection) correspond to the 50^{th} percentile.

The most commonly used percentiles other than the median are quartiles. The first quartile, denoted by Q1, is the 25^{th} percentile; the third quartile , denoted by Q3, is the 75^{th} percentile. The quartiles, including the median, give some indication of the center, spread, and shape of a distribution. The distance between the first and third quartiles is a sample measure of spread that gives the range covered by the middle half of the data. This distance is called the interquartile range(IQR) and is defined as

IQR=Q3-Q1

We should be aware that no single numerical measure of spread ,such as IQR, is very useful for describing skewed distributions. The spreads of two sides of a skewed distribution are unequal . Therefore, it is more informative to also provide the two quartiles Q1 and Q3, along with the median , M. One common rule of thumb for identifying suspected outliers is to single out values falling at least 1:5 X IQR above the third quartile or below the first quartile.

Because Q,M, and Q3 contain no information about the endpoints (e.g., tails) of the data , a fuller summary of the shape of a distribution can be obtained by providing the highest and lowest data values as well. This is known as the five-number summary. The five-number summary of ad distribution consists of the median M, the quartiles Q1 and Q3, and the smallest and the largest individual observations, written in the order Minimum,Q,M,Q3, Maximum.

A popularly used visual representation of a distribution is the boxplot. In a

boxplot: Typically, the ends of the box are at the quartiles, so that the box length is the

Interquartile range, IQR.

- A line within the box marks the median
- Two lines (called whiskers) outside the box extend to the smallest(Minimum)and largest(Maximum)Observations.

When dealing with a moderate number of observations, it is worthwhile to plot potential outliers individually. To do this in a boxplot, the whiskers are extended to the extreme high and low observations only if these values are less than 1.5 X IQR beyond the

Quartiles. Otherwise, the whiskers terminate at the most extreme observations occurring

Within 1.5 X IQR of the quartiles . The remaining cases are plotted individually .

Boxplots can be used in the comparisons of several sets of compatible data.

Based on similar reasoning as in our analysis of the we can conclude that –Q1 and

Q3 are holistic measures , as is IQR. The efficient computation of boxplots or even

Approximate boxplots is interesting regarding the mining of large data sets.

13.8 Graph Displays of Basic Statistical Class Descriptions

Aside from the bar charts, pie chrts, and line graphs discussed earlier in this

Chapter, there are also a few additional popularly used graphs for the display of data

Summaries and distributions. These include histograms, quatile plots, q-q plots, scatter Plots, and curves.

Plotting histograms, or frequency histograms, is a univariate graphical method. A Histogram consists of a set of rectangles that reflect the counts or frequencies of the Classes present in the given data. The base of each rectangle is on the horizontal axis, centred at a “class” mark, and the base length is equal to the class width. Typically, the class width is uniform, with classes being defined as the values of a categoric attribute, or equiwidth ranges of a discretized continuous attribute. In these cases, the height of each rectangle is equal to the count or relative frequency of the class it represents, and the histogram is generally referred to as a bar chart. Alternatively, ranges of no non-uniform width may define classes for a continuos attribute . In this case, for a given class, the class width is equal to the range width, and the height of the rectangle is the class density (i.e., the count or relative frequency of the class, divided by the class width).

A quartile plot is a simple and effective way to have a first look at a unvariate Data distribution. First, it displays all of the data (allowing the user to assess both the Overall behaviour and unusual occurrences). Second, it plots quartile information. The Mechanism used in this step is slightly different from the percentile computation. Let X(i), for(I=1 to n) be the data sorted in increasing order so that X(i) is the smallest Observation and X(n)is the largest. Each observation is paired with a percentage, which Indicates that approximately 100% of the data are below or equal to the value X(i). We

Say “approximately” because there may not be a value with exactly a fraction f i of the

Data below or equal to X(i). Note that the o.25 quartile corresponds to quartile Q1, the

0-50 quartile is the median, and the 0.75 quartile is Q3.

13.9 Review Questions

1 Explain about Class Comparison Methods and Implementation?

3 Expalin about Presentation of Class Comparison Descriptions

4 Discuss Class Description: Presentation of Both Characterization and

Comparison

5 Explain Mining Descriptive Statistical Measures in Large Databases:

6 Breafly explain about Quartiles, Outliers, and Boxplots?

7 Discuss Graph Displays of Basic Statistical Class Descriptions

13.10 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