CHAPTER-12

Analytical Characterization : Analysis of Attribute Relevance

12.1 Introduction

12.2 Methods of Attribute Relevance Analysis

12.3 Review Questions

12.4 References

12. Analytical Characterization : Analysis of Attribute Relevance

12.1 Introduction

“What if am not sure which attribute to include or class characterization and class comparison ? I may end up specifying too many attributes , which could slow down the: system considerably .” Measures of attribute relevance analysis can be used to help identify irrelevant or weakly relevant attributes that can be excluded from the concept description process . The incorporation of this preprocessing step into class characterization or comparison is referred to as analytical characterization or analytical comparison, respectively . This section describes a general method of attribute relevance analysis and its integration with attribute-oriented induction.

The first limitation of class characterization for multidimensional data analysis in Data warehouses and OLAP tools is the handling of complex objects . The second Limitation is the lack of an automated generalization process: the user must explicitly Tell the system which dimension should be included in the class characterization and to How high a level each dimension should be generalized . Actually , the user must specify each step of generalization or specification on any dimension.

Usually , it is not difficult for a user to instruct a data mining system regarding how high level each dimension should be generalized . For example , users can set attribute-generalization thresholds for this , or specify which level a given dimension should reach ,such as with the command “generalize dimension location to the country level”. Even without explicit user instruction , a default value such as 2 to 8 can be set by the data mining system , which would allow each dimension to be generalized to a level that contains only 2 to 8 distinct values. If the user is not satisfied with the current level of generalization, she can specify dimensions on which drill-down or roll-up operations should be applied.

It is nontrivial, howesver, for users to determine which dimensions should be included in the analysis of class characteristics. Data relations often contain 50 to 100 attributes , and a user may have little knowledge regarding which attributes or dimensions should be selected for effective data mining. A user may include too few attributes in the analysis, causing the resulting mined descriptions to be incomplete. On the other hand, a user may introduce too many attributes for analysis (e.g. , by indicating “in relevance to *”, which includes all the attributes in the specified relations).

Methods should be introduced to perform attribute (or dimension)relevance Analysis in order to filter out statistically irrelevant or weakly relevant attributes, and retain or even rank the most relevant attributes for the descriptive mining task at hand. Class characterization that includes the analysis of attribute/dimesnsion relevance is called analytical characterization. Class comparison that includes such analysis is called analytical comparison.

Intuitively, an attribute or dimension is considered highly relevant with respect to a Given class if it is likely that the values of the attribute or dimension may be used to Distinguish the class from others. For example, it is unlikely that the color of an Automobile can be used to distinguish expensive from cheap cars, but the model , make, style, and number of cylinders are likely to be more relevant attributes. Moreover, even within the same dimension, different levels of concepts may have dramatically different powers for distinguishing a class from others.

For example, in the birth_date dimension, birth_day and birth_month are unlikely to be relevant to the salary of employees. However, the birth_decade (i.e. , age interval) may be highly relevant to the salary of employees. This implies that the analysis of dimension relevance should be performed at multi-levels of abstraction, and only the most relevant levels of a dimension should be included in the analysis.

Above we said that attribute/ dimension relevance is evaluated based on the ability of the attribute/ dimension to distinguish objects of a class from others. When mining a class comparison (or discrimination), the target class and the contrasting classes are

Explicitly given in the mining query. The relevance analysis should be performed by Comparison of these classes, as we shall see below. However, when mining class Characteristics, there is only one class to be characterized. That is, no contrasting class is specified. It is therefore not obvious what the contrasting class should be for use in of comparable data in the database that excludes the set of data to be characterized. For example, to characterize graduate students, the contrasting class can be composed of the set of undergraduate students.

12.2 Methods of Attribute Relevance Analysis

There have been many studies in machine learning, statistics, fuzzy and rough set Theories, and so on , on attribute relevance analysis. The general idea behind attribute Relevance analysis is to compute some measure that is used to quantify the relevance of an attribute with respect to a given class or concept. Such measures include information gain, the Gini index, uncertainity, and correlation coefficients.

Here we introduce a method that integrates an information gain analysis technique With a dimension-based data analysis method. The resulting method removes the less informative attributes, collecting the more informative ones for use in concept description analysis.

“How does the information gain calculation work ?” Let 5 be a set of training samples, where the class label of each sample is known. Each sample is in fact a tuple. One attribute is used to determine the class of the training samples. For instance, the Attribute status can be used to define the class label of each sample as either “graduate “ or “ undergraduate “ . Suppose that there are m classes. Let S contain $S_{i;}$ samples of class C_i, for i=1,…., m. An arbitrary sample belongs to class C_i, with probability S_i/S, where S is the total number of samples in set S. The expected information needed to classify a given sample is

$$I(S_{1,} S_2, ...S_m)= - \sum(S_i/S)(\log 2)(S_i/S)$$

An attribute A with values { a_i, $a_2>\cdots>a_{v)}$ can be used to partition 5 into the Subsets { S_i $S_z,\cdots, S_v$}, where S_j contains those samples in 5 that have value a_j of A. Let S; contain S_y samples of class Q. The expected information based on this Partitioning by A is known as the entropy of A. It is the weighted average:

$$E(A)=\sum^{Y}_{j=1}(S_{i\,j}+.......+S_{m\,j}/S)\ I\ (Si\ j,.......S_{m\,j})$$

The information gain obtained by this partitioning on A is defined by

$$Gain(A)=I(S_1,S_2,.......S_m)-E(A)$$

In this approach to relevance analysis, we can compute the information gain for each of the attributes defining the samples in S. The attribute with the highest information gain is considered the most discriminating attribute of the given set. By computing the information gain for each attribute, we therefore obtain a ranking of the attributes. This ranking can be used for relevance analysis to select the attributes to be used in concept description.

Attribute relevance analysis for concept description is performed as follows:

Data Collection: Collect data for both the target class and the contrasting class by query processing. For class comparison, the user in the data-mining query provides both the target class and the contrasting class. For class characterization, the target class is the class to be characterized, whereas the contrasting class is the set of comparable data that are not in the target class.

Preliminary relevance analysis using conservative AOI: This step identifies a Set of dimensions and attributes on which the selected relevance measure is to be Applied. Since different levels of a dimension may have dramatically different Relevance with respect to a given class, each attribute defining the conceptual levels of the dimension should be included in the relevance analysis in principle. Attribute-oriented induction (AOI)can be used to perform some preliminary relevance analysis on the data by removing or generalizing attributes having a very large number of distinct values (such as name and phone#). Such attributes are unlikely to be found useful for concept description. To be conservative , the AOI performed here should employ attribute generalization thresholds that are set reasonably large so as to allow more (but not all)attributes to be considered in further relevance analysis by the selected measure (Step 3 below). The relation obtained by such an application of AOI is called the candidate relation of the mining task.

Remove irrelevant and weakly attributes using the selected relevance analysis measure: Evaluate each attribute in the candidate relation using the selected relevance analysis measure. The relevance measure used in this step may be built into the data mining system or provided by the user. For example, the information gain measure described above may be used. The attributes are then sorted(i.e., ranked)according to their computed relevance to the data mining task. Attributes that are not relevant or are weakly relevant to the task are then removed . A threshold may be set to define “weakly relevant.” This step results in an initial Target class working relation and an initial contrasting class working relation.

Generate the concept description using AOI: Perform AOI using a less Conservative set of attribute generalization thresholds. If the descriptive mining

Task is class characterization, only the initial target class working relation is included here. If the descriptive mining task is class comparison, both the initial target class working relation and the initial contrasting class working relation are included.

The complexity of this procedure is the induction process is perfomed twice, that Is, in preliminary relevance analysis (Step 2)and on the initial working relation (Step4). The statistics used in attribute relevance analysis with the selected measure (Step 3) may be collected during the scanning of the database in Step 2.

12.3 Review Questions

1Explain Analytical Characterization?

2 Methods of Attribute Relevance Analysis?

12.4 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