CHAPTER-15

Mining Multilevel Association Rules fromTransaction Databases

15.1 Multilevel Association Rules

15.2 Approaches to mining multilevel Association Rules

15.3 Checking for Redundant Multilevel Association Rules

15.4 Mining Multidimensional Association Rules from Relational Databases and Data Warehouses

15.5 multidimensional Association Rules

15.6 Mining Multidimensional Association Rules Using Static Discretization of Quantitative Attributes

15.7 Mining quantitative Association Rules

15.8 Bining:

15.9 Finding frequent predicate sets:

15.10 Mining Distance-Based Association Rules

15.11 Review Questions

11.12 References

15.Mining Multilevel Association Rules fromTransaction Databases

IN this section,you will learn methods for mining multilevel association rules,that is ,rules involving items at different levels of abstraction.Methods for checking for redundant multilevel rules are also discussed.

15.1 Multilevel Association Rules

For many applications,it is difficult to find strong associations among data items at low or primitive levels of abstraction due to the sparsity of data in multidimensional space.Strong associations discovered at high concept levels that might represent common sense knowledge.However,what may represent common sense to one user may seem novel to another.Therefore,data mining systems should provide capabilities to mine association rules at multiple levels of abstraction and traverse easily among differect abstraction spaces**.**

15.2 Approaches to mining multilevel Association Rules

"How can we mine mine multilevel association rules efficiently using concept hierarchies?" Let's look at some approaches based on a supprot-confidence framework.In general,a top-down strategy is employed,where counts are accumulated for the calculation of frequent itemsets at each concept level,starting at the concept level1 and working towards the lower,more specific concept levels,until no more frequent itemsets can be found.That is,once all frequent itemsets at concept levels,until no more frequent itemsets can be found.That is, once all frequent itemsets at concept level 1 are found,then the freuent itemsets at level2 are found,and so on,For each level,any algorithm for discovering frequent itemsets may be used,such as Apriori or its variations.

- Using uniform minimum support for all levels(referred to as uniform supprort) The same minimum support threshold is used when mining at each level of abstraction

When a uniform minimum support threshold is used,the search procedure is simplified.The method is also simple in that users are required to specify only one minimum support threshold.An optimization technuque can be adopted,based on the knowledge that an ancestor is a superset of its descendents;the search avoids examining itemsets contaiining any item whose anestors do not have minimum support.

The uniform supprot approach,however,has some difficulties.It is unlikely that items at lower levels of abstraction will occur as frequently as those at higher levels of abstraction.If the minimum support threshold is set too high,it could miss several meaningful associations occuring at low abstraction levels.If the threshold is set too low,it may generate many uninteresting associations occuring at high abstraction levels.This provides the motivation for the following approach.

- Using reduced minimum support at lower levels (referred to as reduced sup-port);Each level of abstraction has its own minimum support threshold.The lower the abstraction level,the smaller the corresponding threshold.

For mining multiple-level associations with reduced support,there are a number of alternative search strategies:

Level-by-Level independent: This is a full-breadth search,where no background knowledge of frequent itemsets is used for pruning.Each node is examined,regardless of whether or not its parent node is found to be frequent.

Level -cross-filtering by single item: An item at the ith level is examined if and only if its parent node at the (i-1)th level is frequent .In other words ,we investigate a more specific association from a more general one.If a node is frequent,its children will be examined;otherwise,its descendents are pruned from the search.

Level-cross filtering by -K-itemset: A A-itemset at the ith level is examined if and only if its corresoponding parent A-itemset at the (i-1)th level is frequent.

15.3 Checking for Redundant Multilevel Association Rules

Concept hierarchies are useful in data mining since they permit the discovery of dnowledge at different livels of abstraction,such as multilevel association rules.However,when multilevel association rules are mined,some of the rules found will be redundant due to "ancestor of "IBM desktoop computer"based on the concept hieerarchy.

15.4 Mining Multidimensional Association Rules from Relational Databases and Data Warehouses

In this section,you will learn methods for mining multidimensional association rules,that is ,rules involving more than one dimension or predicate(e.g.,rules relating what a customer buys as well as the customer's age).These methods can be organized according to their treatment of quantitative attributes.

15.5 multidimensional Association Rules

So far we have studied association rules that imply a single predicate,that is ,the predicate buys,For instance,in mining ourABC company database,we may discover the Boolean association rule"IBM desktop computer" implies "Sonyb/w printer" Which can also be written as

buys(X,"IBM desktop computer:")implies buys(X,"sony b/w printer")

where X is a variable representing customers who purchased items in ABCompany transcations .Following the terminology used in multidimensional database ,we refer to each distinct predicate in a rule as a dimensional.Hence,we can refer to the above rule as a single-dimensional or intradimension association rule since it contains a single distenct predicate(e.g..buys) with multiple occurrences.(i.e..predicate occurs more than once within the rule).Such rules are commonly mined from transactions data.Suppose,however,that rather than using a transactional database,sales and related information are stared in a relational database or data warehouse.Such data stares are multidimensional,by definition.For instance,in addition to keeping track of the items purchased in sales transactions,a relational databae may record other attributesassociated with the items,such as the quantity purchased or the price ,or the branch location of the sale.Additional relational information regarding the customers who purchased the items,such as customer ae,occupation,credit rating,income,and address,may also be stored.Considering each database attribute or warehouse dimension as a predicate,it can therfore be interesting to mine association rules containing multiple predicates,such as

age(x,"20.....29")^occupation(X,"student") implies buys(X,"laptop":)

Association rules that involve two or more dimensions or predicates can be referred to as multidimensional association rules.The above rule contains three predicates(age,occupation,and buys),each of which occurs oly once in the rule .Hence ,we say that it has no repeated predicates.Multidimensional association rules with no repeated predicates are called inter-dimensional association rules.We may also be interested in mining multidimensional association rules with repeated predicates,which contain multiple occurrences of some predicates.These rules are called hybrid.-dimension association rules.An example of such a rule is the folowing,where the predicate buys is repeated:

Age(X,"20..29")^buys(X,"laptop") implies buys(X,"sony b/w printer")

Note that database attributes can be categorical or quantitative.Categorical attributes have a finite number of possible values,with no ordering among the values(e.g., occupation,brandy color).Categorical attributes are also called nomi-nal attributes ,since their values are "names of things"Quantitative attributes are numeric and have an implicit ordering among values(e.g.,age,income,price).Technuques for mining multidimensional association rules can be categorized according to three basic approaches regarding the treatment of quantitative attributes.

In the first approach,quantitative attributes are discretezed using predefined concept hierarchies.This discretization occurs prior to mining .For instance, a con-cept hierarchy for income may be used to replace the original numeric values of this attribute by ranges,such as ."0...20k","21k...30k""31k..40k",and so on .Here,discretization is static and predetermined .The discretized numeric attributes,with their range values,can then be treated as categorical attributes(where each range is considered a category).We refer to this as mining multidi-mensional association rules using static discretization of quantitative attributes.

In the second approach,quantitative attributes are discretized into "bins"based on the distribution of the data.These bins may be further combined during the mining process.The discretization process is dynamic and established so as to satisfy some mining criteria,such as maximizing the confidence of the rules mined.Because this strategy treats the numeric attribute values as quantities rather than as predefined ranges or categories,association rules mined from this approach are also referred to as quantitative association rules.

In the third approach,quantitative attributes are discretized so as to capture the semantic meaning of such interval data.This dynamic discretization procedure considers the distance between data points.Hence ,such quantitative association rules are also referred to as distance-based association rules.

Let's study each of these approaches for mining multidimensional association rules.For simplicity,We confine our discussion to interdimensional association rules.Note that rather than searching for frequent itemsets(as is done for single-dimensional association rule mining),in multidimensional association rule mining we search for frequent"predicate set of predicates(age,occupation,buys)form the multidimensional rule is a 3-predicate set.Similar to the notation used for itemsets,we use the notation LK to refer to the set of frequent k-predicate sets

15.6 Mining Multidimensional Association Rules Using Static Discretization of Quantitative Attributes

Quantitative attributes,in this case,are discretized prior to mining using predefined concept hierarchies,where numeric values are replaced by ranges.categorical attributes may also be generalized to higher conceptual levels if desired.if the resulting task-relavant data are stored in a relational table,then the a priori algorithm requires just a slight modification so as to find all frequent predicate sets rather than frequent itemsets(i.e.,by searching throush all of the relevant attributes,instead of searching only one attribute,like buys').Finding all requent k-predicate sets will require k or k+1 scans of the table.Other stratigies,such as hashing,partitioning,and sampling may be employed to improve the performance.

Alternatively,the transformed task-relevant data may be stared in a data cube.Data cubes are well suited for the mining for multidimensional association rules,since they are multidimensional by definition.Data cubes and their computa-tion were discussed in detail in earlier.To review,a data cube consists of a lattice of cuboids that are multidimensional dta structures.These structures can hold the given task-relavant data,as well as aggregate,group-by information.

Due to the ever-increasing use of data warehousing and OLAP technology,it is possible that a data cube containing the dimensions of interest to the user may already exist,fully materialized."if this is the case,how can we go about finding the frequent predicate sets?" A strategy similar to that employed in a priori can be used,based on prior knowledge that every subset of a frequent predicate set must also be frequent.This property can be used to reduce the number of candidate predicate sets generatedl.In cases where no relevant data cube exists for the mining task,one must be created.

15.7 Mining quantitative Association Rules

quantitative association rules are multidimensional association rules in which the numeric attributes are dynamically discretized during the mining process so as to satisfy some mining creteria,such as maximizing the confidence or compactness 'of the rules mined. In this section,we will focus specifically on ow to mine quantitative association rules having two quantitative attributes on the left-hand side of the rule,and one categorical attribute on the right -hand side of the rule,for example,

Aquan1^Aquan2 implies Acat

Where Aquan1 and Aquan2 are tests on quantitative attribute ranges(where the ranges are dynamically determined),and Acat tests a categorical attribute form the task relevant data.Such rules have been referrded to as two-dimensional quantitative association rules,since they contain two quantitative dimensions.For instance,suppose you are curious about the association relationship between pairs of quantitative attributes,like customer age and income,and the type of television that customers like to buy.

"How can we find such rules"? Let's look at an approach used in a system called ARCS(Association Rule Clustering System),which borrows ideas from image processing.Essentially,this approach maps pairs of quantitative attributes onto a 2-D grid for tuples satisfying a given categorical attribute condition.The grid is then searched for clusters of points,from which the association rules are generated.The following steps are involved in ARCS:

15.8 Bining:

Quantitative attributes can have a very wide range of values denning their domain.just think about how big 2-D grid would be if we plotted age and income as axes,where each possible value of age was assigned a unique position on one axis,and similarly,ech possible value of income was assigned a unique position on the other axis! To keep grids down to a manageable size,we instead partition the ranges of quantitative attributes into intervals.These intervals are dynamic in that they may later be further combined during the mining process.The partitioning process is referred to as binning ,that is,where the intervals are considered "bins."Three common binning strategies are

- **Equiwidth binning**, where the interval size of each bin is the same,
- **Equidepth binning,**where each bin has approximately the same number of tuples assigned to it,and

- **Homogeneity-based binning,**where bin size is determined so that the tuples in each bin are uniformly distributed.

ARCS uses equiwidth binning,where the user inputs the bin size for each quantitative attribute.A 2-D array each possible bin combination involving both quantitative attributes is created.Each array cell holds the corre-sponding count distribution for each possible class of the categorical attribute of the rule right-hand side.By creating this data structure,the task-relevant data need only be scanned once.The same 2-D array can be used to generate rules for any value of the categorical attribute,based on the same two quantitative attributes.

15.9 Finding frequent predicate sets:

Once the 2-D array containing the count distribution for each category is set up this can be scanned in order to find the frequent predicate sets(those satisfying minimum support)that also satisfy minimum confidence.Strong association rules can then be generated from these predicate sets,using a rule generation algorithm.

15.10 Mining Distance-Based Association Rules

The previous section described quantitative association rules where quantitative attributes are discretized initially by methods,and the resulting intervals an then combined.Suchan approach,however,may not capture the semantics of intervals data-since they do not consider the relative distance between data points or between intervals

Table 15.1

price($)	Equiwidth (width $10)	Equidepth (depth 2)	Distance-based
7	[0,10]	[7,20]	[7,7]
20	[11,20]	[22,50]	[20,22]
22	[21,30]	[51,53]	[50,3]
50	[31,40]		
51	[41,50]		
52	[51,60]		

Consider,for example.Table 3.1 which shows data for the attribute priori partitioned according to equiwidth and equidepth binning versus a distance-based partitioning.The distance-based partitioning seems the most intuitive,since it groups values that are close together within the same interval(e.g.,[20,22]).in contrast equidepth partitioning groups distant values together(e.g.,[22,50]),Equiwidth must split values that are close together and create intervals for which there are no data clearly,a distance-based partitioning that considers the density or number of ponts an interal,as well as the "closeness'of pointsin ah-interval,helps produce a meaningful discretization .Intervals for each quantitative attribute can be established clustering the values for the attribute.

A disadvantage of association rules is that they do not allow for approximations of attribute values.Consider the folowing association rule:

Item_type(x,"electronic")^manufacturer(X,"foreign") implies price(X,200)

Where X is a variable describing items at ABCompany .In reality,it is more likely that the prices of foreign electronic items are close to or approximately $200,rather than exactly $20.It would be useful to have association rules that can express such a notion of closeness.Note that the support and confidence measures do not consider the closeness of values for a given attribute.This motivates the mining of distance-based association rules,which capture the semantics of interval data while allowing for approximation in data vallues.A two -phase algorithm can be used to mine distance-based association rules.The first phase employs clustering to find the intervals.or clusters,adapting to the amount of available memory.The second phase obtains distance-based association rules by searching for groups of clusters that occur frequently together.

15.11 Review Questions

1 Expalian about Multilevel Association Rules

2 Discuss the Approaches to mining multilevel Association Rules

3 Expalin about Mining Multidimensional Association Rules from Relational Databases and Data Warehouses

4 Discuss multidimensional Association Rules

5 Expalin Mining Multidimensional Association Rules Using Static Discretization of Quantitative Attributes

15.12 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