CHAPTER-22

Density-Based (methods) (Gred based methods)

22.Density-Based (methods) (Gred based methods)

22.1 Introduction

To discover clusters with arbitrary shape, density-based clustering methods have been developed.These typically regard clusters as dense regions of objects in the data space that are separated by regions of low density (representing noise).

The grid based clustering approach uses a multiresolution grid data structure.It quantizes the space into a finite number of cells that form the grid structure on which all operations of clustering are performed.The main advantage of this approach is its fast processing time, which is typically independent of the number of data objects,yet dependent only on the number of cells in each dimension in the quantized space.

22.2 model-Based Clustering Methods

Model -based clustering methods attempt to optimize the fit between the given data and some mathematical model. Such methods are often based on the assumption that a mixture of underlying probability distributions generates the data.Model-based clustering methods follow two major approaches: a satistical approach or a neural network approach.Examples of each approach are described in this section,

22.3 statistical Approach

conceptual clustering is a form of clustering in machine learning that,given a set of unlabeled objects,produces a classification scheme over the objects.Unlike conventional clustering, which prmarily identifies groups of like objects, conceptual clustering goes one step further by also finding characteristic descriptions for each group,where each group represents a concept or class. Hence, conceptual clustering is a two-step process: first, clustering is-performed, followed by characterization. Here, clusering quality is not solely a function of the individual objects. Rather, it incorporates factors such as the generality and simplicity of the derived concept descriptions.

Most methods of conceptual clustering adopt a statistical approach that uses probability measurements in determining the concepts or clusters.Probabilistic descriptions are typically used to represent each derived concept.

COBWEB is a popular and simple method of incremental conceptual clustering.Its input objects are described by categorical attribute-value pairs.COBWEB creates a hierarchical clustering in form of a classification tree.

COBWEB uses a heuristic evaluation measure called category utility to guide constuction of the tree. Category utility(CU) is defined as

$$\frac{\sum P(C_k)\sum_i\sum_j P(A_i = V_{ij}|C_k)^2 - \sum_i\sum_j P(A_i = V_{ij})^2}{n}$$

where n is the number of nodes, concepts , or "categories " forming a partition { c1,c2,...cn} at the given level of the tree.In other words,category utility is the increase in the expected number of attribute values that can be correctly guessed given a partition.Althrough we do not have room to show the derivation,category utility rewards intraclass similarity and interclass dissimilarity where

- **Intraclass similarity** is the probability p(Ai=vij|ck). The larger this value is , the greater the proportion of class members that share this attribute-value pair, and the more predictable the pair is of class members;

- **Interclass dissimilarity** is the probability p(ck|[Ai=Vij).The larger this value is ,the fewer the objects in contrasting classes that share this attribute-value pair,and the more predictive the pair is of the class.

let's have a look at how COBWEB works, COBWEB incrementally incorporates objects into a classification tree.

"Given a new object," you wonder,"how does COBWEB decide where to incorporate it into the classification tree?" COBWEB descends the tree along an appropriate path, updating counts along the way ,in search of the "best host" or node at which to classify the object.This decision is based on temporarily placing the object in each node and computing the category utility of the resulting partition .The placement that results in the highest category utility should be a good host for the object.

"what if the object does not really belong to any of the concepts repre-sented in the tree so far? what if it is better to create a new node for the given object?"In fact, COBWEB also computes the category utility of the partition that would result if a new node were to be created for the object.This is compared to the above computation based on the existing nodes,The object is then placed in an existing class, or a new class is created for it ,based on the partition with the heghest category utility value,Notice that COBWEB has the ability to automatically adjust the number of classes in a partition.It does need to rely on the user to provide such an input parameter.

The two operators mentioned above are highly sensitive to the input order of the object.COBWEB has two additional operators that helpmake it less sensitive to input order. There are merging and splitting. When an object is incorporated, the two best hosts are considered for merging into a singe class. Further more, COBWEB condsiders splitting the childen of the best host among the existing categories. These decisions are based on category utility. The merging and splitting operators allow COBWEB to perform a bidirectional search-for example, a merge can undo a previous split.

"What are the limitations of COBWEB has a number of limitations. First, it is based on the assumption that probability distibutions on separate attributes are statistically independent of one other. This assumption is, however, not always true since correlation between attributes often exists. Moreover, the probability distribution representation of clusters makes it quite expensive to update and store the dusters. This is especially so when the attributes have a large number of values since their time and space complexities depend not only on the number of attiributes, but also on the number of values for each attribute. Futhermore, the classification tree is not height-balanced for skewed input data, which may cause the time and space complexity to degrade dramatically.

CLASSIT is an extenion of COBWEB for incremental clustering of continuous (or real valued)date. It stores a continuous normal distribution (i.e.,mean and standard deviation) for each individual attribute in each node and uses a modified category utility measure that is an integral over continuous attributes instead of a sum over discrete attributes as an COBWEB. However, it suffers similar problems as COBWEB and thus is not suitable for clustering large database data.

In industry, AutoClass is a popular clustering method that uses Bayesian statistical analysis to estimate the number of clusters.Additional research is needed in the pplication of conceptual clustering methods to data mining.Further references are provided in the bibliographic notes.

22.4 Neural Network Approach

The neural network approach to clustering tends to represent each cluster as an exemplar.An exemplar acts as a "prototype"of the cluster and does not necessarily have to correspond to a particular data exampleor object.New objects can be distributed to the cluster whose exemplar is the most similar,based on some distance measure.The attributes of an object assigned to a cluster can be predicted from the attributes of the cluster's exemplar.

In this section,We discuss tow prominent methods of the neural network approach to clustering.The first is competitive learning ,and the second is self-organizing feature maps,both of which involve competing neural units.

Competitive learning involves a hierarchical architecture of several units(or artificial "neurons") that compete in a "winner-takes-all" fashion for the object that is currently being presented to the system .Figure 8.19 shows an example of a competitive learning system.Each circle represents a unit.The winning unit within a cluster becomes active(indicated by a filled circle.), while the others are inactive (indicated by empty circles).connections between layers are excitatory- a unit in a given layer can receive inputs from all of the units in the next lower level.The configuration of active units in layer represents the input pattern to the next higher layer.The units within a cluster at a given layer compete with one another to respond to the pattern that is outputfrom the layer below.Connections within layers are inibitory so that only one unit in any given cluster may be active.The winning unit adjusts the weights on its connections between other units in the cluster so that it will respond even more strongly to future objects that are the same or similar to the current one.If we view the weights as defining an exemplar,then new objects are assigned to the cluster with the closest exemplar.The number of clusters and the number of units per cluster are input parameters.

At the end of the clustering (or any clustering,in general),each cluster can be thought of as a new "feature" that detects some regularity in the objects.Thus,the resulting clusters can be viewed as a mapping of low-level features to higher-level features.

with self-organizing feature having several units compete for the current object also performs maps(SOMs) ,clustering.The unit whose weight vector is closest ot the input object,the weights of the winning unit are adjusted,as well as those of its nearest neighbors.SOMs assume that the units will eventually take on this structure in space.The organization of units is said to form a feature map.SOMs are believed to resemble processing that can occur in the brain and are useful for visualizing hegh-dimensional data in 2-3-D soace.

The neural network approach to clustering has strong theoretical links with actual brain processing.Further research is required in making it readily applicable to large databases due to long processing times and the intricacies of complex data.

22.5 Outlier Analysis

"what is an outlier?" Very often,there exist data objects that do not comply with the general behavior or model of the data.such data objects,which are grossly different from or inconsistent with the remaining set of data,are called outliers .

Outliers can be caused by measurement or execution error.For example, the display of a person's age as -999 could be caused by a program default setting of an unrecorded age .Alternatively,outliers may be the result of inherent data variability.The salary of the chief executive officer of a company,for instance,could naturally stand out as an outlier among the salaries of the other employees in the firm.

Many data mining algorithms try to minimize the influence of outliers or eliminate them all together.This,however,could result in the loss of important hidden information since one person's noise could hear another person's signal. In other words, the outliers themselves may be of particular interest,such as in the case of fraud detection ,where outliers may indicate fraudulent activity .Thus,outlier detection and analysis is an interesting data mining task , referred to as outlier mining .

Outlier mining has wide applications.As mentioned above,it can be used in fraud detection,for example,by detecting unusual usage of credit cards or telecommunication services.In addition it is useful in customized marketing for identifying the spending behavior of customers with extremely low or extremely high incomes ,or in medical analysis for finding unusual responses to various medical treatments.

Outlier mining can be described as follows:Given a set of n data points or objects,and k,the expected number of outliers,find the top k objects that are considerably dissimilar,exceptional,or inconsistent with respect ot the remaining data.The outlier mining problem can be viewed as two subproblems:(1) define what data can be considered as inconsistent in a given dataset,and (2) find an efficient method to mine the outliers so defined.

The problem of defining outliers is nontrival .If a regression model is used for data modeling , analysis of the residuals can give a good estimation for data "extremeness"

The task becomes tricky,however,when finding outliers in time-series data as they may be hidden in trend,seasonal,or other cyclic changes.When multidimensional data are analyzed,not any particular one,but rather a combination of dimension values may be extreme.For nonnumeric(i.e., categorical data), the definition of outliers requires special consideration.

"What about using data visualization methods/or outlier detection?" you may wonder. This may seem to be an-obvious choice,since human eyes are very fast and effective at noticing data inconsistencies.However,this does not apply to data containing cyclic plots,where apparently oul\tlying values vould be prerfectly valid values in reality.Data visualization methods are weak in detecting outliers in data with many categorical attributes or in data of high dimensionality,since human eyes are good at visualizing numeric data of only two to three dimensions.

In this section ,we instead examine computer-based methods for outlier detection.These can be categorized into three approaches:the statistical approach,the distance-based approach,and the deviation-based approach,each of which are studied here.Notice that while clustering algorithms discard outliers as noise,they can be modified to include outlier detection as a byproduct of their execution .In general ,users must chek that each outlier discovered by these approaches is indeed a "real" outlier.

22.6 Statistical-Based Outlier Detection

The satistical approach to outlier detection assumes a distribution or probability model for the given data set (e.g., a nominal distribution) and then identifies outliers

with respect to the model using a discordancy test.Application of the test requires knowledge of the data set parameters (such as the assumed data distribution),knowledge of distribution parameters(such as the mean and variance),and the expected number of outliers.

"How does the discordancy testing work?" A statistical discordancy test examines two hypotheses: a working hypothesis and an altenative hypothesis.A working hypothisis,H <is a statement that the entire data set of n objects comes from an initial distribution model,F that is,

H:Oi&f where i=1,2,....n

The hypothesis is retained if there is no statistically significant eidence supporting its rejection.A discordancy test verfies whether an objecto;is significantly large(or small) in relation to the distribution F.Different test statistics have been proposed for use as a ediscordancy test,depending on the available knowledge of the data.Assuming that some statistic T has been chosen for discordancy test ,and the value of the statistic for object o; is v; then the distribution of T is constructed.Significance probability SP(vi)=prob(T<vi) is evaluated.If some SP(vi) is sufficiently small,then oj is discordant and the working hypothesis is rejected .An alternative hypothesis,H,which states that o,comes form another distribution model F is chosen sinceo,may be an outlier under one model and a perfectly valid under another.

The alternative distribution is very important in determining the power of the test,that is, the probability that the working hypothesis is rejected when oi,is really an outlier.There are different kinds of alternative distribution.

Inherent alternative distribution:In this case,the working hypothesis that all of the objects come from distribution F is rejected in favor of the alternative hypothesis that all of the objects arise from another distribution,G:

H:oi&G, where i=1,2,..,n.

F and G may be different distribution or differ only in parameters of the same distribution.There are constraints on the form of the G distribution in that it must have potential to produce outliers.For example,it may have a different mean or dispersion,or a longer tail.

Mixture alternative distribution: The mixture altenative states that discordant values are not outliers in the F population,but contaminants from some other population,G.In this case ,the alternative hypothesis is

H:0i&(1-h)F+hG, where i=1,2,..,n.

Slippage alternative distribution: This alternative states that all of the objects(apart from some prescribed small number) arise independently from the initial model F with its given parameters,while the remaining objects are independent observations from a modified version of F in which the parameters have been shifted.

There are two basic types of procedures for detecting outliers :

Block procedures: In this case ,either all of the suspect objects are treated as outliers,or all of them are accepted as consistent.

Consecutive(or sequential) procedures:An example of such a procedure is the inside-out procedure.Its main idea is that the object that is least "likely" to be an outlier is tested first.If it is found to be an outlier,then all of the more extreme values are also considered outliers;otherwise,the next most extreme object is tested,and so on.This procedure tends to be more effective than block procedures.

"How effective is the statistical approach at outlier detection?"A major drawback is that most tests are for single attributes,yet many data mining problems require finding outliers in multidimensional space.Moreover,the statistical approach requires knowledge about parameters of the data set,such as the data distribution.

However,in many cases,the data distribution may not be known.Statistical methods do not guarantee that all outliers will be found for the cases where no specific test was developed,or the observed distribution cannot be adequately modeled with any standard distribution.

22.7 **Distance -Based outlier detection**

The notion of distance-based outliers was introdced to counter the main limitations imposed by statistical methods.

"What is a distance-based outier?" AN object o in a date set 5 is a distance-based (DB) outliers with parameter p and d, that is,DB(p,d), if at least a frastion p of the object in 5 lie at a distance greater than d from o. In other words, rather than relying on statical tests, we can think of distance-based outliers as those object whodo not have "enough" neighbors,where neighbors are defined based on distance from the given object. In comparison with statistical-based methods, distance-based outlier detection generalizes the ideas behind discordancy testing for various standard distributions. Distance-based outlier detection avoids the excessive computation that can be assoiated with fitting the observed distrition into some standard distribution and in selecting discordancy tests.

For many discordancy tests ,it can be shown that if an object o is an outlier according to the given test,then o is also a DB(p,d)outlier for some suitably defined p and d. For example,if object that lie 3 or more standard deviations from the mean are considerd to be outliers, assuming a normal distribution, then this dafinition can be generalized by a DB (0.9988,0.13a) outilers.

several efficient algorithms for mining distance-based outlier have been developed.These are outlined as follow.

22.8 Index-based algorithm:

Given a date set, the index-based algorithm uses multidmensional indexing structures, such as R-trees or k-d trees, to search for neighbors of each object o within radius d around that object. Let M be the maximum number of object within the ^- neighborhood of an outlier. Therefore,once M+1 Neighbors of objects o are found , it is clear that o is not an outlier. This algorithm has a wort-case complexity evaultion takes only the search time into account even though the task of building an indax, in itself, can be computationally intensive.

22.9 Nested-loop alogorithm:

The nested-loop aglorithm has the same computational complexity as the index-based algorith but avoids index structure construction and tries to minimize the number of I/Os. Itdivides the memory buffer space into two halves, and the date set into several logical blocks.By carefully choosing the order in which block are loaded into each half, I/O efficiency can be achieved,

22.10 cell-based algorithm:

To avoid O(n2)coputational compexity, a cell-based algorithm was developed for memory-resident data sets. Its complexity is O(cpowk+n), where c is a constant depending on the number of cell and k is the dimensionality. In this method, the data space is partitioned into cell with a side length equal to d/rootk.Each cell has two layers surrounding it. The first layer is one cell think, whike the second is [2 rootk-1] cells thick, rounded up to the closest integer. the algorithm counts outirers on a cell-by-cell rather than an object-by-object basis. For a given cell, it accumulates three counts-the number of objects in the cell, in the first layer together, and in the cell and both layers together. let's refer to thes' counts as cell_count,cell_+_l_layer_count,and cell_+_2_layers_count,respectively.

"How are outliers determined in this method?" Let M be the maximum number of cutlier that can exist in the d-neghdorhood of an outlier.

- An object O in the current cellis considered an outlier only if cell_+l_layer_count isless than or equal to M. If this condition does not hold,then all of the objects in the cell can be removed from further investigation as they cannot be outliers.

If cell_+2_layers_count is less than or equal to M,then all of the objects in the cell are considered outliers.Otherwise,if this number is more than M,then it is possible that some of the objects in the cell may be outliers.To detect these outliers,object _by _object processing is used where,for each object o in the cell,objects in the second layer of o are examined.For objects in the cell,only those objects having no more than M points in their d_neighborhoods are outliers.The d_neighborhood of any object consists of the object's cell,of -its entire first layer,and some of its second layer.

A variation to the algorithm is linear with respect to n and guarantees that no more than three passes over the data set are required .It can be used for large disk-resident data sets,yet does not scale well for high dimensions.

Distance -based outlier detection requires the user to set both the p and d parameters.Finding suitable settings for these parameters can involve much trial and error.

22.11 Deviation-Based Outlier Detection

Deviation-Based outlier detection does not use statistical tests or distance -based measures to identify exceptional objects-Instead,it identifies outliers by examing the main characteristis of objects in a group.objects that "deviate" from this description are considered outliers.Hence,in this approach the term deviations is typically used to refer-to outliers.In this section,we study two technuques for deviation-based outlier detection.The first sequentially compares objects in a set,while the second employs an OLAP data cube approach.

Sequential Exception Technuque

The sequential exception technique simulates the way in which humans can distinguish unusual objects from among a series of supposedly like objects.It uses implicit redundancy,of the data.Given a set S of n objects,it builds a sequence of subsets,(S1,S2,....Sm) of these objects with 2<m<5 n such that

sj-1 subset to sj where sj subset to s

Dissimilarities are assessed between subsets in the sequence .The technique introduces the following key terms.

- **Exception set:** This is the set of deviations or outliers.It is defined as the smallest subset of objects whose removal results in the greatest reduction of dissimilarity in the residual set**.**
- **Dissimilarity function:** This function does not require a metric distance between the objects.It is any function that,if given a set of objects,returns a low value if the objects are similar to one another.The greater the dissimilarity among the objects ,the higher the value returned by the function .The dissimilarity of a subset is incrementally computed based on the subset prior to it in the sequence.Given a subset of n numbers in the set that is ,

1/nsigma (xi-x)pow-2

where x is the mean of the n numbers in the set,For character strings,the dissimilarity function may be in the form of a patern string(e.g., containing wildcard characters) that is used to cover all of the patterns seen so far.The dissimilarity increases when the pattern covering all of the strings in sj-1 does not cover any string in sj that is not in sj-1

For interested readers,this is equivalent to the greatest reduction in kolmogorov complexity for the amount of data discarded.

Cardinality function: This is typically the count of the number of objects in a given set**.**

Smoothing factor: This is a function that is computed for each subset in the seuence.It assesses how much the dissimilarity can be reduced by removing the subset from the original set of objects.This value is scaled by the cardinality of the set.The subset whose smoothing factor value is the largest is the exception set.

The general task of finding an exception set can be NP-hard (ie., intractable).A sequential approach is computationally feasible and can be implemented using a linear algorithm.

"How does this technuque work?"Instead of assessing the dissimilarity of the current subset with respect to its complementary set ,the algorithm selects a sequence of subsets from the set for analysis.For every subset,it determines the dissimilarity difference of the subset with respect to the preceding subset in the sequence.

"Can't the order of the subsets in the sequence affect the results?" To help alleviate any possible influence of the input order on the results,the above process can be repeated several times,each with a different random ordering of the subsets.The subset with the largest smoothing factor value,among all of the iterations,becomes the exception set.

OLAP Data Cube Technique

An OLAP approach to deviation detection uses daa cubes to identify regions of anomalies in large multidimensional data.This technuque was described in detail in chapte2.For added efficiency,the deviation detection process is overlapped with cube computation.The approach is a form of discovery-driven exploration where precomputed measures indicating data exceptions are used to guide the user in data analysis,at all levels of aggregation.A cell value in the cube is considered an exception if it is significantly different from the expected value,based on a statistical model-The method uses visual cues such as bakground color to reflect the degree of exception of each cell.The user can choose to drill down on cells that are flagged as exceptions .The measure value of a cell may reflect exceptions occuring at more detailed or lower levels of the cube, where these exceptions are not visible from the current level.

The model considers variations and patterns in the measure value across all of the dimensions to which a cell belongs.For example ,suppose that you have a data cube for sales data and are viewing the sales summarized per month.With the help of the visual cues,you notice an increase in sales in December in comparison to all the months.This may seem like an exception in the time dimension .However,by drilling down on the month of December to reveal the sales per item in that month,you note that there is a similar increase in sales for other items during December.Therefore,an increase in total sales in December is not an exception.

If the item dimension is consider .The model considers exceptions hidden at all aggregated group-by's of a data cube .Manual detection of such exceptions is difficult since the search space is typically very large,particularly when there are many dimensions involving concept hierarchies with several levels.

22.12 Review Questions

1 Explain about model-Based Clustering Methods

2 Explain about statistical Approach

3 Explain about Statistical-Based Outlier Detection

4 Explain about Distance -Based outlier detection

5 Explain Nested-loop alogorithm:, cell-based algorithm:

22.13 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