ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

CLUSTER ANALYSIS: A SURVEY

Vikas Kumar*, Ankur Singh Bist
* U.P.T.U.
U.P.T.U.

ABSTRACT

Outlier detection is a fundamental issue in data mining, specifically it has been used to detect and remove anomalous objects from data mining. In this paper, we describe what Cluster Analysis is, their advantages and limitations followed by a study of clustering methods for outlier detection

KEYWORDS: Data matrix (or object-by-variable structure) & Dissimilarity matrix (or object-by-object structure) .

INTRODUCTION

WHAT IS CLUSTER ANALYSYS?

The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. A cluster of data objects can be treated collectively as one group and so may be considered as a form of data compression. Cluster analysis is an important human activity. Early in childhood, we learn how to distinguish between cats and dogs, or between animals and plants, by continuously improving subconscious clustering schemes. By automated clustering, we can identify dense and sparse regions in object space and, therefore, discover overall distribution patterns and interesting correlations among data attributes. Cluster analysis has been widely used in numerous applications, including market research, pattern recognition, data analysis, and image processing. In business, clustering can help marketers discover distinct groups in their customer bases and characterize customer groups based on purchasing patterns. In biology, it can be used to derive plant and animal taxonomies, categorize genes with similar functionality, and gain insight into structures inherent in populations. Clustering may also help in the identification of areas of similar land use in an earth observation database and in the identification of groups of houses in a city according to house type, value, and geographic location, as well as the identification of groups of automobile.

Insurance policy holders with a high average claim cost. It can also be used to help classify documents on the Web for information discovery. Clustering is also called data segmentation in some applications because clustering partitions large data sets into groups according to their similarity. Clustering can also be used for outlier detection, where outliers (values that are “far away” from any cluster) may be more interesting than common cases. Applications of outlier detection include the detection of credit card fraud and the monitoring of criminal activities in electronic commerce. For example, exceptional cases in credit card transactions, such as very expensive and frequent purchases, may be of interest as possible fraudulent activity. As a data

Mining function, cluster analysis can be used as a stand-alone tool to gain insight into the distribution of data, to observe the characteristics of each cluster, and to focus on a particular set of clusters for further analysis. Alternatively, it may serve as a preprocessing step for other algorithms, such as characterization, attribute subset selection, and classification, which would then operate on the detected clusters and the selected attributes or features.

WHAT IS THE IMPORTANCE OF ANOMALY DETECTION?

The importance of anomaly detection is due to the fact that anomalies in data translate to

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

significant, and often critical, actionable information in a wide variety of application domains. For example, an anomalous traffic pattern in a computer network could mean that a hacked computer is sending out sensitive data to an unauthorized destination. An anomalous MRI image may indicate the presence of malignant tumours. Anomalies in credit card transaction data could indicate credit card or identity theft or anomalous readings from a space craft sensor could signify a fault in some component of the space craft. Detecting outliers or anomalies in data has been studied in the statistics community as early as the 19th century. Over time, a variety of anomaly detection techniques have been developed in several research communities. Many of these techniques have been specially developed for certain application domains, while others are more generic. This project tries to provide a technique for classification based dataset.

Difficulties in Finding Outlier

The boundary between normal and anomalous behaviour is often not precise. so identifying anomalous behaviour which lies close to normal region is difficult The malicious action can be done as a normal behaviour. The malicious adversaries often adapt themselves to make the anomalous observations appear like normal, thereby making the task of defining normal behaviour more difficult.

1. In many domains normal behaviour keep evolving, so current notion normal behaviour might not be sufficiently typical in future.
2. The exact notion of outlier is different for different domain. For example, in the medical domain a small deviation from normal (e.g., fluctuations in body temperature) might be an outlier, while similar deviation in the stock market domain (e.g., fluctuations in the value of a stock) might be considered as normal. Thus applying a technique developed in one domain to another is not straightforward.
3. Availability of labeled data for training or validation of models used by outlier detection techniques is usually a major issue.
4. Often the data contains noise which tends to be similar to the actual outliers and hence is difficult to distinguish and remove.

Due to these challenges, the anomaly detection problem, in its most general form, is not easy to solve. In fact, most of the existing anomaly detection technique solves a specific formulation of the problem. The formulation is induced by various factors such as the nature of the data, availability of labeled data, type of anomalies to be detected, and so on. Often, these factor are determined by the application domain in which the anomalies need to be detected. Researchers have adopted concepts from diverse disciplines such as statistics, machine learning, data mining, information theory, spectral theory, and have applied them to specific problem formulations.

MOTIVATION

We started with static data i.e. data do not change with respect to time or we can say there was not any time essence associate with data and for that we developed a method which could only be applied only on that type of data. Therefore, that was not suitable for data streams or dynamic data i.e. data have time essence associated with itself but if we see today we have lots of domains where data changes with respect to time and requires focus. Moreover, we can't apply static method for outlier detection on dynamic data because it may not always produce right result but if we see there are not many methods developed for the dynamic data and we have many important domains like fraud detection in credit card where outlier detection for dynamic data is very challenging and important task. In fact, most of the methods are developed for static data. Applications like, in credit card fraud, we have a challenge to find out that culprit who is doing fraud, and how fast we can find him as an outlier. Even in medical diagnosis, we can use the outlier detection in such a way that anomalous behavior can be detected. Many more applications have motivated us to work with outlier detection techniques for dynamic data. Outlier detection is very challenging field, detecting an outlier as an anomalous behavior

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

in many domains, such as credit card fraud, intrusion detection, medical diagnosis, natural disaster, geographic information system, athlete performance analysis and other data mining tasks is really challenging. In all of these cases we need to detect all of the values that are different from the rest of the data. Inadequacy in methods for dynamic data motivated us to develop method for data streams and we come up with this idea of developing a method for outlier detection for dynamic data so that we can apply this method without worrying of wrong results on data streams. Another factor which also motivated us is that existing methods are slow so we needs fast method which takes less time in finding outliers.

TYPES OF DATA IN CLUSTER ANALYSIS

In this section, we study the types of data that often occur in cluster analysis and how to preprocess them for such an analysis. Suppose that a data set to be clustered contains n objects, which may represent persons, houses, documents, countries, and so on. Main memory-based clustering algorithms typically operate on either of the following two data structures.

Data matrix (or object-by-variable structure):

This represents n objects, such as persons, with p variables (also called measurements or attributes), such as age, height, weight, gender, and so on. The structure is in the form of a relational table, or n-by-p matrix (n objects _p variables):

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

Dissimilarity matrix (or object-by-object structure):

This stores a collection of proximities that are available for all pairs of n objects. It is often represented by an n-by-n table: where d (i, j) is the measured difference or dissimilarity between objects i and j. In general, d (i, j) is a nonnegative number that is close to 0 when objects i and j are highly similar or "near" each other, and becomes larger the more they differ. Since d(i, j)=d(j, i), and d(i, i)=0, we have the matrix in (7.2).Measures of dissimilarity are discussed throughout this section.

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

A CATEGORIZATION OF MAJOR CLUSTERING METHODS

Many clustering algorithms exist in the literature. It is difficult to provide a crisp categorization of clustering methods because these categories may overlap, so that a method may have features from several categories. Nevertheless, it is useful to present a relatively organized picture of the different clustering methods. In general, the major clustering methods can be classified into the following Categories .

Partitioning methods:

Given a database of n objects or data tuples, a partitioning method constructs k partitions of the data, where each partition represents a cluster and k _ n. That is, it classifies the data into k groups, which together satisfy the following requirements: (1) each group must contain at least one object, and (2) each object must belong to exactly one group. Notice that the second requirement can be relaxed in some

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

fuzzy partitioning techniques. References to such techniques are given in the bibliographic notes. Given k, the number of partitions to construct, a partitioning method creates an initial partitioning. It then uses an iterative relocation technique that attempts to improve the partitioning by moving objects from one group to another. The general criterion of a good partitioning is that objects in the same cluster are "close" or related to each other, whereas objects of different clusters are "far apart" or very different. There are various kinds of other criteria for judging the quality of partitions.

Hierarchical methods:
A hierarchical method creates a hierarchical decomposition of the given set of data objects. A hierarchical method can be classified as being either agglomerative or divisive, based on how the hierarchical decomposition is formed. The agglomerative approach, also called the bottom-up approach, starts with each object forming a separate group. It successively merges the objects or groups that are close to one another, until all of the groups are merged into one (the topmost level of the hierarchy), or until a termination condition holds. The divisive approach, also called the top-down approach, starts with all of the objects in the same cluster. In each successive iteration, a cluster is split up into smaller clusters, until eventually each object is in one cluster, or until a termination condition holds.

Density-based methods:
Most partitioning methods cluster objects based on the distance between objects. Such methods can find only spherical-shaped clusters and encounter difficulty at discovering clusters of arbitrary shapes. Other clustering methods have been developed based on the notion of density. Their general idea is to continue growing the given cluster as long as the density (number of objects or data points) in the "neighborhood" exceeds some threshold; that is, for each data point Within a given cluster, the neighborhood of a given radius has to contain at least a minimum number of points. Such a method can be used to filter out noise (outliers) and discover clusters of arbitrary shape.

Grid-based methods:
Grid-based methods quantize the object space into a finite number of cells that form a grid structure. All of the clustering operations are performed on the grid structure (i.e., on the quantized space). The main advantage of this approach is its fast processing time, which is typically independent of the number of data objects and dependent only on the number of cells in each dimension in the quantized space. STING is a typical example of a grid-based method. Wave Cluster applies wavelet transformation for clustering analysis and is both grid-based and density-based.

Model-based methods:
Model-based methods hypothesize a model for each of the clusters and find the best fit of the data to the given model. A model-based algorithm may locate clusters by constructing a density function that reflects the spatial distribution of the data points. It also leads to a way of automatically determining the number of clusters based on standard statistics, taking "noise" or outliers into account and thus yielding robust clustering methods.

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 because one

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

person's noise could be another person's signal. In other words, the outliers 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. "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 because one person's noise could be another person's signal. In other words, the outliers 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.

Statistical Distribution-Based Outlier Detection

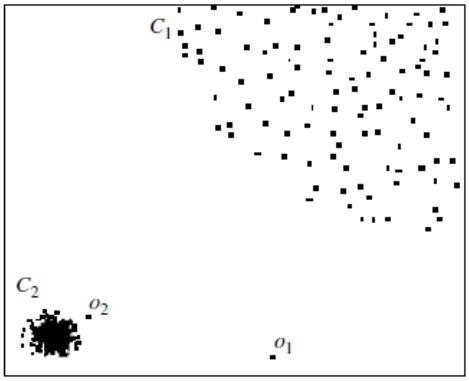
The statistical distribution-based approach to outlier detection assumes a distribution or probability model for the given data set (e.g., a normal or Poisson distribution) and then identifies outliers with respect to the model using a discordance test. Application of the test requires knowledge of the 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.

Distance-Based Outlier Detection

The notion of distance-based outliers was introduced to counter the main limitations imposed by statistical methods. An object, o, in a data set, D, is a distance-based (DB) outlier with parameters pct and dmin, 11 that is, a DB (pct;dmin)-outlier, if at least a fraction, pct, of the objects in D lie at a distance greater than dmin from o. In other words, rather than relying on statistical tests, we can think of distance-based outliers as those objects that do 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 associated with fitting the observed distribution into some standard distribution and in selecting discordancy tests.

Density-Based Local Outlier Detection

Statistical and distance-based outlier detection both depend on the overall or "global" distribution of the given set of data points, D. However, data are usually not uniformly distributed. These methods encounter difficulties when analyzing data with rather different density distributions,



The necessity of density-based local outlier analysis. From [BKNS00].

Deviation-Based Outlier Detection

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

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

CONCLUSIONS

This paper gives a brief overview about Cluster Analysis. There are lots of advancements that are going on in this specific domain. Continuous evolution in this area has added various dimensions in base atoms of concerned area. This study will be helpful for those working in the area Cluster Analysis.

REFERENCES

1. Andoni and P. Indyk. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. Communications of the ACM, 51(1):117–122, 2008
2. F. Angiulli and F. Fassetti. Very efficient mining of distance-based outliers. In M. J. Silva, A. H. F. Laender, R. A. Baeza-Yates, D. L. McGuinness, B. Olstad, Ø. H. Olsen, and A. O. Falc˜ao, editors, CIKM, pages 791–800. ACM, 2007.
3. F. Angiulli and C. Pizzuti. Fast outlier detection in high dimensional spaces. In PKDD '02: Proc. of the 6th European Conf. on Principles of Data Mining and Knowledge Discovery, pages 15–26, London, UK, 2002. Springer-Verlag.
4. S. D. Bay, D. Kibler, M. J. Pazzani, and P. Smyth. The uci kdd archive of large data sets for data mining research and experimentation. SIGKDD Explor. Newsl., 2(2):81–85, 2000.
5. S. D. Bay and M. Schwabacher. Mining distance-based outliers in near linear time with randomization and a simple pruning rule. In 9th ACM SIGKDD Int. Conf. on Knowledge Discovery on Data Mining, 2003.
6. M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. Lof: Identifying density-based localoutliers. In W. Chen, J. F. Naughton, and P. A. Bernstein, editors, Proc. of the 2000 ACM SIGMOD Int. Conf. on Management of Data, May 16-18, 2000, Dallas, Texas, USA, pages 93–104. ACM, 2000.
7. M. Ester, J. Kriegel, H. P.and Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial fatabases with noise. In In Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining. AAAI Press, 1996.
8. C. Faloutsos and K. Lin. FastMap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets. In Proceedings of the 1995 ACM SIGMOD international conference on Management of data, pages 163–174. ACM New York, NY, USA, 1995.
9. A. Ghoting, S. Parthasarathy, and M. E. Otey. Fast mining of distance-based outliers in high-dimensional datasets. 6th SIAM Int. Conf. on Data Mining, April 2005.
10. A. Ghoting, S. Parthasarathy, and M. E. Otey. Fast mining of distance-based outliers in high-dimensional datasets. Data Min. Knowl. Discov., 16(3):349–364, 2008.
11. S. Guha, R. Rastogi, and K. Shim. Cure: an efficient clustering algorithm for large databases. In SIGMOD '98: ACM SIGMOD Int. Conf. on Management of data, pages 73–84, New York, NY, USA, 1998. ACM.
12. Z. Huang. Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Min. Knowl. Discov., 2(3):283–304, 1998.
13. E. M. Knorr and R. T. Ng. Finding intensional knowledge of distance-based outliers. In VLDB '99: 25th Int. Conf. on Very Large Data Bases, pages 211–222, San

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc.

14. H. Kriegel, P. Kroger, and A. Zimek. Outlier Detection Techniques. In Tutorial at the 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2009.
15. J. Laurikkala, M. Juhola, and E. Kentala. Informal identification of outliers in medical data. In The Fifth International Workshop on Intelligent Data Analysis in Medicine and Pharmacology. Citeseer, 2000.
16. M. Mahoney and P. Chan. Learning nonstationary models of normal network traffic for detecting novel attacks. In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 376–385. ACM New York, NY, USA, 2002.
17. M. Mahoney and P. Chan. Learning rules for anomaly detection of hostile network traffic. In Proceedings of the Third IEEE International Conference on Data Mining, page 601. Citeseer, 2003.