Active learning approach to detect outliers from large data sets

Mrs. Shoba., M.E.,
Senior Asst.Prof (CSE Dept)

Rajeswari. M
M.Tech student (CSE Dept)

Christ College of Engineering and Technology
Affiliated to Pondicherry University

Abstract

Outlier detection is the data mining chore whose goal is to segregate the observations which are significantly disparate from enduring data. Outlier detection is an unusual to supervised learning methods, mainly for application in which label information is firm to accomplish or treacherous. A novel approach to outlier detection based on Active learning approach is proposed to spot outlier from the large data set. This tactic helps to discover rare classes or instruct a classifier with lower label cost. Distance based outlier detection is used to detect the outliers from the data sets using outlier weight or score. This technique can be applied to real world applications which include network intrusion, fraud detection, image processing etc.

Index Terms-Data Mining, Active Learning, Outlier Detection.

1. Introduction

Data Mining is the progression of routinely discovering useful information in large data repositories. The data warehouse is used as the source of information for knowledge of data discovery (KDD) systems through a amalgamation of artificial intelligence and statistics-related techniques to find classifications, clusters, and forecasts. The data are then “cleaned" and moved into the warehouse. Rapid advances in data gathering and storage technology have enables organizations to gather immense amounts of data. However, eliciting useful Information has proven extremely demanding. Often, traditional data analysis tools and techniques cannot be used because of the immense size of a data set. Occasionally, the non-traditional nature of the data means that traditional approaches cannot be applied even if the data set is relatively small. A ripening process within the data warehouse moves existing data into old detail data. Typically the data warehouse component architecture has three components:

Data acquisition is back-end software which extracts data from legacy systems and also from peripheral sources in turn resumes the data, and burdens them into the data warehouse. It is often referred to as the target database. The client is front-end software, which allows users and applications to entrance and analyze data in the warehouse. These three components may exist in different platforms.

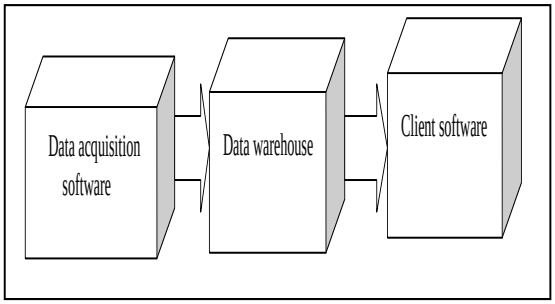

Figure 1: Components of data warehouse

KDD is the process of knowledge extraction from immense of data with the goal of obtaining meaning and therefore perceptive of the data, as well as to obtain novel data.

The process is interactive and iterative, concerning several steps with many decisions being made by the user.

Practical view of the KDD process emphasizing the interactive nature of the process outlines the following basic steps

Data Selection: Where data pertinent to the analysis task are retrieved from the database.

Data Pre-processing: To remove noise and inconsistent data which is called cleaning and integration of data that is combining multiple data sources.

Data Transformation: Where data are transformed or consolidated into appropriate form for mining by performing aggregation operations.

Data Mining: A crucial process where intelligent methods are applied in order to extract data patterns.

Pattern Evaluation: To identify the truly interesting patterns representing knowledge based on some interestingness measures.

Knowledge Presentation: Hallucination and knowledge representation techniques are used to present the mined knowledge.

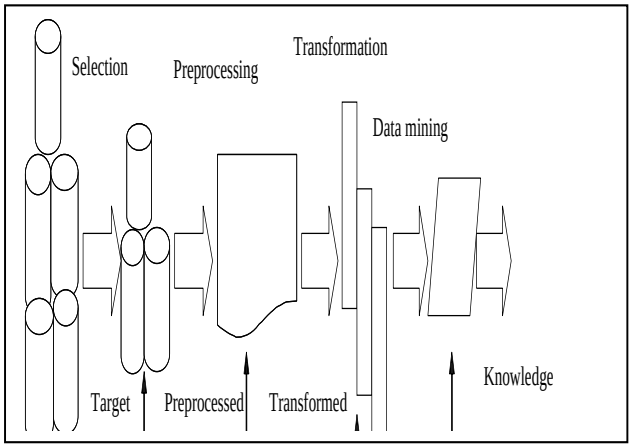

Figure 2: Data mining process

2. OUTLIER DETECTION

Outlier detection is the data mining chore whose goal is to segregate the observations which are significantly disparate from the enduring data .This task has practical applications in several domains such as deceit recognition, invasion discovery, information crackdown, therapeutic analysis [3].

Unsupervised approaches to outlier detection are able to single out each datum as normal or exceptional when no training examples are available. Along with the unsupervised approaches, distance-based methods discern an object as outlier on the basis of the distances to its nearest neighbors .These approaches fluctuate in the way the distance measure is defined, but in broad-spectrum, given a data set of objects, an object can be allied with a load or score, which is, instinctively, a function of its k nearest neighbors distances quantifying the variation of the object from its neighbors.

Many well-known data mining algorithms have intended on the assumption that data are centralized in a single memory ladder. Furthermore, these algorithms are frequently designed to be executed by a single processor.

Outlier detection denote to the problem of finding patterns in data that do not contest to predictable normal behavior. These abnormal patterns are often referred to as outliers, anomalies, harsh annotations, exceptions, faults, defects, aberrations in different

application domains. This problem has been widely researched problem and finds immense use in a wide variety of application domains such as credit card, indemnity, tax deception revealing, interruption detection for fake security, liability detection in safety critical systems, armed supervision for adversary behavior and many other areas [1].

The consequence of outlier detection is due to the fact that outliers in data translate to significant information in a wide variety of application domains. For instance, an irregular traffic pattern in a computer network could mean that a hacked computer is sending out sensitive data to an illegal destination. In civic physical condition data, outlier detection techniques are widely used to detect inconsistent patterns in patient therapeutic records which could be symptoms of a new disease. Likewise, outliers in credit card transaction data could tip out credit card robbery or abuse.

Outliers can also translate to critical entities such as in armed scrutiny, where the occurrence of a curious section in a satellite image of enemy area could indicate enemy troop movement. This has resulted in a huge and highly diverse literature of outlier detection.

2.1 Applications of Outlier Detection

There are many applications to detect the outliers from large datasets in data mining. They are of types:

Intrusion Detection

Intrusion detection refers to detection of cruel activity (break-ins, penetrations, and other forms of computer abuse) in a computer related system. These cruel activities or intrusions are very interesting from a computer security viewpoint. An intrusion is different from the normal behaviour of the system. Outlier detection techniques have been extensively applied for intrusion detection.

Fraud Detection

Fraud detection refers to detection of criminal activities occurring in commercial organizations. The malevolent users might be the genuine customers. The fraud occurs when these users consume the resources provided by the organization in an unauthorized way. The organizations are fascinated in instant detection of such frauds to avoid economic losses.

Medical and Public Health Data

Outlier detection in the medical and public health domains typically work with patient records. The data can have outliers due to several reasons such as abnormal patient condition or instrumentation errors or recording errors. Thus the outlier detection is a very critical problem in this domain and requires high degree of accuracy.

Industrial Damage Detection

Industrial unit suffer damage due to continuous usage and the normal wear and tear. Such damages need to be detected early to prevent further escalation and losses. The data in this field is typically referred to as sensor data since it is recorded using different sensors and collected for analysis. Outlier detection techniques have been extensively applied in this domain to detect such damages**.**

Image Processing

Outlier detection techniques dealing with images are either interested in any changes in an image over time in regions which appear abnormal on the static image. This field includes satellite imagery mammographic image analysis and video surveillance.

Novel Topic Detection in Text Data

Outlier detection techniques in this domain detect novel topics or events or news stories in a collection of documents or intelligence articles. The data in this field is classically high dimensional and very sparse. The data also has a temporal aspect since the documents are collected over time. The outliers are caused due to a new interesting event [3].

3. Active learning approach

Normally outlier detection has many applications such as deception discovery, interruption recognition, and therapeutic system. In the proposed system outliers are detect from the large data sets such as image using active learning approach in a remedial

application. The Distance based outlier detection is based on the colour, texture, edge to detect outlier from the large data sets. Officially, active learning studies the closed-loop event of a novice selecting actions or making queries that influence what data are added to its training set [2].

Examples embrace selecting joint angles or torques to find out the kinematics or dynamics of a robot arm, selecting locations for sensor measurements to identify and locate buried hazardous wastes, or querying a human expert to label a new document in a document classification problem.

Gaussian mixture and support vector machine classifier is used to detect the distinct object from the training data. The main objective of the project is to finding the outlier using active learning method from large data sets.

Embryonic active learning algorithms to optimize both rare class discovery and classification simultaneously is challenging because discovery and classification have conflicting requirements in query criteria [3].

We have proposed an algorithm for active learning to classify a priori undiscovered classes based on adapting two query criteria and choosing classifiers. Classifier learning in the existence of undiscovered classes was achieved by formulating a new model driven by an adaptive mixture of new class seeking and multiclass entropy maximization.

3.1 Nearest neighbour based approaches

Nearest neighbour scrutiny is an extensively used concept in machine learning and data mining in which a data object is analyzed with respect to its nearest neighbours [8]. This concept has been applied for different purposes such as classification, clustering and also outlier detection [5].

The prominent trait of nearest neighbour based outlier detection techniques is that they have an unequivocal notion of propinquity, defined in the form of a distance or similarity measure for any two individual data instances, or a set of instances or a sequence of instances.

While clustering based schemes seize an inclusive view of the data, nearest neighborbased schemes analyze each object with respect to its local neighbourhood. The basic idea behind such schemes is that an outlier will have a neighbourhood where it will rise out, while a normal object will have a neighbourhood where all its neighbours will be exactly like it. The palpable strength of these techniques is that they can work in an unsupervised mode, i.e. they do not presume availability of class labels.

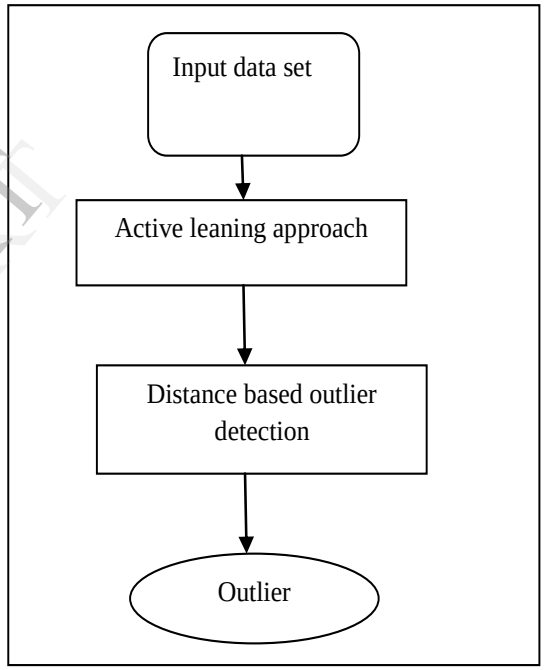

Figure 3: Active learning approach to detect outlier

4. Conclusion and Future work

In this paper, we discussed a new method for outlier detection which is particularly apposite to large data sets. The method works by finding outlier with low label cost and effort. This procedure for outlier detection has compensation over simple density based outliers which cannot triumph over the effects of the large data sets. Outlier detection has been a very important concept in the empire of data analysis. Recently, numerous application domains have realized the direct mapping flanked by outliers in data

and real world anomalies that are of immense interest to an analyst. The techniques discussed in this paper broaden the applicability of outlier detection techniques to high dimensional problems; such cases are most precious from the point of view of data mining applications.

5. References

[1] Fabrizio Angiulli, Senior Member Stefano Basta, Stefano Lodi, and Claudio Sartori "*Distributed Strategies for Mining Outliers in Large Data Sets*", IEEE Trans. Knowledge and Data Eng.,vol 25,No.7,July 2013.

[2] Timothy M. Hospedales, Member, IEEE, Shaogang Gong, and Tao Xiang "Finding Rare Classes: "*Active Learning with Generative and Discriminative Models*", IEEE Trans.Knowledge and data Eng., vol 25, February 2013.

[3] Watson Research Centre, Bianca Zadrozny Instituto de Computacao Universidade Federal Fluminense, John Langford Toyota Technological Institute at Chicago, *"Outlier Detection by Active Learning", IEEE Trans.Knowledge and Engg.*

[4] F. Angiulli, S. Basta, and C. Pizzuti, "Distance-Based Detection and Prediction of Outliers,"IEEE Trans. Knowledge and Data Eng., vol. 18, no. 2, pp. 145-160, Feb. 2006.

[5] F. Angiulli and F. Fassetti, "Dolphin: An Efficient Algorithm for Mining Distance-Based Outliers in very Large Datasets,"Trans.Knowledge Discovery from Data, vol. 3, no. 1, article 4, 2009.

[6] S.D. Bay and M. Schwabacher, "Mining Distance-Based Outliers in Near Linear Time with Randomization and a Simple Pruning Rule,"Proc. Ninth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD),2003.

[7] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly Detection: A Survey,"ACM Computing Survey, vol. 41, no. 3, pp. 15:1-15:58, 2009.

[8] J. He and J. Carbonell, "Nearest-Neighbour-Based Active Learning for Rare Category Detection," Proc. Neural Information Processing Systems, 2007.