A STUDY ON OUTLIER DETECTION FOR LARGE-SCALE DATA

S.Gowri, Dr.Anand,M.Kamarunisha,R.Jothi

Department of computer Applications,Dhanalakshmi Srinivasan College of Arts and Science for Women,,Perambalur , 621 212, Tamilnadu, , India.

Email: gowris1255@gmail.com(Gowri s) Corresponding author: Gowri s

Abstract

For addressing issues in a certain area, data mining is a crucial kind of knowledge discovery. It is also possible to describe data mining as the non-trivial process that automatically harvests the relevant hidden information from the data and takes on the forms of a rule, idea, or pattern. There are many intriguing patterns and regularities hidden in the data that may be discovered by data mining. The categorising of data for the most efficient and effective usage is known as data classification. Computer data may be categorised based on its importance or the frequency with which it is accessed. Many sectors, such as online banking and credit card transactions, rely heavily on the Outlier detection of abnormalities. For categorical datasets, only a few approaches are available to identify anomalies, whereas several methods are available to detect abnormalities in numerical datasets. Somebody has come up with a fresh way. Anomalies are discovered by analysing the most frequently occurring item sets in each record. A-priori property classification generates these outliers in datasets. To differentiate between records with the same frequency records in databases, previous approaches may not work.

Key Terms: Data Mining, Classification, Outlier Detection

1. INTRODUCTION

1.1 DATA MINING

In data mining, enormous amounts of unstructured data are sifted through in order to uncover previously unknown information. A person must be able to put the new information to use. Data mining is the process of sifting through vast amounts of data to uncover important information. Data mining is the process of sifting through vast amounts of data to identify important information. Data analysis tools and methods are used to develop models in order to uncover these patterns and correlations via data mining [1]-[5].

1.2 DATA MINING TASKS

In data mining, there are two primary types of models.

Prediction Methods

Foretell the values of other variables by predicting the values of certain variables.

Description Methods

Find patterns in the data that can be understood by humans.

1.3 DATA MINING PURPOSE

Human analysts or an automated decision support system. Use data mining in scientific and commercial fields that require to examine massive volumes of data in order to detect patterns that they would otherwise be unable to discover One of our most precious assets may be the ability to uncover significant information concealed in raw data.
Using data mining, it is possible to mine previous data and forecast the results of current and future events.

1.3.1 DATAMINING COMMERCIAL VIEWPOINT

First, data mining must be concealed from end-users in order to take centre stage in a company. Data mining methods may be used to build business use cases that are tightly constrained [6]-[10].
There is a lot of information being gathered and stored in the Figures [1] on data mining presented in this article.
Customers' purchases at department and grocery shops through e-commerce
Convenience payments made using a debit or credit card
Computers have become cheaper and more powerful over the years.
For an advantage, provide better, more personalized services (e.g. in Customer Relationship Management)

1.3.2 Mining Large Data Sets – Motivation

Machine learning, artificial intelligence, and classical statistics all have a common ancestor: data mining.
Data frequently contains information that is "hidden" or not easily apparent. Increase the number of analysts who collect data between 1995 and 1999. Four are missing.
It may take weeks for a human analyst to find meaningful information. There is a lot of data that isn't examined at all.

2. TECHNIQUES IN DATA MINING

2.1 Classification

A training set (collection of data) has a set of qualities, one of which is the class.
Find a model based on the values of other attributes for the class attribute.
An important goal is to appropriately classify previously unseen records.
To check the model's correctness, a test set is used. Data sets are often split into training and test sets, with the training set being used to develop the model, and the test set being tested.

- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- Regression [Predictive]
- Deviation Detection [Predictive]

3. DATA WAREHOUSE

Transactional data that has been formatted to facilitate queries and analysis." It's -Ralph Kimbal An integrated, time-variant and non-volatile collection of data to help management's decision-making process. Inmon, W.H. One of the most common types of data warehouses is the enterprise data warehouse (DW, DWH, or EDW) is shown in figure 1. For senior management reporting, such as yearly and quarterly comparisons, it is utilised to hold current and historical data. Data warehouses are used as a source of information for data mining methods. [2]

3.1 Data warehouse application:

- Transformation of the source system and loading (ETL)
- A store of meta data
- User feedback Data warehousing (business)
-

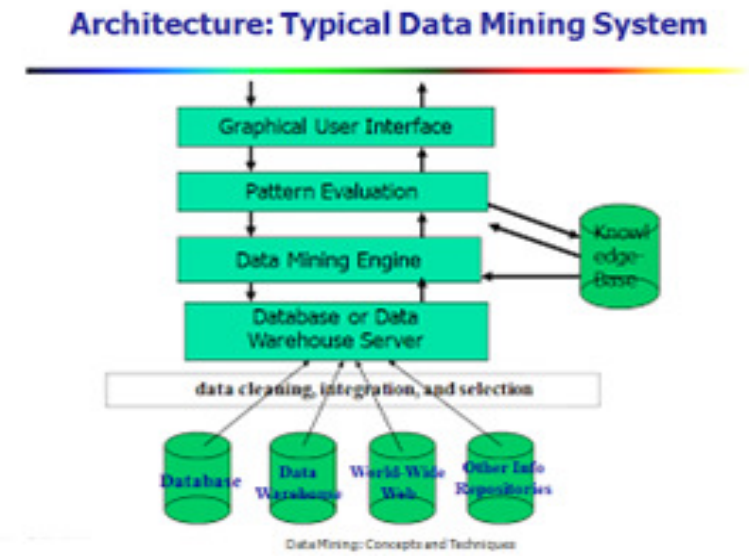

Figure 1: Typical Data Mining system

Knowledge Discovery Edge uses outlier and fraud detection in the architecture described in Figure 2 on standard data mining. [3] Some of Data Mining's best features:

- Subject-oriented
- Mon-volatile
- Integrated
- Time-variant
- There is no virtualization.
- Data warehouses include a number of important properties, including:
- Automated pattern recognition.
- Predicting what could happen.
- The generation of useful data
- Big data sets and databases should be your focus

4. Outlier

Outliers are defined as those individuals who fall outside of the norm. These anomalies occur often, and it is important to be aware of them while working with data. We use the term "outliers" to describe data points that stand out from the rest of the sample in a significant way. This kind of data mining is known as outlier mining, and it is a fascinating endeavour. [3]
As an ongoing research topic, outlier detection deals with the challenge of detecting items in a dataset that do not adhere to well-defined ideas of anticipated behaviour. Anomalies, surprises, aberrant, and other terms for the objects found in the data are all terms used to describe outliers.
Prior to the implementation of a sophisticated data analysis approach, outlier detection may be applied. Additionally, it may be used to detect interest trends, such as the spending habits of a soon-to-be insolvent credit card user. Many practical applications rely on the detection of outliers such as intrusion detection; health system monitoring; and criminal activity detection in e-commerce; and in scientific research for data analysis and knowledge discovery in a wide range of fields. Outlier detection is an essential step in these processes. In terms of techniques, supervised, semi-supervised, and unsupervised are all subcategories of one another. These methods all need some kind of training before to use, but models using the unsupervised approach don't require any training. As a result of this, a training set is necessary in a supervised method.

Density-Based Local Outlier Detection

- The need for local outlier identification based on density.
- The k-distance community
- Distance a person is able to reach
- Outliers in the nearby area (LOF)

Rule-Based Outlier Detection
The training data must be labelled before outlier detection techniques may be implemented supervised or semi-supervised. The Step-by-Step (SS) and Single-Pass (SP) approaches are two effective and efficient algorithms. Only the number of outliers is needed as an input parameter for these methods, and they fully omit the standard parameters for classifying outliers. [5]

5. CLASSIFICATION
A massive quantity of data is being gathered and stored in databases throughout the world today. Every year, the trend is to keep becoming better. Terabytes of data may be found in databases at businesses and research institutes. More than a trillion and a half bytes of data are involved. This information and knowledge is "hidden" in databases, and it is almost difficult to mine for them without the use of automated extraction techniques. Many methods have been developed over the years to extract what are known as "nuggets of knowledge" from big datasets. Some of the approaches include classification, association rule, clustering and so on. " There will be a detailed discussion of categorization in the following section. [6]

5.1 PROBLEM DESCRIPTION
Predicting a certain output from a given input is what classification is all about. A training set of traits and their corresponding outcomes, known as the objective or prediction attribute, is used to predict the result. Predictability may be achieved through discovering (discovering) correlations between the qualities. Techniques for detecting anomalies may be broken down into the following five categories: Statistics, rule-based and distance-based methodologies, as well as profiling and model-based approaches are all included in this section. .

6. CLASSIFICATION TECHNIQUES

6.1 Proximity-Based Methods
Distance-based approaches face the challenge of deciding whether to measure distance or density, and how to prevent excessive time and spatial complexity in the distance computation. To quantify distances between category items, Hamming distance and the Common-Neighbor-Based distance are used, respectively. Neighbor-set generation and outlier mining are two separate processes in the CNB method. An item that has the maximum number of outliers is designated as an outlier. In order to get the desired outcome, the proximity-based technique requires a lot of trial-and-error. The curse of dimensionality also applies to proximity-based approaches when distance or local density metrics are used on the full dimensions. These approaches take a long time and need a lot of storage space, therefore they're not recommended for huge datasets.

6.2 Rule Based Classifiers
Association-rule mining lends the idea of frequent objects to rule-based approaches. Such strategies take into account the most often occurring or least frequently occurring elements in the data collection. Outliers are things with the highest ratings. The Frequent-Item or In Frequent-Item Generating procedures dictate the temporal complexity of both techniques. As the number of characteristics increases, the temporal complexity of the FIB approach grows exponentially. Consequently, this technique is only applicable to low-dimensional datasets. When it comes to the finding of high-level classification rules, rule based classifiers focus on the form if-then. Primarily rules preceding and following rules comprise the rules. Predictor attribute values are referred to as "predictors" in the if and then sections

of the rule. The rule antecedent provides a set of requirements that must be met before the rule can be applied to an example.
Different classification techniques may be used to produce these rules, such as decision trees and sequential covering rule induction procedures.

- Common Neighbor-Based (CNB) Step-by-Step (SS) Single Pass (SP) Frequent Pattern Outlier Factor)

6.3 RECENT TRENDS IN DATAMINING:

Fuzzy Logic
In contrast to "crisp logic," where binary sets have just two values, "fuzzy logic" is a multi-valued logic. Between 0 and 1, fuzzy logic variables have a truth value. Conventional Boolean logic has been modified to include the idea of partial truth into fuzzy logic. An MF is a curve that describes how each point in the input space is assigned a membership value (or the degree of membership) from 0 to 1. Type 1 and Type 2 fuzzy logic are the two main types of fuzzy logic. The constant values of type 1 fuzzy are included. a Type-2 Fuzzy Logic is an extension of Type-1 Fuzzy Logic, in which the fuzzy sets are derived from existing Type-1 Fuzzy sets The grades of membership in a type-2 fuzzy collection are themselves fuzzy. It's possible to have a Type-2 membership grade for any main membership. [8]

7. RESULT AND DISCUSSION

Transactions in a bank's database are now very confidential. Probability-based tests are the most straightforward to assess, since most statistical tests can be translated into probabilities. Even in the setting of multivariate data outlier detection and analysis, density-based models would identify the value as the greatest outlier detect[7].

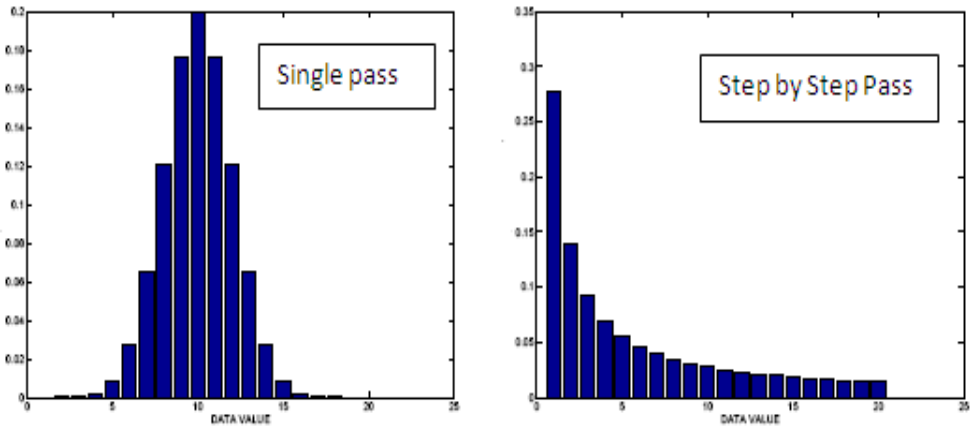

Figure 2: Analysis of Outlier

Figure 2 depicts two means of doing so: A single-pass method is referred to as SS, whereas a step-by-step method is referred to as SS (SP). Outliers are found by using both methods individually. At each SS step, the outlier item with the biggest value is discovered and deleted from the data set, as is customary. Objects are deleted one at a time till the operation is complete. In SP, the outlier factors are calculated just once, and the objects with the biggest values are designated as outliers. However, SS's outlier factors gain from the restricted search area [11-20].

8. CONCLUSION

The outliers are found by looking at the model's deviations. There is no need to estimate the distribution of the data to update an object's outlier factor. Gluttony-based approaches, such as Step by Step (SS) and single pass (SP), may be used to construct two efficient algorithms for outlier detection: SS and SP. Outliers and anomaly candidate sets should also be estimated. We may further minimize the search cost by using this constraint, which was calculated under a fairly plausible estimate about the number of probable outliers. It is important to compare data sets to various algorithms in order to improve the selection of outlier candidates.

REFERENCES

1. Lecture Notes for Chapter 1 Introduction to Data Mining By Tan, Steinbach, Kumar 4/18/2004.Paper presentation via pdf http://www.dataminingarticles.com/data-mining-introduction/data-mining-techniques.

2. Lecture Notes for Chapter 15 Data Warehousing, Introduction and Definition.

3. Data mining Concept and Technique:"JiaweiHa"n*University of Illinois at Urbana-Champaign* "Micheline Kamber." Outlier detection introduction and architecture of typical data mining.

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

5. Information-Theoretic Outlier Detection for Large-Scale Categorical Data Shu Wu, Member, IEEE, and Shengrui Wang, Member, IEEE Transactions on knowledge and data engineering, vol. 25, no. 3, march 2013

6. data mining classification fabriciovoznikaleonardoviana.http://courses.cs.washington.edu/courses/csep521/07wi/prj/leonardo_fabricio.pdf

7. Outlier Analysis Authored by Charu c. Aggarwal ibm t. j. Watson Research Center, Yorktown Heights, NY, USA Kluwer Academic Publishers Boston/Dordrecht/London.

8. S.Kannadhasan, G.Karthikeyan and V.Sethupathi, A Graph Theory Based Energy Efficient Clustering Techniques in Wireless Sensor Networks. Information and Communication Technologies Organized by Noorul Islam University (ICT 2013) Nagercoil on 11-12 April 2013, Published for Conference Proceedings by IEEE Explore Digital Library 978-1-4673-5758-6/13 @2013 IEEE.

9. S.Kannadhasan, M.Shanmuganantham and R.Nagarajan, System Model of VANET Using Optimization-Based Efficient Routing Algorithm, International Conference on Advances in Material Science, Communication and Microelectronics (ICAMCM-2021), Jaipur Engineering College and Research Centre, Jaipur, 19-20 February 2021. Published for IOP Conference Series: Materials Science and Engineering, Vol No: 1119, 2021, doi:10.1088/1757-899X/1119/1/012021

10. Classification and Feature Selection Techniques in Data MiningInternational Journal of Engineering Research & Technology (IJERT)Vol. 1 Issue 6, August – 2012ISSN: 2278-0181 ww.ijert.

[11]singh, d., buddhi, d., & karthick, a. (2022). productivity enhancement of solar still through heat transfer enhancement techniques in latent heat storage system: a review. environmental science and pollution research, 1-34.

[12]haseena, s., saroja, s., madavan, r., karthick, a., pant, b., & kifetew, m. (2022). prediction of the age and gender based on human face images based on deep learning algorithm. computational and mathematical methods in medicine, 2022.

[13]jasti, v., kumar, g. k., kumar, m. s., maheshwari, v., jayagopal, p., pant, b., ... & muhibbullah, m. (2022). relevant-based feature ranking (rbfr) method for text classification based on machine learning algorithm. journal of nanomaterials, 2022.

[14]babu, j. c., kumar, m. s., jayagopal, p., sathishkumar, v. e., rajendran, s., kumar, s., ... & mahseena, a. m. (2022). iot-based intelligent system for internal crack detection in building blocks. journal of nanomaterials, 2022.

[15]chidambaram, s., ganesh, s. s., karthick, a., jayagopal, p., balachander, b., & manoharan, s. (2022). diagnosing breast cancer based on the adaptive neuro-fuzzy inference system. computational and mathematical methods in medicine, 2022.

[16]saroja, s., madavan, r., haseena, s., pepsi, m., karthick, a., mohanavel, v., & muhibbullah, m. (2022). human centered decision-making for covid-19 testing center location selection: tamil nadu—a case study. computational and mathematical methods in medicine, 2022.

[17]kumar, r. r., thanigaivel, s., priya, a. k., karthick, a., malla, c., jayaraman, p., ... & karami, a. m. (2022). fabrication of mno2 nanocomposite on go functionalized with advanced electrode material for supercapacitors. journal of nanomaterials, 2022.
[18]karthick, a., mohanavel, v., chinnaiyan, v. k., karpagam, j., baranilingesan, i., & rajkumar, s. (2022). state of charge prediction of battery management system for electric vehicles. in active electrical distribution network (pp. 163-180). academic press.
[19]bharathwaaj, r., mohanavel, v., karthick, a., vasanthaseelan, s., ravichandran, m., sakthi, t., & rajkumar, s. (2022). modeling of permanent magnet synchronous motor for zero-emission vehicles. in active electrical distribution network (pp. 121-144). academic press.
[20]jayalakshmi, y., subramaniam, u., baranilingesan, i., karthick, a., rahim, r., & ghosh, a. (2021). novel multi-time scale deep learning algorithm for solar irradiance forecasting. energies 2021, 14, 2404.