ISSN:0975-9646

Anomaly Detection and SQL Prepare Data Sets for Data Mining Analysis.

P.V.Balakrishna, M.Tech (CSE)
Vemu Institute of Technology,
P.Kothakota, Chittoor Dist, A.P, India

.**B.Rama Ganesh**, M.Tech, (Phd)
Associate professor CSE Dept, Vemu Institute of Technology,
P.Kothakota, Chittoor Dist, A.P, India.

Abstract-- Anomaly detection has been an important research topic in data mining and machine learning. Many real-world applications such as intrusion or credit card fraud detection require an effective and efficient framework to identify deviated data instances. However, most anomaly detection methods are typically implemented in batch mode, and thus cannot be easily extended to large-scale problems without sacrificing computation and memory requirements. In this paper, we propose an online oversampling principal component analysis (osPCA) algorithm to address this problem, and we aim at detecting the presence of outliers from a large amount of data via an online updating technique. Unlike prior principal component analysis (PCA)-based approaches, we do not store the entire data matrix or covariance matrix, and thus our approach is especially of interest in online or large-scale problems Preparing a data set for analysis is generally the most time consuming task in a data mining project, requiring many complex SQL queries, joining tables, and aggregating columns. Existing SQL aggregations have limitations to prepare data sets because they return one column per aggregated group. In general, a significant manual effort is required to build data sets, where a horizontal layout is required. We propose simple, yet powerful, methods to generate SQL code to return aggregated columns in a horizontal tabular layout, returning a set of numbers instead of one number per row. This new class of functions is called horizontal aggregations. Horizontal aggregations build data sets with a horizontal denormalized layout (e.g., point-dimension, observation variable, instance-feature), which is the standard layout required by most data mining algorithms. We propose three fundamental methods to evaluate horizontal aggregations: CASE: Exploiting the programming CASE construct; SPJ: Based on standard relational algebra operators (SPJ queries); PIVOT: Using the PIVOT operator, which is offered by some DBMSs.

Key words: SQL Code generation, Initial data Analysis, Characteristics of data sample, Main data Analysis, Properties.

INTRODUCTION

ANOMALY (or outlier) detection aims to identify a small group of instances which deviate remarkably from the existing data. A well-known definition of "outlier" is given in "an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism," which gives the general idea of an outlier and motivates many anomaly detection methods. Practically, anomaly detection can be found in applications such as homeland security, credit card fraud detection, intrusion and insider threat detection in cyber-security, fault detection, or malignant diagnosis. However, since only a limited amount of labeled data are available in the above real world applications, how to determine anomaly of unseen data (or events) draws attention from the researchers in data mining and machine learning communities Index Terms—Anomaly detection, Fuzzy logic based method, an anomaly based intrusion detection system

Horizontal aggregation can be performing by using operator, it can easily be implemented inside a query processor, much like a select, project and join. PIVOT operator on tabular data that exchange rows, enable data transformations useful in data modeling, data analysis, and data presentation there are many existing functions and operators for aggregation in Structured Query Language. The most commonly used aggregation is the sum of a column and other aggregation operators return the average, maximum, minimum or row count over groups of rows. All operations for aggregation have many limitations to build large data sets for data mining purposes. Database schemas are also highly normalized for On-Line Transaction Processing (OLTP) systems where data sets that are stored in a relational database or data warehouse.

1. SQL CODE GENERATION:

Our main goal is to define a template to generate SQL code combining aggregation and transposition (pivoting). A

Second goal is to extend the SELECT statement with a clause that combines transposition with aggregation. Consider

the following GROUP BY query in standard SQL that takes a subset L1; . . . ; Lm from D1; . . .; Dp:

SELECT L1; :::; Lm, sum (A)
FROM F
GROUP BY L1; . . . ; Lm;

This aggregation query will produce a wide table with m þ 1 columns (automatically determined), with one group for each unique combination of values L1; . . . ; Lm and one aggregated value per group (sum (A) in this case). In order to evaluate this query the query optimizer takes three input parameters: 1) the input table F, 2) the list of grouping columns L1; . . . ; Lm, 3) the column to aggregate (A). The basic goal of a horizontal aggregation is to transpose (pivot) the aggregated column A by a column subset of L1; . . . ; Lm; for simplicity assume such subset is R1; . . .;Rk where k < m. In other words, we partition the GROUP BY list into two sublists: one list to produce each group (j columns L1; . . . ; Lj) and another list (k columns R1; . . .; Rk) to transpose aggregated values, where fL1; . . . ; Ljg \ fR1; . . .; Rkg ¼ ;.

Each distinct combination of fR1; . . .; Rkg will automatically produce an output column. In particular, if k ¼ 1 then there
are j_R1 ðFÞj columns (i.e., each value in R1 becomes a column storing one aggregation). Therefore, in a horizontal aggregation there are four input parameters to generate SQL code: 1. the input table F, 2. the list of GROUP BY columns L1; . . . ; Lj, 3. the column to aggregate (A), 4. the list of transposing columns R1; . . .; Rk. Horizontal aggregations preserve evaluation semantics of standard (vertical) SQL aggregations. The main difference will be returning a table with a horizontal layout, possibly having extra nulls. The SQL code generation aspect is

Example
In the Fig.1 there is a common field K in F1 and F2.In F2, D2 consist of only two distinct values X and Y and is used to transpose the table. The aggregate operation is used in this is sum (). The values within D1 are repeated, 1
appears 3 times, for row 3, 4 and, and for row 3 & 4 value of D2 is X & Y. So D2X and D2Y is newly generated columns in FH.

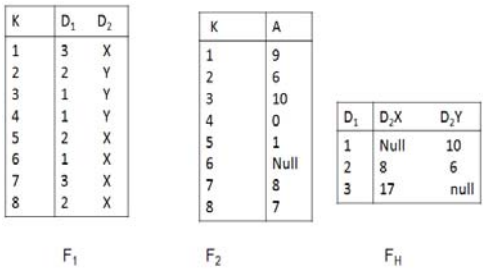


Fig 1.An example of Horizontal aggregation

Commonly using Query Evaluation methods in Horizontal aggregation functions .**Data mining** (the analysis step of the "Knowledge Discovery in Databases" process, or KDD), a field at the intersection of computer science and statistics, is the process that attempts to discover patterns in large data sets. It utilizes methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data preprocessing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.
The term is a buzzword, and is frequently misused to mean any form of large-scale data or information processing (collection, extraction, warehousing, analysis, and statistics) but is also generalized to any kind of computer decision support system, including artificial intelligence, machine learning, and business intelligence. In the proper use of the word, the key term is *discovery*, commonly defined as "detecting something new". Even the popular book "Data mining: Practical machine learning tools and techniques with Java" (which covers mostly machine learning material) was originally to be named just "Practical machine learning", and the term "data mining" was only added for marketing reasons. Often the more general terms "(large scale) data analysis", or "analytics" – or when referring to actual methods, artificial intelligence and machine learning – are more appropriate.
The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection) and dependencies (association rule mining). This usually involves using database techniques such as spatial indexes. These patterns can then be seen as a kind of summary of the input data, and may be used in further *the security expenditures are seen as wasteful because success is too invisible*". However, Schneier assures one that, despite the lack of visible results, the need to secure information still exists. Active attacks attempt to modify system resources or network functionality. Examples of these attacks are message modification, message replay, impersonation and denial of service attacks.analysis or, for example, in machine learning and predictive analytics. For example, the data mining step might identify multiple groups in the data, which can then be used to obtain more accurate prediction results by a decision support system. Neither the data collection, data preparation, nor result interpretation and reporting are part of the data mining step, but do belong to the overall KDD process as additional steps.

2. INITIAL DATA ANALYSIS

The most important distinction between the initial data analysis phase and the main analysis phase, is that during initial data analysis one refrains from any analysis that are aimed at answering the original research question. The initial data analysis phase is guided by the following four questions:

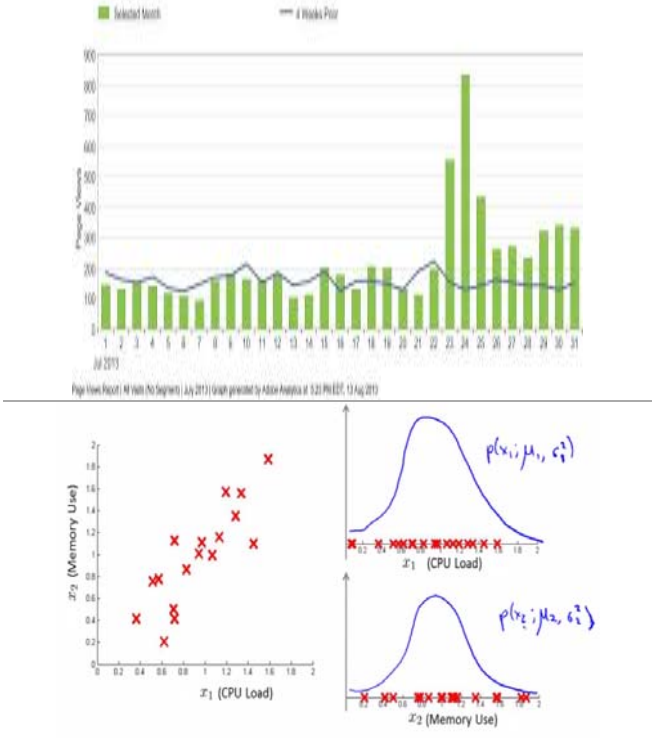
a. Anomaly detection (or **outlier detection**):

The identification of items, events or observations which do not conform to an expected pattern or other items in a dataset. Typically the anomalous items will translate to some kind of problem such as bank fraud, a structural defect, medical problems or finding errors in text. Anomalies are also referred to as outliers, novelties, noise, deviations and exceptions.
In particular in the context of abuse and network intrusion detection, the interesting objects are often not *rare* objects, but unexpected *bursts* in activity. This pattern does not adhere to the common statistical definition of an outlier as a rare object, and many outlier detection methods (in particular unsupervised methods) will fail on such data, unless it has been aggregated appropriately. Instead, a cluster analysis algorithm may be able to detect the micro clusters formed by these patterns.
Three broad categories of anomaly detection techniques exist. **Unsupervised anomaly detection** techniques detect anomalies in an unlabeled test data set under the assumption that the majority of the instances in the data set are normal by looking for instances that seem to fit least to the remainder of the data set. **Supervised anomaly detection** techniques require a data set that has been labeled as "normal" and "abnormal" and involves training a

classifier (the key difference to many other statistical classification problems is the inherent unbalanced nature of outlier detection). **Semi-supervised anomaly detection** techniques construct a model representing normal behavior from a given *normal* training data set, and then testing the likelihood of a test instance to be generated by the learnt model

B.Fuzzy Logic based method'

In recent years fuzzy logic has been adopted in several outlier detection approaches in order to improve results coming from popular outlier detection techniques. Yousri et al. proposed an approach based on fuzzy logic in order to merge results obtained with an outlier detection method, which establishes if a pattern is an outlier and a clustering algorithm which provides results in allocating patterns to clusters. The two results, provided by the two different approaches, are then combined to give a meaure of outlierness. Another fuzzy based approach is proposed by Xue et al. This approach, called Fuzzy Rough Semi Supervised Outlier Detection (FRSSOD), is the combination of two approaches: the Semi-Supervised Outlier Detection method (SSOD) and the Fuzzy Rough C-means clustering (FRCM). The objective of the FRSSOD method is to establish if a pattern under consideration can be considered as an outlier. Finally another fuzzy-based method, called Fuzzy Combination of Outlier Detection Techniques (FUCOT) combines different popular outlier detection methods exploiting the advantages of each of them while overcoming their drawbacks.



An Anomaly-Based System,: a system for detecting computer intrusions and misuse by monitoring system activity and classifying it as either *normal* or *anomalous*. The classification is based on heuristics or rules, rather than patterns or signatures, and attempts to detect any type of misuse that falls out of normal system operation. This is as opposed to signature based systems which can only detect attacks for which a signature has previously been created.

In order to determine what attack traffic is, the system must be taught to recognize normal system activity. This can be accomplished in several ways, most often with artificial intelligence type techniques. Systems using neural networks have been used to great effect. Another method is to define what normal usage of the system comprises using a strict mathematical model, and flag any deviation from this as an attack. This is known as strict anomaly detection. Anomaly-based Intrusion Detection does have some short-comings, namely a high false positive rate and the ability to be fooled by a correctly delivered attack. Attempts have been made to address these issues through techniques used by PAYL and MCPAD.

c. Quality of data

The quality of the data should be checked as early as possible. Data quality can be assessed in several ways, using different types of analyses: frequency counts, descriptive statistics (mean, standard deviation, and median), normality (skewness, kurtosis, frequency histograms, normal probability plots), associations (correlations, scatter plots).

Other initial data quality checks are:

- Checks on data cleaning: have decisions influenced the distribution of the variables? The distribution of the variables before data cleaning is compared to the distribution of the variables after data cleaning to see whether data cleaning has had unwanted effects on the data.
- Analysis of missing observations: are there many missing values, and are the values missing at random? The missing observations in the data are analyzed to see whether more than 25% of the values are missing, whether they are missing at random (MAR), and whether some form of imputation is needed.

D.Quality of measurements

The quality of the measurement instruments should only be checked during the initial data analysis phase when this is not the focus or research question of the study. One should check whether structure of measurement instruments corresponds to structure reported in the literature. There are two ways to assess measurement quality:

- Confirmatory factor analysis
- Analysis of homogeneity (internal consistency), which gives an indication of the reliability of a measurement instrument. During this analysis, one inspects the variances of the items and the scales, the Cronbach's α of the scales, and the change in the Cronbach's alpha when an item would be deleted from a scale.
- Initial transformations

After assessing the quality of the data and of the measurements, one might decide to impute missing data, or to perform initial transformations of one or more variables, although this can also be done during the main analysis phase.

Possible transformations of variables are:

- Square root transformation (if the distribution differs moderately from normal)
- Log-transformation (if the distribution differs substantially from normal)
- Inverse transformation (if the distribution differs severely from normal)
- Make categorical (ordinal / dichotomous) (if the distribution differs severely from normal, and no transformations help)

3. CHARACTERISTICS OF DATA SAMPLE

In any report or article, the structure of the sample must be accurately described. It is especially important to exactly determine the structure of the sample (and specifically the size of the subgroups) when subgroup analyses will be performed during the main analysis phase. The characteristics of the data sample can be assessed by looking at:

- Basic statistics of important variables
- Scatter plots
- Correlations
- Cross-tabulations

a. Final stage of the initial data analysis

During the final stage, the findings of the initial data analysis are documented, and necessary, preferable, and possible corrective actions are taken. Also, the original plan for the main data analyses can and should be specified in more detail and/or rewritten. In order to do this, several decisions about the main data analyses can and should be made:

- In the case of non-normals: should one transform variables; make variables categorical (ordinal/dichotomous); adapt the analysis method?
- In the case of missing data: should one neglect or impute the missing data; which imputation technique should be used?
- In the case of outliers: should one use robust analysis techniques?
- In case items do not fit the scale: should one adapt the measurement instrument by omitting items, or rather ensure comparability with other (uses of the) measurement instrument(s)?

b.Analysis

Several analyses can be used during the initial data analysis phase:

- Univariate statistics
- Bivariate associations (correlations)
- Graphical techniques (scatter plots)

It is important to take the measurement levels of the variables into account for the analyses, as special statistical techniques are available for each level:

- Nominal and ordinal variables
    - Frequency counts (numbers and percentages)
    - Associations
        - circumambulations (cross tabulations)
        - hierarchical loglinear analysis (restricted to a maximum of 8 variables)
        - log linear analysis (to identify relevant/important variables and possible confounders)
    - Exact tests or bootstrapping (in case subgroups are small)
    - Computation of new variables
- Continuous variables
    - Distribution
        - Statistics (M, SD, variance, skewness, kurtosis)
        - Stem-and-leaf displays
        - Box plots

4. MAIN DATA ANALYSIS

In the main analysis phase analyses aimed at answering the research question are performed as well as any other relevant analysis needed to write the first draft of the research report.

a. Exploratory and confirmatory approaches

In the main analysis phase either an exploratory or confirmatory approach can be adopted. Usually the approach is decided before data is collected. In an exploratory analysis no clear hypothesis is stated before analyzing the data, and the data is searched for models that describe the data well. In a confirmatory analysis clear hypotheses about the data are tested.

Exploratory data analysis should be interpreted carefully. When testing multiple models at once there is a high chance on finding at least one of them to be significant, but this can be due to a type 1 error. It is important to always adjust the significance level when testing multiple models with, for example, a bonferroni correction. Also, one should not follow up an exploratory analysis with a confirmatory analysis in the same dataset. An exploratory analysis is used to find ideas for a theory, but not to test that theory as well. When a model is found exploratory in a dataset, then following up that analysis with a confirmatory analysis in the same dataset could simply mean that the results of the confirmatory analysis are due to the same type 1 error that resulted in the exploratory model in the first place. The confirmatory analysis therefore will not be more informative than the original exploratory analysis.

A **data set** (or **dataset**) is a collection of data, usually presented in tabular form. Each column represents a particular variable. Each row corresponds to a given member of the data set in question. It lists values for each of the variables, such as height and weight of an object. Each value is known as a datum. The data set may comprise data for one or more members, corresponding to the number of rows.

Nontabular data sets can take the form of marked up strings of characters, such as an XML file.

5. PROPERTIES

A data set has several characteristics which define its structure and properties. These include the number and types of the attributes or variables and the various statistical

measures which may be applied to them such as standard deviation and kurtosis.

In the simplest case, there is only one variable, and then the data set consist of a single column of values, often represented as a list. In spite of the name, such a univariate data set is not a set in the usual mathematical sense, since a given value may occur multiple times. Usually the order does not matter, and then the collection of values may be considered to be a multiset rather than an (ordered) list. The values may be numbers, such as real numbers or integers, for example representing a person's height in centimeters, but may also be nominal data (i.e., not consisting of numerical values), for example representing a person's ethnicity. More generally, values may be of any of the kinds described as a level of measurement. For each variable, the values will normally all be of the same kind. However, there may also be "missing values", which need to be indicated in some way. In statistics, data sets usually come from actual observations obtained by sampling a statistical population, and each row corresponds to the observations on one element of that population. Data sets may further be generated by algorithms for the purpose of testing certain kinds of software. Some modern statistical analysis software such as PSPP still present their data in the classical data set fashion

a. Classic data sets

Several classic data sets have been used extensively in the statistical literature:

- Iris flower data set - multivariate data set introduced by Ronald Fisher (1936)
- *Categorical data analysis* - Data sets used in the book, *An Introduction to Categorical Data Analysis*, by Agresti are provided on-line by StatLib.
- *Robust statistics* - Data sets used in *Robust Regression and Outlier Detection* (Rousseeuw and Leroy, 1986). Provided on-line at the University of Cologne.
- *Time series* - Data used in Chatfield's book, *The Analysis of Time Series*, are provided on-line by StatLib.
- *Extreme values* - Data used in the book, *An Introduction to the Statistical Modeling of Extreme Values* are provided on-line by Stuart Coles, the book's author. **[Dead link]**
- *Bayesian Data Analysis* - Data used in the book are provided on-line by Andrew Gelman, one of the book's authors.
- The Bupa liver data, used in several papers in the machine learning (data mining) literature.
- Anscombe's quartet Small dataset illustrating the importance of graphing the data to avoid statistical fallacies

CONCLUSION:

Future research will be directed to the following maly detection scenarios: normal data with multiclustering Structure, and data in a extremely high dimensional space. For the former case, it is typically not easy to use linear models such as PCA to estimate the data distribution if there exists multiple data clusters. Moreover, many learning algorithms encounter the “curse of dimensionality” problem in a extremely high-dimensional space. Optimized k-means is significantly faster because of small data set run clustering outside the DBMS.Input to the system is data from multiple tables rather than single table used in traditional horizontal aggregation. Include Euclidean distance computation, pivoting a table to have one dimension value per row. Data manipulating operator Pivot is easy to compute for wide set of values. Pivot is an extension of Group By with unique restrictions and optimization opportunities, and this makes it easy to introduce incrementally on top of existing grouping implementation

REFERENCES

[1] C. Cunningham, G. Graefe, and C.A. Galindo-Legaria. PIVOT and UNPIVOT: Optimization and execution strategies in an RDBMS. In *Proc. VLDB Conference*, pages 998–1009, 2004.

[2] H. Wang, C. Zaniolo, and C.R. Luo. ATLaS: A small but complete SQL extension for data mining and data streams. In *Proc. VLDB Conference*, pages 1113–1116, 2003.

[3] S. Sarawagi, S. Thomas, and R. Agrawal. Integrating association rule mining with relational database systems: alternatives and implications. In *Proc. ACM SIGMOD Conference*, pages 343–354, 1998.

[4] C. Ordonez. Integrating K-means clustering with a relational DBMS using SQL. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 18(2):188–201, 2006.

[5] Perdisci, Roberto; Davide Ariu; Prahlad Fogla; Giorgio Giacinto; Wenke Lee (2009).

[6] Tomek, Ivan (1976). "An Experiment with the Edited Nearest-Neighbor Rule". *IEEE Transactions on Systems, Man, and Cybernetics* .

[7] Yousri, N.A.; Ismal, M.A. Kamel, M.S. "Fuzzy Outlier Analysis a combined Clustering-outlier Detection Approach"

[8] Gao, J.; Cheng, H.B. Tan, P.N. "Semi-supervised outlier detection"

[9] Teng, H. S.; Chen, K.; Lu, S. C. (1990). "Adaptive real-time anomaly detection using inductively generated sequential patterns"