International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 5 | ISSN : 2456-3307

Effective Clustering Approach to Discover Outliers in Voluminous Database using Clustering Approach

M. Veena*[1], Dr. A. Nagarajan[2],
[1]M.Phil (SSP) Research Scholar, Department of Computer Applications, Alagappa University, Karaikudi, Tamil Nadu, India
[2]Assistant Professor, Department of Computer Applications, Alagappa University, Karaikudi, Tamil Nadu, India

ABSTRACT

In recent advancements of the internet have changed the business scenario and plethora of data is available for the decision making and to improve business transactions. Mining interesting data related to server logs is an evolving area and web usage mining caters to the need of the website positioning and marketing strategies. The web server creates log files regarding details about the page, IP address of the user, browser, and operating system used and time/date stamp regarding browsing patterns and this data is mined to extract useful information using web usage mining. The primary objective of this research work is to discover the low hit pages of a website from the log files. The research work mainly focuses on a new technique to find the browsing patterns or the navigational behavior of the users after mining the content of the server log files using a hybrid data mining techniques. The proposed algorithm uses sequential frequent item set mining technique. The proposed methods are evaluated using voluminous benchmarked datasets and from the experimental results, it is founds that the proposed methods are superior with respect to the accuracy, time consumption, memory usage and precision than the existing state of the art algorithms.

Keywords: Web Mining, Sequential Pattern Algorithm, Browsing patterns, Server Log Files

I. INTRODUCTION

Data mining and knowledge discovery from datasets has obtained huge interest in recent years lately. Data mining, the extraction of concealed prognostic information from large databases, is a dominant new technology with immense potential to assist and support companies focus on the most vital information in their data warehouses to improve their business. Knowledge Discovery in Databases (KDD) is an imperative process of identifying valid interesting, previously unidentified but potentially valuable patterns in data. Association rules mining (ARM) is one of the most commonly used techniques in data mining and knowledge discovery and has a wide range of applications related to business, science and many other domains. The ARM is used to arrive at the decisions about plethora of marketing activities like up selling, product placements etc. To discover frequent pattern item set from a given large dataset is define by Association rule mining.

Typical rule structure

A rule structured as left hand side proposition and a right hand side consequent. It is comfortable for the miners to infer interesting and meaningful knowledge from the data. The rule denotes that "if" statement one is true "then" statement two is also true.

Antecedents => Consequents

LHS => RHS

Assume a set of transactions T, the primary goal of the association rule mining is to discover all rules having

(a) Association Rule:

An implication expression of the form X → Y, where X and Y are item sets

(b) Rule Evaluation Metrics:

Association rule is an intentional expression of the form X→Y, where X and Y are disjoint item sets in a transaction. The strength of an association rule is measure in terms of support and confidence parameters.

Support (s):

How frequent a rule is relevant to a given data set? Fractions of transactions that contain both X and Y in a dataset is

Support ≥ minimum support threshold

$$\text{Support} = \frac{\text{Frequency (X, Y)}}{N}$$

Confidence(c):

Establish how frequently items in Y appear in transaction that contains X.

Confidence ≥ minimum confidence threshold.

$$\text{Confidence} = \frac{\text{Frequency (X, Y)}}{\text{Frequency(X)}}$$

The main objective of this research work is to develop a new sequence pattern algorithm to detect the low hit pages using data mining techniques. In generally, Web mining is the application of data mining, discovers user's visiting behaviors, and extracts the user interests using discovered patterns. Similar to all data mining, the process of Web usage mining also comprises of three major steps (i) data pre-processing, (ii) pattern extraction and (iii) analysis. The input log data has to be pre-processed and clean to ensure an appropriate input for the mining algorithms. Different methods in web mining need various input formats and the pre-processing phase usually provides three varieties of output data. Pattern extraction means discovering frequent pattern in the web log data related to the pages visited by the users. The very purpose of web usage mining is to ascertain the useful data from web data or web log files. The result of web usage mining can be utilize for target advertisement, enhancing web design, enhancing satisfaction of customer and personalize websites.

An outlier is a data point, which is considerably different from the remaining data. Outlier can be defined as "An outlier is an examination which deviates so much from the other examinations as to provoke skepticisms that is was generated by a different mechanism". Outliers are often referred to as abnormalities, discordant, deviants, or anomalies in the data mining and statistics literature. In most applications, the data is created by one or more generating processes, which could either reflect activity in the system or observations collected about entities. When the generating process behaves in an unusual way, it results in the creation of outliers. Therefore, an outlier often contains useful information about abnormal characteristics of the systems and entities, which impact the data generation process.

II. LITERATURE REVIEW

The following papers are motivated to propose the association rule mining algorithms.

In 2014 Asieh Ghanbarpour and Behrooz Minaei proposed a DBSCAN-based method to cover multi-density datasets, called EXDBSCAN. This method only gets one parameter from the user and in addition of detecting clusters with different densities, can detect outlier correctly. The results of comparing final clusters of our method with two other clustering methods on some multi-density data sets show our method's performance in such datasets.

In 2018, Pranav Nerurka et al discussed different clustering approaches from the theoretical perspective to understand their relevance in context of massive data-sets and empirically these have been

tested on artificial benchmarks to highlight their strengths and weaknesses.

In 2013, Qinbao Song et al., presented a novel clustering-based feature subset selection algorithm for high dimensional data. The algorithm involves (i) removing irrelevant features, (ii) constructing a minimum spanning tree from relative ones, and (iii) partitioning the MST and selecting representative features. The authors discussed in this proposed algorithm, a cluster consists of features. Each cluster is treated as a single feature and thus dimensionality is drastically reduced. They compared the performance of the proposed algorithm with those of the five well-known feature selection algorithms FCBF, ReliefF, CFS, Consist, and FOCUS-SF on the 35 publicly available image, microarray, and text data from the four different aspects of the proportion of selected features, runtime, classification accuracy of a given classifier, and the Win/Draw/Loss record.

III. PROPOSED METHOD

Over the last decade, data mining became very popular, as the colossal amount of the data collected by the firms related to their activity grows voluminously. The data mining process comprehends the task of data analysis, data extraction and the application of learning procedures that bears to the discovery of hidden knowledge in the data sets. The data mining practice culminates on the creation of a theory/application which may depict a pattern or a relationship among data which should generate meaningful outputs when applied to raw data, from the perception of the end-user.

In our day to day life almost all of our actions are governed by decisions we think valuable for us. Our fore most objectives is to consider the best decisions so that we accomplish the utmost utility at the end of the day, week, month and year. Even if it is a mega-scale problem, the objective remains similar. But choosing the proper decisions maximize the utility value but the risk involved becomes complex and huge. Substantial profits or benefits can be attained, but at the same time the consequences arise will be serious or costly.

For more composite circumstances, many models and methodologies are built based on prior experience faced after knowing the costs/benefits involved on a decision. Using these models we predict about the "best" decision to take on complex situations. Apart from all these, our final decision ought to be the one which was predicted from earlier experience to attain maximum benefits.

The proposed algorithm comprises of many sub procedures and the working principle behind them is illustrated clearly hereunder along with the pseudo codes. The first procedure to prune away the successive duplicate items in transaction rows cleans the dataset at the first level by removing the false hits present in the server log data and saves considerable memory. The first procedure is explained in detail here under.

Table 1. Pseudo code of Remove Noise

PROCEDURE Remove Noise(Dataset D)

INPUT: Sequential Dataset D
OUTPUT: Noiseless Dataset D
BEGIN:

1. Find the total Transactional Rows Ř ∈Đ
2. For all Row Ř ∈Đ do
3. Find the Total Elements Ī ∈ Row Ř
4. For all Elements Ī ∈ Ř do
5. CHECK IF (Elements (Ī) = Elements (Ī +1)) then
6. Remove Elements at (Ī +1)
7. End IF
8. End FOR
9. Return Đ

END PROCEDURE

This procedure helps to remove the false hits present in the weblog dataset as many of the pages will be repeated in succession in the transactional row of the dataset. This has to be removed to provide accurate results.

Let us assume that the sample dataset contains five transaction rows as shown below,

Table 2. Sample pre-processed Log transaction file

Transaction ID	Transaction Data
T1	1,3,4,6
T2	1,3,7,7,7,4,17
T3	1,2,4,7,15,17
T4	1,3,4,6,6,15,17
T5	1,3,3,3,3,4,17

The sample dataset shown in the above table, it comprises of five transaction rows identified as T1 to T5. The average length of the sample dataset is 6.4 and the maximum length of the items is 7.0.

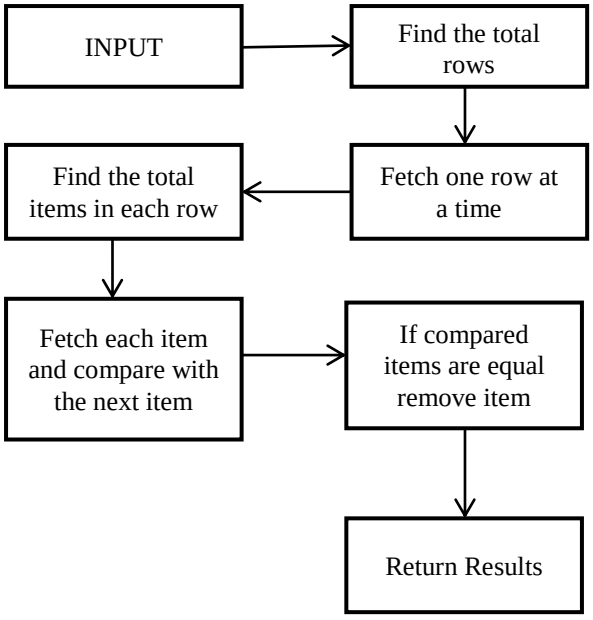

Figure 1. Flow diagram of Remove Noise Procedure

The proposed method mainly focused the "Detect Anomaly in Sequential Pattern Algorithm (DASPAT)" algorithm and its sub procedures. The procedure of the Remove Noise with the pseudo code and the working principle of the procedure are illustrated clearly with pictorial format.

IV. RESULTS AND DISCUSSIONS

Developing tools and technologies related to database expedites the archiving and usage of vast data from business firms, governments and scientific organizations. Downward closure property describes that all supersets of an infrequent item set are infrequent, and all subsets of a frequent item set are frequent. The property offers the algorithms with a formidable pruning technique.

By using this pruning technique in the datasets, once an infrequent item set is discovered by the algorithm, the algorithms stops checking all supersets of that item set. For example, in a dataset with m items, once the algorithm discover an infrequent item set containing j items, there is no necessity to scrutinize and investigate all of its supersets, i.e., $2^{(m-j)}$- 1 item sets.

This method of pruning is only suitable for frequent item set mining but for sequence item set mining this method of pruning alone does not solve the problem. Hence careful techniques should be incorporated in the algorithms while discarding the unpromising items related to support consideration . So most of the algorithms never use stringent methodologies in pruning and this in turn produces huge volume of candidates and mining information from this colossal volume of unpromising candidates increases the scan time, running time and the overall execution time required for the algorithms is large.

The proposed DASPAT algorithm employs sequential pruning factor is reduces the colossal volume of candidates generated to improve the efficiency of the algorithm. This reduction in the candidates reduces the execution time as well as the memory imprints to a greater extent. The proposed algorithm's pruning

strategy is the main miscreant for the large computational cost related to time and the proposed algorithm enhances the speed of execution, reduce the time consumed and memory utilized.

The proposed algorithm is developed with an intention to overcome the afore said problems related to candidate generation, time consumption and memory consumption. The proposed algorithm is experimentally evaluated to showcase the effectiveness and efficiency. The proposed DASPAT algorithm is implemented using .net based tool and executed to test the performance on a dual core 2.66 GHz processor with 1 GB RAM running on windows 7 ultimate platform. The proposed algorithm is executed on benchmarked datasets and compared with many existing state of the art algorithms to prove its effectiveness.

The algorithms are executed on four benchmarked real datasets like Kosarak dataset. It is a very large dataset containing 990 000 sequences of click-stream data from anhungarian news portal. The dataset in its original format can be found at http://fimi.ua.ac.be/data/.

Approaches Compared With Proposed Method

Numerous methods are being employed to detect the outlier present in the transaction datasets, namely association rule based outlier detection method and frequent pattern based outlier detection method. The existing algorithms are modified a little to detect the outliers from the transaction datasets.

GSP Algorithm (Generalized Sequential Pattern algorithm) is an algorithm used for sequence mining [29].

SPADE [22] is an algorithm that is based on lattice theory and applies temporal join operation to find sequential patterns. This algorithm is based on apriori approach and performs better than GSP.

An efficient algorithm called SPAM (Sequential PAttern Mining) [23] that integrates a variety of old and new algorithmic contributions into a practical algorithm. SPAM assumes that the entire database (and all data structures used for the algorithm) completely fit into main memory.

All the aforementioned algorithms and methods are altered a little in such a way to discover the infrequent item sets from the dataset and to find the outliers and compared with the proposed DASPAT algorithm.

This paper mainly compares the proposed algorithm DASPAT with existing algorithms regarding number of candidate generated (i.e.) the probable promising item sets, memory usage and time consumed (i.e.) speed of execution. After the comparison evaluation based on accuracy, precision, correct detection rate and false alarm rate is carried out.

Candidate Comparison on Kosarak Dataset

The proposed algorithm along with the existing algorithms is executed on the Kosarak dataset and the following results are produced with respect to candidates and for a varying support value. The support value is increased steadily to calculate the number of candidates generated by the algorithms and these values are noted and plotted.

Table 3. Candidate generation on Kosarak100K

Dataset – KOSARAK (100000 rows)					
Algorithm	Minimum Support				
	0.30	0.25	0.20	0.15	0.10
GSP	521900	401123	378300	298001	258117
SPADE	419000	321009	298100	236711	192011
FPOF	298009	259000	221999	182811	148310
SPAM	219011	182311	171116	138918	8210
Proposed	132900	119002	105600	7989	4210

From the table it is quite evident that the proposed algorithm DASPAT outscores the state of the art algorithms by a huge margin. The performance of SPAM is slightly largely better than GSP and SPADE and slightly better than FPOF method. The GSP generates huge volume of candidates since it employs level wise pruning technique. The SPAM algorithm which employs two step pruning technique clearly out performs the other algorithms SPADE, GSP and FPOF.

Comparison graph of Candidate on kosarak10K

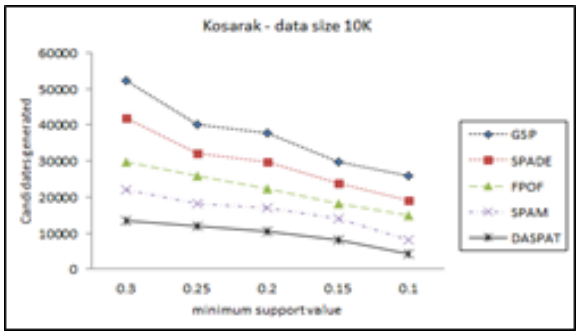

Figure 2. Candidate generated on Kosarak100K

V. CONCLUSION

The proposed algorithm incorporates efficient pruning technique to reduce the volume of candidate generation to a great extent. It is common that the execution gets slower when the memory of the machine is utilized beyond its physical limitation. To overcome this hurdle, the proposed algorithm employs the novel pruning intermittently during generation of candidates and removes the unpromising items and freeing the memory to speeding up the entire process. This research work has shown clearly that the proposed algorithm overcomes the snags present in the state of the art algorithms after evaluations carried out on bench marked datasets.

VI. REFERENCES

[1]. V. Kumar et al., "Clustering using modified harmony search algorithm", International journal of computational intelligence studies, 3(2): pp. 113-133, 2014.

[2]. Du Y.J. and Li H.M., "Strategy for Mining Association Rules for Web Pages Based on Formal Concept Analysis", Applied Soft Computing, Vol. 10, No. 3, pp. 772-783, 2010.

[3]. K. Kianmehr et al., "Association Rules Mining Based Approach for Web Usage Mining", In Proceedings of the IEEE International Conference on Information Reuse and Integration, IEEE Conference Publishing Services, pp. 166-171, August 03-05, 2011.

[4]. Martinez De Pison et al., "Mining Association Rules From Time Seriesto Explain Failures In A Hot-Dip Galvaizing Steel Line", Computers and Industrial Engineering, Vol. 63, No. 1, pp. 22-36, 2012.

[5]. P. Murugavel et al., "Improved Hybrid Clustering and Distance-based Technique for Outlier Removal", International Journal on Computer Science and Engineering (IJCSE).

[6]. R. Pamula et al., "An Outlier Detection Method Based on Clustering". Emerging Applications of Information Technology (EAIT), pp. 253 – 256, 2011.

[7]. Asieh Ghanbarpour and Behrooz Minaei, "EXDBSCAN: An extension of DBSCAN to detect clusters in multi-density datasets", IEEE Intelligent Systems (ICIS), 10.1109/IranianCIS.2014.6802561.

[8]. Pranav Nerurka et al, "Empirical Analysis of Data Clustering Algorithms", 6th International Conference on Smart Computing and Communications, ICSCC 2017, 7-8 December 2017, Kurukshetra, India.

[9]. Qinbao Song et al., "A Fast Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data", IEEE Transactions On Knowledge And Data Engineering Vol:25 No:1, 2013.

[10]. D. Sculley., "Web-scale k-means clustering", In Proceedings of the 19th international conference on World wide web, pp. 1177-1178, 2010.