Fuzzy Cluster Mining Data For MMSI in Bigdata

M.Senthamil Selvi[1], M.Nivetha[2], G.Bharathikannan[3]
[1, 2] Dept of CSE
[2, 3]Assistant Professor,Dept of CSE
[1, 2, 3]Sembodai Rukmani Varatharajan Engineering College

Abstract- *The traditional single minimum support data mining algorithm has some problems, such as too much space occupied by data, resulting in insufficient accuracy of the algorithm, which is difficult to meet the needs of the development of the times. Therefore, an intrusion data mining algorithm based on multiple minimum support is proposed. First, the feature parameters of frequent itemsets of intrusion data are extracted, and the sequence itemsets are divided according to the feature parameters. Then, the data mining features are transformed with the equivalent binary data transformation method, and the multi- support tree structure is optimized according to the data processing results. Data classification mining is carried out with the data tree structure information, and the intrusion data features are deeply mined. Finally, the research of the intrusion data mining algorithm based on the multi-minimum support is completed. Through comparative experiments, it is proved that the accuracy of the intrusion data mining algorithm based on multiple minimum support is 35 % - 75 % higher than that of the traditional single minimum support data mining algorithm.*

Keywords- Multiple minimum support; Data mining; Set function;

I. INTRODUCTION

The rapid development of modern new technology, the development level of information technology has also been continuously improved. In order to further promote the development of society, to effectively integrate and process massive data information and ensure the security of data processing, an intrusion data mining algorithm based on multiple minimum support degrees is proposed to achieve the research goal of real-time and effective security processing and calculation and analysis of massive data streams continuously[1]. In reference [2],a SFLA fuzzy clustering data mining algorithm based on selection and mutation mechanism is proposed. This algorithm first collects intrusion data and extracts features, then completes the research of high-risk intrusion data mining algorithm by using SFLA algorithm of fuzzy clustering. Experimental results show that the algorithm improves the accuracy of intrusion data mining. Reference [3] discusses the definition and theoretical basis of intrusion detection system, divides the functional categories of data mining technology, discusses the problems existing in intrusion detection system, establishes an intrusion detection system model based on data mining technology, and introduces the technical principle of the system model and the latest development of data mining technology.

Probabilistic Graphical Models

Uncertainty is inescapable in real-world applications the graph able to nearly never predict with certainty what's going to happen within the future, and even within the gift and therefore the past, several necessary aspects of the planet aren't determined with certainty. Applied mathematics offers United States the fundamental foundation to model our beliefs concerning the various potential states of the planet, and to update these beliefs as new proof is obtained. These beliefs are often combined with individual preferences to assist guide our actions, and even in choosing that observations to create. Whereas applied mathematics has existed since the seventeenth century, our ability to use it effectively on massive issues involving several inter-related variables is fairly recent, and is due mostly to the event of a framework called Probabilistic Graphical Models (PGMs).

Protein-Protein Interaction Networks

In Protein-Protein Interaction (PPI) networks, the interaction between 2 proteins is mostly established with a likelihood property thanks to the limitation of observation strategies. additionally, it's been verified that the interaction between super molecules A and B can influence the interaction between protein A and another protein C, if A, B and C have some common options. It's been verified that the likelihood of pair wise interaction and correlation among edges will be derived from applied mathematics models. Bunch applied to such related to probabilistic protein-protein interaction network knowledge is useful to find complexes to research the structure properties of the PPI Network.

II. RELATED WORK

K-Means Clustering

It is a partition method technique which finds mutual exclusive clusters of spherical shape. It generates a specific number of disjoint, flat (non-hierarchical) clusters. statically method can be used to cluster to assign rank values to the cluster categorical data[5]. Here categorical data have been converted into numeric by assigning rank value. K-Means algorithm organizes objects into k – partitions where each partition represents a cluster. The existing system start out with initial set of means and classify cases based on their distances to their centers. Next, compute the cluster means again, using the cases that are assigned to the clusters; then, reclassify all cases based on the new set of means. It keep repeating this step until cluster means don't change between successive steps. Finally, calculate the means of cluster once again and assign the cases to their permanent clusters.

DBSCAN Clustering

DBSCAN (Density Based Spatial Clustering of Application with Noise).It grows clusters according to the density of neighborhood objects. It is based on the concept of "density reachability" and "density connect ability", both of which depends upon input parameter- size of epsilon neighborhood e and minimum terms of local distribution of nearest neighbors. Here e parameter controls size of neighborhood and size of clusters. It starts with an arbitrary starting point that has not been visited. The points e-neighborhood is retrieved, and if it contains sufficiently many points, a cluster is started. Otherwise the point is labeled as noise. The number of point parameter impacts detection of outliers. DBSCAN targeting low-dimensional spatial data used DENCLUE algorithm.

OPTICS

OPTICS (Ordering Points to Identify Clustering Structure) is a density based method that generates an augmented ordering of the data's clustering structure. It is a generalization of DBSCAN to multiple ranges, effectively replacing the e parameter with a maximum search radius that mostly affects performance. Min Pts then essentially becomes the minimum cluster size to find. It is an algorithm for finding density based clusters in spatial data which addresses one of DBSCAN‟S major weaknesses i.e. of detecting meaningful clusters in data of varying density. It outputs cluster ordering which is a linear list of all objects under analysis and represents the density-based clustering structure of the data. Here parameter epsilon is not necessary and set to maximum value. OPTICS abstracts from DBSCAN by removing this each point is assigned as „core distance‟, which describes distance to its MinPts point. Both the core-distance and the reachability-distance are undefined if no sufficiently dense cluster w.r.t epsilon parameter is available.

STING

STING (STastical INformation Grid) is a grid-based multi resolution clustering technique in which the embedded spatial area of input object is divided into rectangular cells. Statistical information regarding the attributes in each grid cell, such as the mean, maximum, and minimum values are stored as statistical parameters in these rectangular cells. The quality of STING clustering depends on the granularity of the lowest level of grid structure as it uses a multi resolution approach to cluster analysis. Moreover, STING does not consider the spatial relationship between the children and their neighboring cells for construction of a parent cell. As a result, the shapes of the resulting clusters are isothetic, that is, all the cluster boundaries are either horizontal or vertical, and np diagonal boundary is detected. It approaches clustering result of DBSCAN if granularity approaches 0. Using count and cell size information, dense clusters can be identified approximately using STING.

III. PROPOSED METHODOLOGY

An intrusion data mining algorithm based on multiple minimum support degrees is proposed to achieve the research goal of real-time and effective security processing & calculation and analysis of massive data streams continuously .It also achieves the goal of accurately obtaining data characteristic parameters, effectively checking and entering and exiting intrusive data, then finally achieves the goal of improving the time efficiency of intrusion data mining and reducing the space occupancy of intrusion data mining. At the same time, by optimizing the structure of intrusion data mining tree and combining with the equivalent binary data transformation algorithm, the intrusion data relationship tuples in the data stream are checked to ensure that the calculation of intrusion data mining can be completed efficiently and accurately in the rapidly changing and continuously accumulating network data environment

The system is intended for implementation in the Hadoop environment. It consists of the following functional components: the data collector component that is responsible for the timely and accurate receipt of information about security events from sources of different types; component of the data storage

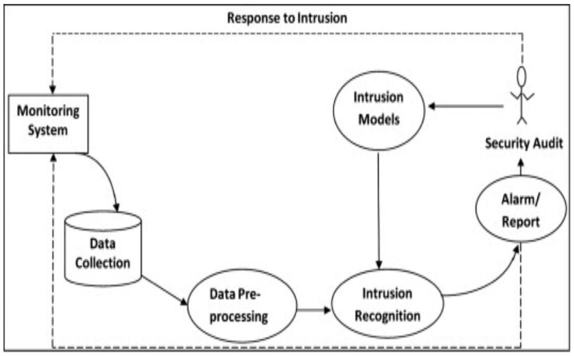


Figure.1 System Architechture

Merits

To ensure that different types of intrusion data feature frequency can be effectively detected and checked in the actual application process.

- Different data items are set up & equivalent binary transformation processing is carried out to effectively reflect the feature attributes and frequency of data mining.

IV. SYSTEM IMPLEMENTATION

SYSTEM MODULE

The System Modules are

- Data Extraction
- Partitioning of item sets
- Data Classification

MODULES DESCRIPTION

Data Extraction

Data extraction is a process that involves retrieval of data from various sources. Frequently, companies extract data in order to process it further, migrate the data to a data repository (such as a data warehouse or a data lake) or to further analyze it. It’s common to transform the data as a part of this process. For example, you might want to perform calculations on the data — such as aggregating sales data — and store those results in the data warehouse. If you are extracting the data to store it in a data warehouse, you might want to add additional metadata or enrich the data with timestamps or geolocation data. Finally, you likely want to combine the data with other data in the target data store. These processes, collectively, are called ETL, or Extraction, Transformation, and Loading. Extraction is the first key step in this process.

Partitioning of item sets

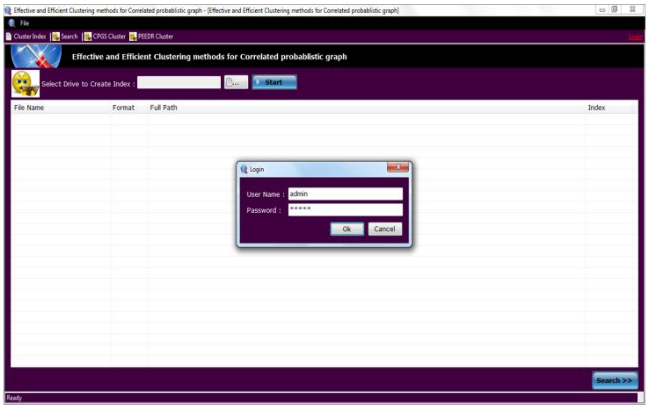
The algorithm uses the concept of partition, local support, local large itemsets, global support and global large itemsets. A partition is defined as any subset of transactions contained in the database D and partitions are non-overlapping. Local support of an itemset X in a partition is defined as the fraction of transactions in the partition, which contains the itemset. If the local support of an itemset X is at least the user-defined minimum support, the itemset is called local large itemset: Global support, global large itemsets, etc. also are defined in the same way, but in the context of global database.

Transformation

Data transformation is the process of converting data or information from one format to another, usually from the format of a source system into the required format of a new destination system. The usual process involves converting documents, but data conversions sometimes involve the conversion of a program from one computer language to another to enable the program to run on a different platform. The usual reason for this data migration is the adoption of a new system that's totally different from the previous one.

Data Classification

Data classification is broadly defined as the process of organizing data by relevant categories so that it may be used and protected more efficiently. On a basic level, the classification process makes data easier to locate and retrieve. Data classification is of particular importance when it comes to risk management, compliance, and data security. Data classification involves tagging data to make it easily searchable and trackable. It also eliminates multiple duplications of data, which can reduce storage and backup costs while speeding up the search process.



Effective and Efficient Clustering methods for Correlated probablistic graph - [Effective and Efficient Clustering methods for Correlated probablistic graph]

Effective and Efficient Clustering methods for Correlated probablistic graph

Select Drive to Create Index :

File Name	Format	Full Path	Index

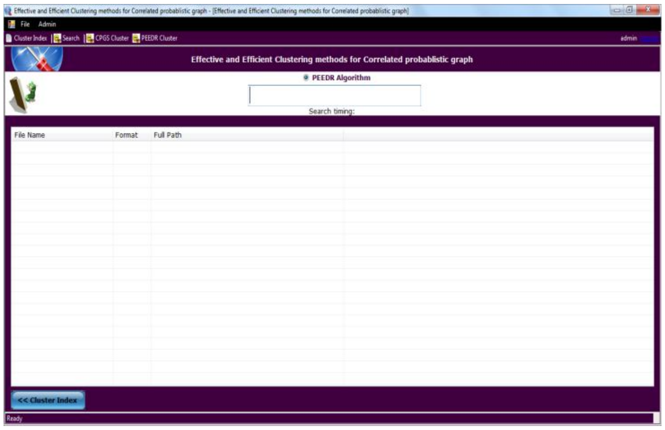
Login

User Name : admin

Password : *****

Ready

Fig A2.1 User Login



Effective and Efficient Clustering methods for Correlated probablistic graph - [Effective and Efficient Clustering methods for Correlated probablistic graph]

Effective and Efficient Clustering methods for Correlated probablistic graph

● PEEDR Algorithm

Search timing:

File Name	Format	Full Path

Ready

Figure A2.2 Algorithm Search

V. CONCLUSION AND FUTURE WORK

Under the background of big data environment, the intrusion data mining algorithm is optimized to ensure that different types of intrusion data feature frequency can be effectively detected and checked in the actual application process. In the traditional intrusion data mining process, the single minimum support algorithm easily causes the data item mining space to occupy too much space, which is not conducive to ensuring the accuracy of data mining. In order to solve the above problems, an intrusion data mining algorithm based on multiple minimum support degrees is proposed. Different data items are set up and equivalent binary transformation processing is carried out to effectively reflect the feature attributes and frequency of data mining. So as to optimize the intrusion data mining algorithm and improve the accuracy of the algorithm. Finally, the comparison experiments show that the accuracy of the intrusion data mining algorithm based on multiple minimum support is obviously improved compared with the traditional single minimum support data mining algorithm, and the occupied space is relatively small, fully satisfying the research goal.

VI. ACKNOWLEDGMENT

I would like to express my special thanks of gratitude to my college management. I would like to give my thanks to guide who gave me the golden opportunity to do this wonderful project on the. In addition, I would also like to thank my parents who helped me a lot in finalizing this project within the limited time frame.

REFERENCES

[1] Ackermann.M.R, Blömer.J, Kuntze.D, and Sohler.C(2014), ‘Analysis of agglomerative clustering’ ,Algorithmica, vol. 69, no. 1, pp. 184–215.

[2] Bonchi.F, Gionis.A, Kollios.G and Potamias.M(2010), ‘K-nearest neighborsin uncertain graphs’, PVLDB, vol. 3, no. 1, pp. 997–1008.

[3] Broschart.A, Schenkel.R, Theobald.M, won Hwang.S and Weikum.G(2007),‘Efficient text proximity search’, in SPIRE, pp. 287– 299.

[4] Cetindil.I, Esmaelnezhad.J, Li.C, and Newman.D(,2012), ‘Analysis of instant search query logs’, in WebDB, pp. 7–12.

[5] Chaudhuri. S, Ganti. V, and Motwani. R(2005), ‘Robust identification of fuzzy Duplicates’, in ICDE, 2005, pp. 865–876.

[6] Chakrabarti. K, Chaudhuri. S, Ganti. V and Xin. D(2008), ‘An efficient filter for approximate membership checking’, in SIGMOD Conference, pp. 805–818.

[7] Ding.B, Jin.R, Liu.L and Wang.H(Jun. 2011), ‘Distance-constraint reachability computation in uncertain graphs’, PVLDB, vol. 4, no. 9, pp. 551–562.

[8] Feng.J, Li.C ,Li.G and Wang.J (2012), ‘Supporting efficient top-k queries intype-ahead search’, in SIGIR, pp. 355–364.

[9] Flynn.P.J, Jain.A.K and Murty.M.N(1999), ‘Data clustering: A review’,ACM Comput. Surv. vol. 31, no. 3, pp. 264–323.

[10] Henzinger.M.R, Marais.H, Moricz.M and Silverstein.C.(1999), ‘Analysis of a very large web search engine query log’ ,SIGIR Forum, vol. 33, no. 1, pp. 6–12.

[11] Hua.M and Pei.J(2010) ‘Probabilistic path queries in road networks: Traffic uncertainty aware path selection’,in Proc. 13th Int. EDBT, New York, NY, pp. 347–358.