International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

Analyzing the classical algorithms on frequent pattern data mining and proposes enhancements on classical algorithms

Inturi kyathi reddy
M.SC(computer science) Chaitanya deemed to be university Kishanpura,hanamkonda, Telangana,INDIA.
Email:inturikyathireddy312@gmail.com
Phone number:9666022922.
Shankaramanchi yadhagiri sharma
M.SC(computer science) Chaitanya deemed to be university Kishanpura,hanamkonda, Telangana,INDIA.
Email:girisharma41@gmail.com
Phone number:7993625022.
Bombok madhukar
M.SC(computer science) Chaitanya deemed to be university Kishanpura,hanamkonda, Telangana,INDIA.
Email:madhubombok08@gmail.com
Phone number:9014171682.

ABSTRACT

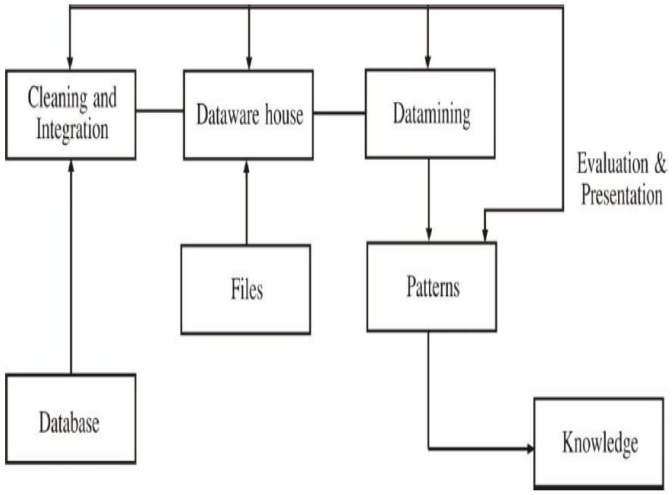
Knowledge mining applications has a vital role in imaging applications, biometric solutions, biological research and diagnostic services. The health care research is highly focused on knowledge extracted from patterns acquired through experiments and observations. The format of a data and its frequent presence is a subject of study started decade ago. The objective of the study is to propose a deep learning based frequent pattern mining algorithm that suits the requirement of big data frame works. Apart from classic pattern mining which trusted on frequency measure and feature relations, a novel approach is needed to process enormous data and predict characteristics from it. This thesis is concentrated on analyzing the classical algorithms on frequent pattern data mining and proposes enhancements on classical algorithms. This work also highlights the need of deep learning in present data centric world. A neural network based solution for feature selection with multilevel data modeling is proposed. The advantage of CNN over ANN is utilized to conclude the study and fulfill the objective. The work is associate with a vast literature survey on data mining and technologies persisted in it. Different algorithms in classification, clustering and supervised learning are discussed. As a predecessor to current method, Nlist and Twig join based FP mining is analyzed. Contributions in term of memory efficiency and processing speed are added to the existing N-list and Twig join method. Evaluations are made to verify the results. The inefficiency of handling big data was often a

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

problem with classical algorithms. The supervised learning to predict data from micro array is also studied as part of thesis. Different kernels were applied in experiments and the results analyzed. Still the problem of kernel selection and optimized result generation is observed. All such experiments involved the human intervention in parameters and threshold settings.

1. INTRODUCTION

Data engineering is a new disciplinary in information technology which handles a large volume of data originated from other sources like science and engineering, medicine and social sciences. The gigantic data sets are generated every day from sales transactions, sales promotions, performance report and feedback. Apart from that remote sensing, scientific experiments and environmental surveillance are some other sources that generate big data. The need of data mining technologies has paved way to many of the versatile algorithms and methods generate efficient data solutions. The first step in every data mining process is the data collection. Data mining studies are closely bound to the disciplinary of Data Analytics. The market study and consumer study of products and services has been common when data mining is introduced to business process. Same way the performance analysis of employees, analysis of medical reports etc are all linked to data mining. The Behavioral analysis is another cutting edge method utilized with knowledge discovery. Solving the tasks with a large number of man hours with in short span of time is the major objective of all data mining and analytical tasks. In traditional computing environment information was recorded in files or catalogues. The early stage of data mining are from the primitive file processing to database management system. Relational database management systems and indexing based methods were the initial technologies for data processing. Methods like online transaction processes were introduced in later stages. Those tools help to analyse and model data in multi-dimensional way which will help to make decisions based on conditions. The high volume data were always been a challenge to data warehousing applications. Often data generated from many online processing tools and techniques are not available in structured format. Various kinds of tools required for data classification, clustering, outlier detection etc. Excreting voluble in vast amount of data is the basis objective of data mining technologies. The knowledge discovery system in data mining process and iterative sequence with different steps from data cleaning to knowledge extraction.



International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

Fig:1.1 Steps in the Process of Knowledge Discovery

The steps in this figure shows the sequential process involved in the process of knowledge discovery. The basic definition of data mining is the process of discovering interesting patterns and knowledge from information repositories or dynamic data streams. Data streams for data mining is often having different format like textual, spatial and numeric data. Each format had led to the development of different branches in data mining like text data mining like pattern mining and image data mining. Some core technologies deep rooted in clustering. Geographical information systems is also depended on data mining technologies through knowledge processing from spatial data. Spatial data and image feature data are examples of high dimensional data models. The selection of technologies in these area is often reflected in the output generation speed. The more sophisticated algorithms must be used while processing multidimensional data.

1.1. DATA CHARACTERISATON

The input data source is always a collection of attributes representing different characteristics of data. These characteristics may be textual or numeric data. There may be a discriminative attribute to represent each instances of a data source. This is often referred as the class attribute. There may be single class attribute or multi class attribute depending on the application scenario. The attribute classification is often based on the manual criteria formulated form other attributes. The error in criteriaselection is corrected by data mining algorithms like expectation maximization, random forest and neural networks. The predicted classes may differ from the original class. The accuracy and precision of such algorithms can be evaluated by analyzing the contrasting classes. Discrimination is a feature connected with data characterization. The discriminative property represents the leniency of instances towards a particular attribute. Discovering these types of knowledge helps in many application like social surveys that to be conducted without any bias.

1.2 CHALLENGES IN DATA MINING

The overall process of data mining is constrained to any factors that depends up on the quality of knowledge discovery. Some of the factors include the data mining methodology, sources of data, user interaction, and social benefits. The methodology in data mining is depended on various factors like

- Variety of knowledge to be discovered
- Learning methodology used - supervised or unsupervised
- Extraction data from multi-dimensional space
- The user interactions within different steps of data mining
- Integrating background data at the time of data mining
- Presentation and visualization of results

Apart from the above issues, there are some issues concerned with scalability and portability of the experimental setup. When gigantic data connected with social networking applications these problems has

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

to be considered. The need of data mining based solutions is also connected with the social impacts of the results. Many of the observations from results having major impact on taking decisions in favour of society.

2. LITERATURE REVIEW

Frequent Pattern Mining Using Record Filter Approach (July, 2010) In today‟s emerging world, the role of data mining is increasing day by day with the new aspect of business. Data mining has been proved as a very basic tool in knowledge discovery and decision making process. Data mining technologies are very frequently used in a variety of applications. Frequent patterns were the itemsets those frequently visited in database transactions at least for the user defined number of times which known as support threshold. Presently a number of algorithms had been proposed in literature to enhance the performance of Apriori Algorithm, for the purpose of determining the frequent pattern. The main issue for any algorithm was to reduce the processing time. Author proposed a new record filter based algorithm which was a variation of the Apriori algorithm and performed fewer database scans than Apriori and utilizes only transaction of specific sizes for the generation of frequent itemsets. As observed by many researchers counting the occurrences of itemsets is a time consuming activity, this paper introduced a new strategy of considering only those transactions whose length was greater than or equal to the length of candidate set was checked, because candidate set of length k , cannot exist in the transaction record of length k-1 , it might exist only in the transaction of length greater than or equal to k. Due to this, proposed approach took very less time for performing computations during mining process. Experiments performed on synthetic datasets. The results explained that proposed approach performd well in terms of execution time and ultimately enhance efficiency as compared to traditional Apriori approach. For the comparative study of classical Apriori and proposed approach, author considered a database of 5000 transactions containing 50 unique items. During this analytical process author considered 1000 transactions to generate the frequent pattern with the support count of 10% and the process was repeated by increasing the transaction gradually.

A Parallel, Distributed Algorithm for Relational Frequent Pattern Discovery from Very• Large Data Sets (January, 2011) Heterogeneity and strong interdependence, which characterize ubiquitous data, required a multi relational approach to be analyzed with WARMR and SPADA. However, relational data mining algorithms did not scale well. Author proposed an extension of a relational algorithm for multilevel frequent pattern discovery, which resorted to data sampling and distributed computation in Grid environments, in order to overcome the computational limits of the original serial algorithm. The set of patterns discovered by the proposed algorithm approximates the set of exact solutions found by the serial algorithm. The quality of approximation depended on three parameters: the

proportion of data in each sample, the minimum support thresholds and the number of samples in which a pattern had to be frequent in order to be considered globally frequent. Author investigated on the third one. Experiments performed by processing both an event log publicly available on ProM web site http://is.tm.tue.nl/_cgunther/dev/prom/ and an event log provided by THINK3 Inc http://www.think3.com/en/default.aspx.

A Frame Work for Frequent Pattern Mining Using Dynamic Function (May, 2011)• Discovering frequent objects (item sets, sequential patterns) is one of the most vital fields in data mining. Apriori algorithm is a standard algorithm of association rules mining. We presented a new research trend on frequent pattern mining in which generate Transaction pair, which provided scalability to massive data sets and improving response time. This framework made pair of transaction instead of item id, so result show more scalable. Author suggested a novel dynamic algorithm for transposed database, mined in transaction pair and found longest common subsequence using dynamic function. Artificial and real-life data sets were tested and result described that proposed FPMDF algorithm was more scalable than Apriori and FP Growth algorithm. Author performed experiment on T40I4D100K dataset, provided by the QUEST generator of data generated from IBM's Almaden lab.

Comparative Analysis of Various Approaches Used in Frequent Pattern Mining (August,• 2011) Frequent pattern mining searched for recurring relationship in a given data set with association rules for interesting k itmesets. Various techniques found to mine frequent patterns with its own pros and cons. Performance of particular technique depended on input data and available resources in different domains like market basket analysis, including applications in marketing, customer segmentation, medicine, e-commerce, classification, clustering, web mining, bioinformatics and finance. This paper presented review of different frequent mining techniques including apriori based algorithms, partition based algorithms, DFS and hybrid algorithms, pattern based algorithms, SQL based algorithms and Incremental apriori based algorithms. Among all of the techniques discussed above, FP- Tree based approach achieved better performed and reduced the computational time. It took less memory by representing large database in compact treestructure. But a word of caution here that association rules should not be used directly for prediction without further analysis or domain knowledge.

Performance Analysis of Distributed Association Rule Mining with Apriori Algorithm (August, 2011) One of the most crucial problems in data mining is association rule mining. It required large computation and I/O traffic capacity. Author considered grid approach to resolve this problem. It offered an effective way to mine for large data sets. Therefore, author implemented distributed data mining with

Apriori algorithm in grid environment. However, usage of grid environment raised some issues about the optimization of the Apriori algorithm, especially the cost of the node to node communication and data distribution. In this paper, an Optimized Distributed Association rule mining approach for geographically distributed data was introduced in parallel and distributed environment and analyzed that this proposed method reduced communication costs. Author implemented experiments on datasets having minimum one million transactions to maximum five million transactions.

Parallel and Distributed Closed Regular Pattern Mining in Large Databases (March,• 2013) Due to huge increase in the records and dimensions of available databases pattern mining in large databases is a challenging problem. Numbers of parallel and distributed FP mining algorithms have been proposed for large and distributed databases based on frequency of item set. Author introduced a novel method called PDCRP-method (Parallel and Distributed closed regular pattern) to discover closed regular patterns using vertical data format on large databases. Conversion of horizontal database to vertical database format needed one database scan. PDCRP method applied in parallel and distributed environment to mine complete set of closed regular patterns based on user given global regularity and support values which minimize I/O cost and worked at each local processor which reduces inter processor communication overhead and getting high degree of parallelism generates complete set of closed regular patterns. Author derived results from experiments, which described PDCRP method is highly efficient in large databases. Author implemented PDCRP method from real (Kosarak) and synthetic (T1014D100K) datasets, available from http://cvs.buu.ac.th/mining/Datasets/synthesis_data/ and UCI Machine Learning Repository (University of California – Irvine, CA), these are used by Almanden Quest research group to develop frequent patterns in mining process.

Mining Efficient Association Rules Through Apriori Algorithm Using Attributes and• Comparative Analysis of Various Association Rule Algorithm (June, 2013) Apriori is the classical and most famous algorithm. Author considered data (bank data) and tried to obtain the result using Weka a data mining tool. Three algorithms tested and got elapsed time by author, named, Apriori Association Rule, PredictiveApriori Association Rule and Tertius Association Rule. According to the result obtained using data mining tool, author declared that Apriori Association algorithm performs better than the PredictiveApriori Association Rule and Tertius Association Rule algorithms. Author implemented experiment on dataset containing six hundred records and eleven attributes.

Distributed Algorithm for Frequent Pattern Mining using Hadoop MapReduce• Framework (2013) With the rapid growth of information technology and in many

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

business applications, mining frequent patterns and finding associations among them requires handling large and distributed databases. As FP-tree considered being the best compact data structure to hold the data patterns in memory there has been efforts to make it parallel and distributed to handle large databases. However, it incurs lot of communication over head during the mining. Author proposed parallel and distributed frequent pattern mining algorithm using Hadoop Map Reduce framework, which helped to derive best performance results for large databases. Proposed algorithm partitioned the database in such a way that, it worked independently at each local node and locally generates the frequent patterns by sharing the global frequent pattern header table. These local frequent patterns merged at final stage. This reduced the complete communication overhead during structure construction as well as during pattern mining. The item set count was also taken into consideration reducing processor idle time. Author used Hadoop Map Reduce framework effectively in all the steps of the algorithm. Experiments were carried out on a PC cluster with five computing nodes which shows execution time efficiency as compared to other algorithms. The experimental result described that proposed algorithm efficiently handles the scalability for very large databases.

3. PROBLEM SPECIFICATION

The bridging of symbolic aspect of biometric data to a feature model is the core challenge in data mining. The challenge in the form of memory efficiency was still pursuing all the present methods. It is observed that modeling different algorithms to work collectively is not been implemented so far, this work focuses the frequent pattern based classification intended for deep understanding and interpretation of patterns. The research issue related with deep insight in data can be resolved with deep learning algorithms that works with the help of neural networks.

4. OBJECTIVE

The ultimate objective of the work is analyzed and experimented using convolution neural networks. Big data arrived from bioinformatics is often be a problem in data representation and multi stage processing. This work clearly indicates a solution for output feature prediction that take part as an integral part of decision making. Accurate prediction is often implemented through the help neural networks. Multi stage identification is the specialty applied in this work. It enables the feature selection with the help of multiple classifiers once after the data get processed through CNN layers. A weight sharing method is applied to find the relevant factors within each layer. A max pooling layer is also introduced to maximize the chance of feature detection. The work has been tested with renowned breast cancer data sets. This work also takes care of dropout rates in both hidden layers and input layers to manage the optimality of experiment. Another feature that is added on present work is the weight penalty correction settings. It helps to maintain the weight penalty within an

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

experimental range. CNN based solutions will take a leading role in future data mining applications. The concept of cloud computing, big data based data warehousing, predictions based decision making etc. are highly depended on memory efferent and resource savvy knowledge mining. The future need of frequent pattern mining is highly applicable in detecting diseases, DNA problems and finding solutions to serious neuro issues. This work open directions to many future enchantments that will further help human kind.

5. DATA MINIG METHODS

The objective of data mining is ultimately connected to the knowledge discovery from gigantic data. The methodologies used for extracting knowledge were developed its previous systems using methods like principle component analysis, vector transformation and other statistical methods. Various data scientists develop methods for knowledge discovery with a set of associated algorithms for data reduction and attribute reduction. The common knowledge discovery process are –

3.1. CLUSTERING

All data objects are considered as tuples. The basic idea behind clustering is to find the similarity measure decides the quality of the clustering proves. Similarity is commonly referred in terms of how close the object are in space. The area of a cluster is decided by a distance measure which is called centroid distance. The distance of each from cluster centroid decides the membership of that community generation and recommendation systems. The effectiveness of clustering depends upon the nature of the data. Ecludian distance and jacked distance are some criteria measures used with different types of clustering applications. Cluster analysis tools based on K-mean, KMedoid and several other methods also have been built in many statistical applications. While dealing with uncertain data some other probabilistic clustering techniques like DB Scan is used. The Clustering techniques has many challenging aspects to be considered when it is adopted for a particular application. The Scalability by which large number of data objects is handled is a primal aspect. Variety of attributes, arbitrary cluster shapes, ability to deal with noise is some such requirements.

3.2. FEATURE SELECTION

Features are the identity of data. The way the features is utilized is important with regard to the applications. The instance learning in this area is a key concept derived for diagnosis purposes and biometric data. The distance between features is a key point while entertained the data source is explored. The data feature distance will definitely distinguish items from its original set. As the distance increases there may be a large number of features representing a data. Still there is chance of many data to be excluded at the time of analysis. Hence the dimensionality reduction is a major technique practiced by researchers. Since features contains information about the target. It has to carefully removed or selected at the time of feature selection. It is

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

better to discriminate the features by its indicative nature. Most discriminative data is better focused after the selection. The classifier algorithms in this regard are the most dependable mechanisms. The selection of kernels is important in classifier building. The Patterns are identified by dimensionally reduction are the main input of the biometric data mining models. The shift from features to patterns is also due to the behavioral coincidence of the data. The evaluation of dataset for pattern recognition can be of two types- Supervised and Unsupervised.

The first method takes data from training set to learn the input. The representation of features from available datasets after necessary sparsity checking will be forwarded to a classifier model. The kernel of the model should be aimed at pattern/feature recognition. The unsupervised learning method refers the input from live data. The training set may be formed from the input data itself in that case. While evaluating features for its importance, a sequence of statistical methods may be required based on the research case. The main statistical methods that can be used are- Pearson Correlation Co-efficient, Chi Square Test, SNR (Signal to Noise Ration) and Mutual information. After applying these, the features listed should be ranked in the order of importance. A cut - off threshold may be decided based on user and use case.

3.3. MULTIVARIATE FEATURE SELECTION

High dimensionality and small sample sizes, and their inherent risk of over fitting, pose great challenges for constructing efficient classifiers in microarray data classification. Therefore a feature selection technique should be conducted prior to data classification to enhance prediction performance. In general, filter methods can be considered as principal or auxiliary selection mechanism because of their simplicity, scalability, and low computational complexity. However, a series, of trivial examples show that filter methods result in less accurate performance because they ignore the dependencies of features. Although few studies are conducted to reveal the relationship of features by multivariate-based methods, the most accepted method to describe the relationships among features is by linear methods. While simple linear combination relationship restricts the improvement in performance. Kernel method is used to discover inherent nonlinear correlations among features as well as between feature attribute and targeted class. The number of orthogonal components is determined by Kernel Fishers Linear Discriminate analysis, in a selfadaptive manner rather than by manual parameter settings. The effectiveness of above method is tested by several experiments and the study has focused to new areas in multi-variate feature selection. Most commonly used classifier are support vector machine and K-Nearest neighbor in multivariate feature based data mining.

3.4. PATTERN RECOGNITION

Pattern recognition is a study discipline that identifies target classes by studying a pattern

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

formulation from datasets. The classical pattern recognition techniques were mainly focused on statistics and decision theory. But with the advent of machine learning and data mining techniques, advanced pattern recognition methods are used when designing practical systems. Many biological data, disease representations and image data provides valuable information from the input. The extraction of information pieces from structured and unstructured databases gives way to pattern recognition technologies. It involves the task of processing unstructured data, such as web-pages, free-form documents and e-mail for extracting named entities such as people, places, organizations and their relationships. The pattern recognition and machine learning have close relationship with each other. As pattern recognition systems shift through a varied format of data, it can be used in many biological and medical applications. The more relevant patterns are mined in the experiment will guide to better decisions. With the help of artificial intelligence
and neural networks, these patterns can be recognized easily.

CONCLUSION

Evolution of frequent pattern algorithms from classical methods to deep learning based algorithms is better exposed in this work. The objective is fulfilled by four novel ideas that include the incorporation of N-lists, micro array, feature relations and convolution networks. Each method is tested with real time data sets and often found tedious and heavy in terms of resource usage. Quality is often compromised in these methods. Introduction of feature centric algorithms has shown better results in real time applications. Features are the core concepts to be mined on every data mining applications. Representation of features is often complicated depending upon the sources of data. Some of the data sets contain data that can be represented in hierarchical format. First two section in the work provide the most efficient methods for relevant feature extraction. The real time problems like friend recommendation, product rating, virility detection etc. contain contexts that need the study on node to node interaction and the improved output on recommendation systems. The gigantic data still makes problem when it is represented and taken for knowledge mining. The multilevel hierarchies get fuzzed up in complicated algorithms. N-list based and Twig join based algorithms performs based on a threshold derived from trails. The nature of data is important in feature selection process. Biometric data is the most required source of feature extraction methods. The micro array based representation of biometric data is difficult to process with classical algorithms. Chapter 5 of this work has given the best approach to process biometric data availed from multiple sources. The relation between features decides which feature is to be selected. The criteria for such selection is decided based on the inter relation between features. The sections explained a series of steps integrated as a custom algorithm. The major

steps involved data binarization, regularization and variance calculation.

REFERENCES

1. Han J, Kamber M, Pei J. Mining frequent patterns, associations, and correlations. Data Mining: Concepts and Techniques. 2nd(edn)., Morgan Kaufmann Publishers Inc.: San Francisco; 2006. p. 227–48. 2. Zarrouk M, Gouider MS. Frequent patterns mining in timesensitive data stream. International Journal of Computer Science Issues. 2012 Jul; 9(4):117–24.

3. Aggarwal CC. Data streams: Models and algorithms. 1st(edn)., Springer US: US. 2007;1–7:61–100.

4. Gruenwald JN. Research issues in data stream association rule mining.ACM SIGMOD Record. 2006 Mar; 35(1):14–19.

5. Zhu Y, Shasha D. StatStream: Statistical monitoring of thousands of data streams in real time.ACM SIGMOD Record. 2005 Jun; 34(2):358–69.

6. Krempl G, Zliobaite I, Brzezinski D, Hullermeier E, Last M, Lemaire V, Noack T, Shaker A, Sievi S, Spiliopoulou M, Stefanowski J. Open challenges for data stream mining research. ACM SIGKDD Explorations Newsletter - Special issue on big data. 2014 Jun; 16(1):1–10. 7. Kohavi R, Mason L, Parekh R, Zheng Z. Lessons and challenges from mining retail e-commerce data. Kluwer Academic Publishers.2004 Oct; 57(1):83–113.

8. Yassir A, Nayak S. Issues in data mining and information retrieval. International Journal of Computer Science and Communication Networks. 2012 Mar; 2(1):93–8.

9. Kanth MR, Loshma G. Parallel multithreaded apriori algorithm for vertical association rule mining. International Journal of Advanced Research in Computer and Communication Engineering. 2013 Dec; 2(12):4729–35.

10. Dandu S, Deekshatulu BL, Chandra P. Improved algorithm for frequent item sets mining based on apriori and FP-tree. Global Journal of Computer Science and Technology Software and Data Engineering. 2013; 13(2):13–16