International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

Improved Association Rule Mining based on UP-Growth Algorithm

Ashwini Patil[1], Poonam Gupta[2]

[1]Pune University, G.H.R.C.E.M., Wagholi, Pune, Maharashtra, India

[2]Professor, Pune University, G.H.R.C.E.M., Wagholi, Pune, Maharashtra, India

Abstract: ***Mining of associations have been considered from different point of views, the use of high-dimensional data sets could still be a hard process. Besides, one should considers that, regardless the heuristic considered to solve the problem, the number of data accesses is high, so a data structure that speeds up the queries is the keystone. In this regard, the goal is to propose a suitable data structure that enables to both handle and fast compute high-dimensional data sets. For that purpose, the desired data structure should allow either to simplify or to reorganize data items in order to reduce the data size and provide a faster access to the stored information. In this paper we introduce new algorithm known as UP-growth which has the ability to mine the high dimensional data in less time complexity.***

Keywords: Association rule mining (ARM), data access, data structure, index compression

1. Introduction

Big data is a milestone in the information age, and brings deep impact on human society. Especially the Internet and electronic storage techniques, we are embracing the age of big data, which involves many critical aspects of our society, such as climate, biology, health, and social science. The available big data sets significantly advance our knowledge, services, and productivity across many sectors of our society. For example, a big medical data set can be used to find the best treatment plan for a given patient; a big traffic data set can improve the related traffic control and reduce congestion. The early version of the big data concept appeared in 2001 in the Gartner report by Laney, and big data was large and complex data sets that current computing facilities were not able to handle. It is characterized by 3Vs (Volume, Velocity, and Variety). Today almost every part of our society is expecting to improve itself using big data.

Approximate nearest neighbor (ANN) search over gigantic data has attracted much attention in the past decades, owing to its good balance between the retrieval performance and computational cost. Representative solutions include tree-based (e.g., k-D tree) and hash-based methods. In practice, visual data is usually represented by high-dimensional features, e.g., SIFT-based bag-of-words feature GIST , etc. Conventional tree-based methods developed in the past decades are known to severely suffer from the high dimensionality of features. Their performances are theoretically shown to degrade to that of linear scan in many cases. Therefore, increasing research efforts have been devoted to hash-based methods. Compared with other alternatives, hashing methods have gained both significant empirical success and theoretic guarantee, which is mainly derived from the so-called "locality sensitive" property.

In this paper, the goal is not to propose new efficient algorithms but a new data structure that could be used by a variety of existing algorithms without modifying its original schema. Thus, our aim is to speed up the association rule mining process regardless the algorithm used to this end, enabling the performance of efficient implementations to be enhanced. The structure simplifies, reorganizes, and speeds up the data access by sorting data by means of a shuffling strategy based on the hamming distance, which achieve similar values to be closer, and considering both an inverted index mapping and a run length encoding compression. In the experimental study, we explore the bounds of the algorithms' performance by using a wide number of data sets that comprise either thousands or millions of both items and records. The results demonstrate the utility of the proposed data structure in enhancing the algorithms' runtime orders of magnitude, and substantially reducing both the auxiliary and the main memory requirements.

Association Rule : An association rule is defined as an implication of the form $X \rightarrow Y$, where X and Y be set of items from a data set. Let $I = \{i1, i2, \ldots, in\}$ be the set of all the items comprised in a data set, and let $T = \{t1, t2, \ldots, tn\}$ be the set of all transactions within that data set. Then, X and Y are subsets of I, i.e., $X \subset I$ and $Y \subset I$. An association rule $X \rightarrow Y$ determines that if X is satisfied, then it is highly likely that Y is also satisfied, X and Y being the antecedent and consequent, respectively.

UP-Growth: *UP-Growth* (*Utility Pattern Growth*), is used for mining high utility itemsets with a set of techniques for pruning candidate item sets. The information of high utility itemsets is maintained in a special data structure named *UP-Tree* (*Utility Pattern Tree*) such that the candidate itemsets can be generated efficiently with only two scans of the database. The performance of UP-Growth was evaluated in comparison with the state-of-the-art algorithms on different types of datasets.

The remainder of this paper is organized as follows. In Section 2, we introduce problem statement. In section 3, we introduce existing system .In section 4, we introduce literature survey, and In Section 5, we provide methodology of proposed system, and in section 6 we provide conclusion.

2. Problem Statement

Mining of associations have been considered from different point of views, the use of high-dimensional data sets could still be a hard process. Besides, one should considers that, regardless the heuristic considered to solve the problem, the number of data accesses is high, so a data structure that speeds up the queries is the keystone. In this regard, the goal is to propose a suitable data structure that enables to both handle and fast compute high-dimensional data sets. For that purpose, the desired data structure should allow either to simplify or to reorganize data items in order to reduce the data size and provide a faster access to the stored information. In this paper we introduce new algorithm known as UP-growth which has the ability to mine the high dimensional data in less time complexity.

3. Existing System

ARM is considered as one of the most important techniques for the extraction of hidden knowledge. Apriori was known as the first algorithm for mining association rules. This algorithm is based on an exhaustive search (with pruning) and divides the ARM problem into two sub problems: 1) obtaining all the existing frequent item sets in the data and 2) extracting all the association rules from the item sets obtained in the previous step. Then FP growth was also another algorithm for mining frequent itemsets which overcomes the drawbacks of Apriori. To solve the problem of memory efficiency we will introduce new concept here in this paper.

3.1Disadvantages of Existing System

1) Existing system takes more time for mining data.
2) Difficult to handle high dimensional data
3) Existing system has problem of memory requirements

4. Literature Survey

- **Jose Maria Luna [1]:** In this paper ARM is considered as one of the most important techniques for the extraction of hidden knowledge. Apriori was known as the first algorithm for mining association rules. This algorithm is based on an exhaustive search (with pruning) and divides the ARM problem into two subproblems: 1) obtaining all the existing frequent item sets in the data and 2) extracting all the association rules from the item sets obtained in the previous step. Then FP growth was also another algorithm for mining frequent itemsets which overcomes the drawbacks of Apriori.
- **H. Gao, S. Shiji, J. N. D. Gupta, and W. Cheng [2]:** In this paper extreme learning machines (ELMs) have proven to be an efficient and effective learning paradigm for pattern classification and regression. However, ELMs are primarily applied to supervised learning problems. Only a few existing research studies have used ELMs to explore unlabeled data. In this paper, author extend ELMs for both semi-supervised and unsupervised tasks based on the manifold regularization, thus greatly expanding the applicability of ELMs. The key advantages of the proposed algorithms are 1) both the semi-supervised ELM (SS-ELM) and the unsupervised ELM (US-ELM) exhibit the learning capability and computational efficiency of ELMs; 2) both algorithms naturally handle multi-class classification or multi-cluster clustering; and 3) both algorithms are inductive and can handle unseen data at test time directly. Moreover, it is shown in this paper that all the supervised, semi-supervised and unsupervised ELMs can actually be put into a unified framework. This provides new perspectives for understanding the mechanism of random feature mapping, which is the key concept in ELM theory. Empirical study on a wide range of data sets demonstrates that the proposed algorithms are competitive with state-of-the-art semi-supervised or unsupervised learning algorithms in terms of accuracy and efficiency.
- **Y. Qian, J. Liang, W. Pedrycz, and C. Dang [3]:** Feature selection (attribute reduction) from large-scale incomplete data is a challenging problem in areas such as pattern recognition, machine learning and data mining. In rough set theory, feature selection from incomplete data aims to retain the discriminatory power of original features. To address this issue, many feature selection algorithms have been proposed, however, these algorithms are often computationally time-consuming. To overcome this shortcoming, in this paper author introduce a theoretic framework based on rough set theory, which is called positive approximation and can be used to accelerate a heuristic process for feature selection from incomplete data. As an application of the proposed accelerator, a general feature selection algorithm is designed. By integrating the accelerator into a heuristic algorithm, they obtain several modified representative heuristic feature selection algorithms in rough set theory. Experiments show that these modified algorithms outperform their original counterparts. It is worth noting that the performance of the modified algorithms becomes more visible when dealing with larger data sets.
- **E. Lo, N. Cheng, W. W. K. Lin, W.-K. Hon, and B. Choi[4]:** In this paper to evaluate the performance of database applications and DBMSs, they usually execute workloads of queries on generated databases of different sizes and measure the response time. This paper introduces MyBenchmark, an offline data generation tool that takes a set of queries as input and generates database instances for which the users can control the characteristics of the resulting workload. Applications of MyBenchmark include database testing, database application testing, and application-driven benchmarking. Author present the architecture and the implementation algorithms of MyBenchmark. Author also present the evaluation results of MyBenchmark using TPC workloads.
- **X. Liu, Y. Mu, D. Zhang, B. Lang, and X. Li[5]:** As we know hashing methods are effective in generating compact binary signatures for images and videos. This paper addresses an important open issue in the literature, i.e., how to learn compact hash codes by enhancing the complementarity among different hash functions. Most of prior studies solve this problem either by adopting time-consuming sequential learning algorithms or by generating the hash functions which are subject to some deliberately-designed constraints (e.g., enforcing hash

functions orthogonal to one another). Author analyze the drawbacks of past works and propose a new solution to this problem. Here idea is to decompose the feature space into a subspace shared by all hash functions and its complementary subspace. On one hand, the shared subspace, corresponding to the common structure across different hash functions, conveys most relevant information for the hashing task. Similar to data de-noising, irrelevant information is explicitly suppressed during hash function generation. On the other hand, in case that the complementary subspace also contains useful information for specific hash functions, the final form of our proposed hashing scheme is a compromise between these two kinds of subspaces. To make hash functions not only preserve the local neighborhood structure but also capture the global cluster distribution of the whole data, an objective function incorporating spectral embedding loss, binary quantization loss, and shared subspace contribution is introduced to guide the hash function learning. They propose an efficient alternating optimization method to simultaneously learn both the shared structure and the hash functions. Experimental results on three well-known benchmarks CIFAR-10, NUS-WIDE, and a-TRECVID demonstrate that our approach significantly outperforms state-of-the-art hashing methods.

- **D. Wegener, M. Mock, D. Adranale, and S. Wrobel [6]:** This paper mentions that the enormous growth of data in a variety of applications has increased the need for high performance data mining based on distributed environments. However, standard data mining toolkits per se do not allow the usage of computing clusters. The success of MapReduce for analyzing large data has raised a general interest in applying this model to other, data intensive applications. Unfortunately current research has not lead to an integration of GUI based data mining toolkits with distributed file system based MapReduce systems. This paper defines novel principles for modeling and design of the user interface, the storage model and the computational model necessary for the integration of such systems. Additionally, it introduces a novel system architecture for interactive GUI based data mining of large data on clusters based on MapReduce that overcomes the limitations of data mining toolkits. As an empirical demonstration we show an implementation based on Weka and Hadoop.
- **UP-Growth: An Efficient Algorithm for High Utility Itemset Mining[7]:** Mining high utility itemsets from a transactional database refers to the discovery of itemsets with high utility like profits. Although a number of relevant approaches have been proposed in recent years, they incur the problem of producing a large number of candidate itemsets for high utility itemsets. Such a large number of candidate itemsets degrades the mining performance in terms ofexecution time and space requirement. The situation may become worse when the database contains lots of long transactions or long high utility itemsets. In this paper, we propose an efficient algorithm, namely *UP-Growth* (*Utility Pattern Growth*), for mining high utility itemsets with a set of techniques for pruning candidate itemsets. The information of high utility itemsets is maintained in a special data structure named *UP-Tree* (*Utility Pattern Tree*) such that the candidate item sets can be generated efficiently with only two scans of the database. The performance of UP-Growth was evaluated in comparison with the state-of-the-art algorithms on different types of datasets. The experimental results show that UP-Growth not only reduces the number of candidates effectively but also outperforms other algorithms substantially in terms of execution time, especially when the database contains lots of long transactions.
- **Fast Algorithms for Mining Association Rules[8]:** the problem of discovering association rules between items in a large database of sales transactions. They present two new algorithms for solving this problem that are fundamentally different from the known algorithms. Empirical evaluation shows that these algorithms outperform the known algorithms by factors ranging from three for small problems to more than an order of magnitude for large problems. They also show how the best features of the two proposed algorithms can be combined into a hybrid algorithm, called AprioriHybrid. Scale-up experiments show that AprioriHybrid scales linearly with the number of transactions. AprioriHybrid also has excellent scale-up properties with respect to the transaction size and the number of items in the database.
- **Statistical comparisons of classifiers over multiple data sets[9]:** While methods for comparing two learning algorithms on a single data set have been scrutinized for quite some time already, the issue of statistical tests for comparisons of more algorithms on multiple data sets, which is even more essential to typical machine learning studies, has been all but ignored. This article reviews the current practice and then theoretically and empirically examines several suitable tests. Based on that, They recommend a set of simple, yet safe and robust non-parametric tests for statistical comparisons of classifiers: the Wilcoxon signed ranks test for comparison of two classifiers and the Friedman test with the corresponding post-hoc tests for comparison of more classifiers over multiple data sets. Results of the latter can also be neatly presented with the newly introduced CD (critical difference) diagrams.
- **Large scale spectral clustering via landmark based sparse representation[10]:** Spectral clustering is one of the most popular clustering approaches. Despite its good performance, it is limited in its applicability to large-scale problems due to its high computational complexity. Recently, many approaches have been proposed to accelerate the spectral clustering. Unfortunately, these methods usually sacri- fice quite a lot information of the original data, thus result in a degradation of performance. In this paper, they propose a novel approach, called Landmark-based Spectral Clustering (LSC), for large scale clustering problems. Specifically, they select p (n) representative data points as the landmarks and represent the original data points as the linear combinations of these landmarks. The spectral embedding of the data can then be efficiently computed with the landmark-based representation. The proposed algorithm scales linearly with the problem size. Extensive experiments show the effectiveness and efficiency of our approach comparing to the state-of-the-art methods.

5. Proposed System

5.1Solving Approach

The goal is to propose a suitable data structure that enables to both handle and fast compute high-dimensional data sets. For that purpose, the desired data structure should allow either to simplify or to reorganize data items in order to reduce the data size and provide a faster access to the stored information. In this paper we introduce new algorithm known as UP-growth which has the ability to mine the high dimensional data in less time complexity.

5.2 Advantages of Proposed System

1) Speed ups the process of mining ARMS
2) Covers the problem of high dataset efficiency.
3) UP-growth, this algorithm is used in turn with new data structure to handle high dimensional item sets in big data.

5.3 Applications

1) Financial Data Analysis.
2) Retail Industry.
3) Telecommunication Industry.
4) Biological Data Analysis.

5.4 Methodology of Work

Fig. 1 shows an overview of the workflow of the methodology considered to store data records into the proposed new data structure. We first describe the shuffling strategy that enables data to be sorted by an HD. Finally, we defines a mapping process that builds a structure by using both an inverted index mapping and a compressing process, which is based on an RLE on the sorted tuples.

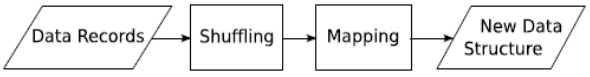

Figure 1: Workflow of the methodology used to store data records into proposed data structure

6. Conclusion

In this paper we are proposing a new data structure which is mainly used by ARM algorithms. Our main aim is to cover the problem of high dataset efficiency. This new data structure speed ups the process of mining ARM. In addition to this we have proposed the new algorithm to named as UP-growth, this algorithm is used in turn with new data structure to handle high dimensional itemsets in big data.

7. Acknowledgment

We would like to thank all the authors of different research papers referred during writing this paper. It was very knowledge gaining and helpful for the further research to be done in future.

References

[1] José María Luna, *Member, IEEE*, Alberto Cano, *Member, IEEE*, Mykola Pechenizkiy, *Member, IEEE*, and Sebastián Ventura, *Senior Member, Speeding-Up Association Rule Mining With Inverted Index CompressionIEEE,2016*.
[2] H. Gao, S. Shiji, J. N. D. Gupta, and W. Cheng, "Semi-supervised and unsupervised extreme learning machines," *IEEE Trans. Cybern.*, vol. 44,no. 12, pp. 2405–2417, Dec. 2014.
[3] Y. Qian, J. Liang, W. Pedrycz, and C. Dang, "An efficient accelerator for attribute reduction from incomplete data in rough set framework," *Pattern Recognit.*, vol. 44, no. 8, pp. 1658–1670, 2011.
[4] E. Lo, N. Cheng, W. W. K. Lin, W.-K. Hon, and B. Choi,"MyBenchmark: Generating databases for query workloads," *Int. J. Very Large Data Bases*, vol. 23, no. 6, pp. 895–913, 2014.
[5] X. Liu, Y. Mu, D. Zhang, B. Lang, and X. Li, "Large-scale unsupervised hashing with shared structure learning," *IEEE Trans. Cybern.*,vol. 45, no. 9, pp. 1811–1822, Sep. 2015.
[6] D. Wegener, M. Mock, D. Adranale, and S. Wrobel, "Toolkit-based high-performance data mining of large data on MapReduce clusters," in *Proc. IEEE Int. Conf. Data Min.*, Miami, FL, USA, 2009, pp. 296–301.
[7] Vincent S. Tseng, Cheng-Wei Wu, Bai-En Shie, and Philip S. Yu," UP-Growth: An Efficient Algorithm for High Utility Itemset Mining"University of Illinois at Chicago, Chicago, Illinois, USA
[8] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," in *Proc. 20th Int. Conf. Very Large DataBases (VLDB)*, Santiago, Chile, 1994, pp. 487–499.
[9] D. Cai and X. Chen, "Large scale spectral clustering via landmark based sparse representation," *IEEE Trans. Cybern.*, vol. 45, no. 8,pp. 1669–1680, Aug. 2015.
[10] J. Demšar, "Statistical comparisons of classifiers over multiple data sets,"*J. Mach. Learn. Res.*, vol. 7, pp. 1–30, Jan. 2006.