A Rule based Data Mining Mechanism: Association Rule Mining

Meenakshi Malik[1]

Shobhit University

Meerut

Mamta[2]

Shobhit University

Meerut

R. P. Agarwal[3]

Shobhit University

Meerut

Abstract

For many years, statistics have been used to analyze data in an effort to find correlations, patterns, and dependencies. However, with advancement in technology, more and more data are available, which greatly exceed the human capacity to manually analyze them. Knowledge discovery in financial organization have been built to evaluate their operation and mainly to support decision making using knowledge as key factor. Before the 1990's, data collected by bankers, credit card companies, department stores and so on have little used. But in recent years, the idea of data mining has emerged with the increase in computational power. Today data mining is primarily used by companies with a strong consumer focus- retail, financial, communication, medical, agriculture and marketing organizations. In this paper, we investigate the use of various data mining techniques for knowledge discovery in agriculture sector. Existing software are inefficient in showing such data characteristics. We introduce different exhibits for discovering knowledge in the form of association rules. Proposed data mining techniques, the decision maker can define the expansion of agriculture activities to empower the different forces in existing agriculture sector. This paper is an outcome of the study on data mining from exiting literature.

Keywords

Data Mining, Association Rule Mining, Agriculture, Knowledge discovery, Decision making.

Introduction

Data mining, the discovery of new and interesting patterns in large datasets, is an exploding field. The current age is often referred to as the information age and in this information age, it is believed that information leads to power and success. Initially, with the advent of computers and means for mass digital storage, everyone started collecting and storing all sorts of data, counting on the power of computers to help sort through this amalgam of information. With the exponential increase in media data on personal computers and the internet, it is critical for end users to efficiently manage metadata to find the information they are looking for. The huge amount of data collected by companies is helpful in determining the relationships among the internal factors such as price, product positioning or staff skills and external factors such as economic indicators, competition and customer demographics. Many softwares have been discovered for mining the data and also the relationships and patterns in stored transaction data was analyzed based on open-ended user queries.[1] Several types of analytical software are available:

statistical, machine learning, and neural networks that mine the data in various ways.

The efficient database management systems have been very important asset for management of a large

corpus of data and especially for effective and efficient database management systems has also contributed to recent massive gathering of all sorts of information. Information retrieval is simply not enough anymore for decision-making. Confronted with huge collections of data new needs to help one make better managerial choices have been created. These needs are automatic summarization of data, extraction of the essence of information stored, and the discovery of patterns in raw data. Data Mining can help us retrieve, organize and manage the exponentially growing media data easily.[2] Due to the rapidly advancing technology in the last few decades, more and more of our everyday life has been changed by information technology. This can be done by decision rules, which incorporate the knowledge about the coherence between data and yield potential. In addition, these rules should give (economically) optimized recommendations.

The connection between information technology and agriculture is and will become an even more interesting area of research in the near future. In this context, IT mostly covers the following three aspects: data collection, analysis and recommendation. This work is based on a dissertation that deals with data mining and knowledge discovery in agriculture from an agrarian point of view.[3] This research led to economically optimized decision rules, but left out some of the details on the used techniques.

2. Data Mining

The various data mining definitions available from literature are:

The data mining system can automatically find and show you new patterns that will lead to fresh insight[4] - Osmar R. Zaïane

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information[5] - Cipolla and Emil T.

Data mining is the process of finding correlations or patterns among dozens of fields in large relational databases[6] - Conner and Louis.

Data mining is the semi-automatic discovery of patterns, changes, associations, anomalies, and other statistically significant structures from large data sets[7] - Robert Grossman.

Data mining is a process that uses a variety of data analysis tools to discover patterns and relationships in data that may be used to make valid predictions[8] - Two Crows, Han, Jiawei, and Micheline Kamber.

3. Issues In Data Mining

One of the key issues raised by data mining technology is not a business or technological one, but a social one. It is the issue of individual privacy. Data mining makes it possible to analyze routine business transactions and glean a significant amount of information about individuals buying habits and preferences.[9] There is in data mining can be categorized as security issues, social issues, user interface issues, mining methodology issues, performance issues, data source issues, data integrity, diplomacy issue and ethical issues.

Data structure: A debated technical issue is whether it is better to set up a relational database structure or a multidimensional one. In a relational structure, data is stored in tables, permitting ad hoc queries. In a multidimensional structure, on the other hand, sets of cubes are arranged in arrays, with subsets created according to category. While multidimensional structures facilitate multidimensional data mining, relational structures thus far have performed better in client/server environments.[9] And, with the explosion of the Internet, the world is

becoming one big client/server environment.

Security and social issues: Security is an important issue with any data collection that is shared and/or is intended to be used for strategic decision-making.[10]

User interface issues: The knowledge discovered by data mining tools is useful as long as it is interesting, and above all understandable by the user.[11]

Mining methodology issues: These issues pertain to the data mining approaches applied and their limitations.[12]

Performance issues: Many artificial intelligence and statistical methods exist for data analysis and interpretation.[13]

Data source issues: There are many issues related to the data sources, some are practical such as the diversity of data types, while others are philosophical like the data glut problem.[14]

Data integrity: A key implementation challenge is integrating conflicting or redundant data from different sources.[15] Clearly, data analysis can only be as good as the data that is being analyzed. For example, a bank may maintain saving accounts on several different databases. The addresses of a single cardholder may be different in each. Software must translate data from one system to another and select the address most recently entered.

Diplomacy issues in data mining: A key problem that arises in any en masse collection of data is that of confidentiality. The need for privacy is sometimes due to law or can be motivated by business interests.[16]

Ethical Issues: The social ethical and legal implications of data mining are examined through recent case law, current public opinion, and small industry-specific examples.[17]

Financial Issue: While system hardware costs have dropped dramatically within the past five years, data mining and data warehousing tend to be self-reinforcing. The more powerful the data mining queries, the greater the utility of the information being gleaned from the data and the greater the pressure to increase the amount of data being collected and maintained, which increases the pressure for faster, more powerful data mining queries. This increases pressure for larger, faster systems, which are more expensive.

4. Association Rule Mining

Association rule mining problem is defined as follows: D ={ t1,t2,…,tn } is a database of transactions. Each transaction consists of I, where { i1,i2,…in}= I is a set of all items. An association rule is an implication of the form A⇒B, where A and B are itemsets, A⊆I, B⊆I, A∩B=∅. In support-confidence framework, each association rule has support and confidence to confirm the validity of the rule. The support denotes the occurrence rate of an itemset in D, and the confidence denotes the proportion of data items containing B in all items containing A in D.

$Sup(i) = Count(i) / Count(DBT)$

$Sup(A \Rightarrow B) = Sup(A \cup B)$

$Conf(A \Rightarrow B) = Sup(A \cup B) / Sup(A)$

When the support and confidence are greater than or equal to the pre-defined threshold, the association rule is considered to be a valid rule.[18] The objective of ARM is to find the universal set S of all valid association rules. The Apriori algorithm is the most well-known association rule algorithm and is used in most commercial products.

Input: Li-1 //Large itemsets of size i – 1

Output: Ci //farmers of size i

Algorithm:

Ci = ø;

for each I _ Li-1

do for each J _ _ Li-1

do if i – 2 of the elements in I and J are equal

Then Ck = Ck U {I U J};

4. Increasing the Efficiency of Association Rules Algorithms

The computational cost of association rules mining can be reduced in the following four ways and in recent years much progress has been made in all these directions.

- by reducing the number of passes over the database
- through parallelization.
- by sampling the database
- by adding extra constraints on the structure of patterns

4.1 Reducing the number of passes over the database

FP-Tree, frequent pattern mining, is another milestone in the development of association rule mining, which breaks the main bottlenecks of the Apriori. The frequent itemsets are generated with only two passes over the database and without any candidate generation process. Only frequent length-1 items will have nodes in the tree, and the tree nodes are arranged in such a way that more frequently occurring nodes will have better chances of sharing nodes than less frequently occurring ones. FP-tree is an extended prefix-tree structure storing crucial, quantitative information about frequent patterns. FP-Tree scales much better than Apriori because as the support threshold goes down, the number as well as the length of frequent itemsets increase dramatically. The candidate sets that Apriori must handle become extremely large and the pattern matching with a lot of candidates by searching through the transactions becomes very expensive. The frequent patterns generation process includes two sub processes: constructing the FT-Tree, and generating frequent patterns from the FP-Tree. The mining result is the same with Apriori series algorithms. To sum up, the efficiency of FP-Tree algorithm account for three reasons. First the FP-Tree is a compressed representation of the original database because only those frequent items are used to construct the tree, other irrelevant information are pruned. Secondly this algorithm only scans the database twice. Thirdly, FP-Tree uses a divide and conquer method that considerably reduced the size of the subsequent conditional FP-Tree.

Every algorithm has his limitations, for FP-Tree it is difficult to be used in an interactive mining system. Another limitation is that FP-Tree is that it is not suitable for incremental mining. Since as time goes on databases keep changing, new datasets may be inserted into the database, those insertions may also lead to a repetition of the whole process if we employ FP-Tree algorithm. Finally association rules are mined from the frequent candidate sets.

4.2 Sampling

Toivonen[19] presented an association rule mining algorithm using sampling. The approach can be divided into two phases. During first phase, a sample of the database is obtained and all associations in the sample are found. These results are then validated against the entire database. To maximize the effectiveness of the overall approach, the author makes use of lowered minimum support on the sample. Since the approach is probabilistic (i.e. dependent on the sample containing all the relevant associations) not all the rules may be found in this first pass. Parthasarathy[20] presented an efficient method to progressively sample for association rules. Chuang et al. explore another progressive sampling algorithm called Sampling Error Estimation, which aims to identify an appropriate sample size for mining association rules. SEE has two advantages. First, SEE is highly efficient because an appropriate sample size can be determined without the need of executing association rules. Second, the identified sample size of SEE is very accurate, meaning that association rules can be highly efficiently executed on a sample of this size to obtain a sufficiently accurate result. Especially, if data comes as a stream flowing at a faster rate than can be processed, sampling seems to be the only choice.

4.3 Parallelization

Association rule discovery techniques have gradually been adapted to parallel systems in order to take advantage of the higher speed and greater storage capacity that they offer. The transition to a distributed memory system requires the partitioning of the database among the processors, a procedure that is generally carried out indiscriminately Cheung et al.[21] presented an algorithm called FDM. FDM is a parallelization of Apriori to (shared nothing machines, each with its own partition of the database. At every level and on each machine, the database scan is performed independently on the local partition. Then a distributed pruning technique is employed. Schuster and Wolff[22] described another Apriori based D-ARM algorithm- DDM. As in FDM, candidates in DDM are generated levelwise and are then counted by each node in its local database. The nodes then perform a distributed decision protocol in order to find out which of the candidates are frequent and which are not. It adopts the count distribution approach and has incorporated two powerful candidate pruning techniques i.e., distributed pruning and global pruning. It has a simple communication scheme which performs only one round of message exchange in each iteration. A new algorithm, Data Allocation Algorithm (DAA), is presented in[23] that uses Principal Component Analysis to improve the data distribution prior to FPM. Parthasarathy et al.[24] have written an excellent recent survey on parallel association rule mining with shared memory architecture covering most trends, challenges and approaches adopted for parallel data mining. All approaches spelled out and compared in this extensive survey are apriori-based. More recently Tang and Turkia proposed a parallelization scheme which can be used to parallelize the efficient and fast frequent itemset mining algorithms based on FP-trees.

4.4 Constraints based Association Rule Mining

Many data mining techniques consist in discovering patterns frequently occurring in the source dataset. Typically the goal is to discover all the patterns whose frequency in the dataset exceeds a user-specified threshold. However, very often users want to restrict the set of patterns to be discovered by adding extra constraints on the structure of patterns. Data mining systems should be able to exploit such constraints to speedup the mining process. Techniques applicable to constraint-driven pattern discovery can be classified into the following groups:

• post-processing (filtering out patterns that do not satisfy user-specified pattern constraints after the actual discovery process);

• pattern filtering (integration of pattern constraints into the actual mining process in order to generate only patterns satisfying the constraints);

• dataset filtering (restricting the source dataset to objects that can possibly contain patterns that satisfy pattern constraints).

Wojciechowski and Zakrzewicz[25] focus on improving the efficiency of constraint-based frequent pattern mining by using dataset filtering techniques. Dataset filtering conceptually transforms a given data mining task into an equivalent one operating on a smaller dataset. Rapid Association Rule Mining (RARM) is an association rule mining method that uses the tree structure to represent the original database and avoids candidate generation process. In order to improve the efficiency of existing mining algorithms, constraints were applied during the mining process to generate only those association rules that are interesting to users instead of all the association rules.

5. Categories of Databases in which Association Rules are Applied

Transactional database refers to the collection of transaction records, in most cases they are sales records. With the popularity of computer and e-commerce, massive transactional databases are available now. Data mining on transactional database focuses on the mining of association rules, finding the correlation between items in the transaction records. One of data mining techniques, generalized association rule mining with taxonomy is potential to discover more useful knowledge than ordinary flat association rule mining by taking application specific information into account.[27] In particular in retail one might consider as items particular brands of items or whole groups like milk, drinks or food. The more general the items chosen the higher one can expect the support to be. Thus one might be interested in discovering frequent itemsets composed of items which themselves form a taxonomy. Earlier work on mining generalized association rules ignore the fact that the taxonomies of items cannot be kept static while new transactions are continuously added into the original database. Empirical evaluations show that this algorithm can maintain its performance even in large amounts of incremental transactions and high degree of taxonomy evolution and is more than an order of magnitude faster than applying the best generalized associations mining algorithms to the whole updated database. Spatial databases usually contain not only traditional data but also the location or geographic information about the corresponding data. Spatial association rules describe the relationship between one set of features and another set of features in a spatial database. The form of spatial association rules is also X⇒Y, where X, Y are sets of predicates and of which some are spatial predicates, and at least one must be a spatial predicate.[26] Temporal association rules can be more useful and informative than basic association rules.

6. Recent Advances in Association Rule Discovery

A serious problem in association rule discovery is that the set of association rules can grow to be unwieldy as the number of transactions increases, especially if the support and confidence thresholds are small. As the number of frequent itemsets increases, the number of rules presented to the user typically increases proportionately. Many of these rules may be redundant.

6.1 Redundant Association Rules

To address the problem of rule redundancy, four types of research on mining association rules have been performed. First, rules have been extracted based on user-defined templates or item constraints. Secondly, researchers have developed interestingness measures to select only interesting rules. Thirdly, researchers have proposed inference rules or inference systems to prune redundant rules and thus present smaller, and usually more understandable sets of association rules to the user. Furthermore, their methods eliminate redundant rules in such a way that they never drop any higher confidence or interesting rules from the resultant rule set. Moreover, there is a need for human intervention in mining interesting association rules. Such intervention is most effective if the human analyst has a robust visualization tool for mining and visualizing association rules.

6.2 Other measures as interestingness of an association

Omiecinski[28] concentrates on finding associations, but with a different slant. That is, he takes a different view of significance. Instead of support, he considers other measures, which he calls all-confidence, and bond. All these measures are indicators of the degree to which items in an association are related to each other. With all confidence, an association is deemed interesting if all rules that can be produced from that association have a confidence greater than or equal to a minimum all-confidence value. The performance results showed that the algorithm can find large

itemsets efficiently. The measure of significance of associations that is used is the chi-squared test for correlation from classical statistics. In the authors still use support as part of their measure of interest of an association. However, when rules are generated, instead of using confidence, the authors use a metric they call conviction which is a measure of implication and not just co-occurrence. In the authors present an approach to the rare item problem. The dilemma that arises in the rare item problem is that searching for rules that involve infrequent (i.e., rare) items requires a low support but using a low support will typically generate many rules that are of no interest. Using a high support typically reduces the number of rules mined but will eliminate the rules with rare items.

6.3 Negative Association Rules

Typical association rules consider only items enumerated in transactions. Such rules are referred to as positive association rules. Negative association rules also consider the same items, but in addition consider negated items (i.e. absent from transactions). Negative association rules are useful in market-basket analysis to identify products that conflict with each other or products that complement each other. Mining negative association rules is a difficult task, due to the fact that there are essential differences between positive and negative association rule mining. The researchers attack two key problems in negative association rule mining:
(i) how to effectively search for interesting itemsets, and
(ii) how to effectively identify negative association rules of interest.
Brin et. al[29] mentioned for the first time in the literature the notion of negative relationships. Their model is chi-square based. They use the statistical test to verify the independence between two variables. To determine the nature (positive or negative) of the relationship, a correlation metric was used. In the authors present a new idea to mine strong negative rules. They combine positive frequent itemsets with domain knowledge in the form of taxonomy to mine negative associations. However, their algorithm is hard to generalize since it is domain dependant and requires a predefined taxonomy. In the authors use only negative associations of the type $X \Rightarrow Y$ to substitute items in market basket analysis.

7. Conclusion

Association rule mining has a wide range of applicability such market basket analysis, medical diagnosis/research, website navigation analysis, homeland security and so on. In this paper, we surveyed the issues related to data mining and how ARM in detail. The conventional algorithm of association rules discovery proceeds in two steps. All frequent itemsets are found in the first step. End users of association rule mining tools encounter several well known problems in practice. First, the algorithms do not always return the results in a reasonable time. It is widely recognized that the set of association rules can rapidly grow to be unwieldy, especially as we lower the frequency requirements. The larger the set of frequent itemsets the more the number of rules presented to the user, many of which are redundant. This is true even for sparse datasets, but for dense datasets it is simply not feasible to mine all possible frequent itemsets, let alone to generate rules, since they typically produce an exponential number of frequent itemsets, finding long itemsets of length 20 or 30 is not uncommon. Although several different strategies have been proposed to tackle efficiency issues, they are not always successful.
Agriculture and information technology are closely interwoven. Making use of those data via IT often leads to dramatic improvements in efficiency. For this purpose, the challenge is to change these raw data into useful information. These techniques can be used for prediction of crop yield and occurrence of pests and bring a tremendous change in the filed of agriculture.

8. References

1. Two Crows, (2005). Introduction to Data Mining and Knowledge Discovery, Third Edition, Two Crows Corporation.
2. Han, Jiawei and Micheline Kamber, (2001) Applications and Trends in Data Mining: Concepts and techniques, San Francisco: Morgan Kaufmann Publishers.
3. Osmar R. Zaïane,(1999). Principles of Knowledge Discovery in Databases, pp 1-15.

4. Data Mining: Issues http://www.anderson. ucla.edu/faculty/jason.frand/teacher/technologies/palace /issues.html.
5. Chris Clifton, Security Issues in Data Mining at http://www.cs.purdue.edu/homes/clifton /cs590m.
6. Helen Vanderberg and Pam Sogard,(1995). Data Mining Fundamentals, MineSet™ 2.5.
7. O. Goldreich, S. Micali and A. Wigderson, (1987). How to Play any Mental Game - A Completeness Theorem for Protocols with Honest Majority., Proceedings of the 19th Annual Symposium on the Theory of Computing (STOC), ACM, pp 218–229.
8. A. C. Yao, (1986). How to generate and exchange secrets, Proceedings 27th Symposium on Foundations of Computer Science (FOCS), IEEE, pp 162–167.
9. Heng Chhay, http://cseserv.engr.scu.edu/ Student WebPages/hchhay/hchhay_FinalPaper.htm.
10. Rokia Missaoui, Roman M. Palenichka, (2005). Effective image and video mining: an overview of model-based approaches", pp 43-52.
11. Umamaheshwaran R, Bijker W, Stein A, (2007). Geoscience and Remote Sensing, IEEE Transactions on Volume 45, Issue 1, pp:246 – 253.
12. Wynne Hsu, Mong Li Lee, Kheng Guan Goh, (2000). Image mining in IRIS: integrated retinal information system, International Conference on Management of Data.
13. Smeaton A F, W. Kraaij and P. Over,(2003). Trecvid- an overview. In Proceedings of Trecvid. USA: NIST.
14. Smeulders, Worring, Santini and Jain R,(2000). Content-based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(12):1349–1380.
15. Spitters M and W Kraaij, (2002). Unsupervised clustering in multilingual news streams. Proceedings of the LREC workshop: Event Modelling for Multilingual Document Linking, pp 42–46.
16. Squire, D. McG., W. Muller, H. Muller, and T. Pun, (2000). Content-based query of image databases: inspirations from text retrieval. In Pattern Recognition Letters, volume 21.
17. Vries, A. de and T. Westerveld, (2004). A comparison of continuous vs. discrete image models for probabilistic image and video retrieval. In Proceedings International Conference on Image Processing (ICIP'04).
18. Kuper, J. and H. Saggion et al, (2003). Intelligent multimedia indexing and retrieval through multi-source information extraction and merging. In 18th International Joint Conference of Artificial Intelligence (IJCAI). Acapulco, Mexico.
19. J. Li, T. Wang, W. Hu, M. Sun and Y. Zhang, (2006). Two dependence Bayesian network for soccer highlight detection, IEEE International Conference on Multimedia & Expo (ICME).
20. X.S. Hua, L. Lu, and H.J. Zhang, (2003). AVE – automated home video editing, ACM Multimedia.
21. A. Ekin, A.M.Tekalp and R. Mehrotr,(2003). Automatic soccer video analysis and summarization, IEEE Trans. on Image Processing, 12(7), pp 796–807.
22. H. Bai, W. Hu, T. Wang, X. Tong and Y. Zhang,(2006). A novel sports video logo detector based on motion analysis, International Conference on Neural Information Processing (ICONIP).
23. M. Xu, N. Maddage, C. Xu, M. Kankanhalli and Q.Tian, (2003). Creating audio keywords for event detection in soccer video, IEEE International Conference on Multimedia & Expo (ICME).
24. P. Viola and M. Jones,(2001). Rapid object detection using a boosted cascade of simple features, IEEE Int'l Conf. on Computer Vision and Pattern Recognition (CVPR).
25. C. Huang, H. Ai, et al., (2005). Vector boosting for rotation invariant multi-view face detection, IEEE International Conference on Computer Vision (ICCV).
26. Y. Li, H. Ai, C. Huang and S.H. Lao,(2006). Robust head tracking with particles based on multiple cues fusion, HCI/ECCV, LNCS 3979, pp 29–39.
27. L. Kennedy,(2006). Lscom lexicon definitions and annotations (version 1.0), DTO Challenge Workshop on Large Scale Concept Ontology for Multimedia, Columbia University Advent Technical Report #217-2006-3.
28. NIST, TREC Video Retrieval Evaluation, at http://www-nlpir.nist.gov/projects/trecvid.
29. J. Cao, Y. Lan et al., (2006). Intelligent multimedia group of Tsinghua University at Trecvid in Proceedings Trecvid.