International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:6, No:1, 2012

Role of Association Rule Mining in Numerical Data Analysis

Sudhir Jagtap, Kodge B. G., Shinde G. N., Devshette P. M

Abstract**—**Numerical analysis naturally finds applications in all fields of engineering and the physical sciences, but in the 21st century, the life sciences and even the arts have adopted elements of scientific computations. The numerical data analysis became key process in research and development of all the fields [6]. In this paper we have made an attempt to analyze the specified numerical patterns with reference to the association rule mining techniques with minimum confidence and minimum support mining criteria. The extracted rules and analyzed results are graphically demonstrated. Association rules are a simple but very useful form of data mining that describe the probabilistic co-occurrence of certain events within a database [7]. They were originally designed to analyze market-basket data, in which the likelihood of items being purchased together within the same transactions are analyzed.

Keywords**—**Numerical data analysis, Data Mining, Association Rule Mining

I. Introduction

BEFORE the advent of modern computers numerical methods and analysis are often depended on hand interpolation in large printed tables. Since the mid 20th century, computers calculate the required functions instead. The interpolation algorithms nevertheless may be used as part of the software for solving differential equations [6].

A. Numerical Analysis

Everyone knows that when scientists and engineers need numerical answers to mathematical problems, they turn to computers. Nevertheless, there is a widespread misconception about this process. The power of numbers has been extraordinary. It is often noted that the scientific revolution was set in motion when Galileo and others made it a principle that everything must be measured. Numerical measurements led to physical laws expressed mathematically, and, in the remarkable cycle whose fruits are all around us, finer measurements led to refined laws, which in turn led to better technology and still finer measurements. The day has long since passed when an advance in the physical sciences could be achieved, or a significant engineering product developed, without numerical mathematics.

Sudhir Jagtap and Kodge B. G. are with Swami Vivekanand Mahavidyalaya, Udgir, Dist. Latur (MS), India, sudhir.jagtap7@gmail.com, kodgebg@gmail.com

Shinde G. N is with Indira Gandhi College, CIDCO, Nanded (MS), India , shindegn@yahoo.co.in

Devshette P. M is with Shri Havagiswami College, Udgir, Dist. Latur (MS), India, p_devshette@yahoo.com

Computers certainly play a part in this story, yet there is a misunderstanding about what their role is. Many people imagine that scientists and mathematicians generate formulas, and then, by inserting numbers into these formulas, computers grind out the necessary results. The reality is nothing like this. What really goes on is a far more interesting process of execution of algorithms. In most cases the job could not be done even in principle by formulas, for most mathematical problems cannot be solved by a finite sequence of elementary operations. What happens instead is that fast algorithms quickly converge to “approximate” answers that are accurate to three or ten digits of precision, or a hundred. For a scientific or engineering application, such an answer may be as good as exact.

Though data mining is a reference to holistic data with relevance, numerical data mining is pertinent to just the numerical aspect of it. When systems, automated and semi-automated alike, are put through extensive screening and analysis, a huge amount of numerical data germinates out of the whole ordeal. Interpreting these numbers becomes a very difficult task for a lot of statistics go into interpretations. The interpretations also involve a lot of organic data, as it is mostly about chemical compounds and drugs mentioned in formulas abiding by organic chemistry [2].

B. Association Rule Mining

In data mining, association rule mining is a popular technique and well researched method for discovering interesting relations between variables in large databases. Piatetsky-Shapiro[3] describes analyzing and presenting strong rules discovered in databases using different measures of interestingness. Based on the concept of strong rules, Agrawal et al.[1] introduced association rules for discovering regularities between products in large scale transaction data recorded by point-of-sale (POS) systems in supermarkets.

For example, the rule {onion, potatos}⇨ burger, found in the sales data of a supermarket would indicate that if a customer buys onions and potatoes together, he or she is likely to also buy hamburger meat. Such information can be used as the basis for decisions about marketing activities such as, e.g., promotional pricing or product placements. In addition to the above example from market basket analysis association rules are employed today in many application areas including Web usage mining, intrusion detection and bioinformatics.

Following, the original definition by Agrawal et al.[4] the problem of association rule mining is defined as: Let $I=\{i_1, i_2, \ldots..i_n\}$ be a set of n binary attributes called items. Let $D= \{t_1, t_2,\ldots t_m\}$ be a set of transactions called the database. Each transaction in D has a unique transaction ID and contains a subset of the items in I. A rule is defined as an implication of

the form X ⇨ Y where X, Y ⊆ I and X ∩ Y = ∅. The sets of items (for short item sets) X and Y are called antecedent (left-hand-side or LHS) and consequent (right-hand-side or RHS) of the rule respectively.

To illustrate the concepts, we use a small example from the supermarket domain. The set of items isI = {milk, bread, butter, beer} and a small database containing the items (1 codes presence and 0 absence of an item in a transaction) is shown in the table to the right. An example rule for the supermarket could be {butter, bread}⇨ {milk} meaning that if butter and bread is bought, customers also buy milk.

To select interesting rules from the set of all possible rules, constraints on various measures of significance and interest can be used. The best-known constraints are minimum thresholds on support and confidence.

The support supp(X) of an item set X is defined as the proportion of transactions in the data set which contain the item set. In the example database, the item set {milk, bread, butter} has a support of 1 / 5 = 0.2 since it occurs in 20% of all transactions (1 out of 5 transactions).
The confidence of a rule is defined as:

$$conf(X⇨Y)=supp(X \cup Y)/supp(X)$$

For example, the rule {milk, bread, butter} ⇨ {butter} has a confidence of 0.2 / 0.4 = 0.5 in the database, which means that for 50% of the transactions containing milk and bread the rule is correct.

Confidence can be interpreted as an estimate of the probability P(Y | X), the probability of finding the RHS of the rule in transactions under the condition that these transactions also contain the LHS [2]. The lift of a rule is defined as:

$$Lift(X⇨Y)=supp(X\cup Y)/supp(Y)*supp(X)$$

or the ratio of the observed support to that expected if X and Y were independent. The rule {milk, bread}⇨{butter} has a lift of 0.2/0.4*0.4 = 1.25.

The conviction of a rule is defined as:

$$Conv(X⇨Y)= 1-supp(Y)/1-conf(X⇨Y)$$

The rule {milk, bread}⇨{butter} has a conviction of 1-0.4/1-0.5=1.2, and can be interpreted as the ratio of the expected frequency that X occurs without Y (that is to say, the frequency that the rule makes an incorrect prediction) if X and Y were independent divided by the observed frequency of incorrect predictions. In this example, the conviction value of 1.2 shows that the rule {milk, bread}⇨{butter} would be incorrect 20% more often (1.2 times as often) if the association between X and Y was purely random chance.

The property of succinctness (Characterized by clear, precise expression in few words) of a constraint. A constraint is succinct if we are able to explicitly write down all Item-sets,that satisfy the constraint.

With reference to the above discussed sections 1.1 and 1.2, we have made an attempt to perform few experiments in MATLAB to analyze the numerical data using association rule mining method.

II. Data, Methods and Material

The data mining tool that extracts Association Rules from numerical data files using a variety of selectable techniques and criteria. The program integrates several mining methods which allow the efficient extraction of rules, while allowing the thoroughness of the mine to be specified at the user's discretion. The program was designed as a tool to assist in the analysis of both the knowledge extracted and the deduction processes by which such a task is undertaken. However, the program can also be used as a straightforward Data Mining tool for the efficient extraction of Association Rules.

The actual knowledge extracted is presented in the form of easy-to-understand rules, while the details of the process, such as time taken and file size considered, are conveniently summarized. The program also allows the results to be displayed through various graphical representations, such as bar charts and line graphs. Such graphics can often help to summarize the knowledge being analyzed by providing a concise conceptualization of the data under scrutiny.

The training numerical data is collected online from a lottery site, processed through required format i.e. in CSV (comma separated value) format. The MATLAB software is used to design and develop the appropriate algorithms for numerical data analysis using association rule mining. Although the type of numerical data which ARMADA could be used to mine are virtually endless, common examples of data sets include;

- POS (Point of Sale) Transaction data
- Medical databases
- Census data
- Statistical data
- Lottery Results (not guaranteed to provide winning lines!)

III. Numerical Analysis using ARM

Association rules mining (ARM) are usually required to satisfy a user-specified minimum support and a user-specified minimum confidence at the same time. Association rule generation is usually split up into two separate steps:

1. First, minimum support is applied to find all frequent item sets in a data.
2. Second, these frequent item sets and the minimum confidence constraint are used to form rules.

While the second step is straight forward, the first step needs more attention.

Finding all frequent item sets in a database is difficult since it involves searching all possible item sets (item combinations) [4].

Selecting criteria to perform mining by is a task that may, at first, appear daunting. There are questions that arise when making the selections before mining which can not be awarded concrete answers. For example, what level of confidence is going to provide a 'useful' set of Association Rules, would the results be just as effective if the Data Sampler was used rather than a full mine, and so on. Here lies the oxymoron that is Association Rule mining. The value of discovering specific

and accurate knowledge from such data mining, is in the unknown quantities of the data set being mined. The simple answer to some of these questions, and more, is that there is no simple answer, at least before mining is undertaken. For this reason, this section is meant as a guide to assist in the process of selecting criteria by which to perform mining, not as a set of concrete rules that must be followed every time in order to produce effective results.

A. Selecting mining criteria

One of the most difficult decisions that must be taken is the selection of the mining criteria, specifically, the two attributes that fall into this category – minimum Support and minimum Confidence. There are no hard and fast rules to selecting suitable values for either, however there are some pieces of information that can help when making a decision.

Firstly, the definition of what Support and Confidence are must be understood. Support is the number of times the items in a rule appear together in a single entry within the entire set. Confidence is the number of times that the LHS of a rule leading to the RHS is true within the data set.

So, if all the items in a rule appeared together 5 times in a data set with only 10 entries, then the support is 5 or 50%. If the LHS of a particular rule led to the RHS in 4 out of those 5 occurrence mentioned, then the confidence is 80%.

The next piece of information to have in mind is that, the lower the values for each of these two criteria, the more rules will be extracted. Therefore, the most rules will be extracted when the values for each are set to 1 (which is the lowest value permitted). Conversely, the higher the values of the two criteria, the smaller the number of rules that will be extracted. Therefore, the least rules will be extracted when the values are set to 100%. The following Fig. 1 request user for numerical data, mining criteria such as minimum confidence and minimum support. The lottery.txt is processed with minimum confidance 30 and minimum support 8%.

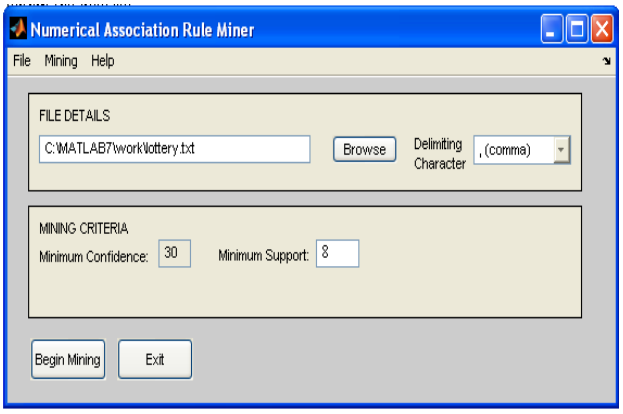

Fig. 1 Preprocessing window.

In practice, the rule of thumb is that to extract all possible rules (sometimes called brute force mining) set the values to the lowest permitted. To extract only rules that apply to every entry in a data set, make the values for the criteria 100%. The latter, however, will usually extract no rules at all.

Using this theory, it would not be unreasonable to suppose that, by setting each criterion to 50%, exactly half of the total number of possible rules would be extracted. However, because of the nature of Association Rule Mining, this does not follow. This is because the rules being mined are dependant upon the items within the data set; the criteria are relative to relationships between items within the data not the overall results. 50% support means items that appear together in 50% of each entry in the data set, not 50% of the total amount of rules that can be extracted.

Although there is no single solution to extracting rules, using the above information, the following can be concluded and hence used as guidelines;

1. If a broad range of rules is required, a low minimum criteria should be selected
2. If a small number of highly occurring rules is required then keep criteria high.

One alternative approach is to begin mining with a very high criteria and, if the rules are not sufficient in number, repeat the process for a slightly lower threshold. This will allow mining to be performed until the number of rules are sufficient to an individual's needs. Of course, the pay-off here is that the mining process must be repeated until this goal is achieved which could be time-consuming.

IV. Results

The post-mining part of the program (Fig. 2) which displays the association rules that have been extracted and a report down the right hand side.

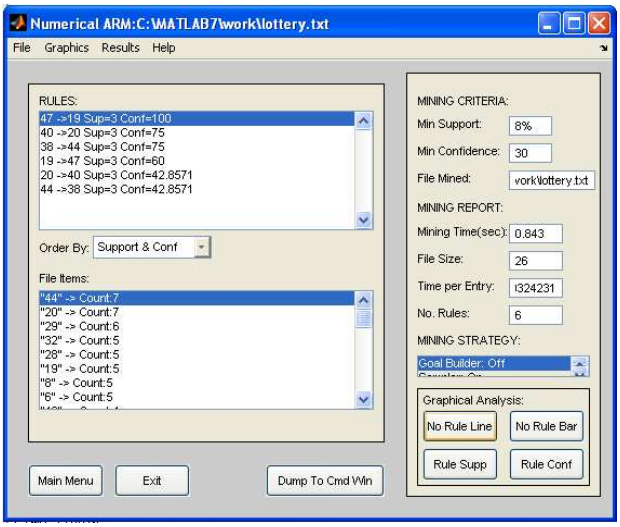

Fig. 2 Extracted rules and file items

The Rules section box displays all of the rules that have been mined using the specified criteria. The rules appear in the format of;

LHS Item(s) -> RHS Item(s) Sup = number Conf = number

The LHS (Left Hand Side), or antecedent, item(s) appear to the left of the '->' symbol. Multiple items are separated with a space. Similarly, the RHS (Right Hand Side), or consequent, items appear to the right of the '->' symbol. The support of the

rule is represented as a numeric value after the 'Sup =' part. The confidence of the rule is represented as a numeric value after the 'Conf =' part.

A. Analyzing the rules

The rules are displayed in a straightforward manner which is relatively self-explanatory. The rules displayed in the Rule Box appear in the format of;

LHS Item(s) -> RHS Item(s) Sup = number Conf = number

An example of such a rule could look like;

1234 2345 3456 -> 4567 Sup=10 Conf=70

This would translate to mean that the items '1234', '2345' and '3456' lead to '4567' with support of 10 and confidence of 70. The displayed rules can also be sorted in order of either their support and confidence (highest top of the list) and by the six of the LHS part of the rule (with one part LHS rules top of the list). The file items list is also provided to give some insight into the data set that has been mined. These items are ordered by their support (highest top of the list). The criteria that were specified are also displayed in the window. Minimum support is displayed as a number alone if it was specified as such, or, if it was specified as a percentage of the data set, then a '%' sign follows the value. Minimum confidence is always expressed as a percentage.

The mining report section displays;

- the total time that mining took
- the file size that was analysed for mining
- the time per entry, which is the time taken to mine each rule
- the number of rules that have been extracted from the mining

B. Rule support analysis

Next, the tool plots a line graph of the support of the rules against the number of rules, starting from the highest support as the left most value down to the lowest at the right most value. This graph often describes a 'waterfall' effect when analysing the rules as the support decreases or a straight horizontal line if the support is constant throughout. (Fig. 3).

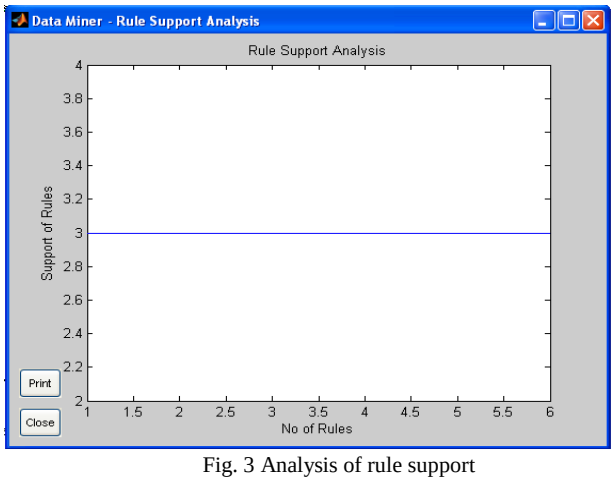

Fig. 3 Analysis of rule support

C. Rule confidence analysis

Finally, the tool plots a line graph of the confidence of the rules against the number of rules, starting from the highest as the left most value down to the lowest as the right most value. Again, this graph often describes a 'waterfall' effect when analyzing the support of the rule or a straight line if the confidence is constant throughout. (Fig. 4).

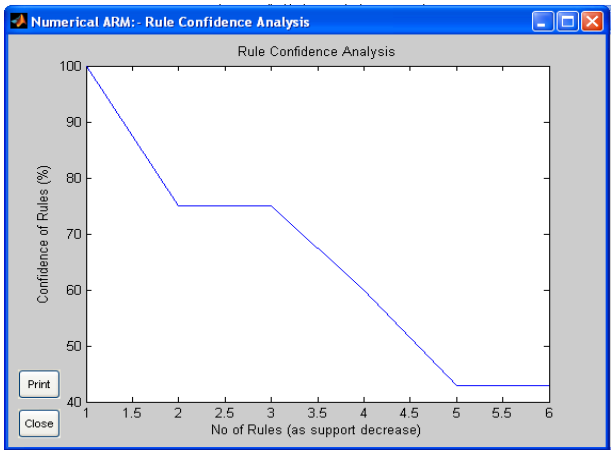

Fig. 4 Analysis of rule confidence

V. Conclusion

The proposed tool that extracts the association rules from numerical data files using a variety of selectable techniques and criteria. The program integrates several mining methods which allow the efficient extraction of rules, while allowing the thoroughness of the mine to be specified at the users discretion. The program also allows the results to be displayed through various graphical representations, such as bar charts and line graphs. Such graphics can often help to summarize the knowledge being analyzed by providing a concise conceptualization of the data under scrutiny.

References

[1] R. Agrawal; T. Imielinski; A. Swami: Mining Association Rules Between Sets of Items in Large Databases", SIGMOD Conference 1993: 207-216
[2] Jochen Hipp, Ulrich Güntzer, and Gholamreza Nakhaeizadeh. Algorithms for association rule mining - A general survey and comparison. SIGKDD Explorations, 2(2):1-58, 2000.
[3] Piatetsky-Shapiro, G. (1991), Discovery, analysis, and presentation of strong rules, in G. Piatetsky-Shapiro & W. J. Frawley, eds, 'Knowledge Discovery in Databases', AAAI/MIT Press, Cambridge, MA.
[4] Tan, Pang-Ning; Michael, Steinbach; Kumar, Vipin (2005). "Chapter 6. Association Analysis: Basic Concepts and Algorithms". Introduction to Data Mining. Addison-Wesley. ISBN 0321321367.
[5] http://www.b3intelligence.com/NumericalDataMining.html
[6] http://en.wikipedia.org/wiki/Numerical_analysis
[7] http://www.mathworks.com/matlabcentral/fileexchange/3016-armada-data-mining-tool-version-1-4