**Available Online at www.ijcsmc.com**

**International Journal of Computer Science and Mobile Computing**

**A Monthly Journal of Computer Science and Information Technology**

**ISSN 2320–088X**
**IMPACT FACTOR: 6.017**

***IJCSMC, Vol. 8, Issue. 3, March 2019, pg.230 – 239***

# Evaluation and Validation of the Interest of the Rules Association in Data-Mining

By
**Ali Yousif Hasan**
Supervised by: **Dr. Sefer Kurnaz**
Altinbaş University, Turkey

***Abstract. The interesting association rules is a special part of knowledge extraction from data. Apriori's support- and rule-based algo-rithms have provided an elegant solution to the problem of rule mining, but they produce too much rules, selecting some rules of no interest and ignoring rules[3] [5]. interesting. Other measures are needed to complete the support and the confidance. In this paper, we review the main measures proposed in the literature and we propose criteria to evaluate them. We then suggest a validation method that uses the tools of statistical learning theory, including VC -dimension. Given the large number of measurements and the multitude of candidate rules, the interest of these tools is to allow the construction of uniform non-asymptotic terminals for all the rules and all the measurements simultaneously.***
***Keywords: association,rules,data mining, validation, evaluation, Dimensions***

## 1. Introduction

The research of associated rules between Boolean attributes is already old, linked to the analysis of 2 × 2 cross-tabulations. As underlined [2], one of the earliest method of finding the rules of association is the GUHHA method initiated by [5], where the basic of support and confidence already appear. Interest in the rules of association has been renewed by the works of (Agrawal, Imielinski and Swami 1993), (Agrawal and Srikant 1994), and then (Srikant and Agrawal 1995) relating to the extraction of association rules from large data that record the content of business transactions.

Since the association-rules increases sumiltanitly with the total of features , it is important to limit the extraction of the most interesting rules. To do this, you must be able to de ne

 

these and identify them, then you have to validate them. We first present the algorithms related to the support and con ance criteria [9] and we show the limits of this approach. We then specify the notion of rule in relation to that of implication or correlation . After having indicated the de nition of the main measures of the interest of the rules we indicate the criteria on the basis of which we can evaluate them . We then propose a synthetic presentation of these measures [10]Finally, we address the problem of validation of rules through the tools of statistical learning and we indicate some perspectives [8].

## 2. Approach Interest of the support and confidence

By a support-confidence approach, we define the extraction algorithms which exhaustively seek the association rules whose support and confidence exceed the thresholds fixed in advance by the user, noted minsupp and minconf.

### 2.1 **Support & Confidence**

Let $N^A$ and $N^B$ are the sum of transactions that continuesly perform the items of A and B, n (AB) the number of those that perform both A and B. Support of a rule is the distribution of activities that achieve them both :

$$Supp\,(A \rightarrow B) = P\,(AB) = \frac{n(AB)}{n}$$

while its con ance is the distribution of activities that realize B, among those that realize A, that is to say the conditional relative frequency of B knowing A:

$$Conf\,(A \rightarrow B) = \frac{P(AB)}{P(A)} = \frac{n(AB)}{n(A)} = 1 - \frac{n\left(A\overline{B}\right)}{n(A)}$$

### 2.2 Extraction algorithms following support and confidence approach

The extraction algorithms linked to the support-confidentiality approach traverse the lattice of the itemsets to search for frequent itemsets, those whose support exceeds minsupp, to deduce the rules of association whose con ance exceeds minconf.
Indeed, the mesh of the itemsets admits a double property which makes very efficient the condition of support during the search for the frequent ones:

- Any subdata of a non-periodic element is uncommon.
- Any subdata of a periodic element is common.

the founding algorithm (Agrawal and Srikant 1994) proceeds in two stages:

1. We look for frequent itemsets, those whose support exceeds minsupp, by scanning the lattice of itemsets in its width and calculating the frequencies by counting in the base, which imposes a pass on the base at each level of the lattice.
2. For each item and frequent X, we keep the only conditions of the type X \ Y → Y, with Y ⊂ X, whose con ance exceeds the threshold minconf.

The rules deduced from the frequent itemsets necessarily have a confidence higher than the support threshold, since Supp (A → B) <Conf (A → B).

 

The efficiency of Apriori decreases in the presence of dense or strongly correlated data. The whole problem of frequent extraction consists of identifying the border between frequent itemsets and non-frequent itemsets in the lattice of itemsets [7]The search can be done in width in the lattice or in depth. In each case, we can proceed by direct counting of the frequency of each item and in the database, or proceed by intersection of the two itemsets which constitute the itemset candidate. Improvements proposed to accelerate the construction of frequent assemblies in certain situations include:

Extraction of a sample of the database that is stored in memory, from which we build the set of frequent itemsets in the sample as well as its negative border consisting of minimal non-frequent itemsets of which all the parts are frequent [10], which limits the risk of incompleteness;

1-progressive decrease of the base: instead of making a pass during the examination of each level of the lattice of the itemsets, we put the whole base in memory and at each

2-Trellis level, transactions are represented by the candidate k-itemsets it contains; a single pass then suffices, but the whole base must be remembered[6].

3-dynamization of the algorithm: one proceeds by levels in the lattice, but at the level k as soon as an itemset has reached the threshold of frequency, one introduces the itemsets candidates of level k + 1 which it contributes to generate, which decreases the number of passes required on the base

| $A \setminus B$ | 0 | 1 | total |
|---|---|---|---|
| 0 | $P(\overline{A}\overline{B})$ | $P(\overline{A}B)$ | $P(\overline{A})$ |
| 1 | $P(A\overline{B})$ | $P(AB)$ | $P(A)$ |
| total | $P(\overline{B})$ | $P(B)$ | 1 |

Tab. 1 Ratings for the disposal of elements A and B

 

| A\B | 0 | 1 | total |
|---|---|---|---|
| 0 | 1− c/l − s/c + s | c/l − s | 1 − s/c |
| 1 | s/c – s | s | s/c |
| total | 1− c/l | c/l | 1 |

Tab. 2 Distribution of A and B according to the support “s”, confidence “c” and lift “l”

The support condition which is the driving force of the extraction process rules out the regulation with a fit support whereas some can have a very strong confidence and present a real interest, the case is common in digital-marketing (the nuggets of the Data Minis ning). If the support is decreased to overcome this disadvantage, the frequent sets are too numerous and the extraction algorithms are asphyxiated.

Finally, the only conditions of support and trust are not enough to ensure the real interest of a ruler. Indeed, a rule A to B whose confidence is equal to the probability of B, ie P (B / A) = P (B) which is the definition of the independence of A and B, bring no information! For example, if P (A) = 80% and P (B) = 90%, the Rule A to B has a support equal to 72% and a confidence of 90% in case of independence.

In summary, one must at least consider other measures of interest in the rules than support and trust, thus favoring an induction bias, hence the importance of realizing to fling to the particular nature of the rules of association if one wants to establish a test bench of the main measures of interest.

Active Algorithms for Association Mining Rules

This part is an extremely long and complicated paper regarding taking a bunch of actions and award association rules in them. as an example, promoting a business may like to boost “What proportion of people which agency bought X to boot bought Y?” Another question may be “What a pair of things are highest amoung of us between ages 10 and twenty 27.” The original resolution would undertake normally.AN thoroughgoing all over finding of the subsets of things along with the count what range exactly convince and please the base conditions that tend to tend to are attempting to find. Considering this viewpoint approach, tho' whether lower in

 

term of capacity-wise (Focusing on storage of mixtures that tend to tend to need) would waste lots of it slow (generating entire potentially available possibilities). Further the document elaborates some unique sets and combinations that start together with a initial data set (earlier which convince a scientist establishes that tend afore tend to require to compare) with establishing all further to the datasets of highest magnitude and area. Also, not entire build working is going to be tackled special fashion. Focusing on the both set in particular document to look especially our faith and assistance.

- Set "X" along with "Y" , associate coorporation rule "X" = "Y" with trust level "c" if filled with dealings particular to action dataset which have "X" comprise "Y".

- Set "X" with "Y", peer coorporation rule "X" = "Y" with supports that this is the case of dealing particular to undertake dataset which has "X ∪ Y".

Above formula focuses on searching out entire linkage patterns which have little all stripped subsistence with a few lesser level of belief. It has been accompolished through initial searching of entire linkages which have lowest assistance (so lowering particularly house of "coorporation rules" which need to be evaluated). Further, afterward check particular datasets regarding people who found with lowest assistence with implementing the "Bayes Rule":

$$\text{The association rule } X \Longrightarrow Y \text{ has confidence } = \frac{support(Y)}{support(X)}$$

A large itemset is one that has bottom support. That is, a minimum of such a large amount of all transactions in our dealing information were transactions on things . AIS discovers entire datasets mostly with thoroughgoing finding (Ensuring entire attainable mixtures with data).

Apriori includes the person generation keystep which originates "k-datasets" with collaborating on (k − 1)-item sets. In continuation, a further overgrown measured which eliminated "k-item sets" which contain a (k − 1)-itemset which doesn't allow

 

bottom assistence. It looks further for dealing “T” that holds contains “k-itemset” in the question, afterward that should hold each set of particular itemset also.

Further it can be reason of the overgrown measures which do not take away either massive “k-itemset”. moreover, from start it has a tendency of further moving along with belief which datasets are often lexicographically instructed, that looks it must always ready to synthethicly visualize that every one massive “k-item sets” are often created with be a part of by getting of the biggest 2 parts from them along with making (k − 1)-item sets.

Apriori with the use of “hash-tree”, specifically tree wherever every node could be hash-table, for storing of item sets.Further with the implementation of the hash, tree is examined from depth “d” to operate upto item “d” within the item set. Further hash results that operate elaborate US that child-pointer to require. Also, lists of leaves store with associate degree item set which were enlightened through regulation.

Entire nodes are leaf nodes from the initial step, however once quantity of enitre combinations in an exceedingly leaf node originated giantly, further reborn deeply in an internalnode.

In additon, for determining either or not associate degree item set “I” to be hold among group action “T”, that have tendency to use a ikon that containsall things in “T”.

“Bitmap (I) ⊆ items (T) iff bitmap (I) & bitmap (items(T)) == bitmap (I)”

“AprioriTID” strives further scale back amount particularly thatof info reads. In addition, assume further that often necessary that have enogh I/O which is actually slower. Also, may even more helpful that tended to negate wish for watching and examining the scan-lock for longer period. Parallel, all transactions were seen in: hT ID with storage of candidate item sets, i. create particularly extra storage economical, also further designate associate ID variety instead to item set, holding “hT ID, Idi”. Unhappily, as negating the wish with stay trying up that things which could be within particular combination known by “ID”, which holds both extension and generator. To conclude this have been sets which could be part ofed for further creating particular item set “generators”. So, hence item sets which have been generated through mixtures on this item set “extensions”.

 

**Assessing of Performance**

It was essential to conduct an evaluation of the algorithms performance and therefore the computational power was important. To ensure that we obtained the nece3ssary computational power, the workstation was equipped with a CPU that was running on AIX 3.2 with a main memory of 62 MB while the clock rate was running at 33 MHz. The workstation was IBM RS/6000 530H and it was to carry out different experiments. The throughput was measured at 2 MB per second while the drive was 3.5" and the size of the SCSI was 2GB. This is the drive that was used to store the data which was stored in the AIX file system. To ensure that the performance is clearly explained, it is necessary to first provide a discussion of the AIS algorithm. After this performance discussion is provided, the next step is to dicusss the SETM performance of the experiments. After discussing this performances, a comparison with the AprioTid and Apriori algorithms is provided. Afterwards, the next step is providing the description of the synthetic datasets that were used in the processes of data presentation as well as the evaluation of the performance. After this information has been discussed, the final step is giving a description of how the combination of Apriori and AprioriTid features can be used to create a AprioriHybrid algorithm. İt was also essential to give a demonstration or explanation of the features of the resulting hybrid.

**The SETM Algorithm**

İt was determined that SQL can be a useful tool in the computation of large datasets and therefore this is the reason why it was decided to utilize it for that purpose. This algorithm has the capability of generating on-the-fly based candidatres by using transactions that have been read or obtained from the database. For this reason, it was possible to use the tool for generatng and counting every itemset that was generated by the AIS Algorithm. One of the setbacks for using SQL in generating candidates is that the SETM algorithm features allow it to separate counting from the process of generating candidates. The saving of the data involves storing the candidate's copy of the itemset together with the generating transactions' TID.

However, the storage takes place in a sequence structure. The itemsets are finally determined after the pass where the aggregating and sorting of the sequence structure takes place. İn order to remember the datasets, the SETM recalls this information by linking the candidate itemsets with the generating transactions' TIDs. İt is essential that the tool does not require subset operations and therefore this data

 

is used in identification of the itemsets and separates it from others in the large itemsets where they are located from the transactions where it has been read. Candidates that do not have minimu datasets that have been determines as not requiring minimum support are deleted. An assumption is made which presumes that a TID order has been used to store the information. Using this assumption, it becomes possible for the SETM to identiufy or sort itemsets in the large itemset by sorting TIDs when the following transactions are being read. The major adnatage is that the SETM only requires a single visitation of the TIDs sequential order so that it can be able to generate other candidates. To facilitate this generation process, the tool uses a merge-join operation that is relational.

The major limitation of this procedure is brough about by the fact that it candidate sets' sizes become large as the processes progresses. İt is so because whenver a candidate is generated, the number of entries for each set is equivalent to the number of transactions where the itemset is located. This is because after attaining the desired itemsets when the time comes for counting the minimum support requirement, it is found that the datasets are not arranged in the correct order. Therefore, it becomes necessary to sort them correctly. İn addition, it also makes it necessary to sort the TIDs before the next pass is initiated.

**The AIS Algorithm**

Similar to the SETM algorithm, this algorithm allows using the on-the-fly to failitate the counting and generation of candidate itemsets. This is done when the database is being scanned. After scanning and reading the dataset, the algorithm determines the transactions' itemsets that were in the previous transaction and are found to be large itemsets. This information is then used in the generation of new candidate itemsets through the extension of these currently determined large datasets with other items that are situated in the transactions in the database. The large itemset which is denoted as l is associated only with large items that are occuring at a later stage in ordering of the items using the lexiographic method[5]. After generating the new candidates after reading of a transaction, the itemsets that have been maintained from the corresponding counts of the entries or the pass if they have been generated from a previous reading of another transaction.

 

**Synthetic data generation**

there was a large range of characteristics for the data after carrying out the process and tehrefore for evaluation of the performance, it was necessary to generate synthetic transactions so that the algorithms could be evaluated efficiently. İt is important to ensure that the generated synthetic transactions have similar characteristics with the retailing environment's transactions. Using this procedure we utilized the designed model that makes the assumption that people tend to purchase itemsets together in the "real world." İt is also assumed that the itemsets in consideration are all individually large itemsets with a maximal value. The model also makes the assumption that it is not necessary to purchase all the items in the dataset at once in every transaction. [2]İn every transaction, it is possible for a customer to purchase more than one large itemset. Therefore, the transactions are used to categorize the itemsets and therefore the transactions may be few but at the same time, the size of the itemsets remains to be large. Consequently, it is necessary to cluster the transactions in mean values.normally, the large itemsets are also clustered around a mean value whereby a large number of items are contained in a few itemsets.

**Conclusion**

Our approach use terribly strict notation, perpetually change our search among sets, itemsets, subsets, candidate sets, large sets, etc. for somebody World Health Organization has not seen this data before, it'd are priceless to work out a definition list within the front of the paper that formally defines every of those terms. it would not are a foul plan to separate the info structure descriptions from the algorithmic program itself. we have a tendency to reached given the running bounds of their knowledge structures and easily used them to research the algorithm's overall performance. Then, they may describe the info structures at liesure in associate degree appendix. this might build the paper significantly less confusing since the reader will specialise in algorithmic program details while not being distracted by hashing and tree traversals.

 

# References

[1]. R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In J.B. Bocca, M. Jarke, and C. Zaniolo, editors, Proceedings of the 20th Very Large Data Bases Conference, pages 487 499. Morgan Kaufmann, 1994.
[2]. R. Agrawal, T. Imielinski, and A.N. Swami. Mining association rules between sets of items in large databases. In P. Buneman and S. Jajodia, editors, ACM SIGMOD Int. Conf. on Management of Data, pages 207 216, 1993.
[3]. J. Azé and Y. Kodrato . A study of the e ect of noisy data in rule extraction systems. In Sixteenth European Meeting on Cybernetics and Systems Research, volume 2, pages 781 786, 2002.
[4]. Y. Benjamini and Y. Hochberg. Controlling the false discovery rate : a practical and powerful approach to multiple testing. J. R. Stat. Soc., B, 57:289 300, 1995.
[5]. Y. Benjamini and W. Liu. A step-down multiple-hypothesis procedure that controls the false discovery rate under independance. J. Stat. Planng Inf., 82: 163 170, 1999.
[6]. C.L. Blake and C.J. Merz. UCI repository of machine learning databases. http://www.ics.uci.edu/~mlearn/MLRepository.html, 1998.
[7]. J. Blanchard, P. Kuntz, F. Guillet, and R. Gras. Mesure de la qualité des rêgles d'association par l'intensité d'implication entropique. Mesures de Qualité pour la Fouille de Données, (RNTI-E-1):33 45, 2004. 24 Lallich et al.
[8]. J. Blanchard, F. Guillet, H. Briand, and R. Gras. Assessing the interestingness of rules with a probabilistic measure of deviation from equilibrium. In ASMDA'05, pages 191 200, 2005.
[9]. C. Borgelt and R. Kruse. Induction of association rules: Apriori implementation. In 15th Conf. on Computational Statistics, 2002.
[10]. S. Brin, R. Motwani, and C. Silverstein. Beyond market baskets: generalizing association rules to correlations. In ACM SIGMOD/PODS'97, pages 265 276, 1997.
[11]. S. Brin, R. Motwani, J.D. Ullman, and S. Tsur. Dynamic itemset counting and implication rules for market basket data. In Joan Peckham, editor, ACM SIGMOD 1997 Int. Conf. on Management of Data, pages 255 264, 1997.

 