Management Science Letters 5 (2015) 167–174

Contents lists available at GrowingScience

Management Science Letters

homepage: www.GrowingScience.com/msl

Loss profit estimation using association rule mining with clustering

Mandeep Mittal[a*], Sarla Pareek[b] and Reshu Agarwal[b]

[a]*Department of Computer Science Engineering, Amity School of Engineering and Technology, Bijwasan, New Delhi, India*
[b]*Apaji Institute of Mathematics and Applied Computer Technology, Banasthali University, Rajasthan, India*

C H R O N I C L E

Article history:
Received September 28, 2014
Accepted 28 December 2014
Available online
January 10 2015

Keywords:
Data mining
Association rule mining
Clustering
Inventory control
Loss Profit
Apriori algorithm

A B S T R A C T

Data mining is the technique to find hidden patterns from a very large volume of historical data. Association rule is a type of data mining that correlates one set of items or events with another set of items or events. Another data mining strategy is clustering technique. This technique is used to create partitions so that all members of each set are similar according to a specified set of metrics. Both the association rule mining and clustering helps in more effective individual and group decision making for optimal inventory control. Owing to the above facts, association rules are mined from each cluster to find frequent items and then loss profit is calculated for each frequent item. Initially, the clustering algorithm is used to partition the transactional database into different clusters. Apriori, a classic data mining algorithm is utilized for mining association rules from each cluster to find frequent items. Later the loss profit is calculated for each frequent item. The obtained loss profit is used to rank frequent items in each cluster. Thus, the ranking of frequent items in each cluster using the proposed approach greatly facilitate optimal inventory control. An example is illustrated to validate the results.

1. Introduction

Data mining is the process to discover previously unknown relationships among the data, especially when the data come from different databases. Businesses can use these new relationships to develop new advertising campaigns or make predictions about how well a product will sell. Data mining techniques, such as classification, association rule mining, sequential pattern mining, and clustering, have attracted attention of several researchers (Zhao & Bhowmick, 2003). Association rules have been broadly used in many applications domains for finding pattern in data. The pattern reveals combinations of events that occur at the same time. One of the best domain is business field, where discovering of pattern or association helps in effective decision making and marketing. The best algorithm for finding association rule is apriori algorithm (Agrawal & Srikant, 1994). Moreover, clustering is the process of organizing objects into groups whose members are similar in some way. Hence, the behavior of the objects is studied by looking at the number of clusters.

*Corresponding author.
E-mail addresses: mittal_mandeep@yahoo.com (M. Mittal)

doi: 10.5267/j.msl.2015.1.004

Broder et al. (1997) defined clusters as maximal connected components of some pair-wise similarity of transactions, thus suffers from the breakdown of the transitivity of pair-wise similarity. Guha et al. (2000) proposed the common neighbors of two transactions as a measure of pair-wise similarity. Wang's et al. (1999) method does not use any notion of pair-wise similarity. They cluster transactions that contain similar items. The difference is that clustering emphasizes on the dissimilarity of clusters. Both the association rule mining and clustering techniques can be used for effective inventory management.

Further, inventory management is mainly about identifying the amount and the position of the goods that a firm has as inventory. Inventory management is imperative as it helps to defend the intended course of production against the chance of running out of important materials or goods. It also includes making essential connections among the replenishment lead time of goods, asset management, the carrying costs of inventory, future inventory price forecasting, physical inventory, available space for inventory, etc. By balancing these competing requirements, a company will discover its optimal inventory levels. For inventory management, many researchers have devoted a great amount of efforts in developing inventory models. Porteus (1986) incorporated the effect of imperfect quality items into the basic economic order quantity model. Rosenblatt and Lee (1986) assumed that the time between the beginning of the production run; i.e., the in-control state; until the process goes out of control is exponential and the defective items can be reworked instantaneously at a cost and concluded that the presence of defective products motivates smaller lot sizes. Later, Lee and Rosenblatt (1987) considered using process inspection during the production run so that the shift to out-of-control state can be detected and restoration made earlier. Salameh and Jaber (2000) developed an inventory model where each order contains a random fraction of imperfect quality items with a known probability distribution. Papachristos and Konstantaras (2006) examined the Salameh and Jaber's (2000) work closely and rectified the proposed conditions to ensure that shortages will not occur. Maddah and Jaber (2008) corrected Salameh and Jaber's (2000) work related to the method of evaluating the expected profit per unit time. Jaggi et al. (2011, 2012, 2013) formulated an inventory model for deteriorating items. Jaggi and Mittal (2011, 2012) developed an inventory model with joint effect of inspection, deterioration, time-dependent demand, inflation and time value of money. Mittal et al. (2014) extended inventory model considering time expressions into association rules. The management of inventory can become more effective, if inventory is classified into categories based on some criteria like ABC classification, loss profit, and cross-selling effect.

Further, for some inventory items, the criteria (such as the price of an item) are derived not only from themselves, but also from their influence on the criteria of other items, usually called the "cross-selling effect" defined by Anand et al. (1997). Thus, items should be classified while considering such relationships. The ABC classification is used for ranking all inventory items on the notion of profit based on historical transactions. However, cross-selling effect is not considered while ranking items in traditional ABC classification. Brijs et al. (1999, 2000) developed a PROFSET model by considering cross-selling effect among items. They calculated the profit of a frequent item-set. However, the PROFSET model does not consider the strength of relationship between items. The PROFSET model does not provide relative ranking of selected items, which is important in classification of inventories. Moreover to calculate the profit of a frequent item-set the maximal frequent item-set had been used. However, the maximal frequent item-set often does not occur as frequently as its sub-sets. Therefore, the PROFSET model cannot be used to classify inventory items. Kaku (2004) classified inventory items based on strength of relationship between items. Kaku and Xiao (2008), further extended inventory classification considering cross-selling effect and ABC classification. They conducted experiments to show that a considerable large part of inventory items change their positions in the ranking list of importance. However, they have not considered that whether and how the strength of relationship with correlated items influences such ranking approach. Xiao et al. (2011) classified inventory items which are correlated each other using the concept of cross-selling effect together with ABC classification and loss profit. They classified items based on loss rule (Wong et al. 2003, 2005). The loss profit of item/item-set is defined as the criterion for evaluating the importance of item, based on which inventory items are classified. They explained that to judge the importance of an item (set), it is not only by looking at the profit it brings in when it is on the shelf, but also the loss profit it may take away when it is absent or stock out. However, they have not classified items in particular clusters.

In this paper, transactional clustering algorithm is used to partition the transactional database into different clusters. Further, apriori algorithm is applied for mining association rules from each cluster to find frequent items. Then, the loss profit is calculated for each frequent item. The frequent items are ranked in decreasing order of loss profit in each cluster. This ranking assists inventory manager to recognize most profitable item in each cluster. Further, an example is illustrated to validate the results.

2. Proposed Work

This paper proposes to calculate lost profit of frequent items in each cluster, which are found by applying apriori along with clustering.

For some inventory items, evaluating the importance of one item comes not only from its own value, but also from its influence on the other items, i.e., the ‘‘cross-selling effect’’ (Anand et al., 1997). Thus, there are more chances of losing sale if cross-selling effect among items is larger. The cross-selling effect among items can be determined by using association rules. Association rule mining aims to find rules of the form: A → B, where A, B are two sets of items. The meaning of the rule is that if the left-hand side A occurs, then the right-hand side B is also very likely to occur. The interestingness of the rules is often measured using support and confidence. The support of a rule is defined as the number of records in the dataset that contain both A and B. The confidence of a rule is defined as the proportion of records containing B among those records containing A. Association rule mining outputs rules with support no less than min_support and confidence no less than min_conf, where min_sup is called the minimum support threshold and min_conf is called the minimum confidence threshold. The two thresholds are specified by users.

Let $I = \{i_1, i_2, i_3, i_4, \ldots, i_m\}$ be a set of items. Now, support of item i_1 is defined as the frequency of its occurrences in total transactions and confidence is defined as conditional probability of purchasing i_2 when i_1 is purchased and is given by formula:

$$\text{Support}(i_1) = \frac{\text{Frequency of } i_1}{\text{Total number of Transactions}} \tag{1}$$

$$\text{Confidence}(i_1 \rightarrow i_2) = \frac{\text{Support of } i_1 \cup i_2}{\text{Support of } i_1} \tag{2}$$

This algorithm was proposed by Agrawal and Srikant (1994). The flowchart of Apriori algorithm is depicted in Fig. 1.

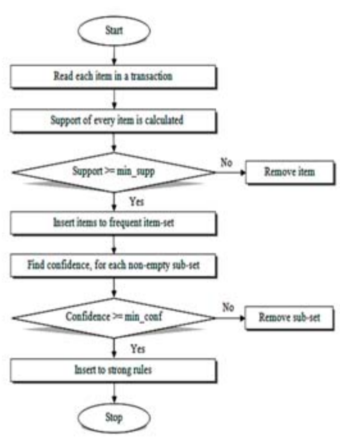

Fig. 1. Flow chart of apriori algorithm

Further, clustering is an important data mining technique that groups together similar transactions. Fast and accurate clustering of transactional data has many potential applications in retail industry, e-commerce intelligence, etc. Here, the term “large items” refers to the items contained in some minimum fraction of transactions in a cluster and is used as similarity measure of a cluster of transactions. The support of an item in cluster C_i is the number of transactions in C_i. Thus, for a minimum support s, an item is large in cluster C_i if its support is at least equal to $s \times C_i$, otherwise item is small. Thus, large items measure similarity in a cluster while small items measure dissimilarity. Two components of cost Ç are to be minimized consists of: the intra-cluster cost and the inter-cluster cost. The intra-cluster cost consists of the total number of small items and the inter-cluster cost measures the duplication of large items in different clusters. This clustering algorithm helps to minimize large items and small items cost. The overview of the clustering algorithm as described by Wang et al. (1999) is shown in Fig. 2.

```
/* Allocation phase */
    (1)  while not end of the file do
    (2)   read the next transaction <t,- >;
    (3)  allocate t to an existing or a new cluster Ci ;
    (4)  write  <t, Ci >;
/* Refinement phase */
    (5)   repeat
    (6)   not_ moved = true;
    (7)   while not end of the file do
    (8)   read the next transaction < t, Ci >;
    (9)   move t to an existing cluster Cj to minimize Cost C;
    (10)  if Ci  ≠ Cj then
    (11) write < t, Cj  >;
    (12)  not_moved=false;
    (13) eliminate any empty cluster;
    (14) until not moved.
```

Fig. 2. The overview of the clustering algorithm

Further, Xiao et al. (2011) ranked items according to their loss profit. The importance of an item is evaluated by considering both the profit it brings plus the loss profit it may take away when it is absent or stock out. The algorithm as proposed by Xiao et al. (2011) can be explained in three steps:

Step 1: Generate the cross-selling profit matrix according to formula:

$$\mathrm{M}_{\mathrm{BA}} = \mathrm{profit(A)confidence(A \rightarrow B)}^2, \tag{3}$$

where M_{BA} indicates the profit loss caused by the cross-selling relationship B→A, which can be read as: the cross-selling profit loss of item B from item A when item B is absent (or stock out).

Step 2: Calculate the loss profits of every item according to formula

$$\text{Total Profit(B)} = \text{profit(B)} + \sum_{\mathrm{A} \neq \mathrm{B}} \mathrm{M}_{\mathrm{BA}} \tag{4}$$

Step 3: Rank all items in terms of loss profit in descending order and do ABC classification.

3. Numerical example

In this section, Xiao et al. (2011) model have been considered to calculate the loss profit of items. Further, loss profit has been calculated for frequent items in each cluster which was not considered by Xiao et al. (2011). Consider the database set D and the inventory item-set, I = {p, q, r, s, t, u, v, y, x} and inventory transaction set, TID = {TID1, TID2, TID3, TID4, TID5, TID6} in Table 1. Each row in Table 1 can be taken as an inventory transaction.

Table 1
An inventory transaction database

TID	Items							
TID1	p	q	r					
TID2	p	q	r	s				
TID3	p	q	r		t			
TID4	p		r			u		
TID5					t		v		x
TID6					t			y	x

Consider an inventory items set, I = {p, q, r, s, t, u, v, y, x} and inventory transactions data base shown in Table 1, and consider the prices of items, p = \$5, q = \$4, r = \$2, s = \$3, t = \$2, y =\$3, u = \$1, v = \$2, x = \$1**.** Now, calculate loss profit for item p, q, r, s, t, u, v, y, and x using Eq. (3) as shown in Table 2.

Table 2
Loss profit of various items

Items	Loss Profit
p	\$46
q	\$32.75
r	\$46
t	\$16.08
s	\$6.08
u	\$2.75
v	\$3.17
x	\$11
y	\$4.17

Therefore, by ranking the items in descending order starting with the largest value of loss profit, we can get a ranking list (p r q t x s y v u). In the above example, Xiao et al. (2011) have not determined the loss profit in particular clusters. In this section, frequent items are determined in each cluster and loss profit is calculated for each frequent item. Further, these frequent items in each cluster are ranked according to descending order of loss profits.

Consider the transaction database of Table 1. Assume that the user-specified minimum support is 60%. A large item must be contained in at least 4 transactions (i.e., 6 ×60%). Consider the clustering $Ç_1$ = {C_1 = {TID1, TID2, TID3, TID4, TID5, TID6}}. We have $Large_1$ = {p, r}, $Small_1$ = {q, s, t, u, v, x, y}. Intra ($Ç_1$) = 7, and Inter ($Ç_1$) =0. So Cost ($Ç_1$) = 7.

Again, consider the clustering $Ç_2$ = {C_1 = {TID1, TID2, TID3, TID4}, C_2 = {TID5, TID6}}. For C_1, a large item should be contained in at least 3 transactions in C_1. Now, $Large_1$ = {p, q, r} and $Small_1$ = {s, t, u}. Similarly, $Large_2$ = {t, x} and $Small_2$ = {v, y}. Hence, Intra ($Ç_2$) = 5, Inter ($Ç_2$) =0, and Cost ($Ç_2$) =5. Thus $Ç_2$ has less cost as compared to $Ç_1$.

Consider the clustering $Ç_3$ = {C_1 = {TID1, TID2}, C_2 = {TID3, TID4}, C_3 = {TID5, TID6}}. We have $Large_1$ ={p, q, r}, $Small_1$ = {s}, $Large_2$ ={p, r}, $Small_2$ = {q, t, u}, $Large_3$ = {t, x}, $Small_3$ = {v, y}. Intra ($Ç_3$) =6, and Inter ($Ç_3$) =2. Hence Cost ($Ç_3$) =8, which is larger than $Ç_2$.

Hence, we will consider cluster $Ç_2$, as it has minimum cost as compared to cluster $Ç_1$ and $Ç_3$. Hence, the transaction database of table 1 is clustered into two clusters consisting of C_1 = {TID1, TID2, TID3, TID4} and C_2 = {TID5, TID6}. Further, we apply apriori algorithm on both clusters. We find item-set {a, b, c} is the most frequent item-set in cluster C_1 and item-set {d, g} is the most frequent item-set in cluster C_2.

Now, we calculate confidence of frequent item-set {p, q, r} of cluster C_1 and {t, x} of cluster C_2 by using equation (2), as shown in Table 3.

Table 3

Confidence of frequent item-set in Cluster C_1 and Cluster C_2

For cluster C_1	
Items	Confidence
$(p \rightarrow q)$	75%
$(p \rightarrow r)$	100%
$(p \rightarrow q \cup r)$	75%
$(q \rightarrow r)$	100%
$(q \rightarrow p)$	100%
$(q \rightarrow r \cup p)$	100%
$(r \rightarrow p)$	100%
$(r \rightarrow q)$	75%
$(r \rightarrow p \cup q)$	75%
For cluster C_2	
$(t \rightarrow x)$	100%
$(x \rightarrow t)$	100%

In cluster C_1, support (p) = 4, support (q) = 3, support (r) = 4 and in cluster C_2, support (t) = 2, support (x) = 2.

For item p, the loss profit (p) is calculated by scanning the transaction database, for each transaction using equation (3),

TID 1:
Loss profit (p, tid_1) = profit (p, tid_1) + profit (q, tid_1).confidence ($q \rightarrow p$) + profit (r, tid_1). confidence ($r \rightarrow p$) = $5 + 4\times1 + 2\times1 \approx \$11$.

TID 2:
Loss profit (p, tid_2) = profit (p, tid_2) + profit (q, tid_2).confidence ($q \rightarrow p$) + profit (r, tid_2). confidence ($r \rightarrow p$) + profit (s, tid_2).confidence ($s \rightarrow p$) = $5 + 4\times1 + 2\times1 + 3\times1 \approx \$14$.

TID 3:
Loss profit (p, tid_3) = profit (p, tid_3) + profit (q, tid_3).confidence ($q \rightarrow p$) + profit (r, tid_3). confidence ($r \rightarrow p$) + profit (t, tid_3).confidence ($t \rightarrow p$) = $5 + 4\times1 + 2\times1 + 2\times1 \approx \$13$.

TID 4:
Loss profit (p, tid_4) = profit (p, tid_4) + profit (r, tid_4). confidence ($r \rightarrow p$) + profit (u, tid_4). confidence ($u \rightarrow p$) = $5 + 2\times1 + 1\times1 \approx \$8$.

Thus, the loss profit of item p using equation (4) is $46.
Similarly, after applying rules and conditions described above, we can determine the loss profit of frequent items in different clusters as shown in Table 4.

Table 4
Loss-profit of frequent items in different clusters

Items	Loss Profit
For cluster C_1	
P	\$46
q	\$32.75
r	\$46
For cluster C_2	
t	\$11
x	\$11

Therefore, by ranking the items in descending order starting with the largest value of loss profit we can get a ranking list of (p r q) in cluster C_1 and (t x) in cluster C_2. Item p has been ranked further than item r as it has larger loss profit in cluster C_1. According to ABC classification, profit of item p = \$20, q = \$12, r = \$8, t = \$4 and x = \$2. Similarly, item t has been ranked further than item x as it has larger loss profit in cluster C_2. Thus, we have applied clustering algorithm to find different clusters of transactions. After that we have applied apriori algorithm on each cluster to find frequent items. Further, we have classified frequent items in each cluster according to loss-profit. Thus, by ranking frequent items in each cluster helps the manager to identify most profitable items in each cluster.

4. Conclusion and Future research

In this paper, clustering algorithm has been applied on transactional database to find different clusters of transactions. Further, apriori algorithm has been applied on each cluster to find frequent items. The frequent items in each cluster have been classified according to loss-profit. The loss profit of item was the total profit that the item may takes away when it is out of stock. A numerical example has been presented to illustrate the utility of the new approach. The inventories can be classified according to three cases: By using traditional ABC classification ranking of frequent items will be (p q r t x), According to loss profit ranking of frequent items ranking list will be (p r q t x). Further, for different clusters ranking of frequent items will be, cluster C_1- p r q, cluster C_2- t x.

In case 1, inventory items are classified using ABC classification, but this case did not consider loss rule for classification. In case 2, inventory items are classified using loss rule, but this case did not consider different clusters for classification. In case 3, inventory items are classified in different clusters using loss rule. Results indicate that a considerable large part of inventory items change their positions when they are ranked according to loss-profit as compared to traditional ABC classification in each cluster. Some items that traditionally do not belong to the A group in each cluster have been moved into the group A by the cross-selling effect to reconfigure their inventory policies, and also some items that traditionally belong to C group in each cluster have been promoted into higher group because of their high values of loss profits and should not be ignored as these were treated before. This approach helps inventory manager to find most profitable items in each cluster, so that he earn profit and easily manage stocks. For future study, it is desirable to extend the proposed model by considering time-varying aspects of databases. Further, an approach based on data mining technique like temporal association rule mining can be proposed to obtain a new ranking list based on loss profit.

References

Agrawal, R., & Srikant, R. (1994). Fast algorithm for mining association rules. *Proceedings of the 20th International Conference on Very Large Data Bases* (pp. 487-499). Chile.

Anand, S.S., Hughes, J.G., Bell, D.A., & Patrick, A.R. (1997). Tackling the cross-sales problem using data mining. *Proceedings of the 2nd Pacific-Asia Conference on Knowledge Discovery & Data Mining* (pp. 331-343). Hongkong.

Brijs, T., Swinnen, G., Vanhoof, K., & Wets, G. (2000). A data mining framework for optimal product selection in retail supermarket data: The generalized PROFSET model. *Proceedings of the 6th ACM*

SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 300-304). New York, USA.
Brijs, T., Swinnen, G., Vanhoof, K., & Wets, G. (1999). Using association rules for product assortment decisions: A case study. *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge discovery & data mining* (pp. 254-260). New York, USA.
Broder, A.Z., Glassman, S.C., Manasse, M.S., & Zweig, G. (1997). Syntactic clustering of the web. *Journal of Computer Networks ISDN Systems*, 29(8), 1157-1166.
Guha , S., Rastogi, R., & Shim, K. (2000). A robust clustering algorithm for categorical attributes. *Information System*, 25(5), 345-366.
Jaggi, C.K., Mittal, M., & Khanna, A. (2012). Effects of inspection on retailer's ordering policy for deteriorating items with time-dependent demand under inflationary conditions. *International Journal of Systems & Science*, 44(9), 1774-1782.
Jaggi, C.K., & Mittal, M. (2012). Retailer ordering policy for deteriorating items with initial inspection and allowable shortage under the condition of permissible delay in payments. *International Journal of Applied Industrial Engineering*, 1(1), 46-79.
Jaggi, C.K., & Mittal, M. (2011). Economic order quantity model for deteriorating items with imperfect quality. *International Journal Revista Invetigacion Operacional*, 32(2), 107-113.
Jaggi, C.K., Goel, S.K., & Mittal, M. (2013). Credit financing in economic ordering policies for defective items with allowable shortages. *International Journal of Applied Mathematics & Computation*, 219(10), 5268–5282.
Jaggi, C.K., Goel, S.K., & Mittal, M. (2011). Economic order quantity model for deteriorating items with imperfect quality and permissible delay on payments. *International Journal of Industrial Engineering Computations*, 2(2), 237-248.
Kaku, I. (2004). A data mining framework for classification of inventories. *Proceedings of the 5th Asia pacific Industrial Engineering & Management Systems* (pp. 450-455). Japan.
Kaku, I., & Xiao, Y. (2008). A new algorithm of inventory classification based on the association rules. *International Journal of Services Sciences*, 1(2), 148-163.
Lee, H.L., & Rosenblatt, M.J. (1987). Simultaneous determination of production cycles and inspection schedules in a production system. *Management Science*, 33(9), 1125-1137.
Maddah B., & Jaber M. Y. (2008). Economic production quantity model for items with imperfect quality: Revisited. *International Journal of Production Economics*, 112(2), 808-815.
Mittal, M., Pareek, S., & Agarwal, R., “Efficient ordering policy for imperfect quality items using association rule mining”. *Encyclopedia of Information Science & Technology 3rd Ed.* (pp. 773-786). United States, Information Science Publishing, 2014.
Papachristos, S., & Kontantaras, I. (2006). Economic ordering quantity models for items with imperfect quality. *International Journal of Production Economics*, 100(1), 148–154.
Porteus, E.L. (1986). Optimal lot sizing, process quality improvement and setup cost reduction. *Operations Research*, 34(1), 137-144.
Rosenblatt, M.J., & Lee, H.L. (1986). Economic production cycles with imperfect production processes. *IIE Transactions*, 18(1), 48-55.
Salameh, M.K., & Jaber, M.Y. (2000). Economic production quantity model for item with imperfect quality. *International Journal of Production Economics*, 64(1), 59-64.
Wang, K., Xu, C., & Liu, B. (1999). Clustering transactions using large items. ACM CIKM *International Conference on Information & Knowledge Management* (pp. 483-490). New York.
Wong, R.C., Fu, A.W., & Wang, K. (2005). Data mining for inventory item selection with cross-selling consideration. *Data Mining & Knowledge Discovery*, 11(1), 81–112.
Wong, R.C., Fu, A.W., & Wang K. (2003). MPIS: Maximal-profit item selection with cross-selling considerations. *IEEE International Conference on Data Mining* (pp. 371-378). Florida, USA.
Xiao, Y., Zhang, R., & Kaku, I. (2011). A new approach of inventory classification based on loss profit. *Expert Systems with Applications*, 38(8), 9382-9391.
Zhao, Q., & Bhowmick, S.S. (2003). Association rule mining: A survey, *Center for Advanced Information Systems, Nanyang Technological University,* Report No. 2003118, Singapore.