International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2020): 7.803

Application of Renewal Process for Finding Mean Number of Occurrence of an Event and also Predicting the Number of Data in which the Event will Happen

Dr. Rajalekshmi. V. G.

Associate Professor, Department of Mathematics, S. D. College, Alappuzha, Kerala, India
rajisreeramam[at]gmail.com

Abstract: ***Pattern mining is one of the important jobs in data mining. For discovering all patterns we have to know the mean number of occurrence of each item in the data. Here we use the renewal theory for finding the mean occurrence of an item and also predicting the number of data in which the event will happen.***

Keywords: Pattern mining, Renewal period, Order of an event, Renewal function, Prediction

1. Introduction

Tracking patterns is a fundamental data mining technique. It involves identifying and monitoring trends or patterns in data to make intelligent inferences about business outcomes. Once an organization identifies a trend in sales data for example, there is a basis for taking action to capitalize on that insight. If it's determined that a certain product is setting more than other for a particular demographic, an organization can use this knowledge to create similar products or services, or simply better stock the original product for this demographic.

Frequent patterns are item sets, subsequences or substructures that appear in a data set with frequency no less than a user-specified threshold. Moreover it is very important in data mining problem. By identifying frequent patterns from a business data we can improve the business by making more frequent items. For finding frequent patterns we have to find the mean number of occurrence of each item in the data base.

In this paper an attempt is made to find the mean number of occurrence of an item in the data base using renewal process. Also by the theory of Renewal process here, explains forecasting the number of data which contain the event.

In the Renewal theory the renewal period is the time between occurrences of successive renewals. Here we consider renewal as the occurrence of a data which contain the event. Renewal period is the total number of transaction data plus one between two consecutive renewals.

2. Related Work

Renewal theorem and its theory can be applied in all fields of life. The theory stochastic renewal process and the renewal theorem have been fundamental to the development of risk – based asset management models [2], in bridge management the renewal theory has been applied [5]. Although the renewal process has been discussed in many mathematical treaties [3, 1] the concepts are not amenable to the engineering community. A conceptually simple and intuitive interpretation of the renewal process with applications are given in [4].

Model description

Here we arrange the transaction data is the data base as T_1, T_2, T_3 ….. (i.e. number data as 1, 2, 3….). Each data T_i is a set of items which is contained in the universal item set 'I'. Consider an item $A \in I$. We have to identify those transactional data which contain A. Finding the transactional data with A is called an event.

Order of the first event is the number of the first transactional data which contain the item and is denoted by n_1. The order of the i^{th} event is the total number of transactional data between $(i - 1)^{th}$ event and i^{th} event including the transactional data containing the i^{th} event. It is denoted by n_i, $i = 1, 2, 3…..$

Corresponding to every transactional data we can define a random variable.
X_i, $i = 1, 2$ …. Such that

$X_i = 1$ if $A \in T_i$
$= 0$ if $A \notin T_i$

The order of the event 'n' is sequence of random variable. The probability distribution of order of event is given by
$P(n = k) = p_k$

The number of transactional data contain the n^{th} event is
$S_n = n_1 + n_2 + ….. + n_n$

This means that out of S_n data there are only 'n' data containing the item A.

Let N_m be the total number of transactional data containing the arbitrary item 'A' among the first 'm' transactional data. This can be considered as a counting process associated with the sequence (S_n) is called the renewal distribution (p_k). It

there are 'n' data contain 'A' out of first 'm' data then N_m = n.

Mean Number of Occurrence of event up to the m^{th} data (Renewal Function)

The Renewal function up to *m* data is the mean number of occurrence of the event up to the m^{th} data.

M(m) = $E(N_m)$

We have $N_m = X_1 + X_2 + X_3 + \ldots + X_m$

Where X_i = 1 if $A \in X_i$

= 0 if $A \in X_i$

$N_m = \sum_{i=1}^{m} X_i$

N_m is the total number of event occurs up to the m^{th} data. Let N_m = n. Then we can write

$$N_m = \sum_{n=1}^{m} 1\,(S_n \leq m) \quad \text{_____(1)}$$

Where $1\,(S_n \leq m) = 1$ if *n* satisfy the condition in the bracket

= 0 otherwise

Now taking expectations on both sides of (1)

$$E(N_m) = \sum_{n=1}^{m} E(1_{(S_n S_m)}) = \sum_{n=1}^{m} F_n(m) \quad \text{_____(2)}$$

F_n denotes the cumulative distribution function of S_n. By using (1) and (2) we can find a recursion equation for the expected number of event occur from *m* data. This we prove by a theorem.

3. Theorem

Let $M_{(m)}$ be the expected number of data containing the event up to the first m data. Then we have recursion equation

$$M(m) = F(m) + \sum_{k=1}^{m} p_k E(N_{m-k})$$

Proof: Let the k^{th} data be the first data which contain the item. Then n_i = k. Therefore the expected number of occurrence of the item in the remaining m – k data are $E(N_{m-k})$. By conditioning the number of a data on n_1 = k.

If the first event occur in the k^{th} data. Then the remaining (n - 1) events must occur in the (m - k) data.

$\therefore n_2 + n_3 + \ldots + n_n \leq m - k$

$1[(n_2 + n_3 + \ldots + n_n) \leq m - k] = 1$

If 'n' satisfy the equality in the bracket

= 0 otherwise

$$E\left({}^{N_m}/_{n_1 = k}\right)$$

$$= E\left(1 + \sum_{p=2}^{m} 1(n_2 + n_3 + \ldots + n_p) \leq m - k \Big/_{n_1 = k}\right) \quad \text{___(3)}$$

Where $1 \leq k \leq m$

From (1) we have

$$\sum_{p=2}^{m} 1\left((n_2 + n_3 + \cdots + n_p) \leq m - k\right) = N_{m-k}$$

Here $1\left((n_2 + n_3 + \cdots + n_p) \leq m - k\right) = 1$

if p satisfies the condition in the bracket.

= 0 otherwise

$\therefore$ equation (3) become

$E(N_m/n_1 = k) = (1 + E(N_{m-1}))p_k$

By sum over k = 1, 2, …m and using the law of total probability

$$E(N_m) = M(m) = F(m) + \sum_{k=1}^{m} p_k \; E(N_{m-k}) \quad \text{----(4)}$$

Predicting the Number of Data in which the Event will happen.

In this context we have to determine in which transactional data the item occur. From 'm' transactional data let the number data in which item contain in N_m data. Among m data the number of the last data contain the item is S_{Nm}. The number of the next data contain the item is S_{Nm+1}. The number of data after the *m* data to reach the $(N_m + 1)^{th}$ event is = S_{Nm+1} - m.

If the m^{th} transactional data contains the item then S_{Nm} = m. Otherwise the m^{th} data arrive after m- S_{Nm} data from a transactional data contains the item.

For Convenience take

$j = S_{Nm+1} - m$, $i = m - S_{Nm}$, $N_{m=k}$

Since N_m = k, we have

$i = m - S_k$

This means that from the first m data after the k^{th} event there are $m - S_k$ data which do not contain the item. S_k means the number of data where the k^{th} event occur.

$i = m - S_k \Rightarrow S_k = m - i$

$S_k \geq k$

$\therefore m - i \geq k$

i.e. $i + k \leq m$.

$j = S_{Nm+1} - m$ and $N_m = k$ we get $S_{k+1} = m + j$.

n_{k+1} order of the $(k + 1)^{th}$ event. Which is the total number of transactional data between the data contain the k^{th} event and the data containing the $(k + 1)^{th}$ event including the data containing the $(k + 1)^{th}$ event.

$n_{k+1} = S_{k+1} - S_k = (m + j) - (m + i) = j + i$

We can represent the total number of data between the m^{th} data and the $(N_m)^{th}$ event including the m^{th} data as a random variable Z(m). Y(m) is the random variable representing the total number of data between the $(N_m + 1)^{th}$ event and the m^{th} data including the data which containing the $(N_m + 1)^{th}$ event.

Then Z(m) = i, N(m) = k, Y(m) = j

The joint distribution of (Z , N_m, Y) is given by

Prob (Z = i, N_m = k, Y = j) = Prob ($S_k = m - i$, $n_{k+1} = j + i$)

$= p_{i+j}$ Prob($S_k = m - i$)

where $i + j \leq m$, $j \geq 1$.

By the law of total probability

$$Prob\,(Z = i, Y = j) = p_{i+j} \sum_{k=1}^{m-i} Prob(S_k = m - i)$$

The probability of the occurrence of the event at the j^{th} transactional data is denoted u_j.

$u_j = \sum_{k=1}^{j} P(S_k = j)$

Then the joint distribution of (Z, Y) is given by

$P(Z, Y) = u_{m-i}\, p_{i+j}$ _____(5)

By the law of total probability, the marginal distribution of the total number of data between the N_m^{th} event and the m^{th} data (including the m^{th} data) as

$$P(Z = i) = u_{m-i} \sum_{j=1}^{\infty} p_{i+j} = u_{m-i}(1 - F(i)) \_\_\_\_\_(6)$$

With $0 \le i \le m$

By the law of total probability the marginal distribution of the total number of data between the m^{th} data and $(N_m + 1)^{th}$ event including the data containing the $(N_m + 1)^{th}$ event is

$P(Y = j) = \sum_{k=0}^{m} u_k p_{m+j-k}$,j≥1 ----------(7)

4. Conclusion

Most of the research works are carried out by collecting data from various sources. Study about the event occurring in data has importance and predicting the occurrence of an event has great application in various data .While activity recognition has been shown to be valuable for pervasive computing application. Less work has focused on techniques for forecasting the future occurrence of activities. This can be applied in other forecasting situation where event prediction is valuable.

By giving some additional property we can use the above model in relational data base. Throughout the last decade, a lot of people have implemented and compared several algorithms to solve the frequent item set mining problem as efficiently as possible. But only a very small model yet has been found using stochastic models. There is great application of stochastic models in data mining. The Renewal process is a way to built queuing models in data mining.

References

[1] Karlin, S. & Taylor, H.M., 1975, *A First Course in Stochastic Processes;* Second Edition. San Diego: Academic Press.
[2] Rackwitz, R., 2001, *Optimizing systematically renewed structures.* Reliability Engineering and System Safety, 73(3) : 269 – 279.
[3] Stephen Breen, Michael S., Waterman and Ning Zhang., *University of Southern California. Renewal Theory for Several Patterns.* J. Appl. Prob. 22, 228 – 234 (1985).
[4] Van der Weide,J.A.M,Pandey,M.D and Van Noortwijk,J.M.,*A Conceptual Interpretation of the Renewal Theorem with applications*,Risk,Reliability and Societal Safety-Aven&Vinnen (eds) 2007.Taylor & Francis group,London, ISBN978-0-415-44786-7
[5] *Van Noortwijk, J.M. & Klatter, H.E., 2004. The use of lifetime distributions in bridge maintenance and replacement modeling.* Computers and Structures, 82(13-14): 1091 – 1099.