Available online at http://www.journalcra.com

International Journal of Current Research
Vol. 9, Issue, 04, pp.48694-48697, April, 2017

INTERNATIONAL JOURNAL OF CURRENT RESEARCH

ISSN: 0975-833X

RESEARCH ARTICLE

STATISTICAL METHOD OF ASSOCIATION FOR DATA MINING: APPLICATION TO CENSUS DATA

[1,*]Richa and [2]Anurag Kulshreshtha

[1]Acedamic Counselor, Department of Statistics, Indira Gandhi Open University, New Delhi, Indian
[2]Indian Institute of Technology, Roorkee, Saharanpur Campus, (U.P.), India

ARTICLE INFO

Article History:
Received 15th January, 2017
Received in revised form
14th February, 2017
Accepted 22nd March, 2017
Published online 20th April, 2017

Key words:

Data Mining,
Association Rule,
Chi Square.

ABSTRACT

In the present scenario Data Mining is an emerging powerful tool for analyzing the data. Data mining is the process of extracting the valid, relevant and useful information from the data. And association belongs to the discovery data mining techniques, which are used to find patterns inside the data. This paper deals with the some of the measures of association for data mining. And the goal of the quest paper is to use the statistical method of association for mining the data. The method is applied on the census data and it is found that the performance of new method is giving satisfactory results.

Citation: Richa and Anurag Kulshreshtha, 2017. "Statistical method of association for data mining: application to census data", *International Journal of Current Research*, 9, (04), 48694-48697.

INTRODUCTION

Association discovery finds rules about items that appear together in an event such as a purchase transaction. Association rules identify collections of data attributes that are statistically related in the underlying data. An association rule is of the form $X \rightarrow Y$ where X and Y are disjoint conjunctions of attribute-value pairs. The confidence of the rule is the conditional probability of Y given X, Pr(Y|X), and the support of the rule is the prior probability of X and Y, Pr(X and Y).

Definition 1: Given a set of items $I = \{I_1, I_2,\ldots,I_m\}$ and a database of transactions $D = \{t_1, t_2,\ldots, t_n\}$ where $t_i = \{I_{i1}, I_{i2},\ldots, I_{ik}\}$ and $I_{ij} \in I$, an 'association rules' is an implication of the form X => Y where X,Y CI are the set of items called itemset and $X \cup Y = \phi$.

Definition 2: The 'supports' for an association rule X=>Y is the percentage of the transactions of the data base that contain. The 'support' of an item (or set of items) is the percentage of transactions in which that item (or items) occurs. The aim of association rule mining is to find interesting and useful patterns in a transaction database. Some of the association rules frequently used in data mining are discussed below.

Corresponding author: Richa
Acedamic Counselor, Department of Statistics, Indira Gandhi Open University, New Delhi, Indian.

Link analysis: It referred to as affinity analysis or association, refers to the data mining task of uncovering relationship among data. The best example of this type of application is to determine association rules. An association rule is a model that identifies specific types of data associations. These associations are often used in the retail sales community to identify items that are frequently purchased together. Many examples the use of association rules in the market basket analysis. The data analyzed consist of information about what items a customer purchases.

Apriori Algorithm: The Apriori algorithm is the most well known association rule algorithm and is used in most commercial products. It uses the following property, which we call the 'large itemset property'. Any subset of a large itemset must be large. Apriori scans the entire database in each pass to count support. Scanning of the entire database may not be needed in all passes.

Partition Algorithm: Partition reduces the number of database scans. It divides the database into small partitions such that each partition can be handled in the main memory. In the first scan, it finds the local large itemsets in each partition. The local large itemsets can be found by using a level-wise algorithm such as Apriori. Since each partition can fit in the main memory, there will be no additional disk I/O (input/output) for each partition after loading the partition into the main memory. In the second scan, it uses the property that

a large itemset in the whole database must be locally large in at least one partition of the database. Then the union of the local large itemsets found in each partition is used as the candidates and are counted through the whole database to find all the large itemsets. Partition favors a homogeneous data distribution. That is, if the count of an itemset is evenly distributed in each partition, then most of the itemsets to be counted in the second scan will be large.

Rule induction: It is one of the major forms of data mining and is perhaps the most common form of knowledge discovery in unsupervised learning systems. It is also perhaps the form of data mining that most closely resembles the process that most people think about when they think about data mining, namely "mining" for gold through a vast database. Rule induction on a data base can be a massive undertaking where all possible patterns are systematically pulled out of the data and then an accuracy and significance are added to them that tell the user how strong the pattern is and how likely it is to occur again. In general these rules are relatively simple such as for a market basket database of items scanned in a consumer market basket.

Previous work

Onkamo *et al.,* (2002) worked for Association analysis for quantitative traits. This ascertainment scheme closely resembles real studies in the sense and there are correlated quantitative traits which have been measured and need to be analysed in the data.

NittayaKerdprasop and KittisakKerdprasop (2003) utilized the approach of incremental data mining, which is dealing with one subset of data at a time. They focused the study on the partitioning of a data into a proper subset for association mining tasks.

Oatley *et al.,* (2004) Link analysis, alternatively referred to as a–nity analysis or association, refers to the data mining task of uncovering relationships among data. They used 'COPLINK Detect' a technique called concept space and to identify such associations from existing crime data automatically. Association rules are used to show the relationships between data items.

Michael Steinbach *et al.,* (2004) introduced support envelopes as a new tool for patterns. The support envelope for a transaction data set and a specified pair of positive integers (m, n) consists of the items and transactions that need to be searched to find any association pattern involving m or more transactions and n or more items.

Mohamad J. Zaki, (2004) presented a new framework for associations. According to him Association rule discovery, a successful and important mining task, aims at uncovering all frequent patterns among transactions composed of data attributes or items. They experimented on several "hard" or dense, as well as sparse databases and confirmed the utility of their framework in terms of reduction in the number of rules presented to the user, and in terms of time.

CristianAflori has given that data mining tasks are predictive and descriptive. He worked on parallel data mining solutions. Which require parallel data mining algorithms and distributed data bases, parallel file systems, parallel input output, tertiary storage management of on line data support for heterogeneous data representations, data security and quality and hardware models.

Richard Kittler and Weidong Wang have given that data mining is a family of methodologies for finding hidden patterns in typically large sets of data to help in explaining the behavior of one or more response. Association Rules is a tool used to look for patterns of coincidence in data. Association rule analysis is useful in discovering patterns of behavior but does not produce a predictive model. Association rules are used to show the relationships between data items.

Statistical method for finding the association

Till now so many different methods of data mining are used for finding out the association, but here a statistical method of association is suggested as the better method for finding out the association. We applied here a statistical method for finding the association between two variables, that is measure of association based on the χ^2 statistic.

As the value of χ^2 is defined as:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Then the measure of association phi square that is *mean square contingency coefficient* is defined as:

$$\phi^2 = \frac{\chi^2}{N}$$

A variation of this measure, suggested by Pearson (1904) and called the *coefficient of contingency*, and is given by:

$$P = \sqrt{\left(\frac{\chi^2 / N}{1 + \chi^2 / N}\right)}$$

This coefficient lies between 0 and 1 as required, and attains its lower limit in the case of complete independence, that is when $\chi^2 = 0$. In general, however, P cannot attain its upper limit, and Kendall and Stuart show that, even in the case of complete association, the value of P depends on the number of rows and columns in the table. To remedy this following function of χ^2 has been suggested:

$$T = \frac{\chi^2 / N}{\sqrt{(r-1)(c-1)}}$$

This again takes the value 0 in case of complete independence, and as shown by Kendall and Stuart, may attain a value of +1 in the case of complete association when

r=c.

An algorithm for the suggested method is given in the Appendix.

Application on the census data

To examine the performance of this statistical method, a huge amount of data is to be taken. The association between Income and expenditure are to be taken here. The data of 10 tehsils

having different number of villages are to be taken as sample. For example first Tehsil is having 674 villages, second is having 315, third is having 560 villages and so on. This data is from the census data of Agra district. The Income and expenditure of first 10 villages are tabulated in the Table 1 for first Tehsil.

Table 1. Income and expenditure table for 10 villages

S.No.	TOT_INC	TOT_EXP
1	73140	73000
2	76318	76318
3	6000	6000
4	0	0
5	242417	242417
6	275	267
7	0	0
8	187800	187800
9	276012	270318
10	186350	140210

Now these income and expenditure are divided into different classes as first income class is from 0 to 5000, second is 5000 to 80000, third class is having income more than 80000. Same three classes are made for expenditure. Now after taking these three classes of income and three classes of expenditure together, it is having total 9 classes as (1, 1), (1, 2), (1, 3), (2, 1), (2, 2), (2, 3), (3, 1), (3, 2), (3, 3). So first village is coming under the class (2, 2), second as (2, 2), and so on. The classes of first 10 villages are shown in the table below:

Table 2. Classes for income and expenditure

S.No.	TOT_INC	Inc class	TOT_EXP	Exp class	Class
1	73140	2	73000	2	(2, 2)
2	76318	2	76318	2	(2, 2)
3	6000	2	6000	2	(2, 2)
4	0	1	0	1	(1, 1)
5	242417	3	242417	3	(3, 3)
6	275	1	267	1	(1, 1)
7	0	1	0	1	(1, 1)
8	187800	3	187800	3	(3, 3)
9	276012	3	270318	3	(3, 3)
10	186350	3	140210	3	(3, 3)

Now according to the different classes all 674 villages of first Tehsil are divided as in the table below:

(Here income of 235 villages are coming in the class (1, 1), income of 2 villages are coming in class (1, 2) and so on.)

Table 3. Two way distribution of no. of villages

Inc-Exp	1	2	3	sum
1	230	7	9	246
2	10	198	31	239
3	8	7	174	189
Sum	248	212	214	674

The above table is showing the frequencies of Tehsil I only. Now after applying χ^2statistic the value of χ^2 is:

$\chi^2 = 979.2928$

And the*mean square contingency coefficient is:*

$\phi^2 = 1.4529$

The *coefficient of contingency* is:
P = .7696

And

T = .7264

After having the value of P = .7696 we can conclude that here the association between income and expenditure is high. Now the table given below is for 10 Tehsils of Agra district.

Table 4. Table for Chi Square Test

Tehsil	No.ofvill	chi sq	phi sq	P	T
1	674	979.2928	1.452957	0.769629	0.726478
2	315	243.1541	0.360763	0.514896	0.385959
3	560	514.9932	0.764085	0.658129	0.459815
4	422	497.6979	0.738424	0.651741	0.589689
5	1197	1316.26	1.952908	0.813235	0.549816
6	1210	1667.143	2.473506	0.843864	0.688902
7	672	935.1305	1.387434	0.762326	0.695782
8	880	1186.995	1.761121	0.798641	0.674429
9	938	1449.224	2.150184	0.826171	0.772508
10	814	1182.421	1.754334	0.798083	0.726303

Conclusion

In this paper we used a well known statistical method of association for the data mining. After applying the method on the census data it can be concluded that it is giving satisfactory results. The main advantage of this method is that this is purely quantitative and with the help of simple algorithm it can be easily implemented for finding association in data mining. It can also be used in Partition Algorithm to reduce the number of variables.

REFERENCES

Alex Berson and Syephen J. Smith, 1997. "Data Warehousing, Data Mining, And OLAP". MC Graw-Hill.

Arun K. Pujari, 2001. "Data Mining Techniques", Published by University Press (India) Limited 3-5-819 Hyderguda.

David Hand, HeiikiMannila and Padhraic Smyth, 2007. "Pricipals of Data Mining", Cambridge, MA: MIT Press.

Dr. Diego Kuonen, 2004. "Data Mining and Statistics: What is the Connection?", Data A Data Mininginistrator Newsletter 30.0, October.

Jiawei Han and Micheline Kamber, 2007. "Data Mining Concepts and Techniques", Elsevier India Pte. Ltd, New Delhi.

ManjushaSusheel Joshi, 2008. "Data mining: Role of TEX files", TU Gboat, Volume 29, No. 3 Proceedings of the 2008 Annual Meating.

Margaret H. Dunham, 2003. "Data Mining Introduction and Advance Topics", Pearson Education (Singapore), Pte. Ltd, India.

Michael Steinbach, Pang Ning Tan, Vipin Kumar, 2004. "Support Envelops : A Technique for Exploring the Structure of Association Pattern", KDD'04 August 22-25, Seattle Washington, USA.

Mohammad J. Zaki, 2004. "Mining Non Redundant Association Rules", Data Mining and Knowledge Discovery, 9, 223, 248.

Nikhil N. Salvithal1, Dr. R. B. Kulkarni, 2013. "Evaluating Performance of Data Mining Classification Algorithm in Weka", *International Journal of Application or Innovation in Engineering & Management (IJAIEM).*

Nitesh V. Chawla, 2005. "Data Mining for Imbalanced Datasets: An Overview", The Data Mining and Knowledge Discovery Handbook, 853-867.

NittayaKerdprasop and KittisakKerdprasop, 2003. "Data Partitioning for Incremental Data Mining, Proceedings of 1st International Forum on Information and Computer Technology January 9-10, Shizuoka University, Hamamatsu, Japan.

Oatley, G.C., B. W. Ewart, J. Zeleznikow, 2004. "Decision Supp ort Systems For Police: Lessons From The Application of Data Mining Techniques To 'soft forensic evidence".

Onkamo, P., V. Ollikainen, O. Sevon, H. T. Toivonen, H. Manila and J. Kere, 2002. "Association analysis for quantitative traits by data mining: QHPM", Ann. Hum. Genet., 66, 19–429 'UniversityCollegeLondon. United Kingdom

Paolo Guidici, 2003. "Applied Data Mining, Statistical Methods for Business and Industry" Wiley publication, Singapore.

Richaand Vineeta Singh, 2012. "Statistical Method of Clusterisation for Data Mining", *Journal: Synchronizing Management Theories and Business Practices,* ISBN: 978-93-82338-08-6

Richaand Vineeta Singh, 2015. "Method of Classification for Data Mining: A Statistical Approach". *International journal of Management Development & Information Technology,* ISSN 0976-8440.

Ritika and Aman Paul, 2014. "Prediction of Blood Donors. Population using Data Mining Classification Technique", *International Journal of Advanced Research in Computer Science and Software Engineering*, Volume 4, Issue 6, June, ISSN: 2277 128X

Ryan S.J.D. Baker and KalinaYacef, 2009. "The State of educational Data Mining in: A Review and Future Visions".

****