Apriori algorithm on Marine Fisheries Biological Data

D. PUGAZHENDI
Madras Research Centre of Central Marine Fisheries Research Institute
Chennai, India
pugalcmfri@gmail.com

Abstract - Data Mining (DM) is the process of analysing data from different vista and gives summary on specific determination. Association rules are rules describing the associations or correlations to bring out the hidden pattern among attributes in data sets. The most widely used algorithm in association technique is Apriori algorithm which is meant for only categorical data analysis. The sample fishery biological data consist of six attributes out of which two are numerical values. As a new attempt, the numerical values were converted to unique nominal values in order to maintain all categorical values. The Apriori algorithm applied on specific criteria such as minimum support and confidence enabled to derive many meaningful patterns on different perspectives. The *taeniopterus* apecies has more associations between the attributes of total_length range between 120 to 150 and month of August, weight of Thirty and sex of Male.

Key words: Data Mining on Fishery Biology, Association on fishery data, length and weight data, Data Mining, Association

I. Introduction

Data mining (DM) is the practice of automatically searching large stores of data to discover patterns and trends that go beyond simple analysis. Data mining uses sophisticated mathematical logarithms to segment the data and evaluate the probability of future events and contributing greatly to business strategies and medical research[3]. Data mining is also known as Knowledge Discovery in Data (KDD)[4]. In real world application, a DM process can be broken into six major phases; business understanding, data understanding, data preparation, modeling, evaluation and deployment as define by the CRISPDM (Cross Industry Standard Process for Data Mining).

Association rule mining is one of the popular data mining technique on groups such as classification and prediction, clustering, outlier deduction, sequence analysis, time series analysis, text mining, web mining and also some new techniques such as social network analysis and sentiment analysis [2]. The mining process are classified as Descriptive mining and Predictive mining. Some of the descriptive the mining process are Clustering, Association and Sequential which brings the essential of the data in the database. Predictive mining is the process of inferring patterns form data to make predictions such as Classification, Regression and Deviation detection[7]. There are many efficient techniques to get rule, although most of the techniques require that the values of the attributes be discrete and solve the issue, the discretize technique used on numeric attribute, but this involves some loss of information (J. Mata, et. al). The association rules under unsupervised category performs on both categorical and numerical dataset. Association rules are used as a method to convey relationships among the attributes in large data sets[6].

The patterns let out with in the form of association rules are measures such as strength by support, confidence and lift. Support is the probability of occurrences in the dataset which contain both A and B denoted by P(AUB); confidence is a key measure to accept the rule that the percentage of occurrences containing A that also contain B denoted by B(B\A) and the lift is ration of confidence to the percentage of cases containing B denoted by lift

$$(A => B) = \frac{P(AUB)}{P(A)P(B)}$$

In addition to above three measures, 20 measures exist and Chi-square, conviction, Gini and leverage are important among them (Association in R). Apriori is a classic algorithm for association rules [6]. Finding rules are much easier than identifying meaning rules so interpretation is a more crucial part.

R is an open source software environment for statistical computing, graphics and data mining technique. Though R software is one among the statistical packages, it is also a best choice for analysis of machine learning by the way of classification, cluster, association, spatial data and data pruning process of large data set. R provides a comprehensive indexing of 4000 package repositories and functions for various domains. KDnuggets a top resource of data mining community is conducted a poll for the year on top Languages for analytics, data mining, data science and revealed that R is used by 61% and followed by Python 39%.[8]

II. Materials and Methods

The biological data published by CMFRI (Central Marine Fisheries Research Institute) for the year 2000 of Madras Fisheries Harbour, Chennai with Trawl Net and different species were taken for analysis. The dataset composed of 1172 instances and 6 attributes namely month, species, total_length, weight, sex and stage. The six attributes of dataset are used in the database, each of which include several different values. They are

i) The month attribute consists of twelve nominal values namely January, February and so on.

ii) Species attribute has 3 values of species names like *sulphurus*, *taeniopterus* and *moluccensis* of Upeneus genus and commonly phoned as goatfish.

iii) Total_length attribute values are under ordinal values of categorical type ranging from 92 to 192 mm

iv) Weight field values range start from 10 and end with 90

v) Sex is the fifth attribute and has only two nominal values of Male and Female

vi) Lastly, the STAGE field identified with the values SECOND, THIRD, FOURTH AND FIFTH.

The numerical values of total length and weight are converted to unique ordinal values. There are only two instances that come under outliers and since the percentage is less than 1, data pruning process was not performed; the two records cannot influence on rule building process.

Apriori Algorithm

The Apriori algorithm is a best choice to analyse categorical data and the algorithm is a most common algorithm for mining frequent itemsets (Iriva tudor..). Apriori algorithm find the frequent itemsets by the sets in which items have minimum support and the subset of frequent itemset must also be a frequent itemset iteratively(k), to generate association rule. The Apriori Algorithm calculates rule to find probabilistic relation between items in frequent item sets [5]. The data taken here for analysis is combination of categorical and numerical values. The Apriori algorithm is generally used to analyse the categorical data and Genetic algorithm is used to analyse the numerical data values. Here we made use of the Apriori algorithm after changing the numerical values to equivalent ordinal values which is under categorical values. Generally, discredization technique is used to analyse the numerical data but here numerical values are converted to the string names. There are different views on genetic algorithms performance, so we drive the with new way of approach towards the same goal using Apriori algorithm which is the most popular among the category. The popular key terms of association rules are as follows.

Itemset: An itemset is a set of items. A k-itemset is an itemset, that contain k number of items.

Frequent itemset: This is an itemset which has minimum support.

Candidate Set: This is the name given to a set of itemsets that requires testing to see if they fit a certain requirement[1].

The key step starts with finding the frequent itemsets and sets of item that have minimum support ie if {AB} is a frequent itemset, both {A} and {B} should be frequent itemset. Iteratively frequent itemsets are found with cardinality from 1 to k and finally the itemsets are used to generate possible association rules. Apriori discovers pattern with frequency above the minimum support threshold (Bhavani.). Strong is a measure to accept the rule by minimum support and minimum confidence, and if both are satisfied, it is treated as strong rule.

The steps involves in Apriori algorithms are,

i) Analyse all the transactions in a dataset for each item support count.

ii) Find minimum support counts and removed as candidates.

iii) The above two steps are repeated to generate consecutive candidates list and the process is stopped when the support count of all item sets are complete.

iv) All the candidates (c1,c2,c3..) itemsets generated with a support count greats then the minimum support count form a set of frequent itemsets.

v) These frequent itemsets will be used to generate strong association rules where association rules satisfies both minimum support and minimum confidence.

In order to utilize this information, it works with the dependency framework by actual and expected occurrences. To determine the reliability of threshold by statistically chi-square, impact and lift are used.

III. Results & Discussion

Since it is a pilot study of Apriori algorithm on marine fisheries data, only modest number of instances were analysed. The main aim of this study is to find the suitability of Apriori algorithm on marine fisheries data and the study has revealed that this algorithm could bring very useful information in the form of association rule.

As default, the rules criteria of support and confidence of .005 and 95% respectively was fixed. While fixing some rhs (right hand side), there are many useful lfh (left hand side) association could be generated and vice-versa. The Apriori algorithm fetched many rules and some of them are mentioned in the following table.

Sl. No	Lfh	Rhs	Support	Confidence	lift
1	{weight=twenty)	=> {stage=second}	0.1169	0.9448	1.4475
2	{stage=fourth}	=> {sex=female}	0.1476	1.0000	1.5585
3	{weight=twentyfive)	=> {stage=second}	0.1536	0.8257	1.2949
4	{Sex=male)	=> {stage=second}	0.3311	0.9238	1.4153
5	{Month=January, stage=second}	=> {species=sulphureus}	0.1102	1.0000	2.5042
6	{species=sulphureus, sex=male}	=> {stage=second}	0.1536	.9836	1.5069
7	{species=taeniopterus, sex=male}	=> {stage=second}	0.1758	0.8766	1.3429
RHS is STAGE=2					
1	{ Length=97 }	=> {STAGE=2}	0005	1.0000	1.532
2	{ Length=107 }	=> {STAGE=2}	0.005	1.0000	1.532
RHS is SEX					
1	{Month=april, weight=15,stage=3}	=>{ sex= female }	0.005	1.000	1.588
2	{ Species= sulphureus, tl=151, stage=2 }	=>{ sex=male }	0.005	1.000	2.79
3	{ Month= December, species= tenipterus, total-length=133, stage=2 }	=>{ sex=male }	0.005	0.85	2.39

The species *moluccensis* of *Upeneus* genus has only 10 instances and did not generate any association rule and therefore we conclude that more number of instances only can generate strong rules.

IV. References.

[1] Anubha Sharma and Nirupma Tivari, “A survey of Association Rule Mining Using Genetic Algorithm”, Interanational Journal of Computer Applications & Information Technology, Vol. I(2), 2012:5-11.

[2] Geoffrey I. Web, “Discovering associations with numeric variables”, KDD San Fransisco, CA, August 2001

[3] Irina Tudor, “Association Rule Mining as a Data Mining Technique”, Universitatea Petrol-Gaze din Ploiesti, Vol. LX(1/2008):49-56.

[4] J. Mata, J. L. Alvarez and J. C. Riquelme, “An Evolutionary Algorithm to Discover Numeric Association Rules”, SAC 2002, Madrisd, Spain.

[5] K. Bhavani and R. Hemalatha, “Role of Association Rule Mining in String and Numerical Data”, International Journal of Computer & Organization Trends, Vol 3(4), 2013, 94-96.

[6] R. Agarwal, T. Imielinski and A. N. Swami, “Mining Association Rules between sets of items in large databases” , ACM SIGMOD, 207-216, In proceedings of the 1993.

[7] V. Umarani, and Dr. M. Punithavalli,”A study on effective mining of association rules from huge databases”, International Journal of Computer Science and Research, Vol. 1 Issue 1, 2010:30-34

[8] http://www.kdnuggets.com/2013/08/languages-for-analytics-data-mining-data-science.html, KDnuggets a top resource of data mining.