Mining bookstore and titanic data by Weka for understanding promotional strategy and predicting survival pattern

Technical Report · June 2018

DOI: 10.13140/RG.2.2.15611.41764

Rajib HOSSAIN Khan
Linnaeus University

Mana Abedi Sohrforouzani
Linnaeus University

Mining bookstore and titanic data by Weka for understanding promotional strategy and predicting survival pattern

Rajib Hossain Khan (rk222id@student.lnu.se)
Mana Abedi Sohrforouzani (ma223ye@student.lnu.se)
Shahrzad Darvishi (sd222im@student.lnu.se)
Marie Claire Ukwishaka (mu222ch@student.lnu.se)

Introduction

Data mining can be described as the process of trawling through data to find previously unknown relationships among the data that may be interesting to the user of that data (Hand, 1998). Although the concept of data mining is not new but quite old, it is still challenged by some scholars who claim that it does not contribute to the business and society in a large scale (Pechenizkiy, Puuronen and Tsymbal, 2005). It is because, people perceive data mining to be a potentially powerful tool, but the real benefit of this tool is yet to be fully recognized (Wang, Hu and Zhu, 2007). In order to become a genuine business intelligence tool for comprehensive knowledge discovery, data mining must be integrated with knowledge management for improving the organizational knowledge (Wang and Wang, 2008). The main purpose of this study is to perform some data mining algorithms and analysis on a bookstore dataset and titanic dataset to understand which books can be promoted together and which passengers survived the tragedy.

Techniques used for data mining

Data mining is a process for extracting knowledge or interesting patterns from existing databases to serve specific purposes (Lee, Hong and Lin, 2005). Pattern recognition technologies as well as statistical and mathematical techniques are employed for performing data mining (Shmueli, Patel and Bruce, 2011). In this particular study, association rule learning, classification technique and clustering algorithm were used for mining the bookstore and titanic dataset.

Association rule learning

Association rule learning is a popular and well-accepted method for discovering interesting relations between variables in large databases (Lekha, Srikrishna and Vinod, 2013). Agrawal, Imieliński and Swami (1993) presented an efficient algorithm that generates significant association rules from a large collection of basket data type transactions. They exemplified such an association rule with the statement “that 90% of transactions that purchase bread and butter also purchase milk”. Borgelt and Kruse (2002) reported that, association rule learning is a powerful method for so-called market basket analysis that aims at finding regularities in the shopping behaviour of customers. The classic algorithm of association rules is Apriori algorithm and it enumerates all of the frequent item sets (Yabing, 2013). Therefore, it is used for mining the bookstore dataset.

Classification technique

Classification is one of the most common learning methods in data mining (Giraud-Carrier and Povel, 2003). Classification is a key data mining technique where database tuples that acts as training samples are analyzed to develop a model of the given data (Fayyad, Piatetsky-Shapiro and Smyth, 1996). There exist a number of classification techniques from the statistics and machine learning communities (Fayyad, Djorgovski and Weir, 1996). The most popular method of classification is the induction of decision trees (Quinlan, 1986). A decision tree is a flow chart type of structure containing internal nodes, leaf nodes and branches (Kamber et al., 1997). J48 is a very simple classifier to make a decision tree but it can give efficient result (Youn and McLeod, 2007). That is the reason why J48 is used for the titanic dataset.

Clustering algorithm

Huang (1998) reported that, partitioning a set of objects into homogeneous groups or clusters is a fundamental operation in data mining. So, the main purpose of clustering method partition is to arrange a set of objects into clusters such that objects in the same cluster are more similar to each other than objects in different clusters according to some defined criteria. The primary use of clustering algorithms is to discover the grouping structures inherent in data. The most attractive property of k-means algorithm in data mining is its efficiency in clustering large data sets. K-means clustering is a very popular clustering technique that is used in numerous problematic situations (Kanungo et al., 2000). For creating homogeneous groups with having common characteristics SimpleKMeans clustering algorithm were used for the titanic dataset.

Bookstore dataset

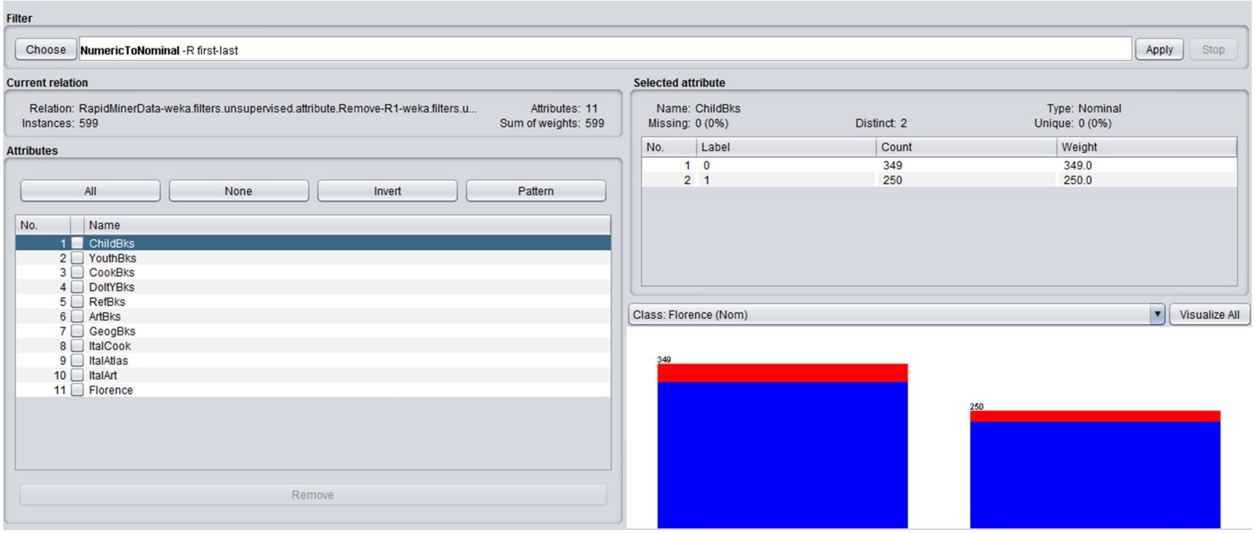
The bookstore dataset conveys sells information of different types of books from a particular bookstore. The dataset contains 12 attributes and all of them are numeric. One of the purposes of this study is to recommend the owner that which books should be promoted together for a specific customer group. In order to accomplish this, Apriori algorithm is applied to the data but it is also required to filter the dataset for the application of this algorithm.

Filtering bookstore dataset

Before filtering the dataset, the attribute ‘ID transaction’ was selected and removed as it may affect the quality of the result. After that, a filter called ‘weka.filters.unsupervised.attribute.NumericToNominal’ is applied to the bookstore dataset to convert the numeric attributes to nominal ones. It is required because Weka did not allow carrying out Apriori algorithm on bookstore data when the attributes are numeric. Figure-1 depicts the result of filtering.

Application of Apriori algorithm on bookstore dataset

Before applying ‘weka.associations.Apriori’ algorithm on the dataset some options were manipulated in order to get the best results. For example, ‘minMetric’ was changed from ‘0.9’ to ‘0.75’. The value of this option is considered as the standard for determining the rules. So, rules with scores higher than this value is then drawn by the algorithm. The value of the ‘numRules’ option is modified from ‘10’ to ‘20’ so that the system is enable to produce more rules rather than concluding with only a few rules, since the dataset is not so large. It is also quite important to switch the value of ‘treatZeroAsMissing’ option from ‘False’ to ‘True’ to avoid getting rules of customers’ non-purchasing behavior. Finally, although the default value (‘Confidence’) of ‘metricType’ option is not changed but it is worthy enough to notice that this is used to rank the rules. Figure-2 visualizes the settings in Weka user interface.



NumericToNominal -R first-last

Relation: RapidMinerData-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.u...
Instances: 599
Attributes: 11
Sum of weights: 599

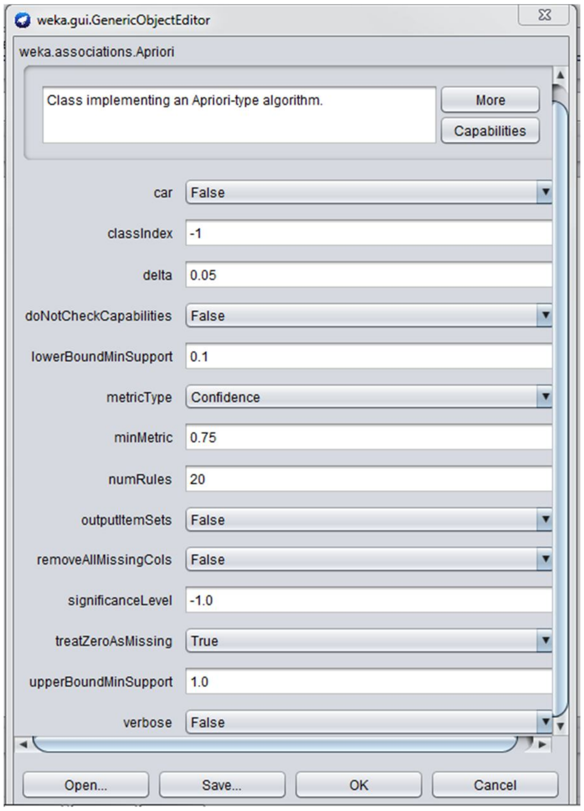
No.		Name
1	□	ChildBks
2	□	YouthBks
3	□	CookBks
4	□	DoItYBks
5	□	RefBks
6	□	ArtBks
7	□	GeogBks
8	□	ItalCook
9	□	ItalAtlas
10	□	ItalArt
11	□	Florence

Name: ChildBks
Missing: 0 (0%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	0	349	349.0
2	1	250	250.0

Class: Florence (Nom)

Figure-1: Bookstore dataset filtered



weka.associations.Apriori

Class implementing an Apriori-type algorithm.

car	False
classIndex	-1
delta	0.05
doNotCheckCapabilities	False
lowerBoundMinSupport	0.1
metricType	Confidence
minMetric	0.75
numRules	20
outputItemSets	False
removeAllMissingCols	False
significanceLevel	-1.0
treatZeroAsMissing	True
upperBoundMinSupport	1.0
verbose	False

Figure-2: Settings for ‘weka.associations.Apriori’

Results of Apriori algorithm for bookstore dataset

Based on the above settings, Weka produces output of Apriori algorithm for the dataset which shows 7 best rules were produced with a minimum support of 0.1 and minimum metric (confidence) of 0.75. The sizes of sets of large itemsets were also displayed and there are three large itemsets that have the required minimum support. Figure-3 illustrates the output. The confidence value for all the seven rules is more than 0.75 which indicates that in more than 75% cases there is a chance that the rules will be found true. Since, the lift value for the generated rules is greater than 1; it further increases the chances of the rules to become true. The lev (leverage) value 0 means independence, but the lev value in this case is more than 0 for all the instances which is also a very good sign (indicates dependency). Finally, the conv (conviction) value 1 also directs to independence but a higher value than 1 means dependency. So, the conv value for the rules is higher than 1 which completely backs the rules produced for the bookstore dataset.

Business model for bookstore dataset

Based on the result of applied data mining algorithm it is suggested to the CEO of the bookstore to promote the following books to the customers that were recommended in Table-1.

Books to be promoted together	Reason
Promote youth books and cook books with child books.	Customers who bought child books also purchased youth books and cook books.
Promote cook books and reference books with child books.	Consumers who procured child books also acquired cook books and reference books. It was also found that book lovers who purchased cook books also bought child books and reference books.
Promote cook books and geography books with child books.	Patrons who bought child books also acquired cook books and geography books.
Promote cook books and art books with child books.	Readers who acquired child books also purchased cook books and art books.
Promote cook books and do it books with child books.	It was found that customers who procured child books also bought cook books and do it books. On the other hand, consumers who purchased cook books also acquired child books and do it books.

Table-1: Recommendations for the promotion of books to the CEO of bookstore

Titanic dataset

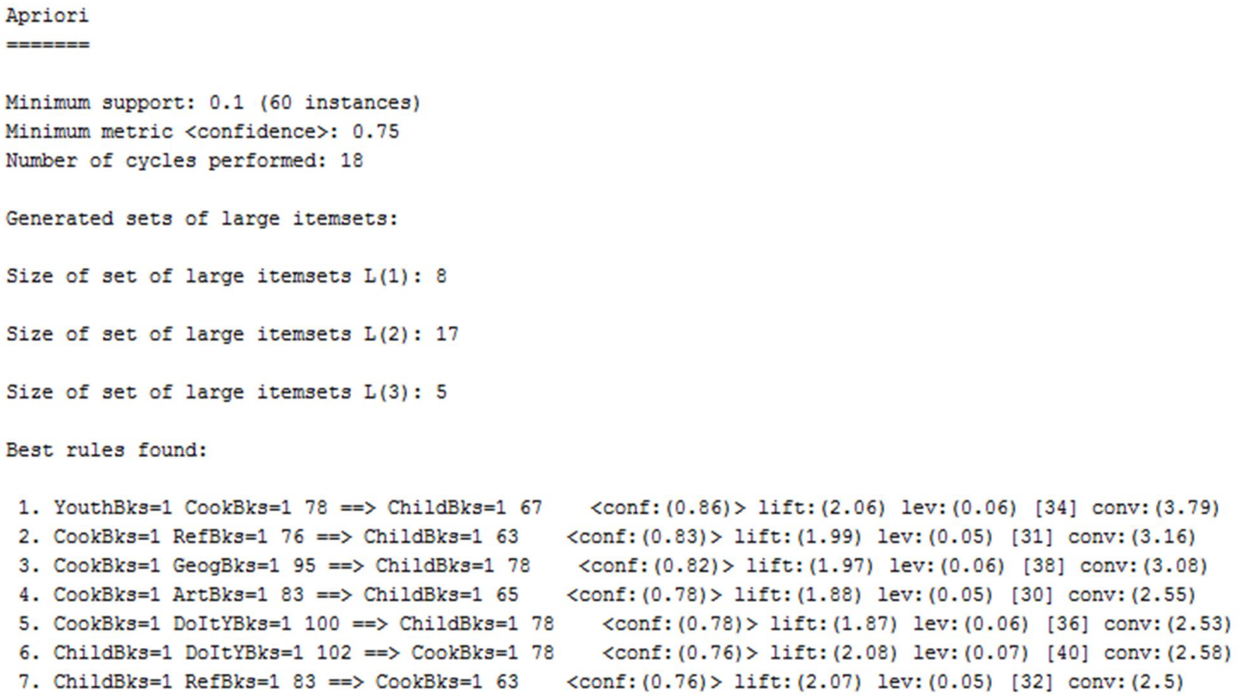
The titanic dataset contains information about the class, age and sex of the passengers who were either able to be survived or not survived the tragedy. The dataset contains five attributes and four of them are nominal. The other attribute that refers to the serial number of the passenger, is numeric but it is removed for the sake of accuracy and quality. The objective of data mining with this particular dataset is to learn which types of passengers were able to be survived the tragedy. In order to attain this, J48 classifier and SimpleKMeans clustering is used.

Application of J48 classifier on titanic dataset

The ‘weka.classifiers.trees.J48’ was applied to the titanic dataset with default settings. The result is discussed as follows.

Output of J48 pruned tree

The result of J48 pruned tree shows that 267 passengers survived the tragedy and most of them were women (251.0) who were either 1st class (145.0) or 2nd class (106.0) passengers but the women who traveled as 3rd class passenger did not survive. There are also 16.0 instances in the total numbers of people survived from the age group ‘child’. Meanwhile, 1049 people did not survive the tragedy and among them most are men. The breakdown of people failed to survive the tragedy is as follows, 175.0, 168.0 and 510.0 adult men from 1st, 2nd and 3rd class respectively along with 196.0 women from 3rd class. Figuer-4 and figure-5 represents the J48 pruned tree. One interesting thing is that in some cases the tree showed results like ‘(175.0/57.0)’, which happened because there are missing attribute values because of misclassification. However, in those results the first figure is the total number of instances reaching the leaf or correctly classified instances while the second figure is the number of those instances that are misclassified.



```
Apriori
=======

Minimum support: 0.1 (60 instances)
Minimum metric <confidence>: 0.75
Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 8

Size of set of large itemsets L(2): 17

Size of set of large itemsets L(3): 5

Best rules found:

 1. YouthBks=1 CookBks=1 78 ==> ChildBks=1 67    <conf:(0.86)> lift:(2.06) lev:(0.06) [34] conv:(3.79)
 2. CookBks=1 RefBks=1 76 ==> ChildBks=1 63    <conf:(0.83)> lift:(1.99) lev:(0.05) [31] conv:(3.16)
 3. CookBks=1 GeogBks=1 95 ==> ChildBks=1 78    <conf:(0.82)> lift:(1.97) lev:(0.06) [38] conv:(3.08)
 4. CookBks=1 ArtBks=1 83 ==> ChildBks=1 65    <conf:(0.78)> lift:(1.88) lev:(0.05) [30] conv:(2.55)
 5. CookBks=1 DoItYBks=1 100 ==> ChildBks=1 78    <conf:(0.78)> lift:(1.87) lev:(0.06) [36] conv:(2.53)
 6. ChildBks=1 DoItYBks=1 102 ==> CookBks=1 78    <conf:(0.76)> lift:(2.08) lev:(0.07) [40] conv:(2.58)
 7. ChildBks=1 RefBks=1 83 ==> CookBks=1 63    <conf:(0.76)> lift:(2.07) lev:(0.05) [32] conv:(2.5)
```

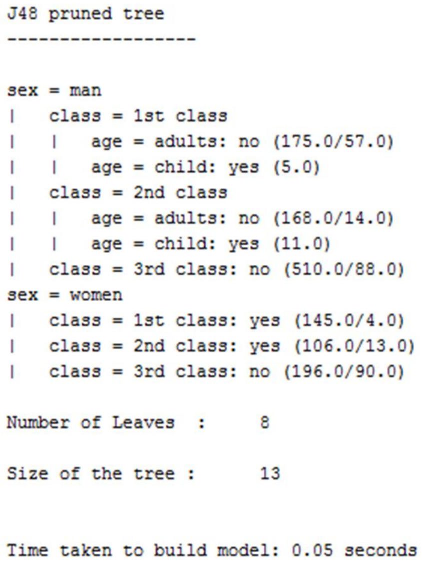
Figure-3: Output of Apriori algorithm for bookstore dataset

Summary of the J48 classifier output

Figure-6 denotes that the percentage of ‘Correctly Classified Instances’ is 79.787% (1050) while about 21.212% (266) of instances were incorrectly classified. The total numbers of instances were 1316. The higher the percentage of ‘Correctly Classified Instances’ the better the model is. So, the resulting model is a good one that is almost 80% accurate. Landis and Koch (1977) reported that the value of Kappa statistic between 0-0.20 is slight, 0.21-0.40 is fair, 0.41-0.60 is moderate, 0.61-0.80 is substantial and 0.81-1 is almost perfect. So, in this case the Kappa statistic value is 0.528 which is also an indication of the acceptance of the derived model since it is a measure of the agreement of prediction with the true class. The resulted error values are not so meaningful for classification tasks.

Detailed accuracy by class and confusion matrix

The confusion matrix at figure-6 shows that J48 classified the titanic dataset based on the attribute ‘survived {yes, no}’. It predicted that 267 people could be able to survive the tragedy while 1049 people could not. The ‘TP Rate’ for class ‘yes’ is 0.501 which entails that when the passengers really survived in 50% of the cases the algorithm predicted that the passengers could survive. On the other hand, the ‘TP Rate’ for class ‘no’ is 0.979 that means while the passengers actually did not survive in almost 98% cases the algorithm predicted that they could not survive. Both ‘TP Rate’ and ‘Recall’ are equivalent to each other. The ‘FP Rate’ for class ‘yes’ is 0.021 which indicates that in 2% cases the algorithm predicted that the passengers could survive while they actually did not survive. In contrast, the ‘FP Rate’ for class ‘no’ is 0.499 that means in almost 50% cases the algorithm predicted that the passengers could not survive but actually they managed to survive. The precision measures the correctness of the prediction. The precision value for class ‘yes’ is 0.936 and for class ‘no’, it is 0.763 which means the predictions are approximately accurate.



```
J48 pruned tree
------------------

sex = man
|   class = 1st class
|   |   age = adults: no (175.0/57.0)
|   |   age = child: yes (5.0)
|   class = 2nd class
|   |   age = adults: no (168.0/14.0)
|   |   age = child: yes (11.0)
|   class = 3rd class: no (510.0/88.0)
sex = women
|   class = 1st class: yes (145.0/4.0)
|   class = 2nd class: yes (106.0/13.0)
|   class = 3rd class: no (196.0/90.0)

Number of Leaves  :     8

Size of the tree :      13


Time taken to build model: 0.05 seconds
```

Figure-4: J48 pruned tree

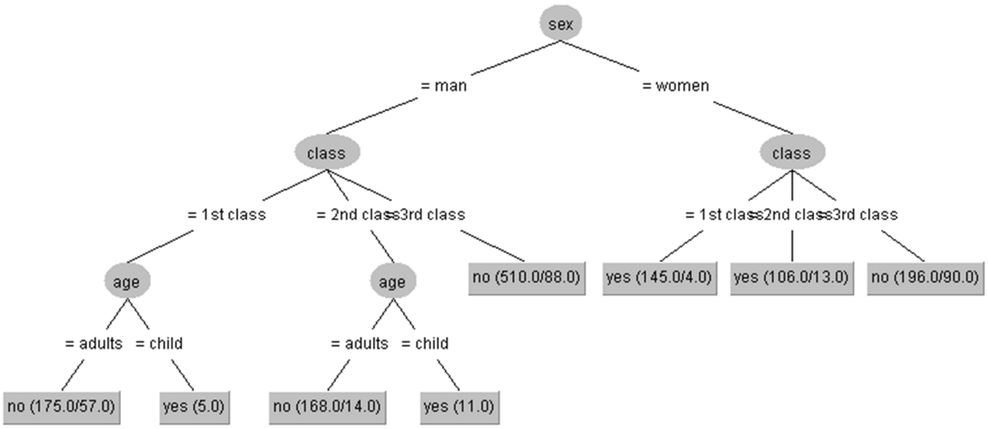
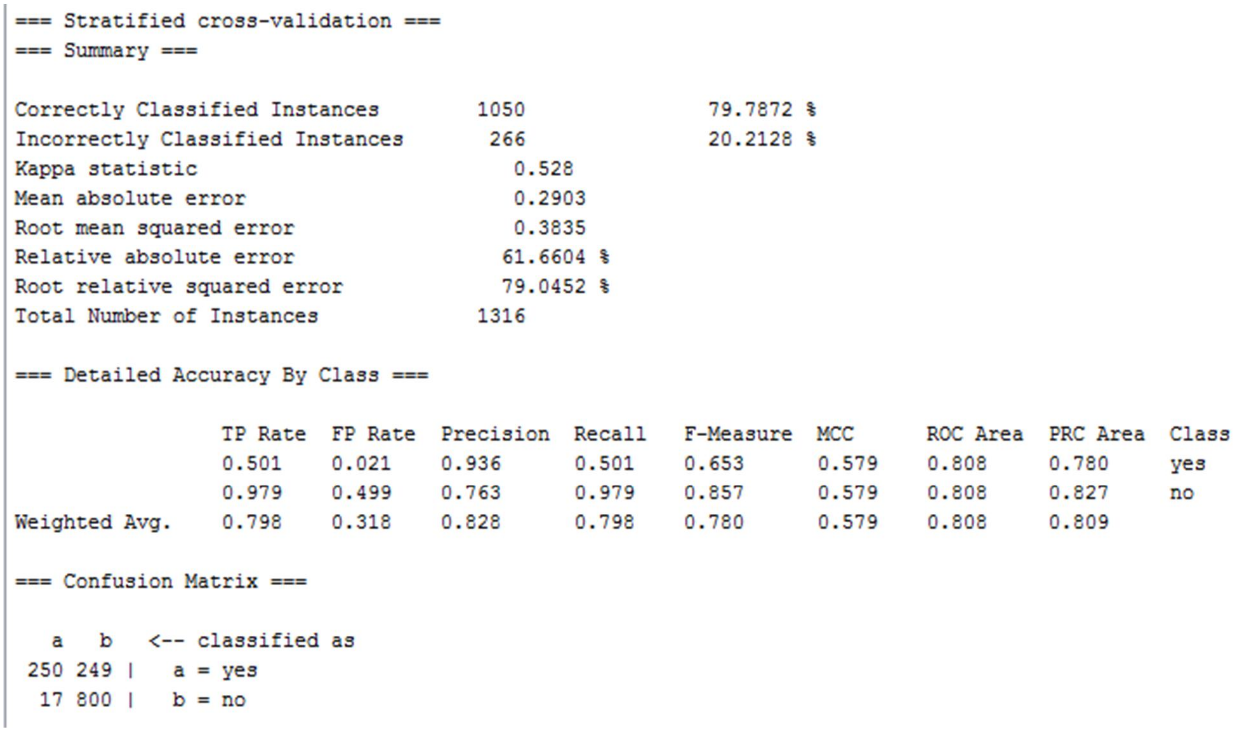


Figure-5: Visualization of the tree

Application of SimpleKMeans clustering on titanic dataset

The ‘weka.clusteres.SimpleKMeans’ was applied on the titanic dataset with the default settings. The results of the data mining are interpreted here.



```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1050               79.7872 %
Incorrectly Classified Instances       266               20.2128 %
Kappa statistic                          0.528
Mean absolute error                      0.2903
Root mean squared error                  0.3835
Relative absolute error                 61.6604 %
Root relative squared error             79.0452 %
Total Number of Instances             1316

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.501    0.021    0.936      0.501    0.653      0.579    0.808     0.780     yes
                 0.979    0.499    0.763      0.979    0.857      0.579    0.808     0.827     no
Weighted Avg.    0.798    0.318    0.828      0.798    0.780      0.579    0.808     0.809

=== Confusion Matrix ===

   a   b   <-- classified as
 250 249 |   a = yes
  17 800 |   b = no
```

Figure-6: Summary of the J48 classifier output

Centroid of the clusters

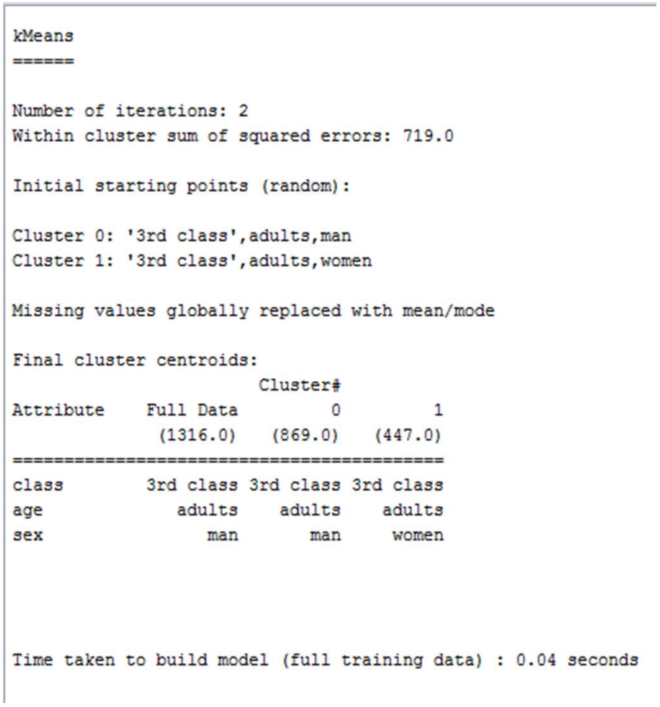
From figure-7 it is evident that there are two clusters. Centroids are used to characterize the clusters. ‘Cluster 0’ has 869.0 instances and the passengers are adult man from 3rd class while ‘cluster 1’ has 447.0 instances and the passengers are adult women from 3rd class.

Classes to clusters

Figure-8 shows that the class attribute is ‘survived’. The algorithm assigned classes to clusters based on the highest value of the class attribute within each cluster. For the instances of ‘cluster 0’ the highest value (694) of the class attribute is ‘no’ which means that most of the passengers of this cluster could not survive. On the contrary, for the instances of ‘cluster 1’ the highest value (324) of the class attribute is ‘yes’ that means most of the passengers of this cluster could survive the tragedy. Therefore, ‘no’ is assigned to ‘cluster 0’ and ‘yes’ is assigned to ‘cluster 1’. The ‘incorrectly clustered instances’ are 298.0 (22.64%) which is quite acceptable.

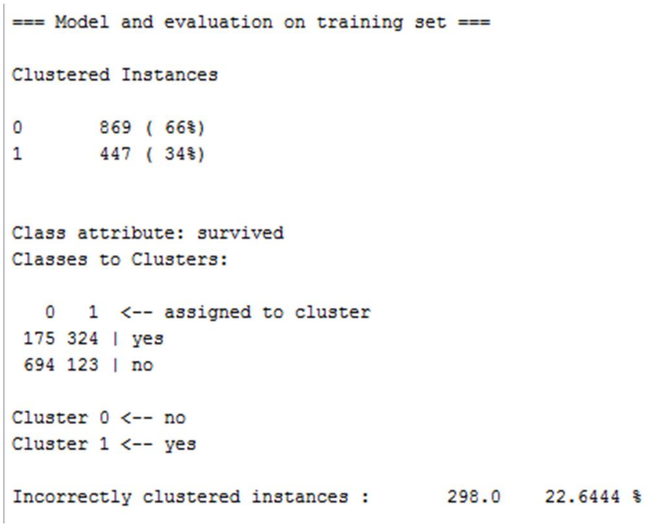
Visualizing cluster assignments

It is also possible to understand the characteristics of each cluster via visualizing cluster assignments. Here, the class attribute was chosen as the x-axis, the clustered attributes were picked as the y-axis and the cluster was selected as the color dimension. The visualization is narrated as survived vs. class, survived vs. age and survived vs. sex.



```
kMeans
======

Number of iterations: 2
Within cluster sum of squared errors: 719.0

Initial starting points (random):

Cluster 0: '3rd class',adults,man
Cluster 1: '3rd class',adults,women

Missing values globally replaced with mean/mode

Final cluster centroids:
                        Cluster#
Attribute    Full Data         0         1
              (1316.0)   (869.0)   (447.0)
==========================================
class        3rd class 3rd class 3rd class
age             adults    adults    adults
sex                man       man     women




Time taken to build model (full training data) : 0.04 seconds
```

Figure-7: Centroid of the cluster

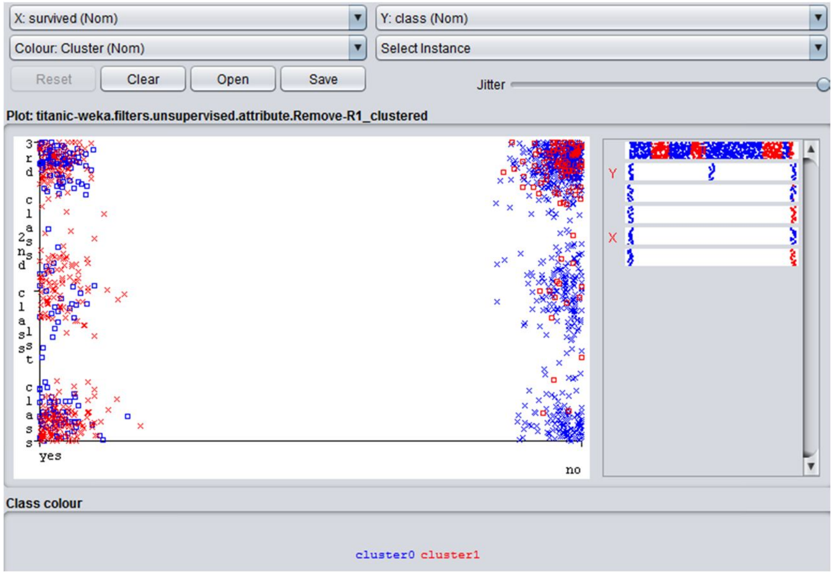


```
=== Model and evaluation on training set ===

Clustered Instances

0       869 ( 66%)
1       447 ( 34%)


Class attribute: survived
Classes to Clusters:

   0   1  <-- assigned to cluster
 175 324 | yes
 694 123 | no

Cluster 0 <-- no
Cluster 1 <-- yes

Incorrectly clustered instances :       298.0    22.6444 %
```

Figure-8: Classes to clusters

Survived vs. class

Figure-9 illustrates that passengers from ‘cluster 1’ mostly survived the tragedy throughout all the classes.



X: survived (Nom)

Y: class (Nom)

Colour: Cluster (Nom)

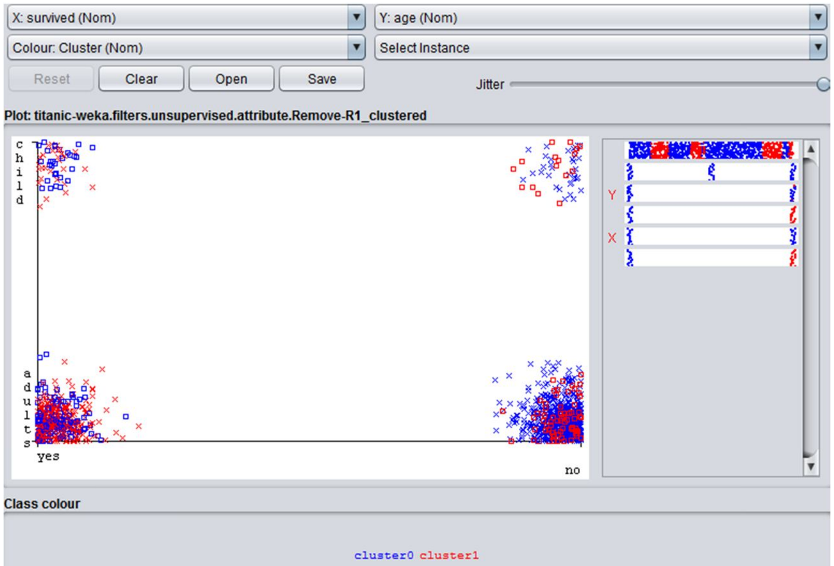
Select Instance

Plot: titanic-weka.filters.unsupervised.attribute.Remove-R1_clustered

Figure-9: Visualizing cluster assignments (survived vs. class)

Survived vs. age

Figure-10 represents that most adult and child passengers from ‘cluster 1’ survived the tragedy in comparison to their counterparts in ‘cluster 0’.



X: survived (Nom)

Y: age (Nom)

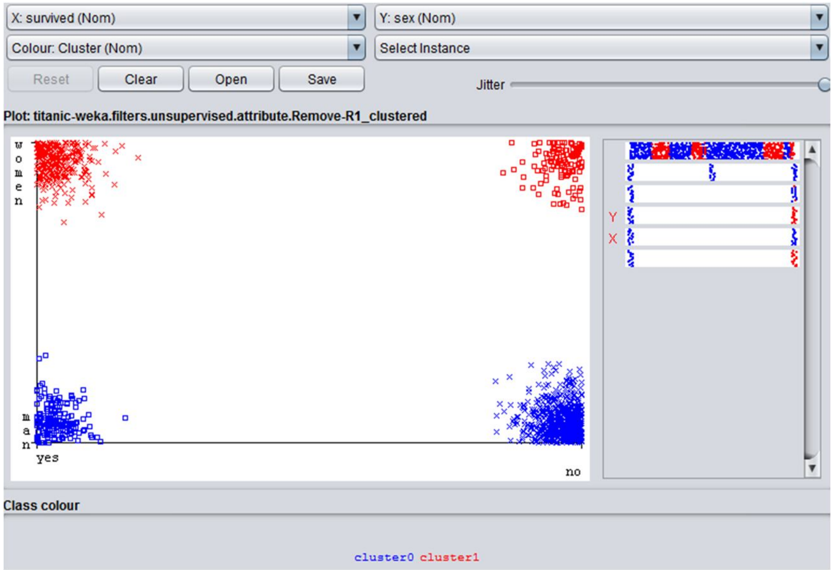
Colour: Cluster (Nom)

Plot: titanic-weka.filters.unsupervised.attribute.Remove-R1_clustered

Figure-10: Visualizing cluster assignments (survived vs. age)

Survived vs. sex

Figure-11 depicts something completely interesting. In ‘cluster 1’ all the instances were women and the density of survived women is greater than those that could not survive. On the other hand, ‘cluster 0’ consists of instances that are men and their survival rate is less when compared with those that could not survive. When the comparison takes place between the women and men who survived the tragedy, then the ratio of survived women passengers from ‘cluster 1’ is greater than the ratio of survived men passengers from ‘cluster 0’.



X: survived (Nom)

Y: sex (Nom)

Colour: Cluster (Nom)

Plot: titanic-weka.filters.unsupervised.attribute.Remove-R1_clustered

Figure-11: Visualizing cluster assignments (survived vs. age)

So, the visual results supports the conclusion drawn (the instances of 'cluster 1' are more likely to be survived) from the output of SimpleKMeans clustering for the titanic dataset.

Decision model for the titanic dataset

On the basis of above analysis, it can be predicted that female passengers mostly survived the tragedy than the male passengers. If we consider the class of the passengers then it can be predicted that the female passengers of the 1st and 2nd class were able to survive the tragedy but those who were 3rd class female passengers did not survive. While considering the age, the survival rate of child was much higher than the adults.

Conclusion

The main purpose of this study was to perform some data mining algorithms and analysis on a bookstore dataset and titanic dataset to recommend which books can be promoted together and to predict which passengers survived the tragedy. For the accomplishment of this purpose, Apriori algorithm is used for the bookstore dataset to generate association rules. Weka produced 7 best rules and based on these rules 5 types of promotional strategy was recommended to the CEO of the bookstore. J48 classifier was used for the titanic dataset and it revealed that the women and child passengers from 1st and 2nd class survived the tragedy. SimpleKMeans clustering was also applied for the titanic dataset and it also determined that the chances of survival are greater for the women than the other type of passengers. However, it would be great to apply other relevant data mining algorithms on these two datasets for the sake of cross validating the results.

References

Agrawal, R., Imieliński, T. and Swami, A., 1993, June. Mining association rules between sets of items in large databases. In *Acm sigmod record* (Vol. 22, No. 2, pp. 207-216). ACM.

Borgelt, C. and Kruse, R., 2002. Induction of association rules: Apriori implementation. In *Compstat* (pp. 395-400). Physica, Heidelberg.

Fayyad, U.M., Piatetsky-Shapiro, G. and Smyth, P., 1996, August. Knowledge Discovery and Data Mining: Towards a Unifying Framework. In *KDD* (Vol. 96, pp. 82-88).

Fayyad, U.M., Djorgovski, S.G. and Weir, N., 1996, February. Automating the analysis and cataloging of sky surveys. In *Advances in knowledge discovery and data mining* (pp. 471-493). American Association for Artificial Intelligence.

Giraud-Carrier, C. and Povel, O., 2003. Characterising data mining software. *Intelligent Data Analysis*, *7*(3), pp.181-192.

Hand, D.J., 1998. Data mining: Statistics and more?. *The American Statistician*, *52*(2), pp.112-118.

Huang, Z., 1998. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data mining and knowledge discovery*, *2*(3), pp.283-304.

Kamber, M., Winstone, L., Gong, W., Cheng, S. and Han, J., 1997, April. Generalization and decision tree induction: efficient classification in data mining. In *Research Issues in Data Engineering, 1997. Proceedings. Seventh International Workshop on* (pp. 111-120). IEEE.

Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C., Silverman, R. and Wu, A.Y., 2000, May. The analysis of a simple k-means clustering algorithm. In *Proceedings of the sixteenth annual symposium on Computational geometry* (pp. 100-109). ACM.

Landis, J.R. and Koch, G.G., 1977. The measurement of observer agreement for categorical data. *biometrics*, pp.159-174.

Lee, Y.C., Hong, T.P. and Lin, W.Y., 2005. Mining association rules with multiple minimum supports using maximum constraints. *International Journal of Approximate Reasoning*, *40*(1-2), pp.44-54.

Lekha, A., Srikrishna, C.V. and Vinod, V., 2013, January. Utility of association rule mining: A case study using Weka tool. In *Emerging Trends in VLSI, Embedded System, Nano Electronics and Telecommunication System (ICEVENT), 2013 International Conference on* (pp. 1-6). IEEE.

Pechenizkiy, M., Puuronen, S. and Tsymbal, A., 2005. Why data mining research does not contribute to business?. In *Proc. of Data Mining for Business Workshop DMBiz (ECML/PKDD'05), Porto, Portugal* (pp. 67-71).

Quinlan, J.R., 1986. Induction of decision trees. *Machine learning*, *1*(1), pp.81-106.

Shmueli, G., Patel, N.R. and Bruce, P.C., 2011. *Data mining for business intelligence: Concepts, techniques, and applications in Microsoft Office Excel with XLMiner*. John Wiley and Sons.

Yabing, J., 2013. Research of an improved apriori algorithm in data mining association rules. *International Journal of Computer and Communication Engineering*, *2*(1), p.25.

Youn, S. and McLeod, D., 2007. A comparative study for email classification. In *Advances and innovations in systems, computing sciences and software engineering* (pp. 387-391). Springer, Dordrecht.

Wang, J., Hu, X. and Zhu, D., 2007. Diminishing downsides of data mining. *International Journal of Business Intelligence and Data Mining*, *2*(2), pp.177-196.

Wang, H. and Wang, S., 2008. A knowledge management approach to data mining process for business intelligence. *Industrial Management & Data Systems*, *108*(5), pp.622-634.