Journal of Banking,
Information Technology and Management
Volume 7 No. 1
January-June 2010

E-Commerce Luxury Products Promotion through WEKA using Naive Bayes Classification Approach

Roma Chauhan & Ritu Chauhan

Abstract

Data mining technique has been developed as a field of basic and applied research in computer science in general and E-Commerce & E-Business in particular. It has been used as an effective tool by researchers and key organizations since long time for the purpose of improving business intelligence process. We obtained the database of affluent people across the globe and further applied classification and visualization technique to determine the buying patterns and do some predictive analysis which can further help in focusing product marketing within certain confined areas of e-commerce. The paper examines the role of how Naive Bayes classification approach can be used to classify the database in set of wealth and age further to promote luxury products within e-commerce value chain.

Key words: *Data Mining, Classification, Naive Bayes, Data Visualization, E-Commerce, Business Intelligence, WEKA*

Introduction

Luxury good manufacturers always look for a significant place to enhance their overall brand visibility across the globe. Choosing the correct product combination is not an easy task. The luxury retailers must satisfy everyone from the first time luxury purchaser to the seasoned super-rich with a wide range of disposable income. The seasoned super-rich want more product and brand availability, access to products that will really make them feel "special". Yet making things little complicated as many such luxury brands and products are difficult to find and simply not very visible. When it comes to luxury goods the high-end brand conscious consumers demand the newest, most up-to-date products to be selected by them for their basket. Luxury producers can go a long way towards improving their prospects by data mining, both for setting up physical stores and e-commerce portals across the globe. According to research survey, there is much that can be done to increase luxury goods sales across the globe. Luxury retailers must work to improve the overall shopping experience by improving brand and product selection, store atmosphere, and customer service.

The preferences of the current generation are changing rapidly. The eager customer does not mind paying extra for better facilities and ambience. The market for luxury products is surely climbing at an astonishing rate as compared to what it was a

decade ago which was almost negligible. The paper illustrates how data mining techniques can be further utilized for the purpose of goods promotion. The challenge here is to provide goods to the high end customers across the globe including regions such as Asia, Europe, the United States, Middle East and others commerce of the products through web portals. The technique is to use data mining techniques to understand billionaire's segregation in these regions.

Drift Towards e-commerce

In the emerging economic globally e-commerce and e-business have increasingly become a necessary components of to making a business strategy, e commerce have changed the most of the business functionality, with development of internet technology and web-based technology distinctions between customers and retailers, retailers and distributors and distributes and factories and factories to number of suppliers being narrow down. The customer buying patterns also depends on change in marketing strategies of companies with change in consumer buying behaviour. With change in consumer buying behaviour the companies also made necessary changes in their marketing strategies. The changes that includes launching of premium products by companies to fulfil requirements of high class consumers. The demand for global luxury online sales is on the increase. Recent reports indicate that the wealthy are almost all online and are pleased with making online purchases.

Website and e-store design seek to achieve more than basic, functional requirements such as providing a conducive and pleasant shopping experience. E-commerce involves a constant flow of innovative means of differentiation that will meet the expectations of the online consumer in order to generate more online traffic and maintain customer's loyalty. The challenge of selling luxury goods online is enormous. The luxury shopping experience is different from the conventional shopping experience. Luxury goods are sensory in nature and their purchase requires a high aesthetical appreciation and the utilization of the human senses of vision, aural, smell and touch. This often requires human and physical store presence, which is absent in the online virtual environment.

The major drawback with Internet is that it also lacks the exclusive and prestigious locations where the luxury stores are situated. Therefore the question of creating a prestigious online atmosphere, replacing the human senses in the virtual environment and matching 'high class' with the 'mass class' of the Internet world is a challenge to be achieved. Online luxury consumers have high expectations and this includes their belief that although the luxury e-boutique is available to the masses, it should be designed to feel right to select only a few.

Applying Data Mining

Well data mining have its dominant goal to access the useful information into a massive amount of data which is useful for decision-makers and to decide the business strategy. Use of the data mining in e-commerce provides a significant way to analyze the customers' demands, to predict the customers' purchasing decisions, through data mining the e-retailer analyze the customer simply by knowing what type of customer is likely to purchase a particular item. Data mining helps in establishing a relationship with the customer without ever seeing his or

her face since customers are not aware data mining process but customers like the items suggested to them that are exactly what they like, within the software and data base which support data mining.

Data mining is also known as knowledge discovery in data base, and is the process to extract the relevant into from the database. Today the volume of data has risen to terabytes. The massive data have kind of hidden information which can be strategically used. But the question is how massive collection of data can be used to draw meaningful conclusions about a particular domain? The solution is data mining which is basically a technique to transfer these data into relevant information. It is commonly dealt in a wide range of profiling practices such as marketing, surveillance fraud detection, human factor related issue and scientific discovery. It is not only used by business organization to collect the business information but it is used in the science and technology to extract information from the enormous data set generated by modern research mrthods and observations.

Implementing Naive Bayes in WEKA

Naive Bayes is one of the most efficient and effective inductive learning algorithms for machine learning and data mining [9]. For a record to be classified based on Bayesian theorem, the categories of the predictor variables are noted and the record is classified according to the most frequent class among the same values of those predictor variables in the training set. A rigorous application of the Bayes theorem would require availability of all possible combinations of the values of the predictor variables. When the number of variables is large enough, this requires a training set of unrealistically large size. The Naive Bayes method overcomes this practical limitation of the rigorous Bayes approach to classification. The research paper explains several challenging problems based on our experiments and analysis in implementing mining techniques to promote products within e-commerce chain. The acceptability of the software was primarily for CSV or ARFF file format. The data was complex to be analyzed.

Classification is a data mining technique which maps data in to redefine groups and class, Naïve Bayes is simple probabilistic classifier which is based upon Bayes theorem classifier with strong independence assumptions

According to Baye's rule:

Given a hypothesis h and data D which bears on the hypothesis:

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}$$

where:

P(h): independent probability of h: prior probability

P(D): independent probability of D

P(D | h): conditional probability of D given h: likelihood

P(h | D): conditional probability of h given D: posterior probability

The basic idea of Bayes rule is that outcome of any event can be predicted based on

its evidence that can be observed, for example to predict the chance of raining we can usually use some evidence such that amount of dark cloud in the area, that means there are predictions that there is dark cloud when it rains. Of course, there can be some more evidences which are related to these events. Generally, it is better to have more than one evidence to support the predication of an event. However, the evidence must be related to an event. For example if we add an evidence earthquake in this event then above model might yield worse performance because raining is not related to the evidences of earthquake

To making an NB model with the independence assumption, we can rewrite the Bayes Rule as follows; (which is Naïve Bayes classification)

The naive Bayes classifier assigns an instance *sk* with attribute values (*A*1=v1, *A*2=v2, …, *Am*=v*m*) to class *Ci* with maximum Prob(*Ci* | (v1, v2, …, v*m*)) for all *i*.

Under the assumption of independent attributes

$$P[(\mathrm{v}_1, \mathrm{v}_2, ..., \mathrm{v}_m) | C_j]$$
$$= P(A_1 = \mathrm{v}_1 | C_j) \cdot P(A_2 = \mathrm{v}_2 | C_j) \cdot .... \cdot P(A_m = \mathrm{v}_m | C_j)$$
$$= \prod_{h=1}^{m} P(A_h = \mathrm{v}_h | C_j)$$

Furthermore, **P(Cj)** ***can be computed by a***

$$\frac{\text{number of training samples belonging to } C_j}{\text{total number of training samples}}$$

Data Set Description

We would be working on dataset Billionaires 1992, dataset which is available for free use. *Fortune* magazine publishes the list of billionaires annually. The 1992 list included 233 individuals or families. Their wealth, age and geographic location (Asia, Europe, Middle East, United States or others) is reported covering number of 233 cases [1][2].

ARFF File Format

% Reference: *Fortune*, September 7,1992. "The Billionaires." pp. 98-138.

% Authorization: free use

% Description: *Fortune* magazine publishes the list of billionaires annually. The 1992 list included 233 individuals or families. Their wealth, age and geographic location (Asia, Europe, Middle East, United States or Other) is reported.

% Number of cases: 233

% Variable Names:

%

% wealth: Wealth of family or individual in billions of dollars

Percentage: Age in years (for families it is the maximum age of family members)

% region: Region of the World (Asia, Europe, Middle East, United States and Others)

%

@RELATION relation

@ATTRIBUTE 'wealth' numeric

@ATTRIBUTE 'age' numeric

@ATTRIBUTE 'region' {"A","E","M","O","U"}

@DATA

Experimental Results

As shown in Figure 2 WEKA visualization techniques allow seeing how disparity lies geographically in context to number of billionaires in a particular region. It shows that highest number of billionaires belongs to European region. After European region it is the United States, Asians, others and finally the Middle East region.

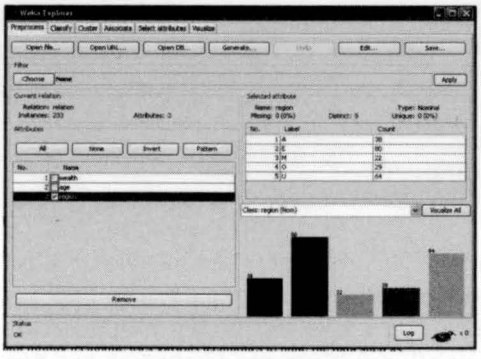

Figure 1. Data Preprocessing

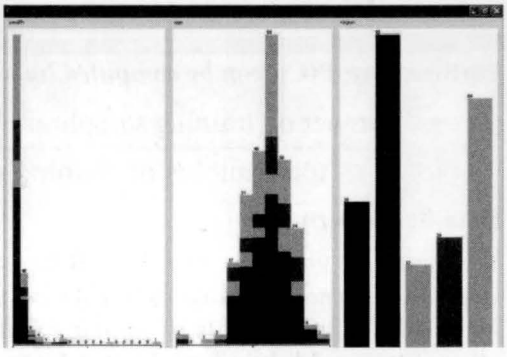

Figure 2 Data Visualization

Experimental Result [3]

=== Run information ===

Scheme: weka.classifiers.bayes.NaiveBayes

Relation: relation

Instances: 233

Attributes: 3

wealth

age

region

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

Naive Bayes Classifier

	Class				
Attribute	A	E	M	O	U
	(0.16)	(0.34)	(0.1)	(0.13)	(0.27)
wealth					
mean	2.6053	2.1844	4.2614	2.25	2.9648
std. dev.	2.1814	1.6268	7.4225	1.2612	3.5636
weight sum	38	80	22	29	64
precision	0.75	0.75	0.75	0.75	0.75
age					
mean	63.4685	63.9912	64.2424	64.0476	64.1935
std. dev.	9.9426	14.6725	19.1665	12.8207	11.8297
weight sum	37	76	22	28	62
precision	1.6667	1.6667	1.6667	1.6667	1.6667

Stratified cross-validation

Summary

Correctly Classified Instances	80	34.3348 %
Incorrectly Classified Instances	153	65.6652 %
Kappa statistic	0.0182	
Mean absolute error	0.301	
Root mean squared error	0.4043	
Relative absolute error	99.4407 %	
Root relative squared error	103.9815 %	
Total Number of Instances	233	

Confusion Matrix

a	b	c	d	e	<— classified as
0	35	0	0	3	a = A
0	75	1	0	4	b = E
0	19	0	0	3	c = M
0	27	0	0	2	d = O
1	53	5	0	5	e = U

Result Analysis

According to the classification result, Figure 3 shows the count classified according to region. The maximum count is available for European region and the lowest for Middle East.

Selected attribute

Name: region Type: Nominal
Missing: 0 (0%) Distinct: 5 Unique: 0 (0%)

No.	Label	Count
1	A	38
2	E	80
3	M	22
4	O	29
5	U	64

Figure 3. Region Classification

Also classification result shows that how mean wealth and age in particular regions are classified. The mean wealth in Middle East is the highest and in European countries is the lowest. But, the count of rich billionaires is the lowest in Middle East. The mean age is the highest in Middle East and the lowest in Asian countries. A low standard deviation indicates that the data points tend to be very close to the mean, whereas high standard deviation as in case of Middle East indicates that the data are spread out over a large range of values.

Conclusion

The result shows that which regions across the globe have majority of wealth so that they can be targeted as proper destinations to market initially the luxury product and to understand accurately the potential customers. Although challenging, but it is definitely possible to reach the desired consumer of the luxurious and prestigious store atmosphere of luxury brands through the Internet virtual environment and it is possible to sell luxury fashion goods online. We further aim to extend our approach and apply other data mining techniques and make the result comparison set.

References

1. *Fortune*, September 7,1992. "The Billionaires." pp. 98-138.
2. http://www.hakank.org/weka/DASL/Billionaires92.arff
3. WEKA Project - http://www.cs.waikato.ac.nz/~ml/
4. Michael Hahsler, *Knowledge Management, Data Warehouses and Data Mining*, http://wwwai.wuwien.ac.at/~hahsler/research/datawarehouse_webster2001/talk/
5. S. Wadhwa and Avneet Saxena, "Knowledge Management based Supply Chain: An Evolution Perspective," *Global Journal of e-Business and Knowledge Management* 2005, Vol. 2, No.2 , pp 13-29.

6. Charles Dennis, David Marsland and Tony Cockett, "Data Mining for Shopping Centres – Customer Knowledge Management Framework," *Journal of Knowledge Management* (2001) 5 (4): 368-374, 1367-3270.
7. Harry Zhang, *The Optimality of Naive Bayes,* Faculty of Computer Science, University of New Brunswick, Fredericton, New Brunswick, Canada http://www.cs.unb.ca/profs/hzhang/publications/FLAIRS04ZhangH.pdf
8. Kim Larsen, "Generalized Naive Bayes Classifiers," Concord, California, citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.109.178&rep...
9. Umeshwar Dayal, "Data Mining Meets E-Business: Opportunities and Challenges," http://dimacs.rutgers.edu/Workshops/MiningTutorial/dayal-slides.pdf
10. Suhail Ansari, Ron Kohavi, Llew Mason, and Zijian Zheng, "Integrating E-Commerce and Data Mining: Architecture and Challenges," WEBKDD'2000 workshop: Web Mining for E-Commerce — Challenges and Opportunities.

Roma Chauhan obtained her masters degree in computer science from Jamia Hamdard, Hamdard University. She is currently Lecture of Computer Science at Computer Science department, Institute of Management Education, Sahibabad, India. Prior to joining at IME she has worked with leading institutes and corporate giants. Her research focuses on business intelligence.

Ritu Chauhan is pursuing her Ph.D. in computer Science from Jamia Hamdard, Hamdard University. In the past she has worked with several academic institutes. During tenure of job she has handled several responsibilities like design of curriculum in data mining and analysis for several data mining techniques.