Original Research Paper

Computer Science

On-line Retail Business Mining for Effective Identification of Potential Customers in E-Commerce

D.Sridevi Assistant Professor,Department of Computer Applications,Valliammai Engineering College,Chennai

Dr.A.Pandurangan Professor& Head,Mathematics Department,Bharath University,Chennai

Dr.S.Gunaekaran Professor &Head, Applied Sciences,Gandhigram Rural University,Dindugal

Dr.A.Kumaravel Professor&Head, IT Department,Bharath University,Chennai

ABSTRACT We tackle the problem of identifying the potential customer for E-commerce through online retail business. The complexity embark upon of prediction becomes challenging especially when the customers are in the remote places, while launching any product in the appropriate market segments, the main issue is to cut cost while customer base is huge which cannot be mentioned easily. Hence in this paper, we recommend a set of data mining algorithms for analysing the pattern of purchases and arriving at optimal model for making recommendations for the possible potential identifiers. Likelihood is measured using Bayes scheme and applied to find the maximum probability attached with each customer's record. This will support the decision making on budget for customer relationship management.

KEYWORDS : E-Commerce, Data Mining, Classifiers, NaiveBayes, SMO, Lazy IB1, Accuracy, Lift curve.

1. INTRODUCTION

Retail and E-commerce are one of the first industries that recognized the benefits of using predictive analytics and started to employ it. In fact understanding of the customer is a first-priority goal for any retailer. In today's competitive business environment understanding of your customer requirement and offering the right products at right time is the key of any successful business. Due to high growth of internet, online shopping is becoming most interesting and popular activities for the consumers. Online shopping is providing a variety of products for consumers and is increasing the sales challenges for e-commerce players.[1]

The Web is one of the most revolutionary technologies that changed the business environment and has a dramatic impact on the future of electronic commerce (EC). The future of EC will accelerate the shift of the power toward the consumer, which will lead to fundamental changes in the way companies relate to their customers and compete with one another. Previous studies in Information Science (IS) literature like The Consumer Behavior towards online shopping of electronics in Pakistan (Adil Bashir 2013), Online Consumer Behaviour (Dr. Bas Donkers 2013), Influencing the online consumer's behavior: the Web experience (Efthymios Constantinides 2010),) Post-purchase behavior (Dibb et al., 2004; Jobber, 2010; Boyd et al., 2012; Kotler, 2011; Brassington and Pettitt, 2013) have proposed various models explaining customer buying behavior. These research models typically derive hypotheses from a literature review. Based on this hypothesis, evaluation of a multi-channel customer choice data can be done. Commerce networks involve buying and selling activities among individuals or organizations. [2]

Getting a deeper understanding of e-commerce networks, web data provides comparative advantages for mass merchants to analyze and reveal important parts of online consuming behavior [2]. Based on the analysis of the retailer's transaction data and a literature review, we derive hypotheses to explain consumer purchasing behavior.

2. BACKGROUND

The E-Commerce industry represents one of the largest industries worldwide. For example, in the United States, it is the second largest industry in terms of both the number of establishments and profits, with $3.8 trillion in sales annually. [3] In addition, this industry is facing similar trends to those affecting other sectors, for instance, the globalization of markets, aggressive competition, increasing cost pressures and the rise of customized demand with high product variants. Manual capture of sales information increases transaction costs and can cause inventory inaccuracies.

This kind of processing involves numerous human interventions at different levels such as order taking, data entry, processing of the order, invoicing and forwarding. The accuracy of the model is questionable and may not be consider few important factors while developing it. To overcome this problem, data mining can be used to analyze big data and develop efficient marketing strategies It is ideal because many of the ingredients required for successful data mining are easily satisfied: data records are plentiful, electronic collection provides reliable data, insight can easily be turned into action, and return on investment can be measured by identifying potential customers. [4].

3. CONSUMER BEHAVIOR IN E-COMMERCE

In the past few years, the development of the World Wide Web exceeded all expectations. Retrieving data has become a very difficult task taking into consideration the impressive variety of the Web. Web consists of several types of data such as text data, images, audio or video, structured records such as lists or tables and hyperlinks. Web content mining can be used to mine text, graphs and pictures from a Web page and apply data mining algorithms to generate patterns used for knowledge discovery [5].

For a successful e-commerce site, reducing user-perceived latency is the second most important quality after good site- navigation quality.

The most successful approach towards reducing user-perceived latency has been the extraction of path traversal patterns from past users buying history to predict future user buying behavior and to fetch the required resources. [6] Vallamkondu & Gruenwald (2003) describe an approach to predict user behavior in e-commerce sites. The core of their approach involves extracting knowledge from integrated data of purchase and path traversal patterns of past users to develop a pricing model which focuses on profits as well as customer satisfaction. [7]

Web sites are often used to establish a company's image, to promote and sell goods and to provide customer support. The success of a web site directly affects the success of the company in an electronic market.

4. DESCRIPTION OF DATA

In this paper, a detailed study based on data mining techniques was conducted in order to extract knowledge in a data set with information about user's history associated to an e-commerce website. These datasets are directly mined from UCI repository [8]. Using an online software, which converts html documents to data tables. The main purpose to web mine data is to apply a set of descriptive data mining techniques to induce rules that allow data analyst working at ecommerce companies make strategic decisions to boost their sales as well as provide effective customer service. The techniques used are NaiveBayes, SMO, Lazy.

4.1. Dataset Description

The online retail data set consisted of 65536 records.

Steps in Data Pre-Processing:

Data cleaning:
1. Removed blank spaces.
2. Removed Noisy data
3. Removed duplicates.

After data cleaning 39,027 records were identified as valid records.
1. Records were grouped based on stock code.
2. Records were grouped based on customer location.

How to find the frequency:
1. Grouped record based on customer code to identify the buying frequency.
2. By this we can identify the potential customers.
3. Stock code and customer code were the main attributes.

Classification:

Various Classification Techniques were applied. Based on the probability the above said classification techniques were applied. Before applying techniques, the record sets were classified.

1. The frequency obtained by taking customer code and location.
2. We classified the total record set into Very Low, Low, Medium, High and Very High.
3. We reduced record set by this classification and very high alone considered for further classification.
4. At the end we reduced the record set to 730 and applied High, Medium and Low.

Identified the more purchase frequency from the following location:

Table 1: High Purchase Frequency

S. NO	Name of the Attribute	Description
1	InvoiceNo	A 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.
2	StockCode	Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.
3	Quantity	The quantities of each product (item) per transaction. Numeric.
4	InvoiceDate	Invoice Date and time. Numeric, the day and time when each transaction was generated.
5	UnitPrice	Unit price. Numeric, Product price per unit in sterling.
6	Customer ID	Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.
7	Country	Nominal, the name of the country where each customer resides.

Table 2: Attributes in the Dataset

S.NO	Location	Frequency
1	Australia	102
2	France	112
3	Germany	98
4	Japan	90
5	Spain	118
6	U.K	210

5. APPLICATION OF VARIOUS ALGORITHMS FOR CLASSIFICATION

Naïve Bayes Algorithm

The Naive Bayes Classifier technique is based on Bayesian theorem and is particularly used when the dimensionality of the inputs is high. The Bayesian Classifier is capable of calculating the most possible output based on the input. It is also possible to add new raw data at runtime and have a better probabilistic classifier. A naive Bayes classifier considers that the presence (or absence) of a particular feature (attribute) of a class is unrelated to the presence (or absence) of any other feature when the class variable is given.

For example, a fruit may be considered to be an apple if it is red, round. Even if these features depend on each other or upon the existence of other features of a class, a naive Bayes classifier considers all of these properties to independently contribute to the probability that this fruit is an apple. Algorithm works as follows, Bayes theorem provides a way of calculating the posterior probability, $P(c|x)$, from $P(c)$, $P(x)$, and $P(x|c)$. Naive Bayes classifier considers that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors except FP rate.[9]

SMO Classifier

SMO Classifier – Sequential minimal optimization (SMO) is an algorithm for efficiently solving the optimization problem which arises during the training of support vector machines. It was invented by John Platt in 1998 at Microsoft Research. SMO [9] is widely used for training support vector machines and is implemented by the popular libsvm tool. The publication of the SMO algorithm in 1998 has generated a lot of excitement in the SVM community, as previously available methods for SVM training were much more complex and required expensive third-party QP solvers. SMO is an iterative algorithm for solving the optimization problem described above. SMO breaks this problem into a series of smallest possible sub-problems, which are then solved analytically. Because of the linear equality constraint involving the Lagrange multiplier, the smallest possible problem involves two such multipliers. The algorithm proceeds as follows: Find a Lagrange multiplier that violates the

Karush–Kuhn–Tucker (KKT) conditions for the optimization problem. Pick a second multiplier and optimize the pair.

Repeat steps 1 and 2 until convergence.

When all the Lagrange multipliers assure the KKT conditions, the problem has been solved. Although this algorithm is guaranteed to converge, heuristics are used to choose the pair off of multipliers so that it can accelerate the rate of convergence.

Lazy Learning

In machine learning, lazy learning is a learning method in which generalization beyond the training data is delayed until a query is made to the system, as opposed to in eager learning, where the system tries to generalize the training data before receiving queries.

The main advantage gained in employing a lazy learning method, such as case-based reasoning, is that the target function will be approximated locally, such as in the k-nearest neighbour algorithm. Because the target function is approximated locally for each query to the system, lazy learning systems can simultaneously solve multiple problems and deal successfully with changes in the problem domain. Lazy classifiers are most useful for large datasets with few attributes.[9]

6. DATA MINING TOOLS DESCRIPTION:

Today's various data mining tools that are available to handle or manage the large number of datasets and also to improve the quality of data, such tools are RapidMiner, Weka, R, scikit-learn, KNIME, Orange, KEEL, Tanagra etc. These data mining tools makes easy for analyst to get the knowledgeable information. Data mining tools are used to predict future trends, behaviours, allowing business to make proactive, knowledge driven decisions [10].

The various Data mining techniques and algorithms have been implemented on these tools to extract the information and also to check their efficiency and accuracy. In this paper, we are going to discuss and compare only three tools among of these that are; RapidMiner, WEKA, and KNIME which are using the same platform(Java).The description of these tools are as follows:

Simple Command line [11]. But Explorer is the main interface of WEKA. WEKA is Java based software and can run in different platforms. With the Java based version, the tool is so revolutionary and used in various application including visualization and algorithm for data analysis and predictive modeling [12].

It is freely available for download and offers many powerful features.

WEKA is Java based open source data mining tool.

It is easy to use for beginners and has the ability of running several learning algorithms and comparing.

Features:
1. It is platform independent.

2. It performs various data mining tasks including: 3.Data pre-processing, Classification rules, regression, Clustering, association rules, visualization, feature selection and improving the knowledge discovery.
4. WEKA has 49 Data pre-processing tools, 76 Classification/ regression algorithms, 8 Clustering algorithms, 3 algorithm for finding association rules, 15 attribute/subset evaluator plus 10 search algorithms for feature selection [13].
5. There are various built in features.
6. There is no programming and coding language required.

Advantages

- Easy to manipulate the data.
- Provide access to SQL databases.
- It provides two options for the user to interact through Explorer and Command line [14].
- Specially used for data mining.
- It provides various machine learning algorithms for data mining tasks.

It supports various standard Data mining tasks that include: Data pre-processing, Clustering and Classification, Regression, Visualization and Feature selection [15].

RapidMiner RapidMiner, previously YALE (Yet Another Learning Environment) was developed at the Technical University of Dortmund in 2001 by Ralf Klinkenberg, Ingo Mierswa and Simon Fischer.After, this software name was changed in 2007 from YALE to RapidMiner and is developed by the company RapidMiner, Germany. RapidMiner is an open source java based system for data mining and provides an integrated environment for machine learning, data mining, text mining ,predictive analysis and business analytics and is mainly used for business and industrial application[10].

RapidMiner is the most powerful, easy to use and intuitive Graphical User Interface for the design of analytic process, that contain several "operators".The operator functions as a single task in their process in which the input is produced by the existing output of the operator[16].

Features

It is platform independent.

It has compatibility with various databases like oracle, MySQL, Excel, SPSS, Microsoft SQL server etc.

It provides Drag and Drop interface to design the analytics process.

It supports and accepts new data drivers.

It provides more than 500 operators for all machine learning procedures, and also combines learning schemes and attributes evaluators of the WEKA learning environment [17].

It allow user to work with different sizes and types of data sources.

Advantages

- It has enormous flexibility.
- It provides the integration of maximum algorithm of such tools.
- Easy to debug the errors.

Disadvantages

Limited partitioning abilities for dataset to training and testing sets.

KNIME Konstanz Information Miner is an open source general data mining tool that is based on the Eclipse platform, developed and supported by KNIME.com.AG. In 2004, the KNIME initially developed by the team of software engineer at the University of Konstanz, Germany and in 2006, the initial version of KNIME was released [18].

KNIME is very powerful tool for analytical task, extracting data and knowledge from the web communities. The KNIME base version already incorporates hundreds of processing nodes for data I/O, pre-processing and cleansing, modeling, analysis and data mining as well as various interactive views, such as scatter plots, parallel coordinates and others[19].

In KNIME, representation of data sources and sinks, mining algorithm, transformations, visualizations, etc defined by set of nodes called "workflow" and each node has its specific input and output ports that depends on the functionality of the node [20].

For both simple and complex data types, KNIME allows revolutionary analysis to discover trends and predict future results. KNIME uses for teaching as well as research which allows to integrate the new algorithm and tools in a simpler manner.

Features

Available to everyone i.e., allow users to use the well- defined node API to add proprietary extensions.
Intuitive user interface.
It is easy to use and handle different functions.
KNIME modules cover a wide variety of functionalities like, I/O, data manipulation, views, hilting etc to better understand your data.
It provides the users to create data flows or pipeline visually, users can selectively execute some or all analysis steps, study the results, prototypes, and collaborative interpretations [14].

Advantages

- The major benefit of this is easy to use plug-in [21].
- It provides data flow process by dragging and dropping new nodes.

Disadvantages

Partitioning ability is limited for dataset [22].
Thus the Weka tool is selected for experiments in finding the potential customers.

7. PROPOSED ARCHITECTURE DIAGRAM F0R EFFICIENT MODEL

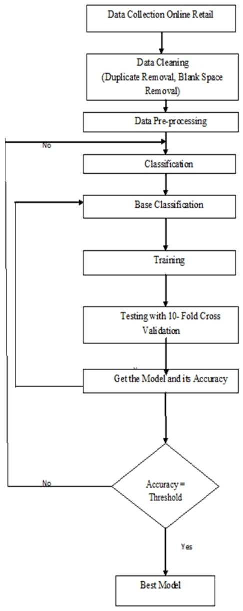

Data Collection Online Retail
Data Cleaning (Duplicate Removal, Blank Space Removal)
Data Pre-processing
No
Classification
Base Classification
Training
Testing with 10- Fold Cross Validation
Get the Model and its Accuracy
No
Accuracy= Threshold
Yes
Best Model

Figure 1: Proposed Architecture diagram for efficient model

SELECTION OF OPTIMAL CLASSIFIER BY LIFT CURVE

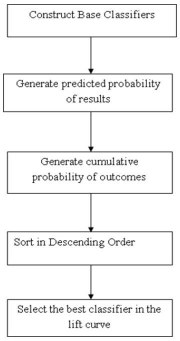

Construct Base Classifiers
Generate predicted probability of results
Generate cumulative probability of outcomes
Sort in Descending Order
Select the best classifier in the lift curve

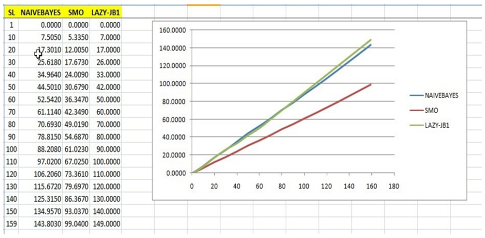
Figure 2: Selection of Optimal Classifier by Lift Curve

8. PREDICTIONS ON TEST DATA

The pre-processed data was uploaded in the Weka Tool for analyzing various classification techniques. Base Classifications like Naive Bayes, SMO and Lazy ID3 etc were applied and finally three above said classifiers are Number of Instances, actual, predicted, error and probability distribution are obtained when logged in Weka. Here the probability distribution was taken and the cumulative probability was calculated and thus the Lift Curve graph was drawn for all the classifiers and best one selected to find the potential customer based on frequency.

9.RESULT

From the lift curve obtained the best classifier in Base classifier is Naive Bayes classifier. This classifier is more powerful for identifying the potential customer.



SL	NAIVEBAYES	SMO	LAZY-JB1
1	0.0000	0.0000	0.0000
10	7.5050	5.3350	7.0000
20	17.3010	12.0050	17.0000
30	25.6180	17.6730	26.0000
40	34.9640	24.0090	33.0000
50	44.5010	30.6790	42.0000
60	52.5420	36.3470	50.0000
70	61.1140	42.3490	60.0000
80	70.6930	49.0190	70.0000
90	78.8150	54.6870	80.0000
100	88.2080	61.0230	90.0000
110	97.0200	67.0250	100.0000
120	106.2060	73.3610	110.0000
130	115.6720	79.6970	120.0000
140	125.3150	86.3670	130.0000
150	134.9570	93.0370	140.0000
159	143.8030	99.0400	149.0000

Figure 3: Lift Curve for Base Classifiers shows that the NaiveBayes is the best classifier.

FUTURE ENHANCEMENTS:

The results can be extended by tuning the parameters of the selected base classifiers.

Moreover, few attributes can be eliminated for feature reduction to enhance the accuracy of the base classifiers. More number of Base Classifiers can be added for designing efficient ensemble for E-Commerce Data Mining.

REFERENCES

[1] Know Your Buyer: A predictive approach to understand online buyers' behavior, By Sandip Pal Happiest Minds, Analytics Practice

[2] Quinlan J. R. (1986). "Induction of decision trees.Machine Learning," Vol.1-1, pp. 81-106.

[3] J. R. Quinlan, "C4.5: Programs for Machine Learning," Morgan Kaufmann Publishers, Inc., 1993.

[4] Ding Xiang-wu and Wang Bin, "An Improved Pre-pruning Algorithm Based on Id3," Jisuanji Yuxiandaihua,Vol.9, pp. 47,2008.

[5] Ming Fan, Xiaofeng Meng translated, "Data mining techniques and concepts", Machinary Industry Press, Beijing, pp. 136-145, Feb., 2004.

[6] N R Srinivasa Raghavan, "Data mining in e-commerce: A survey," Sadhana Vol. 30, Parts 2 & 3, April/June 2005, pp.275–289.

[7] B. Schafer, J.A. Konstan, and J. Reidl, "E-Commerce Recommendation Applications," Data Mining and Knowledge Discovery, Kluwer Academic, 2001, pp. 115-153.

[8] UCI Machine Learning Repository http://archive.ics.uci.edu/ml/datasets/online+retail

[9] S. Bhoomi Trivedi,Ms. Neha Kapadia,INDUS institute of Eng. & Tech,TCET, Kandivali(E),Ahmedabad Modified Stacked generalization withSequential Learning. TCET2012 on IJCA.

[10] K. Rangra, K.L. Bansal , Comparative Study of Data Mining Tools, International Journal of Advanced Research in Computer Science and Software Engineering, 4(6), June 2014.

[11] S. Srivastava, WEKA: A Tool for Data Preprocessing, Classification, Ensemble, Clustering and Association Rule mining, International Journal of Computer Applications, 88(10), February 2014.

[12] P.S. Patel, S.G. Desai , Comparative Study on data Mining tools, International Journal of Advanced Trends in Computer Science and Engineering, 4(2), April 2015.

[13] S.K. David, Amr T.M. Saeb, K.A. Rubeaan, Comparative Analysis of Data Mining Tools and Classification Techniques using WEKA in Medical Bioinformatics, Computer Engineering and Intelligent System, 4(13), 2013.

[14] K. Saravanapriya, A Study on Free Open Source Data Mining Tools, International Journal of Engineering and Computer Science, 3(12), December 2014.

[15] S. Singhal, M. Jena, A study on WEKA tool for Data Preprocessing, Classification and Clustering, International Journal of Innovative Technology and Exploring Engineering (IJITEE), 2(6), May 2013.

[16] S. Sarumathi, N. Shanthi, S. Vidhya, M. Sharmila , A Review: Comparative Study of Diverse Collection of Data Mining Tools, International Journal of Computer, Electrical, Automation, Control and Information Engineering, 8(6), 2014.

[17] M. Vijayakamal, M. Narendhar, A Novel Approach for WEKA & Study On Data Mining Tools, International Journal of Engineering and Innovative Technology (IJEIT), 2(2), August 2012.

[18] A. Jovic, K. Brikic, N. Bogunovic , An Overview of free software tools for general Data mining.

[19] L. Kataria, Implementation of Knime-Data Mining Tool, International Journal of Advanced Research in Computer Science and Software Engineering, 3(11), November 2013.

[20] S. Gunnemann, H. Kremer, R. Musiol, R.Haag, T. Seidl, A Subspace Clustering Extension For the KNIME Data Mining Framework, 2012 IEEE 12th International Conference on Data Mining Workshops.

[21] P. Subathra, R. Deepika, K. Yamini, P. Arunprasad, S.k Vasudevan, A Study of Open Source Data Mining Tools and its Applications, 10(10), 2015.

[22] H. Solanki, Comparative Study of Data Mining Tools and Analysis with Unified Data Mining Theory, International Journal of Computer Applications, 75(16), August 2013.