An Approach Based on Data Mining to Support Management in E-Commerce

Vitor V. de S. Campos, Carlos E. Bueno, Jacques D. Brancher, Fabio T. Matsunaga. Rafael, R. Negrao
State University of Londrina
Londrina – PR, Brazil
valerio@uel.br, carlosebueno@gmail.com, jacques@uel.br, ftakematsu@gmail.com, rafael@uel.br

ABSTRACT

The ability of managers to analyze large volumes of data is not enough to identify all relevant associations and necessary for the decision-making process. Make use of a classification model can generate information that typically a manager could not create without the utilization of this technology. The aim of this work is to reach a classification model, based on data from purchases made by customers through electronic media in an automated manner. Precisely this model presents a set of rules to assist in decision-making applicable to a sale of vehicles, parts and accessories. For the construction of this model, we applied a process of knowledge discovery in databases, in which classification techniques was evaluated in an experiment regarding accuracy, interpretability and learned model to the computing performance. Data mining has been used to find this classifier.

KEYWORDS

E-commerce, Classification Rule, Data Mining Decision Making

1 INTRODUCTION

E-commerce purchases have become a routine for many people due to several factors such as greater convenience in purchasing the product or service, the ease in comparative research, accessible at any time through smartphones or other devices with an internet connection, availability of the shop be open 24x7 days a week, among others. Companies usually carry out their sales through physical stores, are also including the sales system using virtual stores for the trade in their products or services. With this practice, companies seek to not only access to new markets for their goods or services but also the loyalty of their customers who tend to opt for online purchase [1].

Analyze the data from customers already obtained by the company, trying to extract some knowledge is an important task that can assist the process of improvement of the company's sales through e-commerce [2]. The ability of managers to analyze large volumes of such data is not sufficient to identify all relevant associations and necessary for the decision-making process. Making the use of machine learning algorithms, clustering and Association methods can generate information that typically a manager could not create without the use of such technologies [3],[4]. The company regularly stores data about the products or services marketed, and they show the relationship of their customers with the company over time. Through customer relationship data with the business in a given historical period, it is possible to extract significant rules to the decision-making process. Over time the data may change, requiring a new analysis to provide managers a new set of rules that assist in the decision-making process. Thus, it would be helpful to have a ranking model that automated data analysis and produced a set of rules based on the data available at the time of decision-making.

From the application of the classification model will be made available a set of rules that can be easily parseable by humans. For example, the Manager can use the knowledge obtained to find the database of potential customers to shop online store. It is also possible the Manager use this knowledge to allow marketing strategies are set to improve sales, among other possibilities, increasing customer retention or the search for new customers in different markets.

The present work aims to reach a classification model, based on data from purchases made by customers through electronic media in an automated manner. Specially this model gives a set of rules to assist in decision-making applicable to a sale of vehicles, parts and accessories. For the construction of this model, we applied a process of knowledge discovery in databases, in which classification techniques was evaluated in an experiment regarding accuracy, interpretability and learned model to the computing performance.

2 DATA MINING

The Knowledge Discovery in Databases (KDD), is the process of extracting useful knowledge from identifying patterns in data [4]. This process consists of three steps. The first is the preprocessing phase, which involves data cleaning tasks, such as: applying filters, selection and construction of filling missing values, attributes, processing of noises, among other tasks, so that the data can used for extraction of patterns. The second stage is the data mining, in which they extracted defaults of the data processed in the previous step. The last phase is post-processing, which addresses the issues of visualization of results and interpretation of standards.

In the extraction stage designs and patterns, can be used different methods and machine learning techniques [2],[4]. The standard problem of supervised machine learning algorithm input consists of a set of examples S, with N examples *Ti*, i = 1, ..., N, chosen from a domain *X* with a distribution *D* fixed, unknown and arbitrary, of the form $\{(x_1, y_1),..., (x_N, y_N)\}$ for some unknown function $y = f(x)$. The xi is typically fashion vectors $(xi_1, xi_2,..., xi_M)$ with discrete values or numeric. xj refers attribute value *j*, named *Xj*, the example *Ti*. yi values refer to the value of attribute *Y*, often termed class. The y-values in classification problems, as is the case in this work, are typically owned by a discrete set of classes *C*v, v = 1,..., N_{Ci}, i. and y ϵ {C1,..., C_{NCl}}[5].

In this paper, we use four algorithms: two based on rules (JRip and PART) and two decision-tree-based (J48 and REPTree) [3]. Both algorithms offer as output rulesets easily interpretable by humans. The goal of four algorithms is to induce a classifier consisting of decision rules. In this work, we represent a rule *R* built as $R = B \rightarrow H$, where *B* is the body or the rule condition, and *H* is the head of the rule. In a classification rule, the body is a conjunction of attribute tests fashion *Xi op Value*, where *Xi* is the name of an attribute, *op* is an operator belonging to the set $\{=, \neq, <, \leq, >, \geq\}$ and *Value* is a valid value for the attribute *Xi*. *H* takes the form class = *Ci*, where class is the attribute that should be predicted from the domain (class attribute), and $Ci \in C_v$. [5].

In the evaluation phase of the models, the assessment can be quantitative, involving domain experts explored, or qualitative, which depends on the techniques and machine learning methods used. In this paper, we evaluate the quality of rules built. Given a rule $R = B \rightarrow H$ and a data set $S = \{(x_1, y_1),..., (x_N, y_N)\}$, if the rule is a rule of decision, one of the measures used to evaluate the rule's coverage. The coverage of a rule was defined as follows: the examples that satisfy the rule, i.e. whose values present in xi satisfy the conditions in *B*, are covered by *R*; examples that satisfy *B* and *H*, i.e. the values *yi* are equal to class in *H*, are properly covered by *R*; examples that satisfy *B* but not *H* are incorrectly covered by rule; and examples that do not meet *B* are not covered by *R*.

3 CASE STUDY: CAR DEALERSHIP

In this article, we argue that a ranking model that manages in an automated manner a set of rules that help the Manager in the decision-making process, can be used for the task of identifying customers who has a greater tendency to buy via e-commerce. To evaluate the proposal was implemented a case study using a real-world scenario, described below.

This is a dealership of vehicles, parts and accessories that has four systems that support the daily sales activity, managing the relationship with customers and decision-making by managers. The company's computer systems are not interconnected among themselves. So, has the Enterprise Resource Planning (ERP) which brings together all the transactional processes. The Customer Relationship Management (CRM) that stores customer information. The WEB system (e-commerce site), which keeps the information from the sales data from the site, being that sales take effect are then inserted in the ERP. Finally, there is the Data Warehouse (DW) that stores the data from other systems used in the company.

Our focus is to analyze the sales data performed by the WEB system and for this collected from transactions in the years 2013 and 2014 customer data.

First, we analyze which portion of the data would be considered useful to examine the extraction of standards regarding the purchase of products through the e-commerce site. It was identified that some tables and various fields possess unnecessary information and irrelevant to the work, being discarded in the process of selection of attributes. The attributes that contains data about accessories, parts, tires, customers (individual or company), financial information (credit card, bank transfer), place of purchase (identifies if the client that performed the purchase is in the same State of vehicle dealership or not) and the last attribute indicates whether was effected the purchase, have been identified as relevant to the process of obtaining the ranking model. The data associated with these attributes are periodically stored in DW, which allows extracting customer information from analyzes of rules obtained by sorting algorithms. All attribute values were transformed to discrete values zero (0) and one (1) to be used by the classification algorithms

Once selected and transformed the values of attributes, was raised a file in the format accepted by the WEKA [6], tool used in the experiment, containing the eight attributes.

3.1 Data mining algorithms

The parameters used for each of the algorithms based on rules with their default values are described in table 1 and 2.

Table 1. Parameters of the rule-based algorithms used in the experiment

JRip	PART

checkErrorRate= True
debug = False
folds = 3
minNo = 2.0
optmizations = 2
seed = 1
usePruning = True
	binarySplits = False
	confidenceFactor = 0.25
	debug = False
	minNumObj = 2
	numFolds = 3
	reducedErrorPruning = False
	seed = 1
	unpruned = False

Table 2. Parameters of tree-based algorithms used in the experiment

J48	REPTree

binarySplits = False
confidenceFactor = 0.25
debug = False
minNumObj = 2
numFolds = 3
reducedErrorPruning = False
saveInstanceData = False
seed = 1
subtreeRaising = True
unpruned = False
useLaplace = False
	Debug = False
	MaxDepth = -1
	minNum = 2.0
	minVarianceProp = 0.001
	noPruning = False
	numFolds = 3
	seed = 1

In the experiment, we used the validation method 10-fold cross-validation to split the database in training and testing. In this method, the data is divided into a number k of folds. At each iteration, a fold is presented with test data while the remaining k-1 are used as training data. This procedure runs k times, so that in each iteration one of the folds can act as a test. The performance of the classifiers is calculated as the average obtained in k iterations [6]. In the experiment, we used k = 10.

The results obtained were evaluated for accuracy (that indicates the percentage of correctly classified instances), precision (which indicates the percentage of defaults that were correctly classified in a category) and coverage (which shows the percentage of defaults that have been recovered). In addition, were observed the speed of construction of the model (representing the computational cost of learning) and interpretability (clarity and ease of interpretation of the model learned by end-users).

3.2 Results Analysis

The results obtained for the rule-based algorithms are presented in Table 3, the accuracy was around 68%, with little variation among the algorithms used. Rule-based algorithms had a higher proportion of hits, with emphasis on the PART. With worse accuracy was the REPTree algorithm, how values are approximated by them, gives us an indication that any of the evaluated algorithms could be used in practice. Regarding speed, the superior result was J48 virtually instant.

Table 3. Results of the algorithms used in the experiment

Algorithm	Accuracy	Speed (seconds)	Rules
JRip	68,72%	0,1	3
PART	69,09%	0,32	4
J48	67,63%	0,0	11
REPTree	67,27%	0,08	11

On the question of interpretability, the decision tree based algorithms have the advantage of expressing the graphic model or verbatim, by inducing decision trees or only for conversion rules. The analysis of the results of the experiment suggests that the difference between the number of rules generated by rule algorithms and decision tree did not make much of a difference in accuracy of the generated model. This indicates that could lead to the adoption of the model with fewer rules without significant losses.

Considering that, in all the iterations, the values obtained in the comparison of algorithms were very close but not too high, still represents a good outcome, and that the runtime was little for all algorithms, all could be considered for the classification.

4 RELATED WORKS

Data mining techniques are widely used in e-commerce applications due to the importance of commercial information available which may involve historical data, information stored in Data Warehouse (DW) and data are available on the web. All this information is stored in large databases. For example, Hu and Li [8] applied to mining web at tourism website to help e-commerce customer relationship management and marketing network.

Lemos [9] discusses the concepts of data mining emphasizing the use of methods of Artificial Neural Networks and decision trees based on tools WEKA and MATLAB. These methods have been used to assist decision granting of bank credit to new clients based on historical data previously acquired in the banking institutions.

De Jesus et al. [10] sorting and grouping techniques applied to improve the suggestions and recommendations to users based on profile library and history of loans of books.

Alvares and Silva [11] used the geolocation information of a social network to demonstrate the existence of clusters that relates a given region with matters of common concern and the texts published in the area.

Feng et al. [12], in turn, used the ID3 algorithm, to deal with the high volume of customer data and reduce the computational cost, and based on the decision tree generated improve efficiency in the decision-making process in e-commerce.

5 CONCLUSIONS

The present work presents an approach to the automatic construction of a classifier to determine a set of rules that can be used to assist in the decision-making process by the managers of a company. The approach followed the KDD process and was experienced in a real scenario, where the results were analyzed with respect to the accuracy, interpretability of the learned model and computing performance.

The experiment was conducted using the database of the concessionaire of vehicles, parts and accessories and evaluated the performance of four learning algorithms for classifiers. Two of them based on rules and two of them based on the decision tree. In terms of accuracy, all the algorithms presented percentages at 68%, which indicates that all could be considered in the classification of rules for decision-making on the part of managers.

REFERENCES

[1] X. Zhang and J. Zhang, “CRM applications in e-commerce strategy,” in Computational and Information Sciences (ICCIS), 2013 Fifth International Conference on, June 2013, pp. 605–608.

[2] T. Mehenni and A. Moussaoui, “Data mining from multiple heterogeneous relational databases using decision tree classification,” Pattern Recognition Letters, vol. 33, no. 13, pp. 1768–1775, Oct. 2012.

[3] D.-Q. Li, H. Yang, and L.-L. Li, “Research and application of data mining technique in e-commerce,” in E-Business and E-Government (ICEE), 2010 International Conference on, May 2010, pp. 4295–4298.

[4] U. Fayyad, G. Piatetsky-shapiro, and P. Smyth, “From Data Mining to Knowledge Discovery in Databases,” AI Magazine, vol. 17, no. 3, pp. 37–54, 1996.

[5] F. C. Bernardini,. Combinação de classificadores simbólicos utilizando medidas de regras de conhecimento e algoritmos genéticos [tese]. São Carlos: , Instituto de Ciências Matemáticas e de Computação; 2006 [acesso 2015-02-25]. Avaliable in: http://www.teses.usp.br/teses/disponiveis/55/55134/tde-29092006-110806/.

[6] M. Hall, E. Frank, and G. Holmes, “The WEKA data mining software: an update,” ACM SIGKDD Explorations, vol. 11, no. 1, pp. 10–18, 2009.

[7] O. Depren, M. Topallar, E. Anarim, M. K. Ciliz, 2005. An intelligent intrusion detection system (IDS) for anomaly and misuse detection in computer networks. Expert Systems with Applications 29, 713–722.

[8] J.-f. Hu and B.-s. Li, “Research on the application of web data mining technology in tourism e-commerce,” in World Automation Congress (WAC), 2012, June 2012, pp. 1–4.

[9] E. P. Lemos, M. T. A. Steiner, and J. C. Nievola, “Análise de crédito bancário por meio de redes neurais e árvores de decisão: uma aplicação simples de data mining,” Revista de Administração, vol. 40, no. 3, pp. 225–234, 2005.

[10] A. P. Jesus, “Personalização de sistemas Web utilizando data mining: um estudo de caso aplicado na biblioteca central da FURB,” 2008. [Online]. Available: http://www.inf.furb.br/seminco/2004/artigos/104-vf.pdf

[11] L. O. Alvares and R. J. Silva, “Análise espaço-temporal de mensagens do Twitter,” 2013. [Online]. Available: http://www.lbd.dcc.ufmg.br/colecoes/erbd/2013/005.pdf

[12] F. Yang, H. Jin, and H. Qi, “Study on the application of data mining for customer groups based on the modified id3 algorithm in the ecommerce,” in Computer Science and Information Processing (CSIP), 2012 International Conference on, Aug 2012, pp. 615–619.