Advances in Economics and Business 3(8): 329-336, 2015
DOI: 10.13189/aeb.2015.030805

Knowledge Discovery Using an Integration of Clustering and Classification to Support Decision-making in E-commerce

Vitor Campos*, Carlos Bueno, Jacques Brancher, Fabio Matsunaga, Rafael Negrao

Department of Computer, State University of Londrina, Brazil

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 and clustering 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 linking them to clusters, 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 and clustering techniques was evaluated in an experiment regarding accuracy, interpretability and learned a model of the computing performance. Data mining has been used to find this classifier.

Keywords E-commerce, Decision Making, Data Mining, J48, JRip, REPTree. PART, Classification, Cluster

1. Introduction

Purchase products and services using the internet/e-commerce 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 [17].

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 [11]. 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 [2,3]. The company constantly stores data about the products or services marketed. These historical data show which product the customer bought and in what quantity, the frequency of purchases during the year as well as the form of payment used for other data. For example, the client may have used the bank transfer or the credit card.

These data 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 as well as groupings of clients that are associated with the rules found by classifiers.

Over time the data may change, thereby changing the rules and the groupings of customers related to the rules found. Thus, it is necessary a new analysis to provide managers a new set of rules and a new profile of clients associated with the rules that help 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 and clusters based on the data available at the time of decision-making [12].

As the application of the model of classification and clustering will be provided, a set of rules and cluster associated with those rules may easily be parsed 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, create new products based on the information presented on other possibilities, increasing customer retention or the search for new customers in different markets.

The present work aims to reach a classification model linking them to clusters, based on data from purchases made by customers through electronic media in an automated manner. Specifically this model presents a set of rules and the list of rules with the client group to assist in the decision-making process 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 and clustering techniques was evaluated in an experiment regarding accuracy, interpretability and learned a model of the computing performance.

2. Data Mining

The Knowledge Discovery in Databases (KDD), is the process of extracting useful knowledge from identifying patterns in data [3]. 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, which can be supervised and non-supervised [3],[11].

2.1. Supervised Machine Learning

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 $\{(x1, y1),..., (x_N, y_N)\}$ for some unknown function y = f(x). The xi is typically fashion vectors $(xi1, xi2,..., xi_M)$ 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 Cv, v = 1,..., NCi, i. and $y \in \{C1,..., C_{NCI}\}$ [1].

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 $Ci \in Cv$. [1].

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 $S = \{(x1, y1),..., (x_N, y_N)\}$, 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.

In this paper, we use four algorithms: two based on rules (JRip and PART) and two decision-tree-based (J48 and REPTree) [2]. Both algorithms offer as output rulesets easily interpretable by humans. The goal of four algorithms is to induce a classifier consisting of decision rules.

These algorithms will be evaluated in relation to accuracy (that indicates the percentage of correctly classified instances), precision (which indicates the percentage of defaults, which were correctly classified in a category) and coverage (which shows the percentage of defaults that have been recovered).

2.2. Non-supervised Machine Learning

The non-supervised Data mining does not have a defined goal. She usually works through heuristics that assist in finding information in the database to respond to relevant questions, which have not been formulated or presented previously. One of his techniques is known as clustering [7].

Clustering is a data mining technique non-supervised that aims to discover patterns in data objects, generating a database partition containing a cluster of objects with common characteristics among themselves. The analysis of the clusters formed in respect of the implementation of the algorithm should be made by an expert in the area where knowledge is being applied the technique.

Can explain in general terms that a clustering algorithm seeks to minimize the average squared distance between points in a same group (cluster). The use of the algorithm normally is based on passing the parameter "k" representing the estimated number of clusters. When starting the clustering process, the algorithm inserts random values for the centroids, for which the objective is to reduce the distance to objects in each iteration.

The object closest to each centroid will be identified to the group to which it belongs. The method terminates when no centroid is modified. The most widely used calculation to set the distance between the points is the Euclidean distance that defines that distance between two points. It is calculated by the square of the differences between the coordinates of the root.

For the analysis of Clustering algorithm will be used EM

[15]. An algorithm of expectation-maximization (EM) is an iterative method to find a maximum likelihood of parameters in statistical models, where the model depends on latent variables not observed. Iteration EM alternates between performing an expectation step (E), which calculates the reason of verisimilitude by using the current estimate of the parameters and the maximization (M), which calculates the parameters by maximizing the likelihood ratio found in step E. These parameter estimates are used to determine the distribution of the latent variables in the next step and in can decide how many clusters create by cross-validation or can specify a priori how many clusters generate. As cross-validation is used to determine the number of clusters applies the following steps:

a. the number of clusters is set to 1.
b. the training set is divided randomly into ten folds.
c. EM is performed ten times using the ten folds the usual CV way.
d. the log likelihood is averaged over all ten results.
e. If log likelihood has increased the number of clusters is increased by 1 and the program continues to step b.

The result of cluster analysis is written to a table named class indices. The values in this table indicate the class indices, where a value of ' 0 ' refers to the first cluster; a value of ' 1 ' refers to the second cluster, until all clusters are described.

The choice of these classifying and clustering algorithms because allow the use of cross-validation method validation to split the database in training and testing. Other algorithms such as SimpleKMeans, HierarchicalClusterer and Filtered Cluster that offered no such option.

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 dealership of vehicles, parts and accessories that has four systems that support the daily sales activities, a system of managing the relationship with customers and a management system for 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 and bank billet), 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 it 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 and clustering algorithms. The complete description of the attributes is presented in Table 1.

Table 1. Complete description of variables

Variable	Category	Values
State	Nominal	0-1
Parts	Nominal	0-1
Tires	Nominal	0-1
Accessories	Nominal	0-1
Customers	Nominal	0-1
Credit card	Nominal	0-1
Billet	Nominal	0-1
Bought	Nominal	0-1

Once selected and transformed the values of attributes, was raised a file in the format accepted by the WEKA [5], tool used in the experiment, containing the eight attributes.

3.1. Parameters Used in Data Mining Algorithms

The parameters used for each of the algorithms based on rules and based on trees with their default values are shown in Tables 2 and 3 respectively.

Table 2. 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 3. 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

The parameters used for the Simple EM (Expectation Maximisation) algorithm with their default values are shown in Table 4.

Table 4. Parameters of Simple EM algorithm used in the experiment

Simple EM (Expectation Maximisation)

debug = False
displayModelInOldFormat = False
maxInerations = 100
minStdDev = 1.0E-6
numClusters = -1
seed = 100

In the experiment, it was used for both sorting algorithms as grouping 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 and the clustering algorithm is calculated as the average obtained in k iterations [6]. In the experiment, we used k = 10.

The data obtained from running the experiment for classification algorithms will be evaluated as to the accuracy, precision and coverage. 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. Performance Measures for Data Mining Algorithms

To measure performance the concept of sensitivity is often used for evaluation of classifiers. This concept is easily usable for the assessment of any binary classifier. TP is true positive, FP is false positive, TN is true negative and FN false negative. The True Positive rate is the same as sensitivity

a) True Positive Rate: it is simply the ratio of true positives to true positives plus false negatives, it is equivalent to Recall. It can be defined as:

$$\text{Sensibility} = \text{TPR} = \frac{\text{TP}}{\text{TP}+\text{FN}} \qquad (1)$$

b) False Positive Rate: it is simply the ratio of false positives to false positives plus true negatives. It can be defined as:

$$\text{FPR} = \frac{\text{FP}}{\text{FP}+\text{TN}} \qquad (2)$$

c) Precision: The information retrieved from the positive predictive is called precision. It is calculated as the number of instances correctly classified belongs to X divided by the number of cases classified as belonging to class X; that is, the proportion of true positives out all positive results. Can be defined as:

$$\text{precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \qquad (3)$$

d) Accuracy: Accuracy is simply a ratio of (((number of instances sorted)/(total number of instances)) * 100). Technically, it can be defined as:

$$\text{accuracy} = \frac{\text{TP}+\text{TN}}{((\text{TP}+\text{FP}) + +(\text{FN}+\text{TN}))} \qquad (4)$$

e) F-Measure: F-measure is a way to combine the scores of recall and precision in a single measure of performance. The formula for this is

$$\text{F-Measure} = \frac{2{*}\text{recall}{*}\text{precision}}{\text{recall} + \text{precision}} \qquad (5)$$

f) The Confusion Matrix: Table 5 the confusion matrix showed the number of correct classifications as opposed to the predicted classifications for the class.

Table 5. Confusion Matrix

Class	Predicted C_+	Predicted C_-
True C_+	TP	FN
True C_-	FP	TN

To measure the accuracy of the algorithms of classification and clustering algorithm, will be calculated: the number of instances correctly classified belongs to rule X, plus the number of cases in the cluster associated with the rule X, divided by the total number of instances classified as belonging to the rule X, plus the total number of cases of the cluster. It can be defined as:

$$\text{Rule Precision} = \frac{\text{TPR} + \text{IC}}{((\text{TPR} + \text{FPR}) + \text{TIC})} \qquad (6)$$

TPR = True positive ratio
FPR = False positive ratio
IC = Instances in the Cluster
TIC = Total number of Instances in the Cluster

Since the data obtained from running the experiment for the clustering algorithm will be evaluated in association with the data obtained from the algorithms of classification according to the formula 6.

3.3. Application of Data Mining Algorithms

The application proposed here is the association between classification algorithms and clustering algorithm using a precision measurement, defined by the formula 6, which aims to measure the accuracy of classification rules generated by REPTree, JRip, PART, and J48 algorithms according to the sequence of steps described below:

a) Set the cutoff point for the values obtained under same are disposed.
b) Perform the clustering algorithm to obtain the number of clusters.
c) Run the classification algorithm to obtain the rule set.
d) For each rule obtained by implementing the classification algorithm find the Positive True Ratio (TPR) and the False Positive Ratio (FPR) and find the cluster that is associated with the rule.
e) Apply the rule of precision according to the formula 6
f) Repeat steps d-e until you have more rules to evaluate.
g) Eliminate the values that are below the cut-off point and insert in the table above the cutoff point.
h) Repeat steps c-g until you have better algorithms to evaluate

Kumar, & Rathee [9] do a comparative analysis of J48 classification algorithm implementation, with the use of the original data, and the implementation of J48 with information to which the cluster belongs instance by adding clustered attribute in the original data obtained by running the SimpleKmeans algorithm. This work differs from the work presented by Kumar, & Rathee [9] for making an association between classification and clustering algorithms and from this find Association rules that are relevant for taking a decision excluding the rules below certain cut-off

In succinct form Kumar, & Rathee [9] make use of a large database 'Fisher's Iris Dataset' containing 5 attributes and 150 instances to perform an integration of clustering and classification techniques of data mining. They compared results of simple classification technique (using J48 classifier) with the results of integration of clustering and classification technique, based on various parameters using WEKA (Waikato Environment for Knowledge Analysis), a Data Mining tool. The results of the experiment show that the integration of clustering and classification gives promising results with utmost accuracy rate and robustness even when the data set is containing missing values.

3.2. Results Analysis

The results obtained from running the experiment for the rule-based algorithms are presented in Table 6, 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 6. Confusion Matrix

Parameters	JRip	PART	J48	REPTree
Recall/TPR	0.687	0.691	0.676	0.673
FPR	0.394	0.377	0.424	0.486
Precision	0.683	0.691	0.676	0.651
F-Measure	0.685	0.691	0.670	0649
Velocity (s)	0.1	0.32	0.0	0.08

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 results of this experiment show the technique of classification through the measure of accuracy based on true positive, false positive, true negative, false negative as well as the error rate and the number of rules that are generated by each of the algorithms, are shown in Table 7.

Table 7. Performance Evaluation

Parameters	JRip	PART	J48	REP Tree
TP	49	52	44	32
FP	46	43	51	63
TN	140	138	142	153
FN	40	42	38	27
Leaves in tree/rules	3	4	6	6
Size of tree	-	-	11	11
Error Rate	0.404	0.362	0.389	0.405
Accuracy	0.687	0.691	0.676	0.673

The figures 1 and 2 show respectively the tree structure of the REPTree and the rules of model. It is also illustrated in figure 3 the number of selected cluster after run the grouping algorithm. The total construction time of the model was of 5.56 seconds. For the determination of the precision of the rule is used detailed data generated by clustering algorithm where are the number of instances for the value 0, the number of instances for the value 1 and the sum total, as shown in Figure 4.

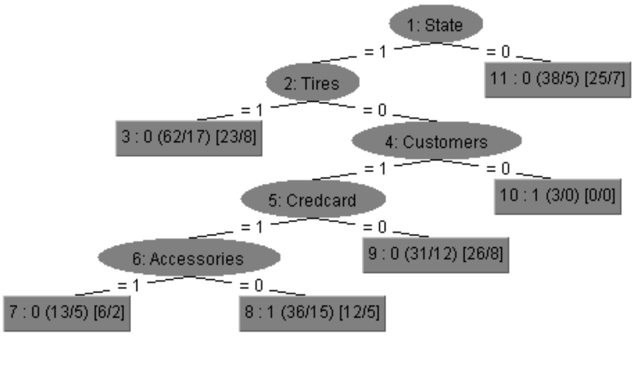
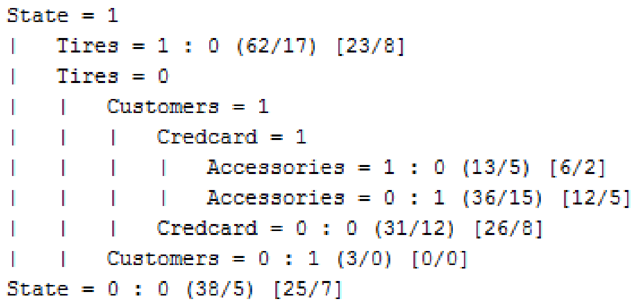
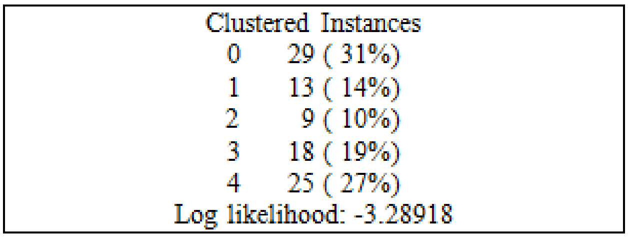


Figure 1. Decision tree for classification of vehicle dealership data.



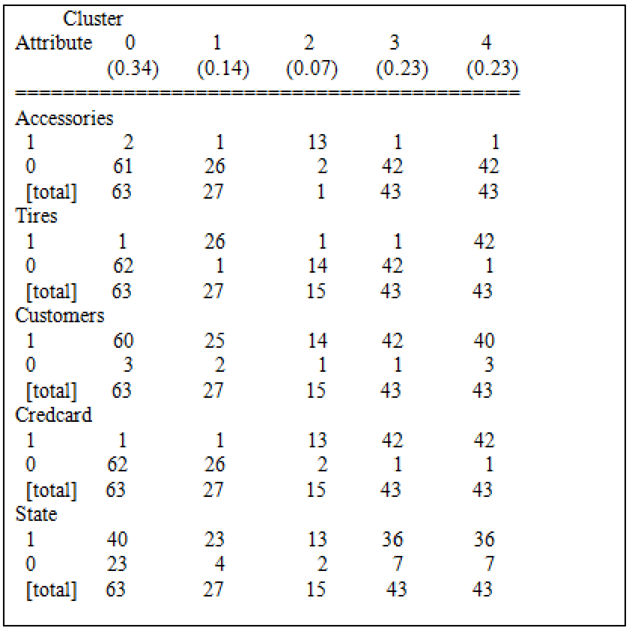
```
State = 1
|   Tires = 1 : 0 (62/17) [23/8]
|   Tires = 0
|   |   Customers = 1
|   |   |   Credcard = 1
|   |   |   |   Accessories = 1 : 0 (13/5) [6/2]
|   |   |   |   Accessories = 0 : 1 (36/15) [12/5]
|   |   |   Credcard = 0 : 0 (31/12) [26/8]
|   |   Customers = 0 : 1 (3/0) [0/0]
State = 0 : 0 (38/5) [25/7]
```

Figure 2. Rules of classification of vehicle dealership data.



```
   Clustered Instances
0      29 ( 31%)
1      13 ( 14%)
2       9 ( 10%)
3      18 ( 19%)
4      25 ( 27%)
Log likelihood: -3.28918
```

Figure 3. Clustered instances of vehicle dealership data.



```
       Cluster
Attribute    0        1        2        3        4
          (0.34)   (0.14)   (0.07)   (0.23)   (0.23)
==================================================
Accessories
 1           2        1       13        1        1
 0          61       26        2       42       42
 [total]    63       27        1       43       43
Tires
 1           1       26        1        1       42
 0          62        1       14       42        1
 [total]    63       27       15       43       43
Customers
 1          60       25       14       42       40
 0           3        2        1        1        3
 [total]    63       27       15       43       43
Credcard
 1           1        1       13       42       42
 0          62       26        2        1        1
 [total]    63       27       15       43       43
State
 1          40       23       13       36       36
 0          23        4        2        7        7
 [total]    63       27       15       43       43
```

Figure 4. Detailed data of the cluster generated by clustering technique

Analyzing the values contained in the figure of rules and values contained in Figure accuracy cluster for the first rule for the REPTree classifier is shown below:

$$\text{Rule Precision} = \frac{(62+23)+42}{((85+25)+43)} = 0{,}830$$

This precision measurement is performed for each of the rules of the REPTree algorithm, JRip, PART, and J48. To check the rules that appear used the cutoff point of 75 percent, dropping below the values the same. Therefore, the rules associated with the tires, accessories, cred card, parts and Billet shown in table 8

Table 8. Performance measures of precision rule

Rule Precision	JRip	PART	J48	REPTree
Rule 1 Tires$_{+}$	-	-	0.913	0,830
Rule 2 Tires$_{-}$	-	-	0.813	-
Rule 3 Acces-sories$_{+}$	-	-	-	0.780
Rule 4 Acces-sories$_{-}$	-	-	-	0.832
Rule 5 Credcard	-	0.980	-	0.850
Rule 6 Client	-	-	-	0.090
Rule 8 Parts$_{+}$	0.820	0.803	0.821	-
Rule 10 Billet	0.831	0,864	0.864	-

In this experiment was presented the results of the Association of 4 classification algorithms and a collation algorithm applied to a dataset for a dealership vehicles containing 275 instances with 8 selected attributes. During the implementation of the experiment was run first the 4 sorting algorithms, then the collation algorithm and lastly was the rule of accuracy to indicate the set of relevant rules and their association with the cluster encountered by clustering algorithm.

The results of this experiment show that the integration of the technique of classification and grouping technique shows an accuracy greater than simply apply sorting algorithms as x and y tables show. In this experiment was measured the accuracy of the rule based on true Positive Rule, false positives from this rule Instances in the Cluster and Total Instances in the Cluster.

Observing and analyzing the data presented in table 8 and in figures 3 and 4 we, for example, could extract the information presented below and who would be assisting in the decision-making process:

- Customers Seek spare parts for their vehicles and are always individuals and in most cases opt for payment by bank transfer. Its location is always of a State other than as is the physical store.;
- Customers seeking a little of each group of the item sold on the site, but a large part is interested in buying tires. The majority of these clients are individuals and prefer to pay with a credit card. More than 90% of them are from a State other than the State of the physical store. The group that performs more purchases is not the majority.
- Customers with an interest in parts and tires, this is the group that performs no purchase. Usually are interested in the purchase of parts, are individuals and prefer payment by bank transfer. Temse-groups with a small amount of customers who typically do not conclude the purchase with the additional information that the State itself are all from the physical store. This analysis by lead us to the conclusion that State's customers should be the focus of attention to increasing sales in the State.

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, [6] to mining web at tourism website to help e-commerce customer relationship management and marketing network.

Lemos, Steiner, & Nievola [10] 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.

Jesus [7] 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.

Yang, Jin, & Qi, [16], 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.

Guidini, Nascimento, Bone, & Alves [4] showed a way of classifying the management styles of organizations using grouping techniques business through the K-Means cluster. He was merely addressing the styles: authoritative, benevolent, consultative and participatory. Conducted a survey on a question, with closed questions answered by 111 controllers of companies of 1.000 Value magazine between the years 2000 to 2004.

Shaw, Subramaniam, Tan, & Welge, [13] argues that with the proliferation of information systems and technology, businesses increasingly have the capability to accumulate huge amounts of customer data in large databases. However, much of the useful marketing insights into customer characteristics and their purchase patterns are largely hidden and untapped. The current emphasis on customer relationship management makes the marketing function an ideal application area to greatly benefit from the use of data mining tools for decision support. A systematic methodology that uses data mining and knowledge management techniques is proposed to manage the marketing knowledge and support marketing decisions.

5. Conclusions

The present work presents an approach to the automatic construction of a set of rules using the database of vehicles, parts and accessories, where accuracy of all classification algorithms presented percentages at 68%, and with the association with the collation algorithm and applying the formula 6, which features the precision of rule, with a cut-off of 75% the average percentage of accuracy was around 89%. Thus, the results presented in in table 8 and in figures 3 and 4 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.

REFERENCES

[1] Bernardini FC. Combining symbolic classifiers using measures of knowledge rules and genetic algorithms [Thesis] São Paulo (SP): University of São Paulo; 2006. (in Portuguese)

[2] Dao-Quan LI, Hua YANG, Li-Li LI. Research and Application of Data Mining Technique in E-commerce. ICEE 2010. Proceedings of the International Conference on E-Business and E-Government; 2010 May 7-9; Guangzhou: IEEE; 2010. p. 4295-8.

[3] Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. AI magazine. 1996; 17(3): 37.

[4] Guidini MB, Nascimento AM, Bone RB, Alves TW. Application of k-means clustering to sort managerial styles. Contextus. 2008; 6(2). (in portuguese)

[5] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. SIGKDD explorations. 2009; 11(1): 10-8.

[6] Hu JF, Li BS. Research on the application of web data mining technology in tourism E-Commerce. WAC 2012: Proceedings of the 2012 World Automation Congress. 2012 Jun 24-28; Puerto Vallarta: IEEE; 2012. p. 1-4.

[7] Jesus, A. Customization of Web systems using data mining: a case study applied at the Central Library of FURB. (in Portuguese) Online available from: http://www.inf.furb.br/seminco/2004/ artigos/104-vf.pdf

[8] Kaufman L, Rousseeuw PJ. Finding groups in data: an introduction to cluster analysis. New Jersey: John Wiley & Sons, Inc. (2009) 368p.

[9] Kumar V, Rathee N. Knowledge discovery from database Using an integration of clustering and classification. International Journal of Advanced Computer Science and Applications (IJACSA). 2011; 2(3): 29-33.

[10] Lemos EP, Steiner MTA, Nievola JC. Bank credit analysis through neural networks and decision trees: a simple application of data mining. R. Adm. 2005; 40(3): 225-34. (in Portuguese)

[11] Mehenni T, Moussaoui A. Data mining from multiple heterogeneous relational databases using decision tree classification. Pattern Recognition Letters. 2012; 33(13): 1768-75.

[12] Rao IR. Data mining and clustering techniques. In: DRTC Workshop on Semantic Web (2003). Online available from: http://www.researchgate.net/publication/265145730_Data_Mining_and_Clustering_Techniques.

[13] Shaw MJ, Subramaniam C, Tan GW, Welge ME. Knowledge management and data mining for marketing. Decision support systems. 2001; 31(1): 127-37.

[14] Silva RJ, Alvares LO. Spatio-temporal analysis of Twitter messages. IX Regional School database. 2013. (in Portuguese) Online available from:
http://www.lbd.dcc.ufmg.br/colecoes/erbd/2013/005.pdf

[15] Witten IH, Frank E. Hall MA. Data Mining: practical machine learning tools and techniques. 3th Ed. Burlington: The Morgan Kaufmann series in data management systems. 2011. 629p.

[16] Yang F, Jin H, Qi H. Study on the application of data mining for customer groups based on the modified ID3 algorithm in the e-commerce. CSIP 2012: Proceedings of the 2012 International Conference on Computer Science and Information Processing; 2012 Aug 24-26; Xi'an, Shaanxi: IEEE; 2013. p. 615-9.

[17] Zhang X, Zhang J. CRM applications in e-commerce strategy. ICCIS 2013: Proceedings of the 2013 International Conference on Computational and Information Sciences; 2013 Jun 21-23; Shiyan, Hubei, China: IEEE; 2013. p. 605-8.