An Efficient Multi-set HPID3 Algorithm based on RFM Model

Priyanka Rani
M.tech

Nitin Mishra
M.tech

Samidha Diwedi Sharma
PhD

ABSTRACT

Data mining is a latest emerging technique, which is mainly used to inspect large database in order to discover hidden knowledge and information about customers' behaviors. With the increasing contest in the retail industry, the main focus of superstore is to classify valuable customers accurately and quickly among the large volume of data. The decision tree algorithm is a more general data classification function algorithm based on machine learning. In this paper the concept of Recency, Frequency and Monetary is introduced, which is usually used by marketing investigators to develop marketing strategies, to find important patterns. Conventional ID3 algorithm is modified by horizontally splitting the sample of customer purchasing RFM dataset and then classification rules are discovered to predict future customer behaviors by matching pattern. The dataset has been accessed from blood transfusion service center and has 5 attributes and 748 instances. The experimental result shows that the proposed HPID3 is more effective than conventional ID3 in terms of accuracy and processing speed.

Keywords

Data mining, ID3, HPID3, RFM, customer classification, Decision tree

1. INTRODUCTION

Data mining is generally thought of as the process of extracting hidden, previously unknown and potentially useful information from databases. Exploiting large volumes of data for superior decision making by looking for interesting patterns in the data has become a main task in today's business environment. Data classification is one of the most widely used technologies in data mining. Its main purpose is to build a classification model, which can be mapped to a particular subclass through the data list in the database. Classification is very essential to organize data, retrieve information correctly and rapidly. At present, the decision tree has become an important data mining method. It is a more general data classification algorithm based on machine learning. There exist many methods to do decision analysis. Each method has its own advantages and disadvantages. In machine learning, decision tree learning is one of the most popular techniques for making decisions in pattern recognition. The basic learning approach of decision tree is greedy algorithm, which uses the recursive top-down approach of decision tree structure. The decision tree as a classification method has the advantage of processing large amount of data quickly and accurately. ID3 algorithm is one of the most widely used mathematical algorithms for building the decision tree. It was invented by J. Ross Quinlan and uses Information Theory invented by Shannon. It builds the tree from the top down, with no backtracking. ID3 algorithm is an example of Symbolic Learning and Rule Induction. It is also a supervised learner which means it looks at examples like a training data set to make its decisions. The key feature of ID3 is choosing information gain as the standard for testing attributes for classification and then expanding the branches of decision tree recursively until the entire tree has been built completely. The main objective of this paper is to predict future customer behaviors from customer purchasing R-F-M dataset using HPID3 algorithm. Firstly the historical data is collected from which we can predict customer behaviors or customer purchasing patterns and then the entire dataset is partitioned horizontally. Using R–F–M attributes of customer transaction dataset classification rules are discovered by HPID3 algorithm in which divide-and-conquer strategy is used and the decision tree is pruned to avoid over fitting problem. The mathematical calculation for entropy and information gain is performed on each attribute in multiple partitions. At the end, the information gains of the same attribute in multiple partitions are added together to find out the total information gain of a particular attribute in the given training dataset. The attribute with the largest information gain is selected as the root of the decision tree and their values as its branches. After that information gain of the remaining attributes is again calculated with respect to selected attribute. This process continues recursively until it comes to a conclusion at the decision tree's leaf node. In this way the proposed system will minimize information content needed and the average depth of the generated decision tree. The dataset is accessed from http://archive.ics.uci.edu/ml/machine-learningdatabases/ blood-transfusion/transfusion data. There are 5 attributes and 748 instances in the dataset. The description of the dataset is given in Table 1.

Table 1: (Description of dataset)

Attribute	Description
R	Months since last blood donation.
F	Total number of donation.
M	Total blood donated in c.c.
T	Months since first donation
class	1 for donating blood & 0 for not donating

1.1 Recency, Frequency and Monetary

The RFM is the most frequently adopted segmentation technique that contains three measures (recency, frequency and monetary) & used to analyze the behavior of a customer and then make predictions based on the behavior in the database. Segmentation divides markets into groups with similar customer needs and characteristics that are likely to exhibit similar purchasing behaviors. Due to the effectiveness of RFM in marketing, some data mining techniques have been carried out with RFM. The most common techniques are: (1) clustering and (2) classification. Clustering based on RFM attributes provides information of customers' actual marketing levels. Classification based on RFM attributes provides valuable information for managers to predict future customer behavior.

1.1.1 Definition:

Recency : Recency is commonly defined by the time period from the last activity to now. A customer having a high score of recency implies that he or she is more likely to make a repeated action.

Monetary: Monetary is commonly defined by the total amount of money spent during a specified period of time. Higher monetary score indicates that customer is more important.

Frequency: Frequency is commonly defined by the number of activities made in a certain time period. Higher frequency score indicates greater customer loyalty he or she has great demand for the product and is more likely to purchase the products repeatedly.

1.2 Advantages of RFM methodology

1. It not only determines the existence of a pattern but also checks whether it conforms to the recency & monetary constraints.
2. It measures when people buy, how often they buy and how much they buy.
3. It allows marketers to test marketing actions to smaller segments of customers, and direct larger actions only towards those customer segments that are predicted to respond profitably.
4. Company can easily classify their customers who are important & valuable.
5. It can provide company's profit in a short time.
6. It acts as a common tool to develop marketing strategies.
7. It facilitates to choose which customers to target with an offer.
8. It helps in identifying significant and valuable customers.
9. It is used to identify customers and analyze customer profitability.
10. Firms can get much benefit from the adoption of RFM, encompassing increased response rates, lowered order cost and greater profit.

2. RELATED WORK

In 2011 Xiaojing Zhou1, Zhuo Zhang1and Yin Lu2 discovered five kinds of Customer Segmentation methods based on vital statistics, lifestyle, way of act and profit in traditional market segmentation and makes some analysis and research about the efficiency and applicability [1]. In 2011 Wei Jianping used rough set technology & RFM model for classification of VIP customers [2]. In 2010 Liu Yuxun, Xie Niuniu developed a more efficient and accurate Improved ID3 Algorithm for solving the problem of decision tree algorithm based on attribute importance [6]. A new method for quickly determining the customer rank was developed by Wei Jianping using RFM model and Rough Set in 2010, which is helpful for enterprises to develop more feasible marketing programs for different customers [7]. For analyzing the main factors like technology and employee's professional skill that affect Customers satisfaction degree, a decision tree model was used by Mei-Ping Xie and Wei-Ya Zhao in 2010 [8]. In 2009 Derya Birant discovered data mining using RFM analysis tasks, including clustering, classification and association rule mining, to provide market intelligence [9].

3. EXISTING ID3 ALGORITHM

ID3 algorithm is one of the most widely used mathematical algorithms for building the decision tree. It was invented by J. Ross Quinlan and uses Information Theory invented by Shannon. It builds the tree from the top down, with no backtracking. The key feature of ID3 is choosing information gain as the standard for testing attributes for classification and then expanding the branches of decision tree recursively until the entire tree has been built completely.

```
Input:
R, a set of attributes.
C, the class attributes.
S, data set of tuples.

Output:
The Decision tree

Procedure:
1. If R is empty then
2. Return the leaf having the most frequent value in data set S.
3. Else if all tuples in S have the same class value then
4. Return a leaf with that specific class value.
5. Else
6. Determine attribute A with the highest information gain in S.
7. Partition S in m parts S(a1), ..., S(am) such that a1, ..., am are the different values of A.
8. Return a tree with root A and m branches labeled a1...am, such that branch i contains ID3(R − {A}, C, S (ai)).
9. End if
```

Basic Concept-

Definition 1
Information:
Assume there are two classes, P and N. Let dataset S contain p elements of class P and n elements of class N. The amount of information, needed to decide if an arbitrary instance in S belongs to P or N is defined as

$$I(p,n) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}$$

Definition 2
Entropy:
Suppose A is an attribute having n different values {a1,a2,…………an} . Using this property, S can be divided into n number of subsets {s1,s2………sv}, If Si contains pi examples of P and ni examples of N, the entropy, or the expected information needed to classify objects in all sub trees Si is

$$E(A) = \sum_{i=1}^{v}\frac{p_i + n_i}{p+n} I(p_i, n_i)$$

Definition 3
Gain:
The encoding information that would be gained by branching on A

GAIN(A) = I (p, n) – E (A)

4. PROPOSED SYSTEM

Firstly the RFM dataset is collected from which we can predict customer behaviors or customer purchasing patterns and then the entire dataset is partitioned horizontally. Using R–F–M attributes of customer transaction dataset classification rules are discovered by HPID3 algorithm in which divide-and-conquer strategy is used and the decision tree is pruned to avoid over fitting problem. The mathematical calculation for entropy and information gain is performed on each attribute in multiple partitions. At the end, the information gains of the same attribute in multiple partitions are added together to find out the total information gain of a particular attribute in the given training dataset. The attribute with the largest information gain is selected as the root of the decision tree and their values as its branches. After that information gain of the remaining attributes is again calculated with respect to selected attribute. This process continues recursively until it comes to a conclusion at the decision tree's leaf node. In this way the proposed system will minimize information content needed and the average depth of the generated decision tree.

Step 1: Data Preparation for RFM Modeling
The first step is to collect RFM dataset from which we can predict customer behaviors, or customer purchasing patterns. Using RFM model we can select customers who are more likely to buy. In this way, we can improve marketing efficiency and maximize profit.

Step 2: Horizontally partitioning the dataset
After preprocessing, the dataset is partitioned horizontally into multiple partitions.

Step 3: Develop Decision Tree using HPID3 algorithm
Using R–F–M attributes of dataset classification rules are discovered by HPID3 algorithm in which divide-and-conquer strategy is used and the decision tree is pruned to avoid over fitting problem. It calculates overall entropy and information gains of all attributes. The attribute with the highest information gain is chosen to make the decision. So, at each node of tree, horizontal partition decision tree chooses one attribute that most effectively splits the training data into subsets with the best cut point, according to the entropy and information gain.

Step 4: Targeting Segment Selection
The next step is to perform profit analysis by selecting the segment with higher profit or higher response rate. Targeting segments are selected so as to maximize profit.

Step 5: Fetching RFM values
Using RFM model RFM values are fetched & we can select customers who are more likely to buy.

Step 6: Pattern matching
After fetching RFM values we get frequent pattern by sequentially matching it in the dataset.

4.1 The Proposed Algorithm:

```
1. Horizontally partition the dataset into multiple parties P1,
P2, …., Pn Parties.
2. Each Party contains R set of attributes A1, A2, …., AR. 3.
C the class attributes contains c class values C1, C2, …., Cc.
4. For party Pi where i = 1 to n do
5. If R is Empty Then
6. Return a leaf node with class value
7. Else if all transaction in T (Pi) has the same class then
8. Return a leaf node with the class value
9. Else
10. Calculate Expected Information to classify the given
sample for each party Pi individually.
11. Calculate Entropy for each attribute (A1, A2, …., AR) of
each party Pi.
12. Calculate Information Gain for each attribute (A1, A2,….,
AR) of each party Pi
13. End If.
14. End For
15. Calculate Total Information Gain for each attribute of all
parties (TotalInformationGain ( )).
16. ABestAttribute ← MaxInformationGain ( )
17.Let V1, V2, …., Vm be the value of attributes.
ABestAttribute partitioned   P1, P2,…., Pn parties into m
parties
   P1 (V1), P1 (V2), …., P1 (Vm)
   P2 (V1), P2 (V2), …., P2(Vm)
          .               .
          .               .
   Pn (V1), Pn (V2), …., Pn (Vm)
18. Return the Tree whose Root is labeled ABestAttribute and
has m edges labeled V1, V2, …., Vm. Such that for every i
the edge Vi goes to the Tree
19. End.
```

5. TEST RESULTS

The proposed algorithm & ID3 algorithm is tested on the blood transfusion service center RFM dataset taken from Hsin-Chu City in Taiwan. There are 5 attributes and 748 instances in the dataset. The testing is done using Weka tools with the development of classification model on the dataset. The experimental result shows that the proposed HPID3 is more effective than traditional ID3 in terms of accuracy and processing speed. The comparison test results between speed, mean absolute error & relative absolute error are given below-

Table-2 (Comparison in processing time)

no_of_instances	id3time(ms)	HPID3_time(ms)
50	21	10
100	29	16
250	33	19
500	39	24
746	46	31

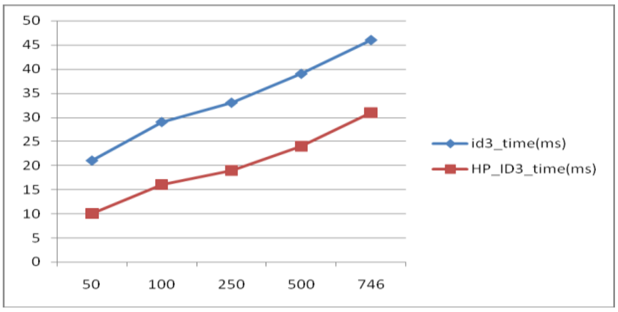


Table-3 (Comparison in mean absolute error)

no_of_instances	ID3_Mean absolute	HP_ID3_Mean absolute
50	0.2857	0.19
100	0.24	0.23
250	0.252	0.24
500	0.224	0.14
746	0.238	0.15

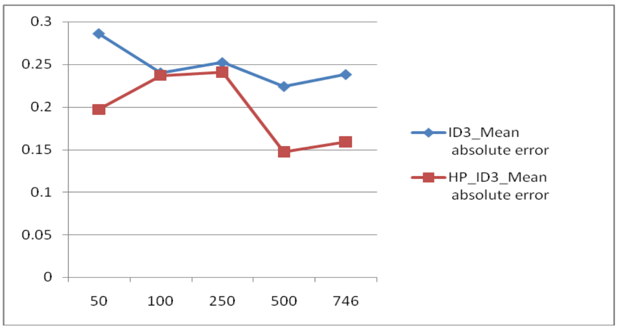
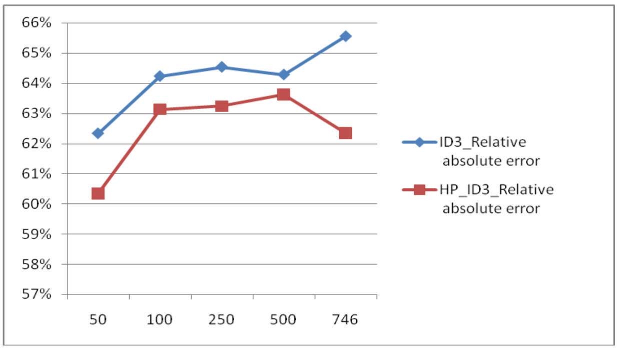


Table-4 (Comparison in relative absolute error)

no_of_instance	ID3_Relative abs. error	HPID3_Relative abs. error
50	62%	60.33%
100	64.22%	63.13%
250	64.53%	63.24%
500	64.28%	63.63%
746	65.55%	62.34%



6. CONCLUSION

In this paper the proposed algorithm is successfully experimented. The conventional ID3 algorithm is modified by horizontally splitting the sample of blood transfusion RFM dataset and then some classification rules are discovered to predict future customer behaviours by matching pattern. The experimental result shows that the proposed HP-ID3 is more effective than traditional ID3 in terms of accuracy and processing speed. The classification performance achieved seems satisfactory so far thus making it useful for use in real applications.

7. REFERENCES

[1] Xiaojing Zhou, Zhuo Zhang, Yin Lu,” Review of Customer Segmentation method in CRM”, 2011 IEEE.

[2] Wei Jianping”Research on VIP Customer Classification Rule Base on RFM Model”, 2011 IEEE.

[3] Xingwen Liu, Dianhong Wang, Liangxiao Jiang, Fenxiong Chen and Shengfeng Gan,” A Novel Method for Inducing ID3 Decision Trees Based on Variable Precision Rough Set”, 2011 IEEE.

[4] Imas Sukaesih Sitanggang, Razali Yaakob, Norwati Mustapha, Ahmad Ainuddin B Nuruddin,” An Extended ID3 Decision Tree Algorithm for Spatial Data”, 2011 IEEE.

[5] Hnin Wint Khaing,” Data Mining based Fragmentation and Prediction of Medical Data”, 2011 IEEE

[6] “Liu Yuxun, Xie Niuniu”,Improved ID3 Algorithm”, 2010 IEEE.

[7] Wei Jianping “Research on Customer Classification Rule Extraction base on RFM Model and Rough Set”, 2010 IEEE.

[8] Mei-Ping Xie, Wei-Ya Zhao,”The Analysis of Customers’ Satisfaction Degree Based On Decision Tree Model”, 2010 IEEE.

[9] Derya Birant,” Data Mining Using RFM Analysis”, Dokuz Eylul University Turkey

[10] Hui-Chu Chang,” Developing EL-RFM Model for Quantification Learner's Learning Behavior in Distance Learning”, Department of Electrical Engineering National Taiwan University, 2010 IEEE.

[11] Ya-Han Hu, Fan Wu, Tzu-Wei Yeh,” Considering RFM-Values of Frequent Patterns in Transactional Databases”, 2010 IEEE

[12] Chaohua Liu,” Customer Segmentation and Evaluation Based On RFM, Cross-selling and Customer Loyalty ”, 2011 IEEE

[13] liu jiale, duhuiying,” Study on Airline Customer Value Evaluation Based on RFM Model”, 2010 International Conference On Computer Design And Appliations (ICCDA 2010)

[14] “Research and Improvement on ID3 Algorithm in Intrusion Detection System”, 2010 Sixth International Conference on Natural Computation (ICNC 2010), 2010 IEEE.

[15] Chen Jin, Luo De-lin, Mu Fen-xiang, “An Improved ID3 Decision Tree Algorithm”, Proceedings of 2009 4th International Conference on Computer Science & Education