IJSTE - International Journal of Science Technology & Engineering | Volume 3 | Issue 10 | April 2017
ISSN (online): 2349-784X

Enhancing the Analysis of Customer Behavior in Supermarket Through Map Reduce

Karthiga S
Assistant Professor
Department of Information Technology
Thiagarajar College of Engineering

Sushma K
PG Student
Department of Information Technology
Thiagarajar College of Engineering

Abstract

Customer behavior analytics have implemented in many systems, though still it's a developing and unexplored market has greater potential for better advancements. One of the major challenges for knowledge discovery and data mining systems stands in developing their data analysis capability to discover out of the ordinary models in the data. Now-a-days data mining became more important due to the arrival of powerful data collection and storage tools. In this paper, a Map Reduce implementation of statistical classifier, C4.5 algorithm has to be proposed. The problem addressed in this paper is Big data processing is complex using traditional database management tools. The objective of this paper is To reduce the average time spend by the customer in supermarket. To analysis customer behavior which helps to turn big data into big value by allowing to predict the buyer behavior to improve their sales. The proposed system is to implementation of C4.5 algorithm using Mapreduce framework.

Keywords: C4.5 algorithm; Hadoop; MapReduce; Supermarket; Recommendation

I. INTRODUCTION

Customer analytics helps to turn big data into big value by allowing the organizations to predict the buyer behavior thereby improving their sales, market optimization, inventory planning, fraud detection and many more applications. A wide range of approaches are available and can be implemented but the one that stands out is the use of decision trees for the purpose of classification that can be efficiently used in customer analytics. To handle various types of data, a variety of decision tree algorithms have been developed over a phase of time with enhancement in performance and ability. One of the well-known decision tree algorithm is C4.5 is C4.5, an extension of basic ID3 decision tree algorithm. Customer analytic is incomplete without visualization of the data. In addition to classification of data using decision trees it is also important to visualize the data so that organizations get a visual aspect of the data in order to understand the variations in customer patterns. C4.5 is an algorithm used to generate a decision tree. The decision trees generated by C4.5 can be used for classification, and for this reason C4.5 is often referred to as a statistical classifier. C4.5algorithm can easily handle missing values. As missing attribute values are not utilized in gain calculations by C4.5. In this paper we are going to collect data of supermarket and store it in a database and by using the collected data the interest of the customer are predicted by using C4.5 algorithm

II. LITERATURE SURVEY

In the paper Customer behavior analysis using big data analytics by khade, they used decision tree algorithm to predict the customer behavior and to analysis the pattern their aim is to implement the customer data visualization using data driven documents which will help to build customer behavior customized graph. Their future plan is to collect the data and make it in cloud hadoop mapreduce framework. In the paper prediction of churn behavior of bank customers using data mining tools by devi prasad have addressed the problem purchasing behavior of customer is predicted using the technique classification tree model C5.0. In the paper Enhancing consumer behavior analysis by data mining techniques by Nan-chenhsieh addressed the problem of customer behavior analysis by using the decision tree. In the paper Analysis and predictions on students behavior using decision trees by vasile predicts the students choice in continuing their education with the post of university studies by using the technique decision tree data mining algorithm C4.5. In the customer behavior analytics and data mining paper by victor the problem addressed is to determine and identify future customers and their behavior by using the apriori algorithm. In the RFID enabled smart billing system by vanitha to reduce the average time a customer spends ata a supermarket by using the technique KNN algorithm. In the customer behavior model using the data mining by milan patel is to predict customer behavior by using the decision tree. In the paper customer relationship management based on decision tree induction by govindu to address the customer classification and prediction.

A. Venn Diagram– Used to Discover Hidden Relationships of the data

It will used to combine multiple segments to locate connections, differences and their relationship. It search for customers who have bought different categories of products and easily to identify cross-selling opportunities.

B. Data Profiling–Used to Identify Customer Attributes in dataset

Select records from your data tree and generate customer profiles that indicate common features and behaviors. Use customer profiles to inform effective sales and marketing strategy.

C. Forecasting – Instance Sequence Analysis

Forecasting enables you to adapt to changes, trends and seasonal patterns. You can accurately predict monthly sales volume or anticipate to the number of orders expected in any given month.

D. Mapping – To Map same type of items

Mapping uses color-coding to indicate customer behavior as it changes across geographic regions. A map divided into polygons that represent geographic regions shows you where your churners are concentrated or where specific products sell the most.

E. Association Rules – It will generate set of rules

This technique is used to find the relationship or similarity patterns across data and generates a set of rules. It automatically selects the rules that are most useful to key business insights: What products do customers purchase simultaneously and when? Which customers are not buying and why? What new cross-selling opportunities exist?

F. Decision Tree – Categorize and Predict Customer Behavior

Decision tree technique is one of the most popular methods for classification in various data mining applications and helps for the process of decision making. Classification helps you to do things like select the right products to suggest to particular customers and predict potential churn. Mostly decision tree algorithms include ID3, C4.5 and CART.
Abbreviations and Acronym
1) KNN- K-Nearest Neighbor algorithm

III. RELATED TECHNOLOGIES

A. Apache Hadoop

Apache Hadoop is an open source software framework. Hadoop consists of two main components: a distributed processing framework named MapReduce and a distributed file system known as the Hadoop distributed file system, or HDFS. One of the most important reason for using this framework in this project is to process a large amount of data and do its analysis which is not possible with other system. The storage is provided by HDFS and the analysis is done by MapReduce. For Mapreduce and distributed file system are best when worked in hadoop platform, the other subprojects provide corresponding services, or build on the core to provide high-level abstractions.

B. Hadoop Distributed File System

The Hadoop Distributed File System (HDFS) is used for storage. In short, HDFS provides a distributed architecture for extremely large scale storage, which can easily be extended by scaling out. When a file is stored in HDFS, the file is separated into uniformed size blocks. The size of block can be customized or the predefined one can be used. In this project, the customer dataset are stored in the HDFS. The dataset contains a lot of customer records with respect to their interest and purchases made by them. Also, the output file containing decision rules is written into HDFS for recommendation.

C. Map Reduce Model

MapReduce is a programming model for processing and generating large data sets with a parallel, distributed algorithm on a cluster. MapReduce works by breaking the processing into two phases: the map phase and the Reduce phase. Every phase has key-value pairs as input and output, the types of which may be chosen by the programmer. The programmer also specifies two functions: the Map function and the Reduce function. The input to our map phase is the raw data of customers. We choose a text input format that gives us each line in the dataset as a text value. The key is the offset of the beginning of the line from the beginning of the file. The output from the map function is processed by the MapReduce framework before being sent to the reduce function. This processing sorts and groups the key-value pairs by key.

IV. METHODOLOGY

The flow of the system is as follows:
1) Customer dataset will be loaded from the HDFS as input for the algorithm.
2) Invoke the instance of C4.5 class.
3) Using the MapReduce framework of Hadoop.

4) Reduce function counts number of occurrences of combination of (index and its value and class Label) and prints count against it.
5) Calculate entropy, information gain and gain ratio of attributes.
6) Process the input dataset from HDFS according to the defined algorithm of C4.5
7) in MapReduce framework.
8) Generate the rules and store it in HDFS.
9) Provide Recommendation according to their interest.

A. Equations

Let C denote the number of classes. In this case, there are two classes in which the records will be classified into. The classes are yes and no. The p(S, j) is the proportion of instances in S that are assigned to jth class. Therefore, the entropy of attribute S is calculated as:

Entropy(S) = - Σ j=1c p(S,j) *log p(S,j) Entropy is calculated for each record of a particular attribute.

Accordingly the information gain by a training dataset T is defined as:

Gain(S,T)= Entropy(S)- Σ v€values(Ts)|T(s,y)/T(s)* log p(S,j) where Values (TS) is the set of values of S in T , Ts is the subset of T induced by S , and TS ,v is the subset of T in which attribute S has a value of v.

V. PROPOSED WORK

MapReduce is a promising parallel and scalable programming model for data-intensive applications and scientific analysis. A MapReduce program expresses a large distributed computation as a sequence of parallel operations on datasets of key/value pairs. The two phases of MapReduce algorithm are, the Map and Reduce phases. The Map phase splits the input data into a large number of fragments, which are evenly distributed to Map tasks across the nodes of a cluster to process. Every Map task takes place in a key-value pair and then generates a set of intermediate key-value pairs. After the MapReduce runtime system groups and sorts all the intermediate values associated with the same intermediate key, the runtime system delivers the intermediate values to Reduce tasks. Each Reduce task takes in all intermediate pairs associated with a particular key and emits a final set of key value pairs. Both input pairs of Map and the output pairs of Reduce are managed by an underlying distributed file system. MapReduce greatly improves program ability by offering automatic data management, highly scalable, and transparent fault tolerant processing. Also, MapReduce is running on clusters of cheap commodity servers an increasingly attractive alternative to expensive computing platforms. Hadoop one of the most popular MapReduce implementations is running on clusters where Hadoop distributed file system (HDFS) stores data to provide high aggregate I/O bandwidth. At the heart of HDFS is a single Name Node a master server that manages the file system namespace and regulates access to files.

Customer analytics helps to turn big data into big value by allowing the organizations to predict the buyer behavior thereby improving their sales, market optimization, inventory planning, fraud detection and many more applications. A wide range of approaches are available and can be implemented but the one that stands out is the use of decision trees for the purpose of classification that can be efficiently used in consumer analytics. A variety of decision tree algorithms have been developed over a phase of time with enhancement in performance and ability. One of the well-known decision tree algorithm is C4.5 is C4.5, an extension of basic ID3 decision tree algorithm. Customer analytic is incomplete without visualization of the data. In addition to classification of data using decision trees it is also important to visualize the data so that organizations get a visual aspect of the data in order to understand the variations in customer patterns.

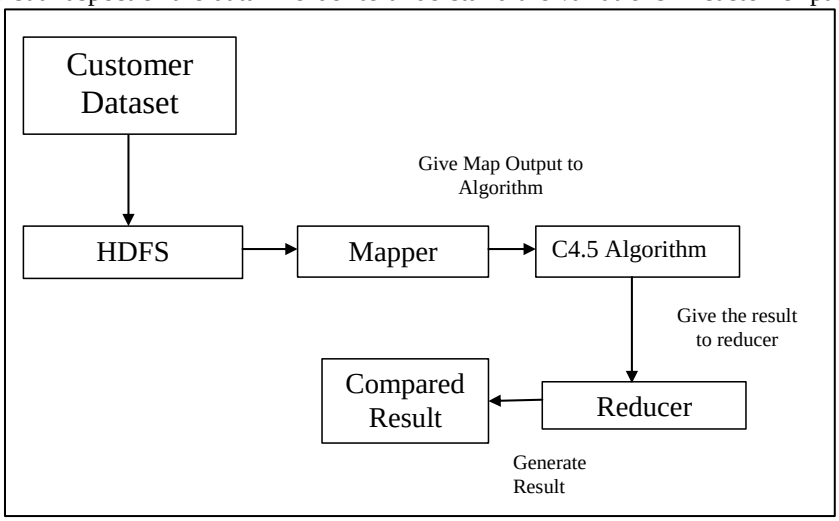
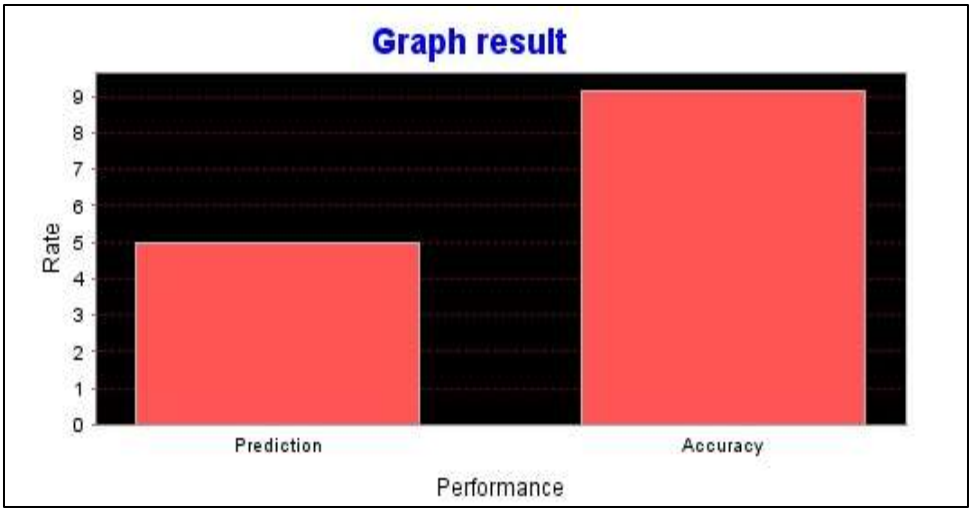


Fig. 4.1: System Architecture

VI. PERFORMANCE EVALUATION



The above graph shows that the rate of accuracy will be twice the rate of prediction. The performance of the prediction and accuracy in this paper is higher than the previous experiments. In previous experiments the prediction and the accuracy rate will be in equal rate. but by using the C4.5 algorithm the predicted data will help to get higher accuracy than the previous. In this experiment by using lower rate of prediction we can get most accurate values, that is by using the lower prediction rate we can recommend items of the customers interest by using the higher accuracy.

VII. CONCLUSION

This paper defines the proposed system for distributed implementation of C4.5 algorithm using Mapreduce framework along with the customer data. The raise in the development of cloud computing and big data, traditional decision tree algorithms cannot fit any more and hence we introduced the mapreduce implementation of C4.5 decision tree algorithm. In future work the data from the supermarket can be collected by using the RFID device and read using the tag, while purchasing the product the reader will automatically read the rate and the quantity of the items and display it in the screen automatically. The data will automatically write in the database. by using the data collected the prediction is done by using C4.5 algorithm and the required item for the customer of their interest will be recommended for the customer.

REFERENCES

[1] Devi Prasad, Madhavi "Prediction of churn behavior of bank customers using data mining tools", Business Intelligence Journal(2014).
[2] Nan-Chen Hsieh Kuo-Chung Chu"Enhancing consumer behavior analysis by data mining techniques ", International Journal of Information and Management Sciences (2015) 79–92.
[3] Vasile Paul Brefelean "Analysis and Predictions on Students' Behavior Using Decision Trees", 29th Int. Conf. on Information Technology Interfaces (2015).
[4] Abhijit Raorane & R.V.Kulkarni "A Source for Consumer Behavior Analysis", Expert Systems With Applications (2015)
[5] Mr. Brijain R Patel, Mr. Kushik K Rana .A Survey on Decision Tree Algorithm for Classification. IJEDR (2014)
[6] Wei Dai and Wei Ji. (2014). A MapReduce Implementation of C4.5 Decision Tree Algorithm. International Journal of Database Theory and Application [Online] 7(1), pp. 49-60.
[7] Angosta, L. Essential Guide to Data Warehousing. Prentice Hall PTR; 1st edition; 2014.
[8] M-S Chen, J. Han and P. S. Yu. "Data Mining: An Overview from Database Perspective". IEEE Trans. On Knowledge And Data Engineering, vol. 8, pp. 866-883, 2015.
[9] Au, W. H. and Chan, K. C. C., Mining fuzzy association rules in a bank-account database, IEEE Transactions on Fuzzy Systems, Vol. 11, 2003.
[10] Baesens, B., Viaene, S., Poel, D., Vanthienen, J. and Dedene, G., Bayesian neural network for repeat purchase modelling in direct marketing, European Journal of Operational Research, Vol. 138, pp.191-211, 2002.
[11] Balakrishnan, P. V. S., Cooper, M. C., Jacob, V. S. and Lewis, P. A., Comparative performance of the FSCL neural net and K-means algorithm for market segmentation, European Journal of Operational Research, Vol. 93, pp.346-357, 1996.
[12] Berry, M. J. A. and Linoff, G. S., Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management, 2nd Edition, John Wiley & Sons, 2004.
[13] Cai, Y., Gercone, N. and Han, J., An attribute-oriented approach for learning classification rules from relational databases, Proceedings of Conference on Data Engineering, pp.281-288, 1990.
[14] Cendrowska, J., PRISM: an algorithm for inducing modular rules, International Journal of Man Machine Studies, Vol. 27, pp.349-370, 1988.
[15] Chan, C.-C. H., Intelligent spider for information retrieval to support mining-based price prediction for online auctioning, Expert Systems with Applications, Vol. 34, pp.347-356, 2008.
[16] Chan, C.-C. H., Online auction customer segmentation using a neural network model, International Journal of Applied Science and Engineering, Vol. 3, pp.101-109, 2005.
[17] Bre_felean VP, Bre_felean M, Ghi_oiu N, Comes C-A. Data mining in continuing education. INTED 2007, International Technology, Education and Development Conference, Valencia, Spain, 2007.
[18] Burlak G, Munoz J, Ochoa A, Hernández JA: Detecting Cheats in Online Student Assessments Using Data Mining. Proceedings of DMIN 2006. p. 204-210

[19] Cohen J, Cohen P, West SG, Aiken LS. Applied multiple regression/correlation analysis for the behavioral sciences. (2nded.) Hillsdale, NJ: Lawrence Erlbaum Associates, 2003.
[20] Cunha MM, Putnik GD, Ávila P. Towards Focused Markets of Resources for Agile / Virtual Enterprise Integration, in Advances in Networked Enterprises: Virtual Organisations, Balanced Automation, and Systems Integration, Kluwer Academic; 2000. p. 15-24.
[21] Gunter S, Bunke H. Evaluation of classical and novel ensemble methods for handwritten word recognition, Proc. IAPR Workshop on Structural and Syntactic Pattern Recognition, Lisbon, 2004.
[22] Heathcote E, Dawson S, Data Mining for Evaluation, Benchmarking and Reflective Practice in a LMS. In Proceedings E-Learn 2005 Vancouver, Canada, 2005.
[23] Witten I H, Frank E. Data mining: practical machine learning tools and techniques, 2nd ed., Morgan Kaufmann series in data management systems, Elsevier Inc.; 2005.
[24] Loing B. ICT And Higher Education, 9th UNESCO/NGO Collective Consultation on Higher Education, Paris; 6-8 April 2005.
[25] Mena J. Investigative Data Mining for Security and Criminal Detection, Elsevier Science, USA; 2003.