Measure Customer Behaviour using C4.5 Decision Tree Map Reduce Implementation in Big Data Analytics and Data Visualization

Nadeem Abdulla  
UG Student  
Department of Computer Engineering  
S.I.E.S Graduate School of Technology

Sreena Gopinathan Nair  
UG Student  
Department of Computer Engineering  
S.I.E.S Graduate School of Technology

Zubair Ahmed Mohammed Gazzali  
UG Student  
Department of Computer Engineering  
S.I.E.S Graduate School of Technology

Anindita Khade  
Assistant Professor  
Department of Computer Engineering  
S.I.E.S Graduate School of Technology

Abstract

Even though there are lots of invented systems that have implemented customer analytics, it’s still an upcoming and unexplored market that has greater potential for better advancements. Big data is one of the most raising technology trends that have the capability for significantly changing the way business organizations use customer behaviour to analyze and transform it into valuable insights. Also decision trees can be used efficiently in the decision making analysis under uncertainty which provides a variety of essential results. Hence to the end of this paper, we propose the MapReduce implementation of well-known statistical classifier, C4.5 decision tree algorithm. The paper additionally mentions why C4.5 is preferred over ID3. Apart from this, our system aims to implement Customer data visualization using Data Driven Documents (d3.js) which allows you to build well customized graphics.

Keywords: Big data analytics, C4.5 algorithm, D3.js, Data visualization, Decision tree, Hadoop, MapReduce

I. INTRODUCTION

Big data is a collection of data that has very large volume, comes from variety of sources like web, business organizations etc. in different formats and comes at us with a great velocity which makes processing complex and tedious using traditional database management tools. It can be termed as a growing torrent. So the major demanding issues in big data processing include storage, search, distribution, transfer, analysis and visualization.

Earlier, the term 'Analytics' indicated the study of existing data to research potential trends and to analyze the effects of certain decisions or events that can be used for business intelligence to gain various valuable insights. Today's biggest challenge is how to discover all the hidden information through the huge amount of data collected from a varied collection of sources. There comes Big Data Analytics into picture. Big data analytics provides us a new world of opportunities. One of them is the customer behaviour analysis which is referred as customer analytics.

Customer analytics helps to turn big data into big value by allowing the organizations to predict the buyer behaviour 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.

Researchers have developed various decision tree algorithms over a period of time with enhancement in performance and ability to handle various types of data. One of the well-known decision tree algorithm is C4.5 is C4.5 [3-4], an extension of basic ID3 decision tree algorithm [5]. Customer analytics 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. Visual diagrams like charts and graphs give statistical and summarized information.

II. LITERATURE SURVEY

A. Traditional Analytical Systems For Customer Behaviour[7]:
In the late 1960s and early 1970s, there were two mainstream approaches to constructing Database Management System’s (DBMS’s). The first approach was based on the hierarchical data model, typified by IMS (Information Management Systems)
Measure Customer Behaviour using C4.5 Decision Tree Map Reduce Implementation in Big Data Analytics and Data Visualization (IJIRST/ Volume 1 / Issue 10 / 047)

from IBM, in response to the enormous information storage requirements generated by the Apollo space program. The second approach was based on the network data model, which attempted to create a database standard and resolve some of the difficulties of the hierarchical model, such as its inability to represent complex relationships DBMSs. However, these two models had some fundamental disadvantages:

1) Complex programs had to be written to answer even simple queries
2) There was minimal data independence
3) There was no widely accepted theoretical foundation.
4) Does not support analysis of unstructured DATA

Many experimental relational DBMS were implemented thereafter, with the first commercial products appearing in the 1970’s and early 1980’s. Relational DBMSs are referred to as second-generation DBMSs. Relational DBMS technology used extensively in the 80’s and 90’s was limited in meeting the more complex entity and data needs of companies, as their operations and applications became increasingly complex. In response to the increasing complexity of database applications, two “new” data models have emerged; the Object-Relational Database Management Systems (ORDBMS) and Object-Oriented Database Management Systems (OODBMS), which subscribes to the relational and object data models respectively. The OODBMS and ORDBMS have been combined to represent the third generation of Database Management Systems.

There is considerable debate between OODBMS and RDBMS proponents as to the adequacies of these applications. The OODBMS proponents claim that RDBMS are satisfactory for standard business applications but lack the capability to support complex applications. The relational supporters claim that relational technology is a necessary part of any real DBMS and that complex applications can be handled by extensions to the relational model.

Recent database trends include the growth of distributed databases and the emergence of object-oriented and hyper-media databases.

B. Dawn Of Big Data Analytics:
Data turns to big data when its volume, velocity, or variety go beyond the abilities of the IT operational systems to gather, store, analyze, and process it. Most of the organizations are capable of handling vast amount of unstructured data using varied tools and equipments but with the rapidly growing volume and fast flood of data, they do not have the capability of mining it and derive necessary insights in a well-timed way.

Big Data is emerging from the realms of science projects at Web companies to help companies like telecommunication giants understand exactly which customers are unhappy with service and what processes caused the dissatisfaction, and predict which customers are going to change carriers. To obtain this information, billions of loosely-structured bytes of data in different locations needs to be processed until the needle in the haystack is found. The analysis enables executive management to fix faulty processes or people and maybe be able to reach out to retain the at-risk customers Big data is becoming one of the most important technology trends that have the potential for dramatically changing the way organizations use consumer behaviour to analyze and transform it into valuable insights.[11]

C. Key concepts of Customer analytics[6]:
The survey on customer analytics revealed the following key concepts:
1) Venn Diagram – Discovering Hidden Relationships
Combine multiple segments to discover connections, relationships or differences. Explore customers that have bought different categories of products and easily identify cross-selling opportunities.
2) Data Profiling – Identify Customer Attributes
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.
3) Forecasting – Time Series 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.
4) Mapping – Identify Geographical Zones
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.
5) Association Rules – Cause/Effect – Basket Analysis
This technique detects relationship or affinity 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?
6) Decision Tree – Classify and Predict Behavior
Decision trees are one of the most popular methods for classification in various data mining applications and assist the process of decision making. Classification helps you do things like select the right products to recommend to particular customers and predict potential churn. Most primarily used decision tree algorithms include ID3, C4.5 and CART.
D. Comparative analysis between ID3 and C4.5 [10]:
1) Model Build Time:
Extraction from dataset to data model is known as Model built time. It is depend upon the number of dataset used in training. And accuracy of program is depending upon the number of dataset, greater number of dataset there will be more accuracy.

<table>
<thead>
<tr>
<th>Size of Data set</th>
<th>ID3</th>
<th>C4.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>0.0156 sec</td>
<td>0 sec</td>
</tr>
<tr>
<td>24</td>
<td>0 sec</td>
<td>0 sec</td>
</tr>
<tr>
<td>35</td>
<td>0.187 sec</td>
<td>0.124 sec</td>
</tr>
</tbody>
</table>

2) Search Time:
Search time is defined as after building model answering time of system is called search time.

<table>
<thead>
<tr>
<th>Size of Data set</th>
<th>ID3</th>
<th>C4.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>0 sec</td>
<td>0 sec</td>
</tr>
<tr>
<td>24</td>
<td>0 sec</td>
<td>0 sec</td>
</tr>
<tr>
<td>35</td>
<td>0.577 sec</td>
<td>0.421 sec</td>
</tr>
</tbody>
</table>

3) Accuracy:
The measurements of a quantity to that quantity’s factual value to the degree of familiarity are known as accuracy.

<table>
<thead>
<tr>
<th>Size of Data set</th>
<th>ID3</th>
<th>C4.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>94.155 %</td>
<td>94.155 %</td>
</tr>
<tr>
<td>24</td>
<td>78.472 %</td>
<td>78.472 %</td>
</tr>
<tr>
<td>35</td>
<td>82.208 %</td>
<td>82.208 %</td>
</tr>
</tbody>
</table>

In other performance parameters like error rate and memory used, C4.5 and ID3 gives the same results. The improvements of C4.5 include:[9]
1) Employ information gain ratio instead of information gain as a measurement to select splitting attributes
2) Not only discrete attributes, but also continuous ones can be handled
3) Handling incomplete training data with missing values
4) Prune during the construction of trees to avoid over-fitting

E. Tools For Data Visualization:
1) Polymaps:
Polymaps is a free JavaScript library and a joint project from SimpleGeo and Stamen. This complex map overlay tool can load data at a range of scales, offering multi-zoom functionality at levels ranging from country all the way down to street view. [12]

2) Flot:
A JavaScript plotting library for jQuery, Flot is a browser-based application compatible with most common browsers — including Internet Explorer, Chrome, Firefox, Safari and Opera. Flot supports a variety of visualization options for data points, interactive charts, stacked charts, panning and zooming, and other capabilities through a variety of plugins for specific functionality. [12]

3) D3.js:
A JavaScript library for creating data visualizations with an emphasis on web standards. Using HTML, SVG and CSS, bring documents to life with a data-driven approach to DOM manipulation — all with the full capabilities of modern browsers and no constraints of proprietary frameworks. [12]

4) SAS Visual Analytics:
SAS Visual Analytics is a tool for exploring data sets of all sizes visually for more comprehensive analytics. With an intuitive platform and automatic forecasting tools, SAS Visual Analytics allows even non-technical users to explore the deeper relationships behind data and uncover hidden opportunities. [12]
III. RELATED TECHNOLOGIES

A. Apache Hadoop:
Apache Hadoop[13] is an open source software framework [16]. 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[2]. 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. Although Hadoop is best known for MapReduce and its distributed file system, the other subprojects provide complementary services, or build on the core to provide high-level abstractions. [1]

B. Hadoop Distributed File System:
The Hadoop Distributed File System (HDFS)[15] is the storage component. 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 divided into evenly sized blocks. The size of block can be customized or the predefined one can be used. In this project, the customer dataset is stored in HDFS. The dataset contains a lot of customer records with respect to purchases. Also, the output file containing decision rules of is written into HDFS.

C. MapReduce:
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. Each 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. [1]

Java code for the map function and the reduce function for this implementation is written for overriding the default map and reduce function provided by hadoop framework. The programming logic for the respective is based on C4.5 algorithm.

IV. METHODOLOGY

The flow of the system is as follows:
1) Loading the customer dataset from HDFS as input for the algorithm.
2) Invoke the instance of C4.5 class.
3) Using the MapReduce framework of Hadoop, Map function is invoked which checks whether this instance belongs to Current Node or not. For all uncovered attributes it outputs index and its value and class label of instance.
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 decision tree data mining in MapReduce framework.
7) Generate the decision rules and store it in HDFS.
8) Accept the new test data from web UI.
9) Access the rules and based on it, decide the category of the new data.
10) Provide visualization of the dataset from HDFS on the Web UI in the form of bar graphs, pie charts etc. using D3.js.

A. C4.5 Algorithm:
C4.5[3-4] is an algorithm used to generate a decision tree developed by Ross Quinlan. C4.5 is an extension of Quinlan’s earlier ID3 algorithm. 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. C4.5 algorithm uses information gain as splitting criteria. It can accept data with categorical or numerical values. To handle continuous values it generates threshold and then divides attributes with values above the threshold and values equal to or below the threshold. C4.5 algorithm can easily handle missing values. As missing attribute values are not utilized in gain calculations by C4.5.[8]

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 j-th class. Therefore, the entropy of attribute S is calculated as:

\[ \text{Entropy}(S) = - \sum_{j=1}^{C} p(S, j) \log p(S, j) \]

Entropy is calculated of each record of a particular attribute.

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

\[ \text{Gain}(S, T) = \text{Entropy}(S) - \sum_{v \in \text{Values}(T_S)} \frac{|T_{S,v}|}{|T_S|} \times \log \frac{|T_{S,v}|}{|T_S|} \]

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 .[9]

After calculating entropy and information gain from the above formulas ,GainRatio and SplitInfo are calculated as per the formulas given in the algorithm below[9].
Algorithm 1 C4.5(T)

Input: training dataset T; attributes S.
Output: decision tree Tree.

1: if T is NULL then
2: return failure
3: end if
4: if S is NULL then
5: return Tree as a single node with most frequent class label in T
6: end if
7: if |S| = 1 then
8: return Tree as a single node S
9: end if
10: set Tree = {}
11: for a ∈ S do
12: set Info(a, T) = 0, and SplitInfo(a, T) = 0
13: compute Entropy(a)
14: for v ∈ values(a, T) do
15: set T_{a,v} as the subset of T with attribute a = v
16: Info(a, T) += \frac{|T_{a,v}|}{|T_a|} Entropy(a_v)
17: SplitInfo(a, T) += -\frac{|T_{a,v}|}{|T_a|} \log \frac{|T_{a,v}|}{|T_a|}
18: end for
19: Gain(a, T) = Entropy(a) - Info(a, T)
20: GainRatio(a, T) = \frac{Gain(a, T)}{SplitInfo(a, T)}
21: end for
22: set a_{best} = \arg\max_a \{GainRatio(a, T)\}
23: attach a_{best} into Tree
24: for v ∈ values(a_{best}, T) do
25: call C4.5(T_{a,v})
26: end for
27: return Tree

Fig. 3: Algorithm Description of C4.5

B. Data Visualization using D3.js:
D3.js is a JavaScript library for manipulating documents based on data. D3 helps you bring data to life using HTML, SVG, and CSS. D3’s emphasis on web standards gives you the full capabilities of modern browsers without tying yourself to a proprietary framework, combining powerful visualization components and a data-driven approach to DOM manipulation.[14]

1) Key features of D3.js[14]:
   1) Bind arbitrary data to DOM
   2) Create interactive SVG bar charts
   3) Generate HTML tables from data sets
   4) Variety of components and plugins to enhance capabilities
   5) Built-in reusable components for ease of coding
Measure Customer Behaviour using C4.5 Decision Tree Map Reduce Implementation in Big Data Analytics and Data Visualization

Fig. 4: Input Dataset in csv Format

Fig. 4. indicates the .csv format of the dataset used in visualization which involves various attributes such as Date, Time, Location, Category, Price, Payment and Purchase Decision.

Fig. 5: Customer data Visualization of Sales per day

Fig. 5. shows the data visualization on the input dataset. Here the bar graph generated using D3.js library indicate the sales of various products on the selected day

V. CONCLUSION

This paper defines the proposed system for distributed implementation of C4.5 algorithm using MapReduce framework along with the customer data visualization. With the rise in 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. Visualization done using D3.js is fast and reusable because it uses traditional HTML elements along with Scalable Vector Graphics (SVG). In future works, the use of fast and real time database systems like Apache HBase or MongoDB can be incorporated with this system. In addition to this, we can use distributed refined algorithms like ForestTree implemented in Apache Mahout to increase efficiency and scalability.

ACKNOWLEDGMENT

We are highly indebted to our Project guide for the guidance and constant supervision in completing our report successfully. We also thank our college to provide us with an opportunity to learn an emerging technology like big data.

REFERENCES

Measure Customer Behaviour using C4.5 Decision Tree Map Reduce Implementation in Big Data Analytics and Data Visualization


