A Technical Seminar report on

PERFORMING CUSTOMER BEHAVIOUR USING BIG DATA ANALYTICS

A Seminar Report Submitted to JNTU Hyderabad in partial fulfillment of the academic requirements for the award of the degree.

Bachelor of Technology

In

Computer Science and Engineering

Submitted by

VASIREDDY UJWALA

(18H51A05L7)

Under the esteemed guidance of

P CHANDRASHEKHAR REDDY

(Associate Prof. CSE)

Department of Computer Science and Engineering

CMR COLLEGE OF ENGINEERING & TECHNOLOGY

(An Autonomous Institution under UGC & JNTUH , Approved by AICTE, Permanently Affiliated to JNTUH, Accredited by NAAC with 'A+' Grade.)

KANDLAKOYA , MEDCHAL ROAD, HYDERABAD - 501401.

2018- 2022

CMR COLLEGE OF ENGINEERING & TECHNOLOGY

KANDLAKOYA, MEDCHAL ROAD, HYDERABAD – 501401

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the technical seminar report entitled **"PERFORMING CUSTOMER BEHAVIOUR USING BIG DATA ANALYTICS"** being submitted by **VASIREDDY UJWALA *(18H51A05L7)*,** in partial fulfillment for the award of **Bachelor of Technology in Computer Science and Engineering** is a record of bonafide work carried out his/her under my guidance and supervision.

The results embodied in this project report have not been submitted to any other University or Institute for the award of any Degree.

P Chandrashekhar Reddy
Associate Prof. CSE
Dept. of CSE

Dr.K.Vijaya Kumar
Professor and HOD
Dept. of CSE

Submitted for viva voce Examination held on

External Examiner

Acknowledgment

With great pleasure I want to take this opportunity to express my heartfelt gratitude to all the people who helpedin making this project work a grand success.

I am grateful to **P Chandrasekhar Reddy,** Associate. Prof. CSE, Dept of Computer Science and Engineeringfor his valuable suggestions and guidance during the execution of this project work.

I would like to thank **Dr. K. Vijaya Kumar**, Head of the Department of Computer Science andEngineering, for his moral support throughout the period of my study in CMRCET.

I am highly indebted to **Major Dr. V.A. Narayana** , Principal CMRCET for giving permission to carryout this project in a successful and fruitful way.

I would like to thank the Teaching & Non- teaching staff of Department of Computer Science and Engineering fortheir cooperation

Finally, I express my sincere thanks to **Mr. Ch. Gopal Reddy**, Secretary, CMR Group of Institutions, forhis continuous care. I sincerely acknowledge and thank all those who gave support directly and indirectlyin completion of this project work.

VASIREDDY UJWALA
18H51A05L7

ABSTRACT

Consumer behaviour study is a new, interdisciplinary and emerging science, developed in the 1960s. It's main sources of information came from economics, psychology, sociology, anthropology and artificial intelligence. If a century ago, most people will be living in small towns, with limited possibilities to leave the community, and few ways to satisfy their needs, now, due to the accelerated evolution of technology and the radical change of life style, consumers begin to have increasingly diverse needs.

The customer relationship management (CRM) is a business methodology used to build long term profitable customers by analyzing customer needs and behaviors. The customer behavior is analyzed by choosing important attributes in the customer database. The customers are then segmented into groups according to their attribute values. The rules are generated using rule induction algorithms to describe the customers in each group. These rules can be used by the entrepreneur to predict the behavior of their new customers and to vary the attraction process for existing customers. In this paper a new rule algorithm has been proposed based on the concepts of rough set theory. Its performance has been compared with LEM2 (Learning from Examples Module, version 2) algorithm, an existing rough set based rule induction algorithm. Real data set of the customer transaction is used for analysis. Recency(R), Frequency (F), Monetary (M) and Payment (P) are the attributes chosen for analyzing customer data. The proposed algorithm on average achieves 0.439% increase in sensitivity, 0.007% increase in specificity, 0.151% increase in accuracy, 0.014% increase in positive predictive value, 0.218% increase in negative predictive value and 0.228% increase in F-measure when compared to LEM2 algorithm.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

The internet is the phenomenal innovation of the information system era. With billions of internet users across the globe generate huge amount of data like never before. There are millions of web sites which cater to various demands of its users. E-commerce, search engines, online shopping, banking, trading etc. are all accessible from any parts of the world. This has brought a new dimension to the user behaviour pattern. For example, there are multiple airline sites through which one can book an air ticket to any destination. While these facilities make the consumer life simpler, but creates a competitive environment for the service providers. Since there are multiple options available for customers, it is very difficult to sustain the customer base as with a slightest inconvenience can force customer to switch to a different service provider. Therefore, it is utmost important for the service providers to understand the customer demand, choices, preferences etc. and provide them high quality of services.

The term "Big Data" describes the accumulation and analysis of vast amounts of information. But Big Data is much more than a big amount of data. It is also the ability to extract meaning: to sort through big volumes of numbers and find the hidden patterns, unexpected correlations, and surprising connections that can be used in different industries like medical field, security and protection field or marketing field. In marketing field, companies that adopt "data-driven decision making" enjoy significantly greater productivity than those that do not. So, by using Big Data rapid deployment solutions they can lower the complexity of implementation projects, and hence project risks, while accelerating time to value. As a result, Big Data is an extraordinary knowledge revolution that is sweeping almost invisible through business, academia, government, healthcare, and everyday life. It already ensures safety and independence for older people, enables us to provide a healthier life for our children, conserves precious resources like water and energy, and peers into our own individual genetic makeup. Big Data is the perfect instrument to study today's consumer behavior.

Customer relationship management (CRM) technology is a mediator between customer management activities in all stages of a relationship (initiation, maintenance and termination) and business performance. It helps industries to gain insight into the behavior of customers and their value so that the enterprise can increase their profit by acting according to the customer characteristics. It is classified into operational and analytical. Operational CRM refers to the automation of business processes whereas analytical CRM refers to the analysis of customer characteristics and behaviors. Analytical CRM helps the entrepreneur to discriminate their customers and decide their marketing activities accordingly. It consists of four ideologies namely customer identification, customer attraction, customer retention and customer development. Customer identification is the process in which the customers are grouped and their characteristics are analyzed. Customer attraction is the process in which the customers buy for the next time by providing customer service, coupon distribution, direct mailing and discounts. Customer retention is the process in which the customer's needs are satisfied by introducing new products and rectifying their complaints. Customer development involves in expansion of transaction intensity, transaction value and individual customer profitability. Customer identification is the most important phase in analytical CRM because once the customer is identified correctly; he can be retained and developed further. The customer identification phase consists of customer segmentation and target customer analysis. Customer segmentation involves in segmenting customers into predefined number of customer groups. Target customer analysis involves in analyzing customer behavior or characteristics in each customer group. It helps the entrepreneur to vary the attraction process for existing customers and to predict new customer's behaviors. Data mining techniques are good at extracting and identifying useful information

and knowledge from enormous customer databases, and for making different CRM decisions. The application of data mining techniques in CRM is an emerging trend in the global economy.

Data mining is a collection of techniques for efficient automated discovery of previously unknown, valid, novel, useful and understandable patterns in large databases. These patterns are used in an enterprise's decision making process. The tasks that can be performed in data mining are clustering, association rule mining, rule induction and classification. Clustering is an unsupervised classification used to group data with similar characteristics. It produces clusters for the given input data where data in one cluster is more similar when compared to data in other clusters. Association rule mining produces dependency rules which will predict the occurrence of an attribute based on the occurrence of other attributes in the data base. Rule induction belongs to supervised learning where data are already clustered into groups and it generates rules by finding regularities in the data in each cluster. Rules are in the form of If-Then condition. If part is called as antecedent and Then part is called as consequent. Antecedent contains conditional variables and consequent contains single decision variable. The conditional variables are the attributes in the given data and the decision variable is the cluster number assigned to the data using clustering algorithm. Rules generated for a cluster constitute the rule set for that cluster. Each data in the cluster should be described by at least one rule in the rule set of that cluster. This property of rule induction algorithm is called completeness. Each rule in the rule set of a cluster should be satisfied only by the data in that cluster. Rule set for a cluster should cover all the data within that cluster and no rule should be satisfied by any data in other clusters. This property of rule induction algorithm is called consistency. The data's satisfied by rules are not mutually exclusive because a data can be described by any number of rules. Classification on the other hand also generates rule for describing data in each cluster. The main difference between classification and rule induction is that the classification rules are mutually exclusive which means each data in the database is described by exactly only one rule. Rule induction is used to describe the characteristics of the data rather than classification rules because in real data set, each data has to be described by all of its possible combinations of attributes value which means only one rule for each data is not sufficient. Clustering and rule induction of data mining technique is used for customer segmentation and target customer analysis of customer identification phase in CRM.

In this paper, an improved rule induction algorithm based on rough set theory has been developed to generate rules for clustered customer's data. The proposed algorithm has been compared with LEM2, a rough set based approach. The rest of the paper is organized in the following: In Section 2 we describe the overview of customer relationship management, clustering algorithms, rule induction algorithms and LEM2 algorithm. In Section 3 we propose an improved rule induction algorithm based on rough set approach. In Section 4 we compare the prediction results obtained using rule induction algorithms. Finally in Section 5 we conclude the best rule induction algorithm according to the criteria chosen for comparison.

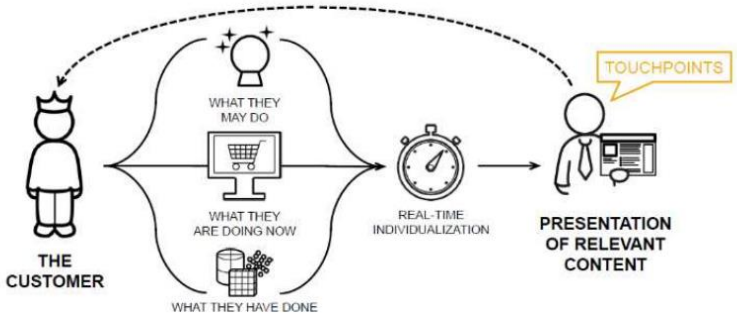

Fig.1. Improving Consumer Experience

2. LITERATURE REVIEW

2.1 Traditional Analytical Systems For Customer Behaviour

In the late 1970s, there were two approaches for constructing Database Management System's (DBMS's). The first approach was based on the hierarchical data model, typified by (Information Management Systems) 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 like the complex programs had to be written to answer even simple queries. Also there was minimal data independence.

Many experimental relational DBMS were implemented thereafter, with the first commercial products appearing in the 1970's and early 1980's. Relational DBMS 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 had 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.

2.2 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 companies to help telecommunication giants understand exactly which customers are happy with their service and what processes caused the dissatisfaction, and predict which customers are going to change the service. To obtain this information, billions of loosely-structured bytes of data in different locations needs to be processed until the required data is found out. This type of analysis enables executive management to fix faulty processes or people and may be able to reach out to retain 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 customer behaviour to analyze and transform it into valuable insights.

2.3 Key concepts of Customer analytics

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.

2.4 Tools for data visualization

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 multizoom functionality at levels ranging from country all the way down to street view.
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.
D3.js: A JavaScript library for creating data visualizations with an emphasis on web standardsUsing 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.
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.

3. MODEL, CONCEPTS

The proposed model figure 2.0 uses theory, survey (Primary data), statistical model and big data. The theory is based on quality of service/product and the service usages. Quantitative research methodologies have been used in building the model.

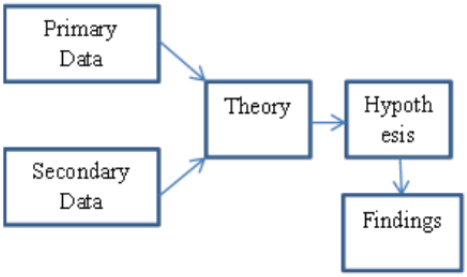

Fig.2. Quantitative Research to establish Big Data factors for behaviour prediction

3.1 Quantitative research

Quantitative research is required to establish numerically that big data is useful in predicting the behaviour of customers. Survey has been conducted with the big data users to gather big data usages for e-commerce through questionnaire sessions. Participants were selected from Big Data users from e-commerce industries. Using statistical analysis hypotheses have been evaluated using IBM SPSS©.

3.2. Dependent Variable vs. Independent Variable

In table the dependent and independent variables have been shown. The goal is to evaluate the impact of independent variables on the dependent variables.

Independent Variable	**Dependent Variable**
Transaction information	i) Service Usages
Feedback	
Browsing information	ii) Quality of Service

Table 1. Independent vs. Dependent variables

3.3. Theory

3.3.1 Quality of product/service

Quality of service or product (Valarie A. Zeithaml, A. Parasuraman, Arvind Malhotra, 2002) has been discussed as an important aspect in the web based services. For example, the transaction on a particular service is less because of less demand of the service or may be that the web service quality is not up to the mark. Therefore, to understand the root cause it is important to analyse the quality of service. If the quality is good, then the numbers of transactions decide the demand of the service or the product under study. Quality of service is further supported by the theory of quality of product information (Chung-Hoon Park and Young-Gul Kim, 2003).

3.3.2 Service Usages

Online service usages (Susan M. Keaveney, Madhavan Parthasarathy, 2001) theory indicates that the online service usage is one of the factors that determine the customer switching behaviour.

This is crucial in favour of online usage and therefore use of big data in studying customer behaviour. With

higher usages of online services, high volume of logs will be produced which can be processed using big data to find out the customer access pattern on a service or product.

3.4 Model

In this model key factors such as blogs, surfing patterns, and transaction information are being fed into big data. The data format is unstructured and the volume is huge which cannot be processed through traditional data processing tools. Using Big Data these factors are tested against the theory of i) Quality of service and ii) Service usages.

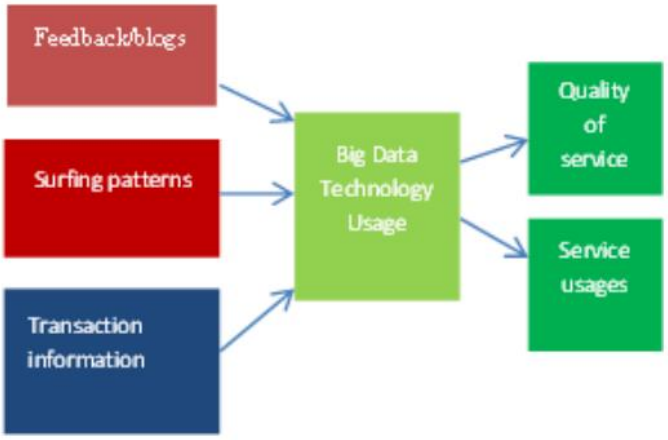

Fig.3. Model for studying key factors of Consumer Behavior using Big Data

3.4 Concepts

3.4.1 Feedback through blogs:

The social media has enabled a new medium through which customers express their opinion about a product or service. If a service/product does not meet the user expectations, users do not hesitate to express it over blogs. The data out of blogs are unstructured and the volume is huge.

Surfing patterns: Understanding surfing patterns is one of the major factor for improving the quality of service. In big data context, big data should be able to identify the demands of a particular service. For example, if a customer is searching a travel destination and how frequently and how many customers are looking for the particular destination can predict the preference of customer. Service demand is discussed (A. Dan et al., 2004).

3.4.2 Transactions: Online transactions are very common. Customers buy and sell using online web sites. Enormous amount of data is generated out of these transactions. Big data has the capabilities to process these data.

3.4.3 One Sample T Test: One sample t test has been chosen in this model as the full population information is not available and to make sure the sample selected comes from a particular population. Big Data is evolving and it is new in the industry. It is having its own class. Therefore one sample T test is justifiable to determine that the sample is selected from a population of known mean (μ).

3.4.4 Customer behaviour prediction: Using big data technology if the quality of service and service usages are identified, then determining the behavioural aspect of customer can be established.

4. EVOLUTION OF CUSTOMER BEHAVIOUR

How do people take purchase decisions today?

The purchasing decision process began to be studied about 300 years ago by Nicholas Bernoulli (in 1783 he introduced the terms of expected utility and marginal utility in the economic theory), followed by John von Neumann and Oskar Morgenstern (they introduced the terms of risk and uncertainty, and in 1944 they published a fundamental article for microeconomics "Theory of Games and Economic Behavior"). They created a mathematical model in order to determine the utility gained after a consumer activity, people being considered pure rational beings (consumers tried only to satisfy self-interest).

Recent research shows that there are numerous factors that influence the purchase decision, besides the rational ones, like social, cognitive and emotional factors. By taking these factors into consideration when modelling the purchasing decision process, a new, interdisciplinary and emerging science appeared in 1960: the study of consumer behavior. It is a complex science that includes information from economics, psychology, sociology, anthropology and artificial intelligence.

Until 1960, the economic perspective of consumer behavior and the models that described it relied on the assumption that all consumers are always rational in their purchases, so they will always buy the product that will bring the higher satisfaction. In this regard, three types of models were developed. Between 1700 and 1930, the Economic Model was used to describe consumer behavior which involved the rational perspective based on Neoclassical Economic Theory. In the next 20 years, the behavior perspective began to be applied which was based on the Learning Model, and after that the cognitive perspective which was based on the Psychoanalytic and Sociological Model.

During this time, people had a conservative behavior because they were buying the same products, consumer behavior being an emergent phenomenon that has evolved along with human development. In prehistoric times this behavior was shown in a very limited way, people being organized in small family groups and having a single concern: survival. Much later, the social skills began to develop that finally led to the emergence of money, social status, wealth and ultimately shaped the consumer behavior.

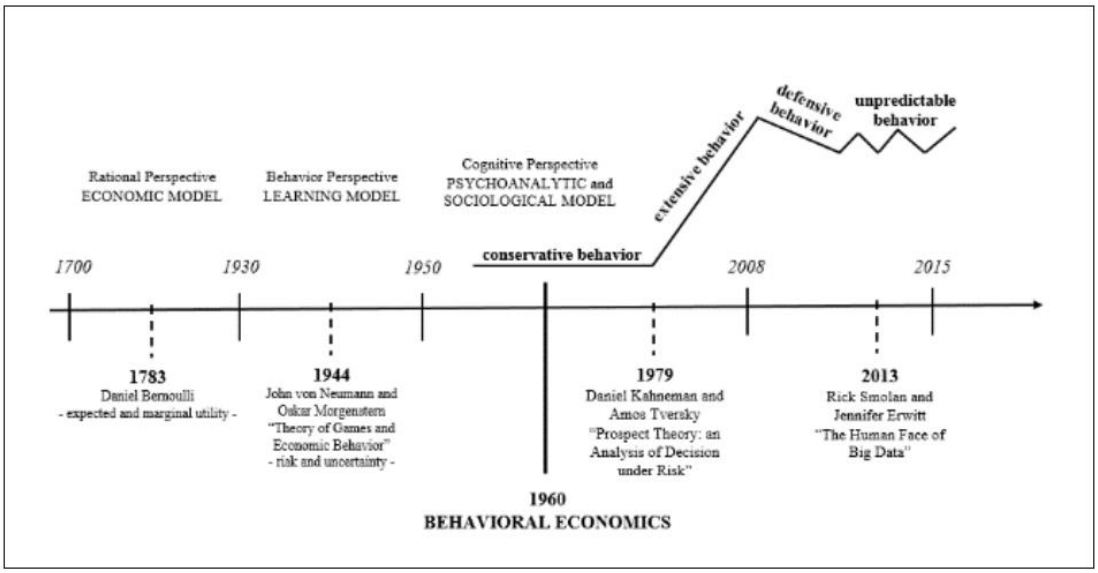

Fig.4. Evolution of Consumer Behaviour

The main cause that determined the researchers to study the consumer behavior is the diversification of needs. In the same time, looking back a century ago, a strong connection can be observed between the moment when the population began recording a strong upward trend and the science of studying consumers behavior appeared.

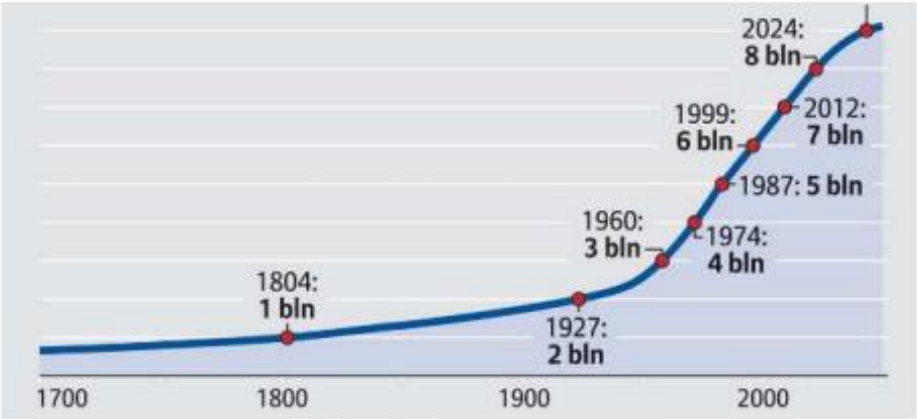

Fig.5. Population Growth on Earth

This correlation can be explained by the fact that a growing population means more needs, more products and more suppliers. Also, the life expectancy has doubled in the last century (at the beginning of 20th century the life expectancy was about 30- 40 years, while in 2008 it changed to 70 years), with the same result in the change of consumer behavior: more needs to be satisfied and a consumer behavior more complex.

Also, the middle of 20th century is the moment when travel started to become accessible to all, due to the large-scale production of machinery and commercial aircraft. By travel, people had the opportunity to discover other cultures and habits and as a result their needs started to diversify. Whereas in the past most people lived in small towns, with limited possibilities to leave their community and few variations in needs, now, due to technical improvements, consumers began to have increasingly more diverse needs.

For half a century, people developed an extensive behavior, buying increasingly more products and increasingly more diverse. One of the most important paper written during this period is “Prospect Theory: an analysis of decision under risk”, written by Daniel Kahneman and Amos Tversky, which proposes a new model for studying consumer behavior. In this paper, the decision making is viewed as a choice between prospects or gambles. The authors developed the new theory from the assumption that the Theory of Expected Utility (which was not challenged for over 250 years), had some flaws regarding the moment when the choice is made by the consumer. They thought that the utility is not only dependent on the actual value of a persons wealth, but also on the evolution of his or her wealth.

The Prospect Theory is the most important model used at the end of the 20th century, but there were also other models created in that period of time: Nicosia model (1966), Engel, Blackwell and Miniard model (1968), Howard Sheth model (1969), Webster and Wind model (1972), Hobbes model (1984) and Veblen model (1994).

The year 2008 represents another important moment in world’s history that influenced the consumer behavior. The economic and financial crisis that spread all over the world led consumers to think twice before buying a product. Because consumers were buying fewer products, their behavior began to be a defensive one. People began to use the internet on a larger scale in order to search products and to compare their price and characteristics. Online marketing began to have a decisive role in the buying process, so new techniques were developed in order to predict the consumer behavior, one of them being Big Data.

Today consumers face an offer too diverse, being assaulted by marketing messages. Because of that, the opportunity cost for a product has significantly increased, making the decision process more and more complicated. According to studies, consumers may ignore the opportunity cost when they don't have to choose from more than 8 products. When the number of choices increases, the consumers become indecisive, and sometimes even give up to the buying process.

The changes in consumer behavior have had strong influences on all enterprises throughout time, a decisive moment being the mid-1970s when a significant macroeconomic change on the law of supply and demand had happened: if by that time the markets were driven by vendors, their control was taken over by buyers both in terms of influence and bargaining power.

Companies understood that the consumer behavior is an emergent phenomenon that has evolved with human development and they became more interested in studying the behavior of their daily consumer. As a result, today's companies are empowered by the final consumer who wants instant value, mobile functionality and userfriendly services. Today, people are more informed (57% of the buying process is completed before a first interaction with sales), socially networked (53% of customers abandoned an in-store purchase due to negative online sentiment) and less loyal (59% of customers are willing to try a new brand to get better customer service).

As a conclusion, the main factors that shaped consumer behavior are:

- demographic changes (the growth of population and life expectancy, had the same result in consumer behavior: more needs to be satisfied);
- evolution of technology (because people now have more ways to travel, they discovered other cultures and life styles, so their needs became more diverse);
- multiplicity (because more and more variables are integrated in every day activities – for example the movie industry has evolved from a one-dimensional to a multi-dimensional experience – also the buying act needs to become a complex experience);
- hyper efficiency (the space-time efficiency is also a daily problem, so people need faster and cheaper ways to satisfy their needs);
- risk and stress (people have too many options to choose from in order to satisfy their needs).

5. CUSTOMER BEHAVIOUR ANALYSIS IN BIG DATA ERA

4.1 Connotation and Characteristics of Big Data Analysis

The Big data refers to the large amount of data involved in the data collection, and conventional analysis tools cannot complete the acquisition, processing and collation of data in a short period of time. Different from sample analysis, big data analysis is to process all data comprehensively, comprehensively and professionally and obtain effective information from it. It is generally believed that big data analysis has five characteristics: Huge amount of data, ultra-fast calculation speed, diversified data types, low value density and high information authenticity. At present, big data analysis has been widely used in all walks of life and has gradually become an indispensable productive force, promoting the efficient allocation and utilization of means of production and rapidly promoting the improvement of social production efficiency.

4.2 For Consumer Behavior Characteristics in Big Data Era

4.2.1 Consumers' Behavior Choice Is More Rational

The advent of the big data era has changed the way consumers obtain product information, and the information they know is more sufficient and accurate. In the traditional market model, consumers mostly know a certain product or brand through advertisements and lack other information support, which will restrict consumers' rational decision making. In the era of big data, consumers can fully grasp the product information through massive analysis data, deeply understand the product attributes, and continuously upgrade from the situational involvement of products to long-term involvement. Therefore, consumers will continuously generate positive internal perception and promote the occurrence of consumer purchasing behavior. Nowadays, there are many consumers with high product involvement in the market. Such buyers will use network information search and comparison, as well as other user evaluations, to comprehensively evaluate factors such as product cost performance, brand advantages and their own needs, and finally make more rational purchase decisions.

4.2.2 Consumers' Demand Continues to Escalate

The popularity of big data is slowly changing the behavior of online and offline traders. Consumers have higher and higher requirements for choosing online shopping. They not only need the function and quality of products, but also satisfy their pleasure and experience of online shopping, that is, they pay more and more attention to personalized services provided by merchants. The value of trading activities includes the use value of the product itself and the purchase experience value. Furthermore, sometimes the utility brought by the experience to consumers plays a decisive role in the purchase decision. Big data promotes personalized marketing of e-commerce, while consumers are demanding more and more innovative and personalized services.

4.2.3 Consumers' Trust in the Commercial Functions of Social Media has Increased

Nowadays, the commercial functions of social media are continuously explored and utilized, and the commercial value is increasingly prominent. And innovative business models appear in social media and are gradually accepted and recognized by consumers. Through social media, enterprises can master more and more comprehensive personal information of consumers, so as to be able to accurately analyze their personal preferences, habits and other information, so as to better meet the deep needs of consumers or tap the potential needs of consumers. Enterprises can analyze people's habits, beliefs and preferences through social media, and can be accurate to a certain extent, thus forming an almost invasive intimate relationship with consumers and better meeting the deep needs of consumers. And consumers also expect their needs to be paid more and more attention, discovered and satisfied, so they also trust and support the commercial promotion of social media more and more.

6. CONSTRUCTION OF CUSTOMER BEHAVIOR MODEL UNDER BIG DATA ENVIRONMENT

5.1 AISAS Model Analysis Framework Based on Big Data Analysis

In the theory of consumer behavior research, AISAS model proposed by Dentsu Company is more suitable for analyzing consumer behavior choices in the era of network economy. According to the theory, consumers go through five stages from coming into contact with product or service information to finally completing the purchase behavior: A (Action), I (Interest), S (Search), A (Action), S (Share). It is developed according to the traditional AIDMA mode. On the one hand, both of them describe a series of behavioral changes in the process of consumer selection. On the other hand, the difference is that in AISAS mode, two "s" with network characteristics-search and share have been added, which reflects the importance of search and share in the Internet era, instead of unilaterally transmitting information and inputting ideas to consumers.

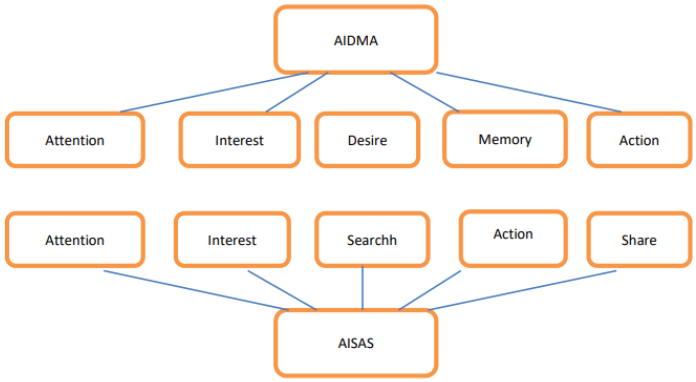

Fig.6. Comparison of Consumer Behavior Models

Factors that affect consumer behavior are usually divided into two aspects: Stimulation of external factors and internal perception. External factors mainly include product promotion, marketing methods, product price, sales volume, brand, user evaluation, etc., which will stimulate consumers' internal perception. On the one hand, big data analysis can help enterprises to better understand consumer demand, determine clear and targeted market strategies, and create more competitive advantages; On the other hand, the results of big data analysis are also used to improve and optimize external factors that affect consumer behavior, guide consumers to make optimal decisions and maximize utility.

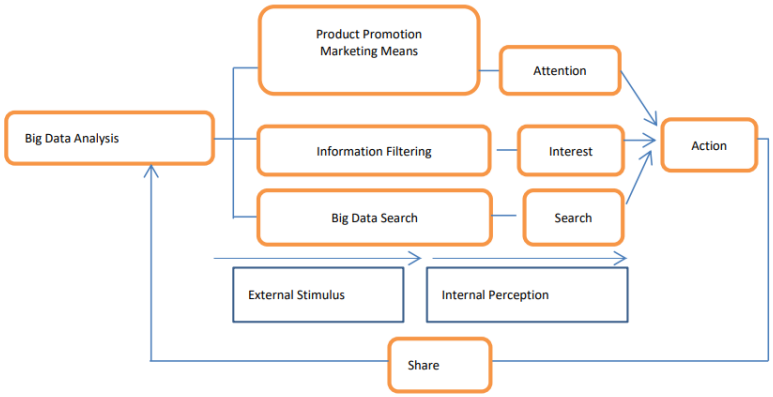

Fig.7. Impact mechanism of Big Data Analysis on Consumer Buehavior

This paper combines big data analysis with AISAS model to build a consumer behavior analysis framework to explain how big data affects consumer behavior, as shown in Figure 2. The impact of big data on external factors produces external stimulation to consumers through product promotion, marketing methods, information screening, data search and other links, and acts on attention, interest, search and other behaviors, affecting consumers' internal perception and finally making purchase decisions.

When the purchase behavior is over, the sharing of purchase experience has also become an important source of information for further big data analysis, which goes back and forth and continuously affects consumers' behavior choices.

5.2 Research Assumptions

Based on this, the consumer behavior model in the big data environment can be expressed by the following formula: $Y=AX1(X2)+BX2+CX3+\varepsilon$ （1）

In Equation (1), X1, X2 and X3 respectively represent the external stimulus factors, internal perception and consumption experience sharing of consumer behavior; Y represents the purchasing behavior of consumers; ε represents the error matrix; A, B and C respectively represent the influence coefficients of influencing factors on consumer behavior. So, how does big data affect consumer behavior through the penetration of these influencing factors? How much is the correlation between consumer behavior and influencing factors? In view of these problems, this paper puts forward the following assumptions:

H1: Big data analysis is conducive to improving product promotion paths and marketing methods, and can increase consumers' attention.

H2: Diversification of types of big data analysis is conducive to consumers' comparative analysis and screening out products of interest.

H3: Big data mining helps consumers to conduct all-round information search on target products, generate more rational internal perception and make final consumption decisions.

H4: After the consumer's purchasing behavior is completed, the sharing of product consumption experience will form a new big data supplement, further affecting the consumer's purchasing behavior and forming a virtuous circle.

H5: Good internal perception will positively affect consumers' purchase intention.

5.3 Questionnaire Design and Result Analysis

Based on the above analysis, this paper conducts a questionnaire survey on consumers by means of questionnaires. The survey objects include employees of enterprises, housewives, college students and other groups with strong consumption ability. A total of 380 questionnaires were distributed and 352 valid questionnaires were recovered.

First of all, we analyzed the reliability and validity. Validity analysis refers to the effectiveness of measurement, which refers to the degree to which the means and tools in the questionnaire survey can accurately measure the object to be measured. In KMO test, the higher its value, the more it shows that the survey object is the object to be studied, that is, the more the results of the survey scale can show the real characteristic validity to be measured.

The following results are obtained by using SPSS software:

Indicators	Action	Attention	Interest	Search	share
α	0.907	0.911	0.878	0.892	0.921
KMO	0.812	0.774	0.769	0.795	0.804

Table 2. Reliability and Validity Analysis

As can be seen from the above table, the α coefficient values of each variable in the model are greater than 0.8, indicating that the internal consistency is high and the reliability meets the conditions. KOM values are above 0.7, and the availability level is also high. After we carry out regression analysis on the hypothesis model, the verification results are as follows:

Assumption	Coefficient	T value	Conclusion
H1	0.635	9.728	Support
H2	0.783	14.304	Support
H3	0.867	12.626	Support
H4	0.846	14.572	Support
H5	0.857	12.435	Support

Table 3. Analysis Table of Model Verification Results

As can be seen from the above table, the five assumptions in the model are all valid. Big data will act on the whole process of consumers' purchase decision-making and affect consumers' internal perception through external stimulation. At the same time, consumers' subsequent consumption experience sharing will continuously optimize and update the basic database of big data analysis. In addition, we can also see that the screening of relevant information and data search have a significant impact on consumer behavior.

7. CUSTOMER RELATIONSHIP MANAGEMENT

Customer segmentation gives a quantifiable way to analyze the customer data and distinguish the customers based on their purchase behavior. It is the process of dividing customers into homogeneous groups on the basis of common attributes. It is typically done by applying some form of cluster analysis to obtain a set of segments. In this way the customers can be grouped into different categories for which the marketing people can employ targeted marketing and thus retain the customers. Target customer analysis is used to analyze the customers in each cluster or segment so as to predict the new customer to the appropriate cluster. The customers are segmented and then rules are generated to describe them. These rules can be used to classify the new customers to the appropriate cluster who have similar purchase characteristics. The customer identification is followed by customer attraction which motivates each segment of customers in different way. Customer retention and customer development deals with retaining the existing customers and maximizing the customer purchase value respectively.

The attributes which describe the purchasing behavior of the customers are first chosen before customer segmentation because it requires a comprehensive understanding of enterprise customers. RFM model is used to identify and represent the customer characteristics by three attributes namely Recency (R), Frequency (F) and Monetary (M). R indicates the interval between the time that the latest consuming behavior happens and present. F indicates the number of transactions that the customer has done in a particular interval of time. M indicates the total value of the customer's transaction amount in a particular interval of time. In RFM method, K-means clustering algorithm and LEM2 are used to obtain the classification rules. According to, customers with the same pattern of purchasing are only clustered and RFM is used to calculate the value of each cluster. Tsai and Chiu (2004) in proposed a market segmentation methodology based on product specific variables such as items purchased and the associative monetary transactional history of customers and they used RFM to analyze the relative profitability of each customer's cluster. In customer behavior is identified using RFM model and grey correlation model is used for customer targeting. Yeh et al. (2009) in extended the traditional RFM model by including two parameters, time since the first purchase and churn probability.

In RFM analysis along with K-means clustering is used to study customer's fluctuations over different time frames. In customer lifetime value (CLV) is calculated using RFM. In, WRFM (Weighted RFM) is used instead of RFM. In this weights were assigned to R, F, and M depending on characteristics of the industry. Stone (1995), suggested for placing the highest weight on the Frequency, followed by the Recency, with the lowest weight on the Monetary measure. In Chuang and Shen (2008), Monetary had the most value and Recency had the least value.

The attributes chosen to describe the customer behavior and the weightage of the attributes will differ from domain to domain. Here, the RFMP model which has four attributes R, F, M and P with equal weights is used. RFMP model is the modified RFM model where the payment details of the customers are considered. P indicates the average time interval between payment and purchase date. Payment detail of the customer is an important attribute because any two customers with same R, F, M value but different P value cannot be treated equally by the company. The customers are segmented using their consuming behavior via RFMP attributes. This ensures that the standards which cluster customer value are not established subjectively, so that the clustering standards are established objectively based on RFMP attributes.

8. ALGORITHMS

7.1 CLUSTERING ALGORITHM

The customers are segmented using clustering based on their important attributes like R, F, M and P. It is an unsupervised classification where there are no predefined classes. The data in the data set is assigned to one of the output class depending upon its distance to other data. The data within each class forms a cluster. The number of clusters is equal to the number of output classes. The clustering technique produces clusters in which the data inside a cluster has high intra class similarity and low inter class similarity. The similarity is measured in terms of the distance between the data. For a numerical dataset, the distance between two data can be calculated using Euclidean, Manhattan and Minkowski distance.

Euclidean distance is given by

$$d(x,y) = squareroot(\sum_{i}^{n} |(x_i - y_i)|\text{^}2)$$

Manhattan distance is given by

$$d(x,y) = \sum_{i}^{n} |(x_i - y_i)|$$

Minkowski distance is given by

$$d(x,y) = (\sum_{i}^{n} |(x_i - y_i)|\text{^ p})\text{^ }(1/p)$$

In the above equations, n indicates the number of attributes in the given data, x and y are the data in the data set, d(x, y) is the distance between data x and y. In Minkowski distance if p=1 it is similar to Manhattan and if p=2 it is similar to Euclidean. In Euclidean distance the variation in one attribute is different from the variation in another attribute but in Manhattan distance the sum of the variation in each attribute is considered. In our real data set all the attributes R, F, M and P are equally weighted, so the variation in all the attributes is to be equally treated. Thus in this case Manhattan distance is used instead of Euclidean distance.

Clustering is mainly classified into hierarchical and partitioning algorithms. The hierarchical algorithms are further sub divided into agglomerative and divisive. Agglomerative clustering treats each data point as a singleton cluster and then successively merges clusters until all points have been merged into a single cluster. Divisive clustering treats all data points in a single cluster and successively breaks the clusters till one data point remains in each cluster. Partitioning algorithms partition the data set into predefined k number of clusters. K-means algorithm is one of the most commonly used clustering algorithms. It is a partitioning clustering algorithm which partitions the database D of n objects into a set of k clusters. The output differs when the initial centers for clusters are varied. The distance between objects in same cluster is less when compared to the distance between objects in different cluster. Each object is placed in exactly one of the k non-overlapping clusters. The steps in K-means algorithm are as follows:

1. Initialize centers for k clusters randomly

2. Calculate distance between each object to k-cluster centers using the Manhattan distance formula given by Equation 1

3. Assign objects to one of the nearest cluster center

4. Calculate the center for each cluster as the mean value of the objects assigned to it

5. Repeat steps 2 to 5 until the objects assigned to the clusters do not change

7.2 RULE INDUCTION ALGORITHMS

The rule induction algorithms are used to generate rules to describe the characteristics of the customers in each segment. Decision trees (DT), artificial neural networks (ANN), genetic algorithms (GA) and rough set theory (RST) are used to produce rules. DT is a flow-chart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent cluster number. ANN is a large number of highly interconnected processing elements (neurons) that uses a mathematical model, computational model or non-linear statistical data modeling tools for information processing to capture and represent complex input/output relationships. GA, which were formally introduced in the United States in the 1970s by John Holland at University of Michigan, are search algorithms applied to solve problems on a computer based on the mechanics of natural selection and the process of natural evolution. In DT, too many instances lead to large decision trees and decrease classification accuracy rate.

In ANN, number of hidden neurons, number of hidden layers and training parameters need to be determined, and has long training times. Moreover, ANN served as black box which leads to inconsistency of the outputs, is a trial-and-error process. GA also has some drawbacks such as slow convergence, a brute computing method, a large computation time and less stability. With respect to rough set theory, the advantages are they do not require any preliminary or additional parameter about the data, less expensive or time to generate rules, ability to handle large amounts data, yield understandable decision rules and stable. It can be used to make decisions in any underlying business. In the experimental results of, accuracy rate 1is more in LEM2 when compared to DT and ANN.

RST introduced by Pawlak in 1982 is a knowledge discovery tool that can be used to help induce logical patterns hidden in massive data. Some of the applications of RST in the field of knowledge discovery are dimensionality reduction, clustering, rule induction and discretization. The concept LERS (Learning from Examples using Rough Sets) was developed for rule induction. The basic algorithms based on LERS are LEM1, LEM2 and AQ. LEM1 algorithm computes global covering of attributes for producing rules. LEM2 algorithm on the other hand computes local covering and then converts into a rule set, so it gives better results compared to LEM1.

AQ algorithm developed by R.S.Mickalski generates cover for each concept by computing stars and selecting from them single complexes to the cover. In the worst case the time complexity of computing conjuncts of partial stars is O(nm) where n is the number of attributes and m is the number of data in the data set. So for large data set, AQ is not efficient when compared to LEM2. LEM2 of LERS is most frequently used since it gives better resul. Extensions of LEM2 are MLEM2 and LEM3.

MLEM2 extends LEM2 capability by inducing rules from data with both symbolic and numerical attributes including data with missing attribute values. It produces the rules sets with the smallest number of rules but needs an additional tool to simplify conditions using numerical attributes. LEM3 is based on incremental learning of production rules from examples so the memory space requirement is minimal but it uses the same rule generating procedure of LEM2.

The variable precision rough set model (VPRS), introduced in, is a generalization of the original rough set data analysis in the direction of relaxing the strict boundaries of equivalence classes. It assumes that rules are only valid within a certain part of the population, and it is able to cope with measurement errors. In, extension of the rough set theory based on the dominance principle is dealt. This method is mainly based on substituting the indiscernibility relation by a dominance relation in the rough approximation of decision classes. However, the

decision rules induced from the lower approximations of the Dominance-based Rough Set Approach (DSRA) are sometimes weak in that only a few objects support them. For this reason, a variant of DSRA, called VC-DRSA, has been proposed in. It allows some inconsistency in the lower approximations of sets by a parameter called consistency level. It is more general than the classic functional or relational model and is more understandable for users because of its natural syntax and because it considers the inconsistency of real-life. The problem domain considered in the paper has complete and consistent data, so the algorithms based on LERS has been concentrated and the LEM2 algorithm has been taken for comparison.

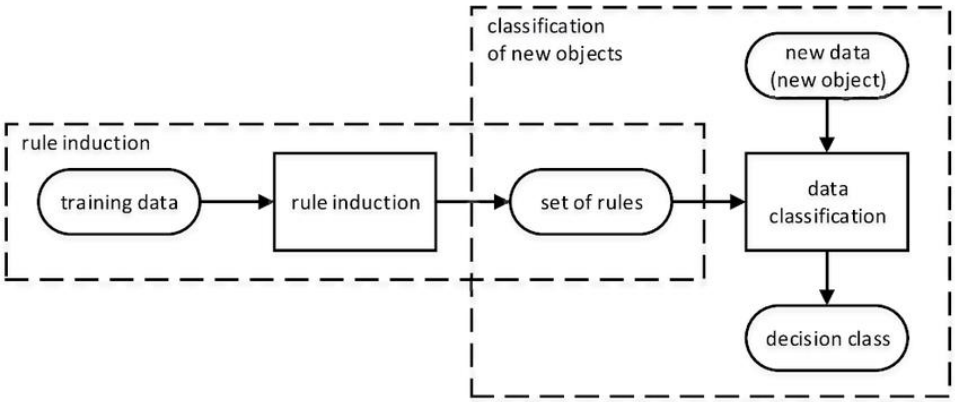

Fig.8. Rule Induction

7.3 LEM2 ALGORITHM

It is a rule induction algorithm based on rough set theory. It is used to find regularities hidden in the data and express in terms of rules. The clustering algorithm output is given as an input so that rules are generated for each cluster. Rules are in the form of

if (attribute-1, value-1) and (attribute-2, value-2) and ... and (attribute-n, value-n) then (decision, value)

In the database, each row is called as a case and each column is called as an attribute. Attributes are independent variables and decision is a single dependent variable. Here, Recency, Frequency, Monetary, Payment are attributes and cluster number is the decision variable.

The set of all cases labeled by same decision value is called a concept. A case x is covered by a rule r if and only if every condition (attribute-value pair) of r is satisfied by the corresponding attribute value for x.

A concept C is completely covered by a rue set R if and only if for every case x from C there exists a rule r from R such that r covers x. R contains set of rules for each decision value. R is complete if and only if every concept from the data set is completely covered by R. A rule r is consistent if and only if for every case x covered by r, x is a member of the concept C indicated by r. R is consistent if and only if every rule from R is consistent with the data set. Rule induction produces complete and consistent rule set.

A block of an attribute-value pair t = (a, v), denoted [t], is the set of all examples that for attribute a have value v. A concept, described by the value w of decision d, is denoted [(d, w)], and it is the set of all examples that have value w for decision d. Let B be a concept and let T be a set of attribute-value pairs. Concept B depends on a set T if and only if

$$\phi \neq [T] = \bigcap_{t \in T} [t] \subseteq B$$

Set T is a minimal complex of concept B if and only if B depends on T and T is minimal.

Let τ be a nonempty collection of nonempty sets of attribute-value pairs.

Set τ is a local covering of B if and only if the following three conditions are satisfied:

1. each member of τ is a minimal complex of B,

2. $\cup \in \tau = T\ [T]\ B$ and

3. τ is minimal (τ has the smallest possible number of members)

For each concept B, the LEM2 algorithm induces production rules by computing a local covering τ . Any set T, a minimal complex which is a member of τ , is computed from attribute-value pairs selected from T (G) of attribute value pairs relevant with a current goal G, i.e., pairs whose blocks have nonempty interaction with G.

The initial goal G is equal to the concept and then it is iteratively updated by subtracting from G the set of examples described by the set of minimal complexes computed so far. Attribute-value pairs from T which are selected as the most relevant,

i.e., on the basis of maximum of the cardinality of [t] ∩ G, if a tie occurs, on the basis of the small cardinality of [t]. The last condition is equivalent to the maximal conditional probability of goal G given attribute-value pair t. For a set X, $|X|$ denotes the cardinality of X [15].

The procedure of LEM2 is as follows:

```
begin
    G:=B;
    τ := φ;
    while G ≠ φ
    begin
        T := φ;
        T (G) := {t | [t] ∩ G ≠ φ};
        while T = φ or [T] ⊃ B
        begin
            select a pair t ∈ T(G) such that |[t] ∩ G| is maximum; if a tie occurs, select a pair t ∈ T(G) with the
            smallest cardinality of [t];
        if another tie occurs, select first pair;
        T := T ∪ {t};
        G := [t] ∩ G;
        T(G) := {t | [t] ∩ G ≠ φ};
        T(G) := T(G) – T;
        end {while};
        for each t ∈ T do
                if [ T – {t} ] ⊆ B then T := T – {t};
        τ := τ ∪ {T};
        G:= B - ⋃_{T∈τ} [T];
    end {while};
    for each T ∈ τ do
        if ⋃_{S∈τ–{T}} [S] = B then τ := τ – {T};
end {procedure}
```

The algorithm is run exactly $|d|$ times, where $|d|$ is the number of decision classes. The number of decision classes indicates the number of clusters produced by K-means algorithm. The while loop ($G \neq \varphi$) is performed at most n times because we may have the whole set as the upper approximation to every decision class. Here n is the number of objects in the training set. To select a pair $t \in T(G)$ as the best one, we have to iterate n * m times so that all possible pairs of attributes and values are examined. Here m is four which indicates the number of attributes in the training set. T contains m elements at most and τ contains n elements at most. So the computational complexity of for loop (for each $t \in T$) is m*n. Therefore the total computational complexity of LEM2 is equal to O ($|d|$ * n * (n * m) * (m * n)) which is simplified as O ($|d|$ * m2 * n3).

8. PROPOSED ALGORITHM

In LEM2 algorithm, the rules generated for each cluster is complete and consistent but it doesn't produce all the consistent rules in a cluster because once a consistent rule is discovered, the objects satisfying that rule is eliminated and rules are discovered for the rest of objects.

Due to this the number of rules produced for a particular cluster becomes less and consequently the chances of predicting the customer to the correct cluster becomes less. In order to overcome this disadvantage, the proposed rule induction algorithm produces all the consistent rules and complete rules for the objects in the cluster. Target cluster is the cluster for which rules are generated. Remaining clusters are the clusters other than target cluster. A block of an attribute-value pair t = (a, v), denoted [t], is the set of all examples that for attribute a have value v.

A block of n attribute-value pair t1 = (a1, v1), t2= (a2, v2), and so on, tn = (an, vn) denoted [t1, t2, . . . tn], is the set of all examples that for attribute a1 have value v1, for attribute a2 have value v2, and so on, an have value vn. A block of size 1 has one attribute – value pair. A block of size n has n attribute – value pairs. For a set X, $|X|$ denotes the cardinality of X. The procedure for improved rule induction algorithm is as follows:

```
begin
        U - Set of all objects in the data set
        B - Set of all objects in the target cluster
        C := U – B (set of all objects in U but not in B)
        G := B;

        temp := ϕ ;
        while G ≠ ϕ
        begin
                T(G) := {t | [t] ∩ B ≠ ϕ and | [t] ∩ B | ≠ | [t] ∩ G | and [t] ∩ C = ϕ }
                for each pair t ∈ T(G) do
                        temp := temp ∪ { [t] ∩ B};
                G := B – temp;
        end {while};
end{procedure}
```

In the while loop of G ≠ φ , find [t] having block size 1 and then block size 2 and so on until block size m. Here m indicates four which denotes the number of attributes in the data set. In each iteration of while loop, G contains set of objects whose T(G) contains set of all attribute – value pairs which satisfies the following three conditions:

1. $[t] \cap B \neq \varphi$
2. $|[t] \cap B| \neq |[t] \cap G|$
3. $[t] \cap C = \varphi$

The first condition chooses the pairs whose blocks have nonempty interactions with B. The last condition

chooses the pairs whose blocks have empty interactions with C. The first and last conditions are required to satisfy the consistency property of rule generating algorithm. The second condition chooses the pairs whose cardinality of blocks satisfied in B is not equal to cardinality of blocks satisfied in C.

This condition is required so that the rules generated are not redundant. The covering property of rule generating algorithm is satisfied by choosing $G \neq \varphi$ as the while loop condition. In each iteration of while loop, rules are generated as the attribute-value pairs of t.

The algorithm is run exactly $|d|$ times, where $|d|$ is the number of decision classes. The number of decision classes indicates the number of clusters produced by K-means algorithm. The while loop ($G \neq \varphi$) is performed at most n times because we may have the whole set as the upper approximation to every decision class. Here n is the number of objects in the training set.

To calculate T (G), all the objects are examined so the computation is n. To select a pair $t \in T(G)$ as the best one, we have to iterate n * m times so that all possible pairs of attributes and values are examined. Here m is the number of attributes in the training set. Therefore the total computational complexity of proposed algorithm is equal to O ($|d|$ * n * n * (n * m)) which is simplified as O ($|d|$ * m * n3).

The proposed algorithm is an improved algorithm in terms of time complexity because LEM2 algorithm computation is m times more than the proposed algorithm. RAM usage for both LEM2 and proposed algorithm are same since both requires the entire clustered output values to be in main memory for analysis.

9. EXPERIMENTAL RESULTS

Real data set of the customer transaction is used for the clustering and rule induction algorithms. The data set is collected from a fertilizer manufacturing company. It consists of 12,028 records of customer transaction for a period of three months for 3278 customers. In each transaction, party id, date of purchase, amount of purchase and payment of purchase are used to define R, F, M and P values. For each distinct party id, R is calculated as the interval between the last purchase and present, F is calculated as the number of his/her transaction records, M is calculated as the sum of his/her purchase amount and P is calculated as the average time interval (in terms of days) between his/her payment date and his/her purchase date for each transaction in the data set. The data set now has only four attributes namely R, F, M and P for 3278 customers. The values of R, F, M and P are normalized as given below:

For normalizing R or P

1. The data set is sorted in descending order of the R or P

2. Divide the data set into five equal parts of 20% record in each

3. Assign numbers 1,2,3,4,5 to first, second, third, fourth, fifth part of records respectively

For normalizing F or M

1. The data set is sorted in ascending order of the F or M

2. Divide the data set into five equal parts of 20% record in each

3. Assign numbers 1,2,3,4,5 to first, second, third, fourth, fifth part of records respectively

After normalization, the values of R, F, M and P are from 1 to 5. The normalized data set is now used by k-means clustering algorithm to segment the 3,278 customers into three groups or clusters. The number of actual cluster required is given by the business people. This number is determined by them according to the number of different scheme to be introduced as their promotional activity. Here the company requires three clusters so we segment the customers into three clusters. As a result, cluster1 contains 1,114 customers, cluster2 contains 1,064 customers and cluster3 contains 1,100 customers. LEM2 and proposed rule induction algorithms are used to generate rules for training data (two-third in each cluster). The test data (remaining one-third in each cluster) is given as input for LEM2 and proposed rule induction algorithm to predict the cluster value according to their generated rules for training data. The training and testing data are mutually exclusive. In training data, cluster1 contains 743 customers, cluster2 contains 709 customers and cluster3 contains 733 customers. In test data, cluster1 contains 371 customers, cluster2 contains 355 customers and cluster3 contains 367 customers. The performance criteria for prediction using rule induction algorithms are false positive (FP), false negative (FN), true positive (TP), true negative (TN), sensitivity, specificity, accuracy, precision, positive predictive value (PPV), negative predictive value (NPV), F-measure.

False Positive (FP) is the number of objects that don't belong to a cluster but are allocated to it. False Negative (FN) is the numbers of objects that belongs to a cluster but are not allocated to it. True Positive (TP) is number of objects that are correctly predicted to its actual cluster. True Negative (TN) is the number of objects that get predicted to a cluster but actually don't belong to. Sensitivity is also called as true positive rate or recall. Sensitivity relates to the test's ability to identify positive results. It measures the proportion of actual positives which are correctly identified as such. Specificity relates to the ability of the test to identify negative results. It measures the proportion of negatives which are correctly identified. Accuracy is defined as proportion of sum of TP and TN against all positive and negative results. Positive predictive value or

precision is defined as proportion of the TP against all the positive results (both TP and FP). Negative predictive value is defined as proportion of the TN against all the negative results (both TN and FN) [11].The F-measure can be used as a single measure of performance of the test. The F-measure is the harmonic mean of precision and recall [24]. The formulas are given below: In this system, user is able to draw a signature by using a mouse. This technique involves two stages namely, registration and verification. In the registration stage, user is asked to draw a signature using mouse and the system extract signature either by enlarge or scale down the signature and rotates if needed. This information stored in the database. The sample output of Syukri algorithm he same pattern on the two dimensional grid. Recall based authentication techniques are classified into three main categories namely; pure recall based, cued recall based and hybrid recall based. The overview of pure recall based, cued recall based and hybrid based techniques.

$$\text{Sensitivity or recall} = \frac{TP}{(TP + FN)}$$

$$\text{Specificity} = \frac{TN}{(TN + FP)}$$

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FN + FP + TN)}$$

$$\text{Positive Predictive Value or Precision} = \frac{TP}{(TP + FP)}$$

$$\text{Negative Predictive Value} = \frac{TN}{(TN + FN)}$$

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

It is observed that the k-means clustering algorithm produces nearly 1000 customers in each cluster. So, LEM2 and proposed rule induction algorithms are repeated numerous times where training data (two-third) and test data (one-third) are randomly chosen from the data set such that training and testing data are mutually exclusive. For many runs, it produces the previously seen run value.

So the twenty runs which produces different values are presented in the Table 1. The performance criteria for prediction are calculated for all the twenty cases. The Table 1 shows the false positive, false negative, true positive, true negative produced by the rule induction algorithms for all the twenty cases.

The objective of the rule induction algorithm is to minimize false positive, false negative and to maximize true positive and true negative. From the Table 1 it is observed that the proposed rule induction algorithm has minimum FP, minimum FN, maximum TP and maximum TN for all the twenty cases when compared to LEM2.

Case	False Positive		False Negative		True Positive		True Negative	
	LEM2	Proposed	LEM2	Proposed	LEM2	Proposed	LEM2	Proposed
1	1	0	17	10	1076	1083	2185	2186
2	0	0	14	10	1079	1083	2186	2186
3	1	1	13	7	1080	1086	2185	2185
4	0	0	10	6	1083	1087	2186	2186
5	0	0	14	10	1079	1083	2186	2186
6	1	1	5	2	1088	1091	2185	2185
7	1	1	11	7	1082	1086	2185	2185
8	0	0	8	6	1085	1087	2186	2186
9	0	0	16	10	1077	1083	2186	2186
10	0	0	15	10	1078	1083	2186	2186
11	1	0	12	8	1081	1085	2185	2186
12	0	0	11	7	1082	1086	2186	2186
13	0	0	9	3	1084	1090	2186	2186
14	1	1	10	7	1083	1086	2185	2185
15	0	0	17	9	1076	1084	2186	2186
16	0	0	15	8	1078	1085	2186	2186
17	0	0	8	3	1085	1090	2186	2186
18	1	1	10	6	1083	1087	2185	2185
19	1	0	14	7	1079	1086	2185	2186
20	0	0	11	8	1082	1085	2186	2186

Table 4. FP, FN, TP and TN for rule induction algorithms

Sensitivity, specificity, accuracy, PPV, NPV and F-measure are calculated using formula 5 to 10 respectively for each algorithm in all the twenty cases. The output is tabularized in Table 2 and Table 3. The objective of the rule induction algorithm is to maximize sensitivity, specificity, accuracy, PPV, NPV and F-measure. From the Table 2 and 3, it is observed that the proposed rule induction algorithm has equal or maximum value than LEM2 in all the twenty cases. The proposed algorithm on average achieves 0.439% increase in sensitivity, 0.007% increase in specificity, 0.151% increase in accuracy, 0.014% increase in positive predictive value, 0.218% increase in negative predictive value and 0.228% increase in F-measure when compared to LEM2 algorithm. The percentage increase in each performance criteria might seems to a smaller value but in real data set where customers are in terms of thousands not in hundreds the proposed algorithm has significant improvement than LEM2. For example, the average accuracy obtained using LEM2 is 99.622% and that of proposed algorithm is 99.773%. LEM2 accuracy for 3278 customers is 3265(i.e. 99.622*3278/100) and that of proposed algorithm is 3270(i.e. 99.773*3278/100). Here five more customers are predicted correctly using proposed algorithm when compared to LEM2.

Case	Sensitivity		Specificity		Accuracy	
	LEM2	Proposed	LEM2	Proposed	LEM2	Proposed
1	0.98445	0.99085	0.99954	1.00000	0.99451	0.99695
2	0.98719	0.99085	1.00000	1.00000	0.99573	0.99695
3	0.98811	0.99360	0.99954	0.99954	0.99573	0.99756
4	0.99085	0.99451	1.00000	1.00000	0.99695	0.99817
5	0.98719	0.99085	1.00000	1.00000	0.99573	0.99695
6	0.99543	0.99817	0.99954	0.99954	0.99817	0.99909
7	0.98994	0.99360	0.99954	0.99954	0.99634	0.99756
8	0.99268	0.99451	1.00000	1.00000	0.99756	0.99817
9	0.98536	0.99085	1.00000	1.00000	0.99512	0.99695
10	0.98628	0.99085	1.00000	1.00000	0.99543	0.99695
11	0.98902	0.99268	0.99954	1.00000	0.99604	0.99756
12	0.98994	0.99360	1.00000	1.00000	0.99665	0.99787
13	0.99177	0.99726	1.00000	1.00000	0.99726	0.99909
14	0.99085	0.99360	0.99954	0.99954	0.99665	0.99756
15	0.98445	0.99177	1.00000	1.00000	0.99482	0.99726
16	0.98628	0.99268	1.00000	1.00000	0.99543	0.99756
17	0.99268	0.99726	1.00000	1.00000	0.99756	0.99909
18	0.99085	0.99451	0.99954	0.99954	0.99665	0.99787
19	0.98719	0.99360	0.99954	1.00000	0.99543	0.99787
20	0.98994	0.99268	1.00000	1.00000	0.99665	0.99756
Average	0.98902	0.99341	0.99982	0.99989	0.99622	0.99773

Table 5. Sensitivity, specificity and accuracy for Rule Induction Algorithm

LEM2 produces only the minimal set of rules whereas proposed rule induction algorithm produces all the possible set of consistent rules to describe the records in the cluster. Thus the proposed algorithm characterizes the customers in each cluster clearly by producing all the consistent rules but eliminates redundant or duplicate rules. Since the number of rules to describe the customer is increased, the prediction accuracy is also improved. This statement is proved experimentally by comparing the performance measure. Hence the chances of judging a customer wrongly is reduced and allotting scheme to the customer is done correctly, which help the business to improve their customer life time value.

Though the proposed algorithm produces more rules than LEM2, the computation complexity is m times less than LEM2 algorithm where m indicates the number of attributes considered for analysis. True Positive, True Negative, False Positive and False Negative are the parameters required to calculate the performance criteria measures of prediction.

The complexity of calculating these parameters are linear with respect to the number of generated rules. The number of rules generated for each cluster or segment is very less when compared to the number of customers dealt. So this calculation complexity is negligible when compared to the rule induction algorithm complexity. Thus, the proposed algorithm is an improved algorithm in terms of cost benefit analysis.

Case	PPV		NPV		F-measure	
	LEM2	Proposed	LEM2	Proposed	LEM2	Proposed
1	0.99907	1.00000	0.99228	0.99545	0.99171	0.99540
2	1.00000	1.00000	0.99364	0.99545	0.99355	0.99540
3	0.99907	0.99908	0.99409	0.99681	0.99356	0.99633
4	1.00000	1.00000	0.99545	0.99726	0.99540	0.99725
5	1.00000	1.00000	0.99364	0.99545	0.99355	0.99540
6	0.99908	0.99908	0.99772	0.99909	0.99725	0.99863
7	0.99908	0.99908	0.99499	0.99681	0.99449	0.99633
8	1.00000	1.00000	0.99635	0.99726	0.99633	0.99725
9	1.00000	1.00000	0.99273	0.99545	0.99263	0.99540
10	1.00000	1.00000	0.99318	0.99545	0.99309	0.99540
11	0.99908	1.00000	0.99454	0.99635	0.99402	0.99633
12	1.00000	1.00000	0.99499	0.99681	0.99494	0.99679
13	1.00000	1.00000	0.99590	0.99863	0.99587	0.99863
14	0.99908	0.99908	0.99544	0.99681	0.99495	0.99633
15	1.00000	1.00000	0.99228	0.99590	0.99216	0.99587
16	1.00000	1.00000	0.99318	0.99635	0.99309	0.99633
17	1.00000	1.00000	0.99635	0.99863	0.99633	0.99863
18	0.99908	0.99908	0.99544	0.99726	0.99495	0.99679
19	0.99907	1.00000	0.99363	0.99681	0.99310	0.99679
20	1.00000	1.00000	0.99499	0.99635	0.99494	0.99633
Average	0.99963	0.99977	0.99454	0.99672	0.99430	0.99658

Table 6. PPV, NPV, F-measure for Rule Induction Algorithms

10. CONCLUSION

Customer relationship management is a technology which helps the entrepreneur to improve their business volume by improving customer relationship. The customer identification is the important phase in CRM. It involves in segmenting the customers and analyzing their behavior for further customer attraction, retention and development. In this paper clustering technique in data mining has been used for customer segmentation and rule induction is used for describing customer behavior in each segment. The entrepreneur can employ different benefit schemes for customer in different clusters or segments. So, classifying a customer to the cluster plays an important role in CRM. For a good rule induction algorithm, the customer's behavior in each cluster should be correctly characterized so that the new customers are predicted to the appropriate cluster. The performance evaluation criteria are chosen based on the prediction accuracy of rule induction algorithm. The proposed algorithm on average achieves 0.439% increase in sensitivity, 0.007% increase in specificity, 0.151% increase in accuracy, 0.014% increase in positive predictive value, 0.218% increase in negative predictive value and 0.228% increase in F-measure when compared to LEM2 algorithm. It has been proved that the time complexity of LEM2 is m times more than the proposed algorithm where m indicates the number of attributes chosen for analysis. Thus, it has been evident from the results that the proposed algorithm is an improved rule induction algorithm which produces better performance in prediction and has less computation when compared to LEM2 algorithm.

11. REFERENCES

[1] A. Berry and G. S. Linoff, Data Mining Techniques for Marketing, Sales and Customer Relationship Management. New Jersey: Wiley Publishers, 2008.

[2] A. J. Berson, S. Smith, and K. Thearling, Building Data Mining Applications for CRM. New York: McGraw-Hill Edition, 2000.

[3] A. K. Beynon and M. J. Peel, Variable precision rough set theory and data discretisation: An application to corporate failure prediction, Omega, vol. 29, no. 6, pp. 561-576, 2001.

[4] J. Blaszczynski, S. Greco, and R. Slowinski, Multi-criteria classification – A new scheme for application of dominance-based decision rules, European Journal of Operational Research, vol. 181, no. 3, pp. 1030-1044, 2007.

[5] M. Bottcher, M. Spott, D. Nauck, and R. Kruse, Mining changing customer segments in dynamic markets, Expert Systems with Applications, vol. 36, no. 3, pp. 155-164, 2009.

[6] O. C. Chan, Incremental learning of production rules from examples under uncertainty: A rough set approach, International Journal of Software Engineering and Knowledge Engineering, vol. 1, no. 4, pp. 439-461, 1991.

[7] O. H. Cheng and Y.-S. Chen, Classifying the segmentation of customer value via RFM model and RS theory, Expert Systems with Applications, vol. 36, no. 3, pp. 4176-4184, 2009.

[8] P. W. T. Ngai, Li Xiu, and D. C. K. Chau, Application of data mining techniques in customer relationship management: A literature review and classification, Expert Systems with Applications, vol. 36, no. 2, pp. 2592-2602, 2009.

[9] S. Huang, E. Chang, and H. Wu, A case study of applying data mining techniques in an outfitter's customer value analysis, Expert Systems with Applications, vol. 36, no. 3, pp. 5905-5915, 2009.

[10] T. Japkowicz and M. Shah, Evaluating Learning Algorithms: A Classification Perspective. Cambridge: Cambridge University Press, 2011.

[11] Z. Pawlak, Rough set, International Journal of Computer and Information Sciences, vol. 11, no. 5, pp. 341-356, 2016.

[12] Z. Pawlak, Rough Sets: Theoretical Aspects of Reasoning about Data. Netherlands: Kluwer Academic Publishers, 2015.

[13] Z. Pawlak, A. Skowron, Rudiments of rough sets, Information Sciences, vol. 177, no. 1, pp. 3-27, 2010.