ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

A Framework for Analyzing the Road Accidents in Data Mining using Rule Mining

Gaganpreet Kaur, Geetika Gandhi

Research Scholar, Dept. of Computer science Engineering, RIMT- Institute of Engineering & Technology, Mandi Gobindgarh, India

Assistant Professor, Dept. of Computer science Engineering, RIMT- Institute of Engineering & Technology, Mandi Gobindgarh, India

ABSTRACT: Road accident is one of the crucial areas of research in India. A variety of research has been done on data collected through police records covering a limited portion of highways. The analysis of such data can only reveal information regarding that portion only; but accidents are scattered not only on highways but also on local roads. A different source of road accident data in India is Emergency Management research Institute (EMRI) which serves and keeps track of every accident record on every type of road and cover information of entire State's road accidents. In this paper, we have used data mining techniques to analyze the data provided by EMRI in which we first cluster the accident data and further association rule mining technique is applied to identify circumstances in which an accident may occur for each cluster. The results can be utilized to put some accident prevention efforts in the areas identified for different categories of accidents to overcome the number of accidents.

KEYWORDS: Data Mining; Road Accidents; Association Rule Mining.

I. INTRODUCTION

Data Mining is a process that discovers the knowledge or hidden pattern from large databases. DM is known as one of the core processes of Knowledge Discovery in Database (KDD). It is the process that results in the discovery of new patterns in large data sets. It is a useful method at the intersection of artificial intelligence, machine learning, statistics, and database systems. It is the principle of picking out relevant information from data. It is usually used by business intelligence organizations, and financial analysts, to extract useful information from large data sets or databases DM is use to derive patterns and trends that exist in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems.
The goal of this technique is to find accurate patterns that were previously not known by us. So, the overall goal of the DM process is to extract information from a data set and transform it into an understandable structure for further use. Organizations like retail stores, hospitals, banks, and insurance companies currently using mining techniques.
DM is primarily used today by companies with a strong consumer focus - retail, financial, communication, and marketing organizations. It enables these companies to determine relationships among "internal" factors such as price, product positioning, or staff skills, and "external" factors such as economic indicators, competition, and customer demographics.

1.1 Over view of data mining

The amount of data stored in computer files and databases is growing at a phenomenal rate. At the same time user of these data are expecting more sophisticated information from them. Data mining is the analysis step of the "knowledge Discovery in databases" process, or KDD). It is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use[17].Data mining is the analysis step of the "knowledge discovery in databases" process, or

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

KDD[18].There are two kind of data mining: predictive and descriptive. These two types have sub types, firstly, predictive such as classification, regression, time series and prediction. Descriptive like as Clustering, summarization, association rules and sequence discovery. The term is a misnomer, because the goal is the extraction of patterns and knowledge from large amounts of data, not the extraction (mining) of data itself.

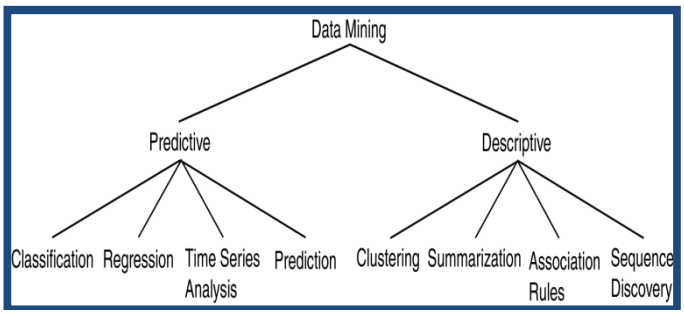

Fig.1. Technique of Data Mining [1]

In data mining communities, there are 3 kinds of mining: data mining, text mining and web mining.[3]Data mining mainly deals with structured data organized in a database, while text mining mainly handles unstructured data or text. Web mining lies in between semi structured and unstructured data. The mining data may vary from structured and unstructured. The basic task of KDD is to extract knowledge (or information) from lower level data (databases).[20]. There are several formal definitions of KDD all agree that the intent is to harvest information by recognizing patterns in raw data. Let us examine definition proposed by Fayyad, Piatetsky-Shapiro and Smyth, "Knowledge Discovery in Databases is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.[21].During the preprocessing stage the data is cleaned. This involves the removal of "outliers" if appropriate; deciding strategies for handling missing data fields; accounting for time sequence information, and applicable normalization of data.[22].The data mining component of KDD often involves repeated iterative application of particular data mining methods. "For example, to develop an accurate, symbolic classification model that predicts whether magazine subscribers will renew their subscriptions, a circulation manager might need to first use clustering to segment the subscriber database, then apply rule induction to automatically create a classification for each desired cluster."[23].

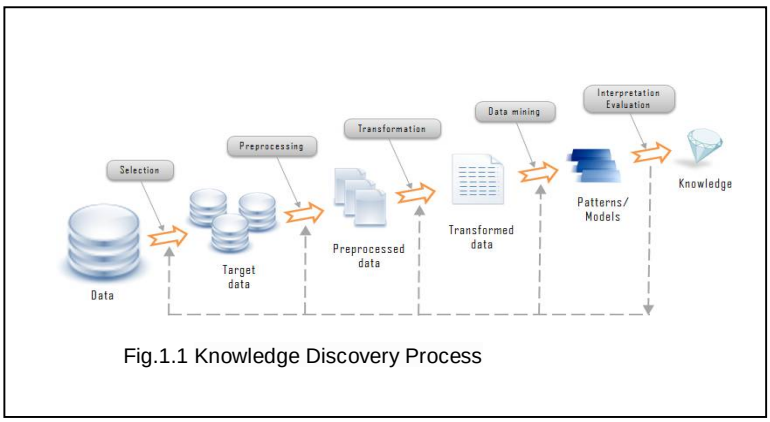

Fig.1.1 Knowledge Discovery Process

1.2 Road accidents in data mining

Road and accidents are uncertain and unsure incidents. In today's world, traffic is increasing at a huge rate which leads to a large numbers of road accidents. The highway safety is being compromised and there are not enough safety factors by which we can analyze the traffic collisions before it happens. A method is proposed by which we can preprocess the accidental factors. Young drivers tend to be more daring and are unable to avoid a crush when they face one. They tend

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

to be more daring after drinking alcohol at night and this causes them to lose control of the car. Drunk driving will not only risk a person's own life but may also cause an incident life to be lost. Number of factors contributes to the risk of collision, including vehicle design, speed of operation, road design, road environment, and driver skill, impairment due to alcohol or drugs, and behavior, notably speeding and street racing. Worldwide, motor vehicle collisions lead to death and disability as well as financial costs to both society and the individuals involved. Road injuries occurred in about 54 million people in 2013[14]. This resulted in 1.4 million deaths in 2013, up from 1.1 million deaths in 1990[15]. About 68,000 of these occurred in children less than five years old. [15] .Almost all high-income countries have decreasing death rates, while the majority of low-income countries have increasing death rates due to traffic collisions.[14] Middle-income countries have the highest rate with 20 deaths per 100,000 inhabitants, 80% of all road fatalities by only 52% of all vehicles.[14] While the death rate in Africa is the highest (24.1 per 100,000 inhabitants), the lowest rate is to be found in Europe. [16]Road and traffic accidents defined by a set of variable .the major issue are analysis of accidents data is its varied nature. The diverse must be considered analysis of the data. So those researchers are used clustering analysis. The clustering analysis is a important technique .cluster analysis are useful to various task. Dr. R. Geetha Ramani. S. Shanthi2 [4] used predicated model technique. In this paper technique algorithm are applied on the like random tree c4.3 tree and j4.3. In this paper researcher are discussions about classifier and predication technique to data mining. In this paper predication of road accident patterns related to pedestrian characteristics. This classifier is voided using cross validation with k folds and evaluated using the accuracy measures: precision and recall and roc. In this paper random tree classifier are given to better result as compared to decision stump [13]. Seoung-hun Park and Young-gunk Ha is used imbalance technique and maps reduce algorithm .imbalance data means data that have a huge difference between the obverted sizes from one data set. So researcher are solved the problem to used sampling technique. There are two type sampling are: over sampling and under sampling .over sampling is to use all observation value in a big class and increase size of observation value in a small class and use this value. Under sampling are those who used lost data. Other problem are occur the researcher are data processing[12] .in this case training set of data make multiple feature .it take time so much so that researcher are solve the problem in map reduce algorithm map reduce algorithm are used to big processing technique.

II. ASSOCIATION RULE MINING

Association Rule Mining [6] is a DM function that discovers probability of co-occurrence of items in a collection. Association rule mining, one of the most important and well researched techniques of DM, was first introduced in [6]. It aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories.The formal statement of ARM problem was firstly stated in [6] by Agrawal. Let I is item set of m distinct attributes, $I = \{I_1, I_2, \ldots, I_m\}$ and D is database (transaction set), $D=\{T_1,T_2,\ldots T_N\}$, where $T \subseteq I$ and there are two item sets X and Y, such that $X \subseteq T$ and $Y \subseteq T$, then association rule, X=>Y holds where $X \subset I$ and $Y \subset I$ and $X \cap Y = \emptyset$. X is called antecedent while Y is called consequent; the rule means X implies Y. There are two important basic measures for association rules, support(s) and confidence(c). Since the database is large and users concern about only those frequently purchased items, usually thresholds of support and confidence are predefined by users to drop those rules that are not so interesting or useful. The two thresholds are called minimal support and minimal confidence respectively. The two basic parameters of Association Rule Mining (ARM) are: support and confidence.

Support(s) of an association rule is defined as the percentage/fraction of transactions in D that contain $X \cup Y$. Support(s) is calculated by the following formula:

$$\sup(X \cup Y) = \frac{count(X \cup Y)}{count(D)} \qquad \text{Eq no.....} (1)$$

Confidence of an association rule is defined as the percentage/fraction of the number of transactions in D that contain X also contains Y. IF component is the Antecedent and THEN component is called consequent. It is calculated by dividing probability of items occurring together by probability of occurrence of antecedent. Confidence is a measure of strength of the association rules. Confidence(c) is calculated by the following formula:

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

$$conf(X => Y) = \frac{\sup(X \cup Y)}{\sup(X)} \text{Eq no.}\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

ARM is to find out association rules that satisfy the predefined sup_{min} and $conf_{min}$ from a given database [6]. Objective of ARM is to find universal set S of all valid association rule.

The problem is usually decomposed into two sub problems.

- One is to find those item sets whose occurrences or support exceed a predefined threshold (sup_{min}) in the database, those item sets are called frequent or large item sets.
- The second problem is to generate association rules from those large item sets with the constraints of $conf_{min}$.

Algorithms for Association Rule Mining

The first sub-problem can be further divided into two sub-problems: candidate item sets generation process and frequent item sets generation process. We call those item sets whose support exceed the support threshold as large or frequent item sets, those item sets that are expected or have the hope to be large or frequent are called candidate item sets. Various algorithms are proposed for ARM:

1. Apriori Algorithm

Apriori involves a phase for finding patterns called frequent item sets. A frequent item set is a set of items appearing together in a number of database records meeting a user-specified threshold. Apriori employs a bottom-up search that enumerates every single frequent item set. This implies in order to produce a frequent item set of length; it must produce all of its subsets since they too must be frequent. This exponential complexity fundamentally restricts Apriori-like algorithms to discovering only short patterns.

2. *FP-tree Algorithm*

FP-tree-based algorithm is to partition the original database to smaller sub-databases by some partition cells, and then to mine item sets in these sub-databases. Unless no new item sets can be found, the partition is recursively performed with the growth of partition cells. The FP-tree construction takes exactly two scans of the transaction database. The first scan collects the set of frequent items, and the second scan constructs the FP-tree. Many other approaches have been introduced in between with minute changes. But main among them and which are basis for new upcoming algorithms is Apriori and FP-tree Algorithm.

III. PROBLEM FORMULATION

There are several major data mining techniques which have been developed and used in various data mining projects. In the proposed work, k –means performance will be enhanced by using hybrid approach for better result to show the effect of noise on the performance of various clustering techniques. Clustering may be applied on database using various approaches, based upon distance, density hierarchy and partition. Clustering is being widely used in many application including medical, finance etc. our purpose is to study how a particular clustering technique is responsive to the noise in the term of time. Apriori algorithm minimum support is needed to generate the large item set from candidate set in which not so required candidate item sets are pruned by utilizing user defined minimum support threshold. Moreover, in Apriori Algorithm, if the frequencies of items vary a great deal, we will encounter two problems. First of all, if minimum support is set too high, those rules that involve rare items will not be found. Secondly, to find rules that involve both frequent and rare items, min support has to be set very low. This may cause combination explosion because those frequent items will be associated with one another in all possible ways. So, Apriori is utilizing hit and trial method to find the required number of rules.

IV. PROPOSED WORK

The research involves exploring various data classification

1. Collection of raw data and then apply filtering techniques to make that raw data into structured format: Filtering techniques like replace Missing Value Filter

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

2. Enhanced K Means Clustering algorithm

1. The size of cluster is fixed and the output of the first phase forms initial clusters. Here, the input array of elements is scanned and split up into sub –arrays, which represent the initial clusters.

2. The cluster size varies and the output of this phase is the finalized clusters. Initial clusters are inputs for this phase. The centered of these initial clusters are computed first, on the basis of which distance from other data elements are calculated. Furthermore, the data elements having less or equal distance remains in the same cluster otherwise they are moved to appropriate clusters. The entire process continues until no changes in the clusters are detected.

3. Enhanced Apriori algorithm

The main improvement of our algorithm is to optimize the frequent single items and those items co-occurrence with them.

The data structure Bit table is also used horizontally and vertically to calculate the token array and count supports, respectively. Token array and the corresponding computing method are proposed. By computing the token, those item sets that co-occurrence with representative item can be identified quickly. The frequent item sets, including representative item and having the same support as representative item, can be identified directly by connecting the representative item with all the combinations of items in its subsume token Thus, the cost for processing this kind of item sets is lowered, and the efficiency is improved

V. RESULTS

This paper selects data sets of traffic accident to test the efficiency of the hybrid clustering algorithm and the k-mode and rule mining algorithm. Results been carried out to demonstrate the performance efficiency of the hybrid k-mode algorithm with improved apriori in this paper.

No of clusters	K-Modes clustering	Hybrid clustering
2	380	160
3	192	120
4	150	100
5	240	140
6	160	110

Table 1: Execution Time comparison of K-Modes and Proposed Hybrid Clustering

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

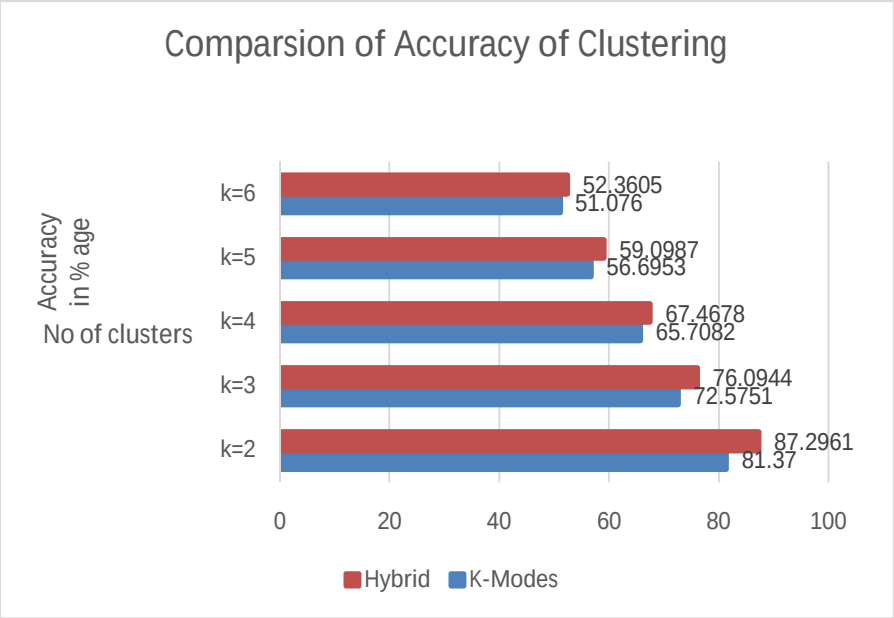

Fig 2: showing the comparison of clustering accuracy between kmeans and hybrid clustering algorithm

Min Support	Apriori	Improved Apriori
0.3	1544	94
0.4	995	75
0.5	655	68
0.6	807	91

Table3: Execution Time comparison of Apriori and Proposed Technique

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

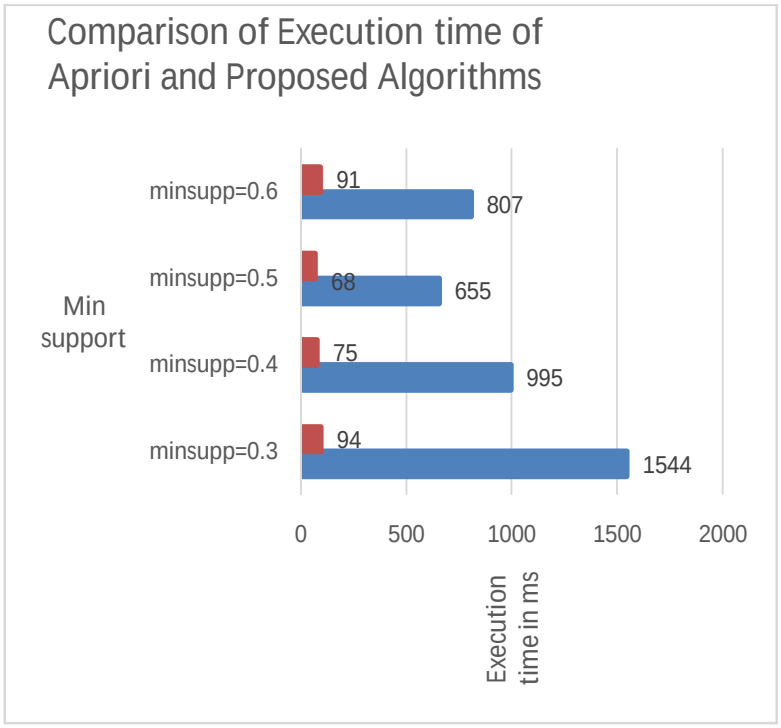

Fig 3: showing the comparison of execution time of Apriori and improved apriori algorithm

Min Support	Apriori	Improved Apriori
0.4	80.5	86.2
0.5	82.4	88
0.6	93	97.7
0.7	95.6	99.6

Table 4: Accuracy comparison of Apriori and Proposed Technique

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

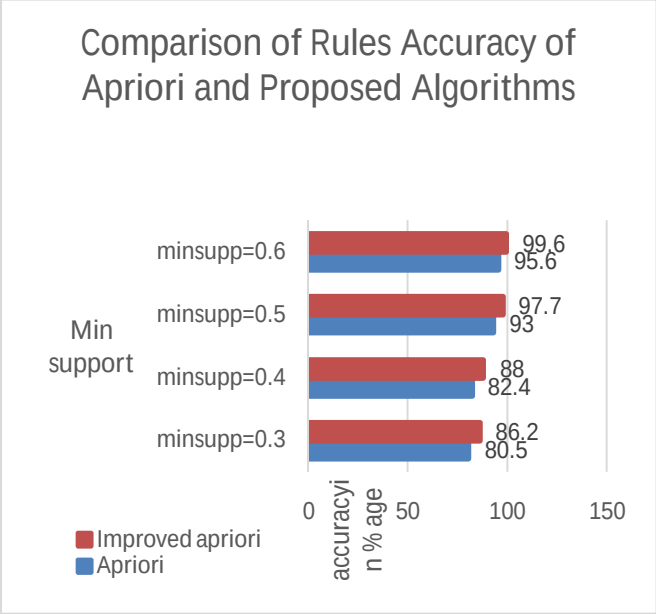

Figure 4: Showing the Accuracy comparison of Apriori and Proposed Technique

VI. CONCLUSION

To overcome the above issues, in this paper, we proposed a framework for analysing accident patterns for different types of accidents on the road which makes use of Hybrid clustering and improved association rule mining algorithm. Comparing and analysing the results of proposed technique with K means clustering and Apriori algorithm on the basis of clustering time, accuracy and association rule mining time.

ACKNOWLEDGMENT

The paper has been composed with the kind assistance, guidance and support of my department who have helped me in this work. I would like to thank all the people whose encouragement and support has made the fulfillment of this work conceivable.

REFERENCES

1. Seoung-hun Park and Young-guk Ha, "Large Imbalance Data Classification Based on Map Reduce for Traffic Accident Prediction", Eighth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing, pp45-49,2014
2. Evangels Gakis1, Dionysius Kehagias1 and Dimitrios Tzovaras1,"Mining Traffic Data for Road Incidents Detection" ,IEEE 17th international conference on Intelligent Transportation System(ITSC) October 8-11,2014 Qingdao, Chine.
3. An Shi, Zhang Tao, Zhang Xining, Wang Jian,"Evolution of Traffic Flow Analysis under Accidents on Highways Using Temporal Data Mining ", Fifth International conference On Intelligent System Design And Engineering Apllication,2014, pp-454-457.
4. http://en.wikipedia.org/wiki/Web_mining [5 3]Performance Evaluation of Clustering Algorithms, R.C Mishra, Professor, CSE Department M.D.U Rohtak, Haryana, India, International Journal of Engineering Trends and Technology (IJETT) - Volume4 Issue7- July 2013.
5. Survey Paper on Clustering Techniques, Navneet Kaur, International Journal of Science, Engineering and Technology Research (IJSETR) Volume 2, Issue 4, April 2013.
6. Omer Adel Nasser and Dr. Nedhal A. Al saiyd ,"The integrating between web usage mining and data mining techniques" 2013 5th International conference on computer science and information technology (CSIT)

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

7. Prashant vats and Manju Mandot "A comparative analysis of various cluster detection techniques for data mining" 2014 international conference on electronic systems, signal processing and computing technologies.
8. Aminder kaur and pankaj kumar "Effect of noise on the performance of clustering techniques" 2010 international conference on networking and information technology.
9. Chintan R. vinegar, Nirali N. Madhak, Trupti M. Kondinariya and jayesh N. Rathod"Web usage mining: A Review on process, Methods and techniques".
10. S. Shanthi, R.geethaRamani "feature relevance analysis and classification of road traffic accident data through data mining techniques",2012 pg no 24-26
11. Dr. R. Geetha Ramani1, S. Shanthi2."Classifier Prediction Evaluation in Modeling Road Traffic Accident Data",IEEE International Conference on Computational Intelligence and Computing Research,pp 1-4,2012
12. Global Burden of Disease Study 2013, Collaborators (22 August 2015). "Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013.". Lancet (London, England) 386 (9995): 743–800.*PMID 26063472*
13. GBD 2013 Mortality and Causes of Death, Collaborators (17 December 2014)."Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013."
14. Global status report on road safety 2013: Supporting a decade of action (PDF) (in English and Russian). Geneva, Switzerland: world health organization WHO. 2013.ISBN 978 92 4 156456 4. Retrieved 3 October 2014.
15. "Data Mining Curriculum". ACM SIGKDD. *2006-04-30*. Retrieved 2014-01-27. *Clifton, Christopher (2010).* "Encyclopedia Britannica: Definition of Data Mining". Retrieved 2010-12-09
16. Fayyad, Osama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic (1996). "From Data Mining to Knowledge Discovery in Databases" (PDF). Retrieved 17 December 2008.
17. Han, Jiawei; Kamber, Michelin (2001). Data mining: concepts and techniques. Morgan Kaufmann. p. 5. ISBN 978-1-55860-489-6. Thus, data mining should have been more appropriately named "knowledge mining from data," which is unfortunately somewhat long..
18. Fayyad, U.; Simoudis, E.; "Knowledge Discovery and Data Mining Tutorial MA1" from Fourteenth International Joint Conference on Artificial Intelligence (IJCAI-95) July 27, 1995 www-aig.jpl.nasa.gov/public/kdd95/tutorials/IJCAI95-tutorial.html
19. Fayyad, U.; Piatetsky-Shapiro, G; Smyth, P; "From Data Mining to Knowledge Discovery: An overview" in Advances in Knowledge discovery and Data Mining. Fayyad, U.; Piatetsky-Shapiro, G; Smyth, P; Uthurusamy, R. MIT Press. Cambridge, Mass. 1996 pp. 1-36
20. Fayyad, U. "Data Mining and Knowledge Discovery: Making Sense Out of Data" in IEEE Expert October 1996 pp. 20-25
21. Simoudis, E. "Reality Check for Data Mining" in IEEE Expert October 1996 pp. 26-33