A Survey on Predictive Data Mining

[1]V.Parimala
[1]Assistant Professor,
[1]Department of Computer Science,
[1]Dr.Umayal Ramanathan College for Women, Karaikudi, India.

Abstract : In this paper, the author focuses towards data mining problems and the need for new techniques and algorithms. In this broad area, to extract the knowledgeable data many techniques were followed. Their importance was also discussed. Among the various tasks to extract the useful knowledgeable data, the paper focuses towards predictive analysis in data mining. They have the capability to compare past success and failures and use those results to predict the future outcomes. A detailed survey of various predictive models and some survey papers are also analyzed. Various data mining techniques and tools, that are used in the extraction of knowledgeable data was also presented.

***IndexTerms* - Data Mining, Knowledgeable data, Data mining techniques, Classification, Clustering, Prediction**

I. INTRODUCTION

On looking into the current trends in technology [15], there was a flood of data from various fields, especially in banking, telecom and other business transactions. Apart from this, data was also generated from research experiments, medical and personal data, Surveillance video and pictures, Digital media, astronomy, biology, physics, etc. Some data were also generated from the web. On looking into such a repository of data, there is a need to extract the knowledgeable data.

In this context of extracting knowledgeable data, researchers face various problems that need to be addressed. The knowledgeable data in the sense, it includes novel, useful and understandable patterns in data. The technique is applicable to situations having the following characteristics: the place where knowledge based decisions is required, changing environment, sub-optimal current methods, having accessible, sufficient and relevant data.

In order to discover the data various steps were included, that involves

- understanding the problem
- Data preparation
- Modelling
- Evaluation
- Deployment.

To carry out all these steps, multiple tasks are available, that can be utilized relevant to the problem in the particular area. They were

- Classification
- Clustering
- Associations
- Visualization
- Summarization
- Deviation detection
- Prediction
- Link Analysis.

Data Mining [11] is nothing but the application of various methods to large and complex database. Its main motive is to reduce the randomness and determine the hidden pattern. In order to reveal these patterns, various mining tools and methodologies were implemented. The main use of data mining in all these fields was its automated prediction of trends and behaviors and automated discovery of previously unknown patterns.

Various data mining techniques are available to extract the knowledgeable and unknown patterns. They were Artificial neural networks, Decision trees, Genetic algorithm, nearest neighbor method, etc.

With the large volume of data [13], retrieved from business and from various fields most of the organizations were aiming to obtain analytic solutions. This is mainly to achieve better decision making. It is categorized into three areas as, Descriptive analytics, predictive analytics and Prescriptive analytics.

Descriptive Analytics: They summarize the raw data and make it easily interpretable to human beings. The approach is mainly useful, since they allow us to learn from past and understand their influence on future outcomes.

Predictive Analytics: These analytics have the capability to predict, what may happen in the future? This approach is mainly based on probabilities. One of main application of predictive analytic approach is in financial services. It is used to determine the probability of customers making future credit payments on time.

Prescriptive Analytics: The approach prescribes a number of actions that move towards a solution. They quantify the effect of future decisions and provide possible outcomes, before making decisions. They go beyond descriptive and predictive analytics. This approach use a combination of techniques and tools such as business rules, algorithms, machine learning, etc.

Among the various categories of analytics [12], our research focuses towards predictive analysis. There are various approaches available that includes,

- Regression Analysis
- Choice Modelling
- Rule Induction.
- Network/Link Analysis.

The needed parameters to do predictive analysis were the expected output, target variable, historical data and the predictors.

II. MOTIVATION

There is a wide variety of data around us in the world, which need to be reformed into useful and knowledgeable data. It is a promising field in the area of research, having numerous applications. Since it is used in various fields, there exists the need to use those vast varieties of data in a useful manner. Even though various techniques are available to extract the data, the area of research is in need of new techniques and algorithms to carry out the process. Hence, our research is focused in this area to extract knowledgeable data.

III. LITERATURE REVIEW

Chonho Lee [1], in 2017 examined the challenges in designing algorithms and systems for healthcare analytics and applications, which is followed by a survey on various relevant solutions. They also discussed next-generation healthcare applications, services and systems, which are related to big healthcare data analytics. Next-generation healthcare systems were expected to integrate various types of EHR data and provide a holistic data-driven approach to predict and pre-empt illnesses, improve patient care and treatment, and ease the burden of clinicians by providing timely and assistive recommendations. They too discussed several applications that were likely to attract attention and interest in the near future.

Fatimetou Zahra Mohamed Mahmoud [2], in 2017, described the purpose of predictive analytics use in many industries and how it is used as a solution to solve many problems in different industries.The predictive analytics have many benefits such as reduce and prevent risk, save time, cost and better management of resources in addition to the ability to take better strategic decisions based on fact not on intuition. Furthermore, the research showed challenges to get real, sufficient and clean data to test models that were developed. It also illustrated the weaknesses in the research area such as the focus on the development of models only, the wrong choice of models variables and algorithms which affect the final results of predictions, and what can be improved in the future research.

Kavya.V, Arumugam.S [3], in 2016 made a review on predictive analytics in data mining. The main process of data mining is to collect, extract and store the valuable information. In advanced analytics, Predictive analytics was one of the branch which is mainly used to make predictions about future events. The two main objectives of predictive analytics were Regression and Classification. It is composed of various analytical and statistical techniques that were used for developing models which predicts the future occurrence, probabilities or events. It deals with both continuous changes and discontinuous changes. It provides a predictive score for each individual to determine, or influence the organizational processes which pertain across huge numbers of individuals, like in fraud detection, manufacturing, credit risk assessment, marketing, and government operations including law enforcement.

Mennatallah El-Assady [4], in 2014, presented two original approaches for visual-interactive prediction of user movie ratings and box office gross after the opening weekend, as designed and awarded during VAST Challenge 2013. Their approaches were driven by machine learning models and interactive data exploration, respectively. They considered an array of different training data types, including categorical/discrete data, time series data, and sentiment data from social media. Their two approaches were only first steps towards visual-interactive prediction, but have delivered improved prediction results as compared to baseline non-interactive prediction, and served as starting points for other predictive applications. Furthermore, they derived an abstract workflow for predictive visual analytics. They also discussed promising challenges for future research in visual-interactive predictive analysis, including design space, evaluation, and model visualization.

Mohammad Ahmad Alkhatib et al [5], in 2015, provided a deep analysis on research in the field of healthcare data analytics, as well as highlighted some of guidelines and gaps in previous studies. Their study has focused on searching relevant papers about healthcare analytics by searching in seven popular databases such as google scholar and springer using specific keywords, in order to understand the healthcare topic and conducted their review. Their paper too has listed some data analytics tools and techniques that have been used to improve healthcare performance in many areas such as: medical operations, reports, decision making, and prediction and prevention system. Their systematic review showed an interesting demographic of fields of publication, research approaches, as well as outlined some of the possible reasons and issues associated with healthcare data analytics, based on geographical distribution theme.

Nishchol Mishra, Dr.Sanjay Silakari [6], in 2012, made a survey on trends, applications, oppurtunities and challenges of predictive analytics. Predictive analytics uses data-mining techniques in order to make predictions about future events, and make recommendations based on those predictions. The process involves an analysis of historic data and based on this analysis to predict the future occurrences or events. A model may also be created to predict using Predictive Analytics modeling techniques. The form of these predictive models varies depending on the data they are using. Among them Classification & Regression are the two main objectives of predictive analytics. Predictive Analytics is composed of various statistical & analytical techniques that were used to develop models which will predict future occurrence, events or probabilities. Predictive analytics was able to deal not only with continuous changes, but discontinuous changes as well. Classification, prediction, and to some extent, affinity analysis constitute the analytical methods employed in predictive analytics.

Pooja Mittal, Nasib Singh Gill [7], in 2013, made an analytical survey on predictive data mining approaches on clinical dataset. The clinical dataset processing was one of the effective and most sensitive areas which were studied under an expert environment. Their paper discusses KDD, data mining with reference to clinical expert system analysis, different applications and the approaches that can be used for the predictive data mining in same area. The scope of their paper is confined to the prediction of a person disease, based on symptoms dataset. The strength of data mining approaches in diverse clinical applications was also analyzed.

Sakshi Rungta, Vanita Jain, Akanksha Utreja [8], in 2015, presented a system to analyse user stories incorporating the data of energy and health demands of four countries – namely India, China, United States of America and Brazil; for the past 30 years. They depicted them graphically using Business Intelligence and finally predicted the future trend of the parameters. The correlation between various entities was then found using Pearson's coefficient. Finally they predicted the values of 30-40 years ahead and predicted the emerging trends in the form of Power View charts.

Shakuntala Jatav, Vivek Sharma, in 2018 [9], proposed an algorithm for predictive data mining approach in medical diagnosis. In their paper they had analyzed prediction systems for Diabetes, Kidney and Liver disease using more number of input attributes. The data mining classification techniques, namely Support Vector Machine (SVM) and Random Forest (RF) were analyzed on Diabetes, Kidney and Liver disease database. The performance of these techniques was then compared, based on precision, recall, accuracy, f_measure as well as time. As a result of study they proposed an algorithm that is designed using SVM and RF algorithm and their experimental result shows the accuracy of 99.35%, 99.37 and 99.14 on diabetes, kidney and liver disease respectively.

CONCLUSION

In this paper the authors made a survey of various data mining techniques and tools. Among the vast repository of data, our motive is to extract knowledgeable data. Various methods exist for the extraction of data. With the development of various technologies, there is a need to find unique techniques and algorithms to extract useful information. In this broad area of research, our focus is on predictive data mining. The interesting feature of the technique is its capability to predict future outcome by analyzing the past events. In recent years, various soft computing techniques were also used for the extraction of knowledgeable patterns. On considering all these issues the paper is focused towards predictive data mining. As a future work, a new algorithm is to be devised to predict future outcome in the application area of medical diagnosis.

REFERENCES

1. Chonho Lee, Zhaojing Luo, Kee Yuan Ngiam, Meihui Zhang, Kaiping Zheng, Gang Chen, Beng Chin Ooi and Wei Luen James Yip," Big Healthcare Data Analytics: Challenges and Applications". © Springer International Publishing AG 2017 S.U. Khan et al. (eds.), Handbook of Large-Scale Distributed Computing in Smart Healthcare, Scalable Computing and Communications, DOI 10.1007/978-3-319-58280-1_2
2. Fatimetou Zahra Mohamed Mahmoud, " The Application Of Predictive Analytics:Benefits, Challenges And How It Can Be Improved". International Journal of Scientific and Research Publications, Volume 7, Issue 5, May 2017 549 ISSN 2250-3153
3. Kavya.V, Arumugam.S," A Review On Predictive Analytics In Data Mining". International Journal of Chaos, Control, Modelling and Simulation (IJCCMS) Vol.5, No.1/2/3, September 2016
4. Mennatallah El-Assady, Wolfgang Jentner, Manuel Stein, Fabian Fischer, Tobias Schreck and Daniel Keim, "Predictive Visual Analytics – Approaches for Movie Ratings and Discussion of Open Research Challenges". Manuscript received 31 Mar. 2014; accepted 1 Aug. 2014; date of publication xx xxx 2014; date of current version xx xxx 2014.
5. Mohammad Ahmad Alkhatib, Amir Talaei-Khoei, Amir Hossein Ghapanchi," Analysis of Research in Healthcare Data Analytics". Australasian Conference on Information Systems Al
2015, Sydney Healthcare Data Analytics
6. Nishchol Mishra, Dr.Sanjay Silakari," Predictive Analytics: A Survey, Trends, Applications, Oppurtunities & Challenges". Nishchol Mishra et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 3 (3) , 2012, 4434- 4438. Issn:0975-9646.

7. Pooja Mittal, Nasib Singh Gill, " Study and Analysis of Predictive Data Mining Approaches for Clinical Dataset". International Journal of Computer Applications (0975 – 8887) Volume 63– No.3, February 2013

8. Sakshi Rungta, Vanita Jain, Akanksha Utreja, " Data Mining Engine using Predictive Analytics". International Journal of Computer Applications (0975 – 8887) Volume 121 – No.5, July 2015.

9. Shakuntala Jatav, Vivek Sharma, "An Algorithm For Predictive Data Mining Approach In Medical Diagnosis". International Journal of Computer Science & Information Technology (IJCSIT) Vol 10, No 1, February 2018.

10.https://data36.com/predictive-analytics-101-part-1/

11.https://data-flair.training/blogs/data-mining/

12.https://eluminoustechnologies.com/blog/2018/01/10/approach-to-predictive-analytics/

13.https://halobi.com/blog/descriptive-predictive-and-prescriptive-analytics-explained/

14. https://www.healthcatalyst.com/three-approaches-to-predictive-analytics-in-healthcare

15.https://www.kdnuggets.com/data_mining_course/x1-intro-to-data-mining-notes.html

16. https://www-users.cs.umn.edu/~kumar001/dmbook/index.php