International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

Predictive Algorithm for Trend Detection

Manjunath H N[1], Nayana M[2], Nehataj Anjum Banu P[3], Ramya K[4]

[1] Department of Computer science ENGG, Autonomous under VTU, BMS college of ENGG, Bull temple road, Basavanagudi, Banglore, India

[2] Department of Computer Science ENGG, Autonomous under VTU, BMS college of ENGG, India

[3] Department of Computer Science ENGG, Autonomous under VTU, BMS college of ENGG, India

[4] Department of Computer Science ENGG, Autonomous under VTU, BMS college of ENGG, India

Abstract: ***Trend Detection is the method for the analysis of the different forms of data in various fields such as machine learning, data mining, image processing/analysis. In this process data can be structured or unstructured, in order to work on different variants in data forms clustering is used for the trend detection. Clustering can be defined as a process of grouping entities of similar type into separate groups or it is just method of partitioning of a data set into subsets, so that the data in each subset according to specific defined distance measure. And we predict using the predictive algorithm techniques or methods in "Trend Detection" for collected information in different fields like social media blogs and we collect the major issues and the data that is what's happened previously for that related issues and the present data and we predict using prediction method called Naive Bayes. Our primary aims the accuracy and the quality of recommendations of the predictive component. The idea of trend detection is different in various fields. Survey on each area gives different option for trend detection. The word carried by C.Anuradha1, T.Velmurugan [1] presents a case study on data mining in educational field. Educational data mining is important as it is crucial task to convert the raw data into information to improve learning practices and the teaching methods. The results of study exhibit techniques for feature selection removes the irrelevant and redundant attributes in order to increase the accuracy and efficiency of the algorithm that is used for clustering which is mainly used to improve student performance. Article TrendLearner: Early prediction of popularity trends of user generated content identifies new research problem [2]. Firstly the work is carried out to tackle the problem of early detection of popular trends in UGC. The result explains the extensive experimental evaluation performed by comparing the method with state of the art. It proposes a solution TrendLearner for predicting the popularity for different kinds of data which enabled automated process.***

Keywords: Clustering, Trend detection, Prediction, Grouping

1. Introduction

Nowadays it is very difficult for students to decide which academic project to take up, this project is developed to make student's job easy. Students can give the topic or domain name which they are interested in or input can be taken from any source and user will be suggested with "n" number of solutions or title of the project which can be taken up.

Process of data analysis includes requirement analysis, collecting data, filtering and processing of data, applying algorithms and data results. Data mining is a technique that is being used for data analysis. It focuses on structuring of data and identifying information for prediction. Data can be in several stages from raw data to processed data. Data mining is used in several areas because the data can be processed in timely manner with appropriate algorithms and other technique which can reduce storage space n time. The areas where data mining is served are as follows: Research areas like mathematics, cybernetics, marketing, biotechnology, aerospace etc.

It is interesting task of applying data mining methods in educational sector. It unwrap the data into meaningful information that could be used for knowledge and learning gains. Detailed review [3] of the various disparate entities serves as the bridge for the gap in educational sector.

An extensive set of experiments are presented in paper Empirical Analysis of Predictive Algorithms for Collaborative Filtering[4] regarding the algorithms for predicting the performance and recommender systems. The result indicates wide range conditional factors for the analysis.

In this project Naive Bayes classifiers are also used, Naive Bayes is a method which can be used for building the models that will assign labels to all problem instances. Naive Bayes requires small set of training data for estimating the parameters that are necessary for the classification process. The potential for learning from previous experience in order to predict future events is the intellectual and important task. Process of learning by algorithms has led to huge amounts of research in constructing algorithms that can be used for prediction. [5] In this project, we work on developing these algorithms for trend detection. Algorithms are designed such a way that trends are refined gradually as data is available. Special attention is given to detect sharp changes in trends. The result shows the usefulness of the trend and the results of the trend in future.

Understanding the behavior of amino acids of proteins can be done in two ways. Giving clarity about the type of acids and how it can response with immune system and antibodies which are chemically synthesized. How proteins will improve the immune system is matter of concern. Here this work is about predicting the nature of proteins and amino acids [6]. Heart disease is one of the most critical and threatening disease. It can be caused irrespective of time and age. Due to poor medical decisions and medicines this disease cannot be tackled easily. This paper [7] tells us how

to predict the heart attack and its symptoms. Various algorithms likes k-means are used to collect data and analyze them.

Now a day's any sort of business process require a huge investments. We can find a variety of business today ranging from small to big. Not everyone has sufficient money in hand to start a venture. In such a case banks provides credits, loans to customer who is in need. This paper [8] is about predicting the type of customer and the type of risk associated with it. Here it makes use of 3 models to analyze and predict the risk associated in providing the credit and to get the best results. The research work by R. Camilleri, D.A. Howey [9] explains the method for developing fast and accurate simulation tool for overall temperature distribution prediction It is shown how the temperature distribution can be improved by controlling the flow distribution. Robots are becoming popular in all the fields in the recent times. The speed with which it travels n performs is most important. Moving robots are exclusively used in most of the places. This paper [10] is about finding the speed and velocity with which the robots move. The wheel slippage is given more importance. This paper is to predict the path and the wheel slippage. The objective of this paper[11] is to analyze the damaged parts by using data mining tools. In the combination with Ahanpishegan Co. the analysis has been carried out. First through interviews with managers and employees the remarkable points about percentage of damaged parts will be found out. The prediction algorithms are applied using a merged database, analyzing outliers, avoiding the constant variables and results will be compared. The purpose of this paper is to enhance the authenticity, accessibility and maintainability.

[12] This paper contains information about a method that was developed for hepatitis diagnosis. In this paper genetic algorithm and development search methods are used for growth and naive Bayes classifier is used for purpose of classification. Input is Hepatitis data set, outputs are accuracy and time. The purpose of this paper[13] is to implement a classification method to guide students in guessing their success in admission in an engineering branch by predicting the success of previous years. An algorithm named fast correlation base filter is used, which is very good in taking off the duplicate and features that are not at all relevant from the dataset, thus the calculation time is decreased and predictive efficiency is increased.

This paper[14] prompts a framework for researching the popularity dynamics of end user uploaded videos, characterization of the popularity dynamics will be given and suggests a model that takes the key features of this dynamics. Study of the popularity dynamics is done using a dataset which keeps track of the number of times that videos are viewed to a sample of recently uploaded videos over the first eight months of their life time. We suggest a model which accurately captures the popularity of collections of recently uploaded videos as they grow old.In this paper[15] we prompt a novel method to distinguish and predict the time progression of popularity in end user created information. To capture the patterns of behavior that has display overtime a function has been defined. The results will be showing that the information can be classified accurately based on the popularity and guess how popular the content will be in the future.

2. Body of the Paper

2.1 Purpose

If you only look at one or two aspects of current trends it can be easy to get the wrong impression and it is difficult to analyze the current trend to work on. Projects/Works on Trends in any field keeps changing with increasing or decreasing its value. One will always want to choose or work on the field where there is a increasing necessity and it is open with many different options.

Many use trend detection process to find out how they're doing compared to competitors. Of course, most competitors won't just share this type of information.

It is important to work on deciding the domain or topic before we actually take up the project. To simplify the work of finding out the current projects or challenges on different domains trend detection plays an crucial role.

2.2 Problem Statement

The domains are clustered using the clustering algorithms. The current updates on the clustered domain are displayed for the reference. The current challenges on the clustered domain are exposed. Predictive algorithms are applied on the selected sub domain to provide the existing reviews, projects on the said domain and based on that predict the coming up projects, developments and other add-on.

2.3 Survey Details

2.3.1 Clustering Algorithms

The continuous flow of bits of information in the ordered manner is called as data stream. It is one of the important field in the research department of data mining[16]. The clustering is the powerful tool in the data mining. The development of clustering of data stream algorithms is quite difficult because of it produces unwanted signals and the outliers, therefore here mentioned some of algorithms and techniques to get rid of unwanted signals and the outliers.

2.3.2 Hierarchical clustering

For clustering of Hesitant Fuzzy [17] Sets such as aggregation operations the agglomerative hierarchical clustering algorithms is used for uncertain data and the algorithm defines the every given HFSs as a separate cluster in the first stage, and then compares with the every pair of the HFSs by using the weighted Hamming distance or the weighted Euclidean distance algorithms. At finally the two clusters with smaller distance are jointed. And the above thing is repeated time and again until the appropriate number of clusters is achieved.

2.4 K-Means – Clustering algorithm used

The conventional k-means clustering algorithm usually uses the statistical methods and probabilistic validation

approaches for clustering and the newly generalized k-means[18] clustering algorithm is a part of conventional k-means clustering algorithm and is developed by without mentioning the previous exact cluster number. To view the performance of the students in the colleges the k-means clustering algorithm[19] is used to maintain the structured set of the data such as there gained grades or marks, results, and other student related personal information year by year and it is not only to create the structured form of data and it is also designed a model to analyze the each and every student's results.

2.5 Naïve Bayes - Prediction Algorithm Used

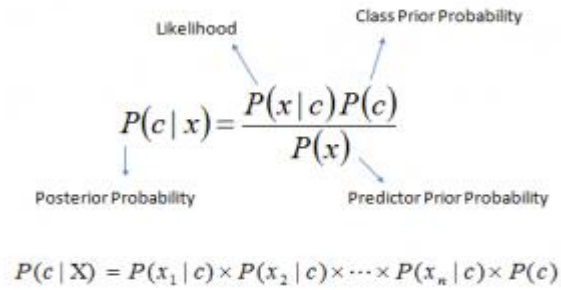

Figure 1

Equation Above gives the basic working:

Posterior Probability - P(c|x) for class (c, target)

Given predictor (x, attributes)

- Prior Probability - P(c)
- Likelihood - P(x|c) (The probability of predictor)
- Prior probability of predictor - P(x)

Example:
Weather Training data
Let Corresponding Target Variable be "Play" (Possibilities of Playing should be suggested)
Now, Classification to be made based on weather condition whether players will play or not.
Steps for performing the same are as follows:
Step 1: Data set to be converted into a frequency table
Step 2: Find the Probabilities like Overcast Probability and Probability of Playing to Create Likelihood table
Overcast probability = 0.29
Probability of playing is 0.64.
Step 3: Use Naive Bayesian equation for calculating the posterior probability of each class.
The class which has highest probability is the result of prediction.

Problem: Players will play if weather is sunny. Is this statement is correct?
Solution:
P(Yes | Sunny) = P(Sunny | Yes) * P(Yes) / P (Sunny)
We have P (Sunny |Yes) = 3/9 = 0.33 and P(Sunny) = 5/14 = 0.36
P(Yes)= 9/14 = 0.64
Now, P (Yes | Sunny) = 0.33 * 0.64 / 0.36 = 0.60, which has higher probability.

In the present trend, the User Generated Content (UGC) such as blogs, wikis, chats, tweets and video and audio clips etc. that was made by the user had got the major place in terms of prediction of popular trends. In this paper[20] describes the prediction of videos i.e. whether the video will spreads or not and how many likes and views will video received that is published in the web.

Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

Frequency Table		
Weather	No	Yes
Overcast		4
Rainy	3	2
Sunny	2	3
Grand Total	5	9

Likelihood table				
Weather	No	Yes		
Overcast		4	=4/14	0.29
Rainy	3	2	=5/14	0.36
Sunny	2	3	=5/14	0.36
All	5	9		
	=5/14	=9/14		
	0.36	0.64		

Figure 2

3. Conclusion & Future Work

In this study, it is explained how predictive algorithms can be applied successfully in estimating of performance variables and critical events. And the main objective of this prediction is that the results should be in accurate. We demonstrate the benefits of our method on a publicly available collection of data sets.

Future work can apply possible ways to unify the mechanism to enrich our understanding of applicability of predictive algorithms for trend detection. Here we also provide estimation on, how long a trend can exist. A graphical representation gives a clear picture on the past experience, current process, future enhancements.

References

[1] C.Anuradha1, T.Velmurugan. "Feature Selection Techniques ToAnalyse Student Acadamic Performance Using Naive Bayes Classifier."The 3rd International Conference on Small & Medium Business 2016.January 19 - 21, 2016, Nikko Saigon Hotel, Hochiminh, Vietnam.
[2] FlavioFigueiredo,JussaraM. Almeida ,MarcosA. Gonçalves ,FabricioBenevenuto. "TrendLearner: Early prediction of popularity trends of user generated content." Elsevier Inc, Information Sciences 349–350 (2016) 172–18.
[3] AshishDutt, SaeedAghabozrgi, MaizatulAkmalBinti Ismail, and HamidrezaMahroeian. "Clustering Algorithms Applied in Educational Data Mining."International Journal of Information and Electronics Engineering, Vol. 5, No. 2, March 2015.
[4] John S. Breese David Heckerman Carl Kadie."Empirical Analysis of Predictive Algorithms for Collaborative Filtering." Microsoft Research Redmond, WA 98052-6399 { breese,heckerma, carlk} @microsoft. Com
[5] R. Vilalta, C. V. Apte, J. L. Hellerstein,S. Ma, S. M. Weiss. "Predictive Algorithms In The ManagementOf Computer Systems." Ibm Systems Journal, Vol 41, No 3, 2002.

[6] Thomas P. Hopp . "Protein Antigen Conformation: Folding Patterns And Predictive Algorithms; Selection Of Antigenic And Immunogenic Peptides." AnnaliSclavoCollanaMonogr, 1984, 1(2), Pp. 47-60.Proceedings Of First International Conference, Siena, Italy, 29-30 October 1984.
[7] SanavarBangi, PoojaGadakh, PradnyaGaikwad, PratikshaRajpure. "Survey paper on Prediction of Heart Disease Using Data Mining Technique."International Journal of Recent Trends in Engineering & Research (IJRTER), Volume 02, Issue 03; March - 2016 [ISSN: 2455-1457].
[8] AnchalGoyal, RanpreetKaur. " Loan Prediction Using Ensemble Technique." International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 3, March 2016.
[9] R. Camilleri, D.A. Howey,M.D. McCulloch. "Predicting the Temperature and Flow Distribution in a DirectOil-Cooled Electrical Machine with Segmented Stator." IEEE Transactions on Industrial Electronics · January 2015
[10] Mohamed Krid, FaizBenamar, and Roland Lenain."A new explicit dynamic path tracking controller using Generalized Predictive Control."International Journal of Control, Automation and Systems.
[11] Golriz Amooee1, BehrouzMinaei-Bidgoli, MaliheBagheri-Dehnavi.. "Comparison Between Data Mining Prediction Algorithmsfor Fault Detection." IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 6, No 3, November 2011.
[12] Karthikeyan.T, Thangaraju.P,"Genetic Algorithm based CFS andNaïve Bayes Algorithm to Enhance the Predictive Accuracy."Indian Journal of Science and Technology, Vol.8, No.26, 2015 pp.1-8.
[13] MitalDoshi, Setu K Chaturvedi,"Correclation Based FeatureSeleciton(CFS) Technique to Predict Student Performance", Int.Journal of Computer Networks & Communications, Vol.6, No.3,2014,pp.197-206.
[14] Y. Borghol, S. Mitra, S. Ardon, N. Carlsson, D. Eager, andA. Mahanti. "Characterizing and modelling popularity ofuser-generated videos". Performance Evaluation, 68(11):1037–1055,Nov. 2011.
[15] Shuxin Ouyang, Chenyu Li," A Peek into the Future: Predicting the Popularity of Online Videos". 2169-3536 (c) 2016 IEEE.
[16] Shifei Ding, Fulin Wu, JunQian, HongjieJia, Fengxiang Jin. "Research on data stream clustering algorithms."Springer Science+Business Media Dordrecht 2013.
[17] XiaoluZhanga and ZeshuiXu."Hesitant fuzzy agglomerative hierarchical clustering algorithms."International Journal of Systems Science, 2013.
[18] Yiu-Ming Cheung. "k_-Means: A new generalized k-means clustering algorithm." Elsevier B.V, Pattern Recognition Letters 24 (2003) 2883–2893.
[19] Oyelade, O. J, Oladipupo, O. O, Obagbuwa, I. C. "Application of k-Means Clustering algorithm forprediction of Students' Academic Performance." (IJCSIS) International Journal of Computer Science and Information Security, Vol. 7, No. 1, 2010.
[20] Flavio Figueiredo," On the Prediction of Popularity of Trends and Hits for User Generated Videos". WSDM'13, February 4–8, 2013, Rome, Italy. Copyright 2013 ACM 978-1-4503-1869-3/13/02