Perspectives in Communication, Embedded-Systems and Signal-Processing (PiCES) – An International Journal
ISSN: 2566-932X, Vol. 3, Issue 1, April 2019

Comparison of Classification Modelling Algorithms in Web Usage Mining

Srujani J
Dept. of Computer Science & Engineering, New Horizon College of Engineering, Bangalore, India,
srujanijagannatha@gmail.com

Priti Badar
Dept. of Computer Science & Engineering, New Horizon College of Engineering, Bangalore, India,
priti_badar@yahoo.co.in

***Abstract:* Web Usage Mining (WUM) includes identification of patterns used and has various empirical approaches. It has evolved into a strong area of analysis in data mining specialization due to critical ethics. It is composed of three stages such as Pre-Processing, Pattern Discovery, Pattern Analysis. Here an experimental differentiation among supervised learning algorithms: Decision Tree Classifier, Naive Bayes Classifier, K Nearest Neighbour Classifier and Support Vector Machine (SVM) is discussed. Classification is one among the mining methods which is concerned about the web dominion. It is used to envision definite class of a particular data set in order to categorize the data to predefined classes. The classifier is a purpose which is used to depict new data to already defined group or category. This paper compares the various classification modelling techniques used to classify web users.**

***Keywords:* Web Usage Mining; Pattern Discovery; Pattern Analysis; Data Mining; Supervised Learning; Classification; SVM; Web Log Mining; Web Log Records.**

I. INTRODUCTION

The approach of data mining that is used to find the utilization samples from web information, which better understands and distributes the requirements of web applications, is known as web usage mining (WUM). It contains three stages such as Pre-Processing, Pattern Discovery and Pattern Analysis. WUM is also known as web log mining (WLM) that decides the usefulness facts about the customer behaviour samples and usage of website related data to ease for different design areas of website. The solution method of analysing WUM contains logs along with few web servers which are matchless in the globe. The log information of users is selectively dissimilar methods like substitute servers, client position and server position. Data Pre-processing results in better management of web log records that consist of cleaning the records, recognition of sessions and customers, position completion of path and transaction identification.

Detecting patterns includes application of various data mining approaches to cleaned information that belongs to statistical learning, interrelation, clustering and sample subsequent. The last step of web usage mining is analysing the patterns which are obtained from the web log records and are strained away. It includes investigation of data as it is occurred in the form of data cubes or structured query language (SQL) to perform the OLAP operations. WUM approaches depend on the information composed from three important things namely web clients, web servers and proxy servers. Web servers are prosperous and most important origin of information. It gathers huge quantity of data in weblog records and also in records that are stored in database used by them. Most of the ISP's provide proxy server facilities to customers which enhance the speed of navigation across caching. The information used by the customers can be recorded by client position with the help of java applets, javascript or browsers. It keeps away the customer problems of session detection along with caching problems.

A method which classifies the information across various groups is known as classification. In case of WUM, it creates a profile for the web users who belongs to a particular group. Here the information is distributed based on few attributes such as how many hours in a day the data was used by the web users. It needs selection and removal of attributes which classify the dataset of a particular group or category. This method is performed with the help of supervised inductive ML algorithms namely decision tree classifier, KNN classifier, naive-bayes classifier and SVM. The machine learning (ML) algorithms distribute web users based on a criterion and allocates it to one group across a set of already defined categories.

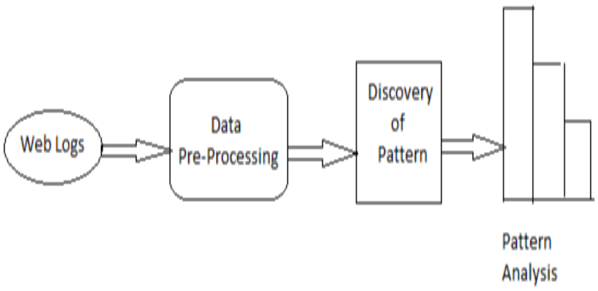


Fig 1. Stages of Web Usage Mining

II. CLASSIFICATION MODELLING ALGORITHMS

This section consists of various supervised inductive machine learning algorithms that are used to perform classification of items i.e., web users to a particular group or category.

A. Decision Tree Classifier

Decision tree classifier is broadly utilized and is a experimental approach that is characterized by inference of supervised information. It constitutes of a strategy which distributes the categorical information based on different features. Decision tree is also used for processing a huge quantity of information that can be used in data mining approaches. For the construction of a decision tree, parameters or the domain knowledge is not necessary. This classifier is abundant and suitable for analytic discovery of knowledge. The characterization of knowledge acquired in the form of a tree is instinctive and simple to perceive. They need less information whereas remaining classifiers need normalized information, duplicate attributes and empty results. The amount required for predicting the data with the help of decision tree is logarithmic which is fed to the tree. It manages various output problems. The performance of decision tree is better even though the assumptions were breached by information from where it was taken.

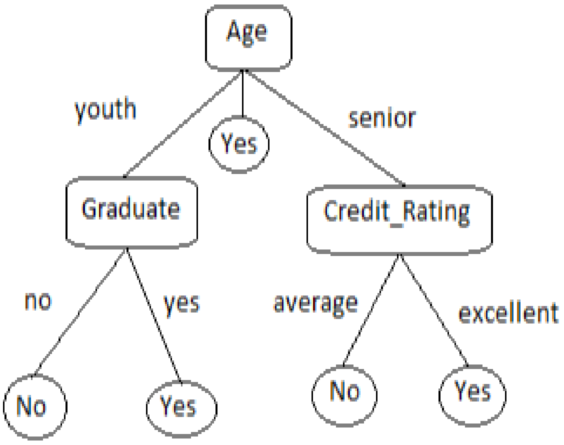
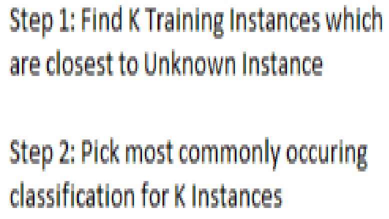


Fig 2. Example of Decision Tree Classifier

B. KNN Classifier

KNN Classifier is an approach for arranging dataset depending upon the nearest training samples in the distinctive area. It is a lazy and instance based classifier where the purpose of it is to estimate locally and the estimation is delayed till the classification. KNN is an easy approach that reserves every occurrence and distributes new occurrences depending upon the correspondence measures. This algorithm is used from 1970 for various approaches such as statistical measures and recognition of samples etc. It keeps the complete training data for perceiving and allocating to every problem a label constituted by many class of its closest neighbour in training data. Here every pattern has to be distributed equally among its surrounding patterns. For an unknown pattern classification, it predict based on the closest neighbour pattern classification. When an unknown pattern and a training data are given, then the distance among unknown pattern and remaining patterns in training data is estimated. K closest neighbour is also known as Case-Based Reasoning, Instance-Based Learning and Lazy Learning.



Step 1: Find K Training Instances which are closest to Unknown Instance

Step 2: Pick most commonly occuring classification for K Instances

Fig 3. K nearest neighbour algorithm

C. Naïve Bayes Classifier

Naive bayes classifier is straightforward algorithm which presumes that classification features are not dependent and does not contain any relation among them. Most of the researchers state that following belief of independence is not applicable in every case due to which remaining approaches are designed to improve the performance. This technique is based upon the probability condition and Max Likelihood appearances. NB classifier is stated as bayes theorem along with probability condition presumption in which all attributes from A_1,.......A_n belonging to group Care not dependent among remaining attributes of C. C is parent node and A_1, ...A_n are child nodes. As A_1, ...A_n are not dependent among remaining attributes of C, we have

$$P\ (A_i \mid C, A_j) = P\ (A_i \mid C)$$

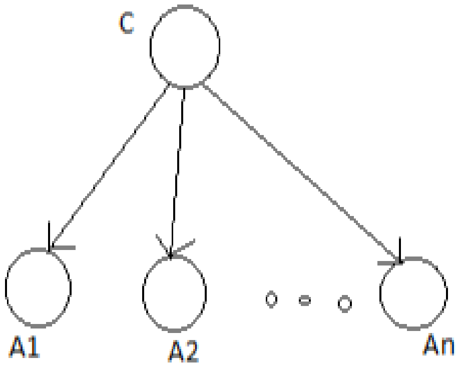


Fig 4. Naive Bayes network

D. Support Vector Machine

Support vector machine (SVM) are efficient of supplying more accomplishment in case of classifying accuracy compared to other information classifying approaches. It is utilized over a large extent in real globe issues namely categorization of text, recognition of hand-written values, voice recognition, classifying images, identification of things and classifying the data. This technique is invariably higher-level when compared to remaining supervised machine learning approaches. In case of few datasets, accomplishment of SVM is too delicate based on how the amount feature and kernel features are fixed. The users have to perform considerable cross execution to check the favourable feature setting. The time and space efficiency of this method was reduced by applying many approaches. The reasonable pattern in the SVM training data is support vector; hence before teaching classifier the support vector (SV) has to be extricated where there is an increase in space and time complexity of classifier.

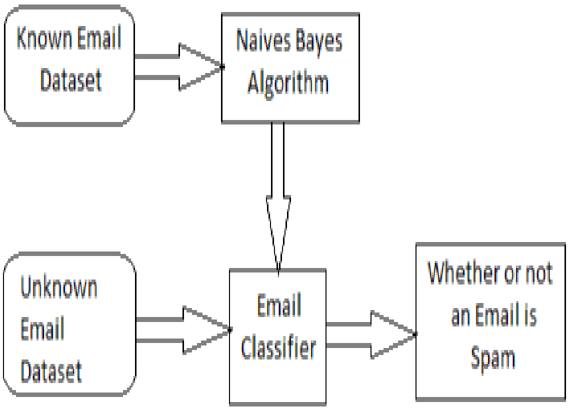
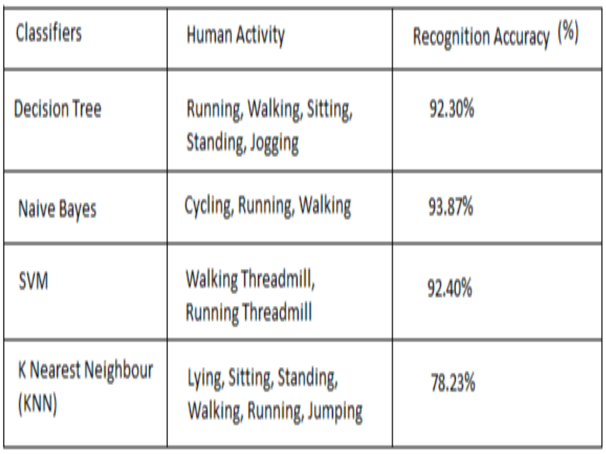


Fig 5. Example of naïve bayes classifier

III. DISCUSSIONS

The comparison between naive bayes, decision tree, k nearest neighbour and support vector machine is as follows. Naive bayes calculates the possibility of training dataset to appear in a particular category. It presumes that every variable is not dependent on others. In naive bayes accomplishment is based upon envision behaviour of the attributes. It is suitable for less quantity of training dataset. Whereas decision tree split-up the information to various data sets this relates the attributes in a better way. It is necessary to trim the tree due to over-fitting. In decision tree there is no need to cover-up or systemize the input. KNN algorithm is strong enough to handle the clattered training data. This algorithm is powerful in case of huge training dataset. The attribute 'K' value has to be estimated which contains the count of closest neighbours. The execution amount is more as the distance from every query to the training dataset has to be estimated. Whereas the support vector machine algorithm has an attribute which solves the over-fitting issue. Here selecting the hyper attributes of support vector machine results in efficient generalization accomplishment.

The above is an example of recognition accuracy obtained for each classifier based on the different types of human activities. From this we can observe that decision tree classifier and support vector machine classifier has more accuracy compared to others. Hence these two classifiers obtain better accuracy in classifying the data and results as the best classifiers.



Classifiers	Human Activity	Recognition Accuracy (%)
Decision Tree	Running, Walking, Sitting, Standing, Jogging	92.30%
Naive Bayes	Cycling, Running, Walking	93.87%
SVM	Walking Threadmill, Running Threadmill	92.40%
K Nearest Neighbour (KNN)	Lying, Sitting, Standing, Walking, Running, Jumping	78.23%

Fig 6. Example of Accuracy obtained for each classifier

IV. CONCLUSIONS

This paper includes comparison on different algorithms used to classify web users in web usage mining. All techniques use different methods to classify which results in better accuracy compared to other methods. In this paper, few issues regarding the classification of data items are also discussed. Hence, as we observed following two classifiers namely decision tree classifier and support vector machine have more accuracy than other algorithms.s

ACKNOWLEDGMENTS

I am really thankful to my teacher Ms. Priti Badar Assistant Professor of Computer Science and Engineering department, NHCE, for her constant support extended towards me during the course of this technical paper. And her help and advice instilled me to complete the paper on time.

REFERENCES

[1] Durgesh K. Srivastava, Lekha Bhambhu, "Data Classification using Support Vector Machine", Journal of Theoretical and Applied Information Technology, JATIT, 2009.

[2] S. Karthika and N. Sairam, "A Naive Bayesian Classifier for Educational Qualification", Indian Journal of Science and Technology, vol 8, july 2015.

[3] M. Aldekhail, "Application and Significance of web usage mining in 21st century: a literature review", International Journal of Computer Theory and Engineering, vol 8, No 1, February 2016.

[4] M. Akhil Jabbar, B L Deekshatulu and Prithi Chandra, "Classification of Heart Disease using KNN", International Conference on Computational Intelligence: modelling techniques and applications, India, 2013.

[5] Anitha Talakokkula, "A Survey on web usage ming, applications and tools", Computer Engineering and Intelligent Systems, vol 6, no 2, 2015.

[6] Mehak, Mukesh Kumar, Naveen Agarwal, "Web usage mining: an analysis", Journal of emerging technologies in web intelligence, vol 5, no 3, august 2013.

[7] Saloni Agarwal, Veenu Mangat, "Application areas of web usage mining", fifth international conference on advanced computing and communication technologies, India, 2015

[8] Rich Caruana, Alexandru Niculescu Mizil, "An Empirical Comparison of Supervised learning algorithms", 23rd international conference on machine learning, Pittsburgh, 2006.

[9] Sadegh Bafandeh Imandoust, Mohammad Bolandraftar, "Application of KNN approach for predicting economic events: theoretical background, International journal of engineering research and applications, vol 3, Issue 5, sept-oct 2013.