

Comparison of Various Machine Learning Classifiers For Clickstream Analysis

Manpreet Dhindsa¹, Roopam Kansal²

School of Computer Science and Engineering

Lovely Professional University

¹manpreet.23551@lpu.co.in, ²roopam.23828@lpu.co.in

Abstract- Growth of internet and its commercialization from past decades make it important to separate and compete with other similar websites especially the E-commerce based websites which sell products directly to the customers. In order to classify the webpages based on feature extraction, five machine learning classifiers have been compared to evaluate the performance in which decision tree gives more accuracy in true classification. The diverse Classifiers compared are: Support Vector Machines (SVM), K-nearest neighbor, Artificial neural network, Naïve-Bayes and Decision trees. Average accuracy for true classification by all classifiers is in between 75% to 99.5% in which decision tree gives almost above 99% percent accuracy in true classification.

Index Terms: Machine learning, click-streaming.

INTRODUCTION

The method of examining data from diverse viewpoints and briefing it into valuable information is the concept of data mining. It is one of the most used analytical tools for examining data. With this, clients can investigate information from numerous measurements or points, classify it, and review the associations recognized. Actually, data mining is the technique for discovering connections or examples among many fields in extensive relational databases. It provisions the alteration of data to information, knowledge and perception, a cycle that is considered in every association.

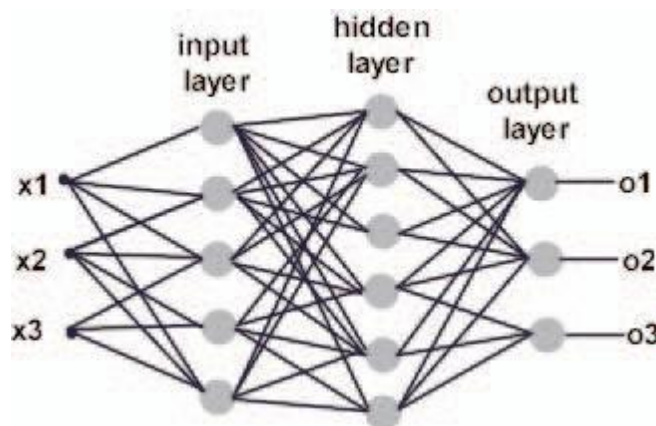
Click Streaming:

As indicated by the specialists, clickstream examination is another thought for estimating the website page significance. A web spider is able to get to the log documents of the server. The utilization data related to client is recorded as web logs. Dissecting web log records for extricating helpful examples is called web utilization mining. Web Usage mining is explained

as the finding of client get to designs from at least one Web server. A server log is a vital hotspot to perform web use mining since it records the perusing conduct of website guest. Clickstream is expressed as an arranged grouping of site page saw by a client that is a session of went by website pages. Pages are introduced one by one of every a column at any given moment [1]. Finish click succession contain the data everything being equal and dissects what number of URLs the client took after by the query output [2].

Artificial neural network

Multi-Layer Perceptron is type of a feed-forward as well as supervised category of neural network. It comprises of an input layer, an output layer along with at least one hidden layers located in the middle of input-output layers. An instance of multi-layer perceptron structure is shown in Figure.



A basic MLP structure

For each of the neurons of input layer, input values are multiplied to the weights. After that, the input is shifted to further layer by using activation function. In this training features are treated as inputs in which corresponding targets are set in the form of 0s and 1s. A trained neural network is further used for testing the query feature set.

K-nearest neighbor

In K-nearest neighbor, the classifier classifies the input test data into the class which has maximum count of votes among its k neighbors. The nearest neighbors are attained during a test

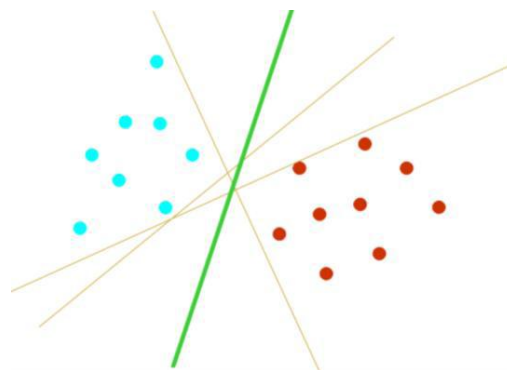
by a criterion called Euclidian distance. In KNN, the entire training data is saved by the classifier. When testing data is provided to the classifier, it searches for its k closest data points and labels the data as a different set that contains maximum number of its k neighbors. The equivalents of k vary according to the user.

Decision trees

The Decision Tree is a non-parametric method of recognizing patterns that is significantly utilized in classifications. It has the attributes of brisk learning and a capacity to achieve exceptionally exact outcomes. It is one of the most prevalent and utilized strategies for inductive inference. This strategy is very powerful to noisy data present in the preparation information. The fundamental concept in DT learning is to discover the quality or highlight that best characterizes the information. A tree is worked from the root node that characterizes the component from which the tree-branching begins other nodes. An obscure information point is grouped by beginning at the root hub and descending the part of the tree that relates to the estimation of that specific quality. This procedure is then rehashed for the sub tree whose root is presently the new node.

Support Vector Machine

The rootsof SVM (Support Vector Machine) had been popularized by Vapnik et al [12] for binary classification purpose. In themost basic form, provided data points which are expressed as p-dimensional vectors, it attempts to look for a hyperplane that differentiates these points among two-class data with maximal margin. Figure below shows two-class data that can be differentiated by various linear classifiers, but only one of themis maximum margin classifier.



Optimal separating hyperplane**Naïve-Bayes**

Naive Bayes classifier was introduced by Thomas Bayes. This particular method learns from the data and predicts classes such that each class has certain probability based on Bayes Theorem [12].

Naïve Bayes formula can be expressed as Eq. (2)

$$P(B|A) = \frac{P(Y) \prod_{i=1}^q P(A_i|B)}{P(B)} \dots (2)$$

where $P(B|A)$ is probability for A vector in Y class whereas $P(Y)$ is initial probability of Y class.

LITERATURE SURVEY:

Jishang Wei et. al. (2012) exhibited a visual cluster investigation way to deal with examine important web clickstream information. Their approach maps the heterogeneous clickstreams in a 2D space, images delegate information tests, and empowers client guided information investigation and clustering. The visual investigation approach enables examiners to reveal client personal conduct standards, take in the example socioeconomics and comprehend the interrelationships between the examples and their socioeconomics.

Gang Kou et. al. (2012) introduced a various factor progressive clustering calculation for extensive scale content accumulations that consolidates space learning, for example, client surfing and retrieval history. This calculation can productively allot a cluster demand of website pages to its nearest group within the sight of a substantial accumulation of content information and manage uproarious information without influencing the calculation of the centroid of clusters.

Joshi et. al. (2012) Propose an entire preprocessing method to preprocess the web log

information to removing the required data adequately. In this paper every one of the means of preprocessing are clarified. Preprocessing has three stages Data Cleaning, User ID and Session distinguishing proof. Information Cleaning expels unimportant information from log record, clients are distinguished based on IP Address and client id and sessions are recognized for every client by taking edge estimation of time.

Kansaraet. al. (2013) proposed a system which is grouping of clustering and classification for anticipating client future requirements. Characterization is utilized to distinguish potential clients from Web Log information and clustering is utilized to aggregate those potential clients with comparable intrigue and consequences of the two procedures order and grouping are utilized to anticipate Users future requirements.

Anupma Surya et. al. (2013) manages Clickstream examination, sub sessions, sessions, and log record that are keys to dissect the clickstream information. Computation of the significance of website page presents a considerable lot of the difficulties. Click analysis is utilized for the extraction of the data about client conduct from the server log document.

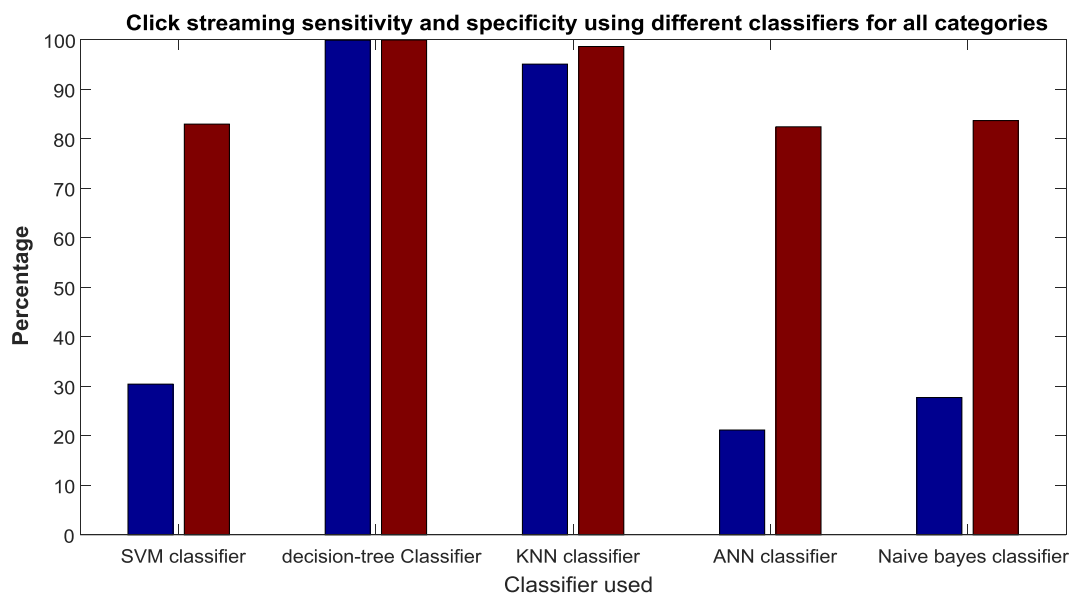
Qiang Su et. al. (2015) devoted to discovering clients' attention designs in view of their click stream investigation. A novel unpleasant pioneer clustering is put forward and tested. This calculation can be utilized to set of clients with comparable surfing practices and decide their attention designs. In view of datasets gathered from a genuine case site, the parameters of the calculation are enhanced, and the adequacy of the calculation is approved.

COMPARISON OF CLASSIFIERS

Average accuracy of category prediction using different classifiers				
SVM Classifier	Decision Tree	KNN Classifier	ANN Classifier	Naïve-Bayes Classifier

71.009	99.9700	98.055	75.82	78.18
--------	---------	--------	-------	-------

Parameters	Average sensitivity and specificity of category prediction using different classifiers				
	SVM Classifier	Decision Tree	KNN Classifier	ANN Classifier	Naïve-Bayes Classifier
Sensitivity	30.44	99.94	95.10	21.17	27.73
Specificity	82.98	99.98	98.66	82.44	83.69



CONCLUSION

The study shown above concludes that out of all the five classifiers- Artificial neural network, KNN Classifier, Decision tree, SVM Classifier and Naïve Bayes classifier reviewed in this paper, Decision tree yields the best result in terms of accuracy, specificity and sensitivity. Due to its 99% accuracy, decision tree can be used in applications such as clickstream analysis in order to make the prediction of webpages most required the users.

REFERENCES

- [1] P. I. Hofgesang and W. Kowalczyk, “Analyzing Clickstream Data: From Anomaly Detection to Visitor Profiling”, ECML/PKDD. Discovery Challenge 2005
- [2] S. Pandey and C. Olston. User-centric web crawling. In Proc. WWW, pages 401–411, 2005
- [3] M. Chau, X. Fang, and O. R. L. Sheng. Analysis of the query logs of a web site search engine. In American Society for Information Science and Technology, 2005
- [4] S. G. O. Guducu, and M. T. Ozsu, “ Incremental click -stream tree model: Learning from new users for web page prediction,” Journal of Distributed and Parallel Databases, Springer, vol. 19, no.1, 2006
- [5] J. Wei, Z. Shen, N. Sundaresan and K. L. Ma, "Visual cluster exploration of web clickstream data," 2012 IEEE Conference on Visual Analytics Science and Technology (VAST), Seattle, WA, 2012, pp. 3-12
- [6] Gang Kou, Chunwei Lou, "Multiple factor hierarchical clustering algorithm for large scale web page and search engine clickstream data" Published in: Annals of Operations Research August 2012, Volume 197, Issue 1, pp 123–134
- [7] Losarwar, Vijayashri and Joshi, Dr. Madhuri (2012) “Data Preprocessing in Web Usage Mining”
- [8] Kansara, Akshay and Patel, Swati (2013) “Improved Approach to Predict user Future Sessions using Classification and Clustering”, ISSN: 2319-7064, Volume 2, Issue 5
- [9] A. Surya and D. K. Sharma, "A comparative analysis of clickstream as web page importance metric," 2013 IEEE Conference on Information & Communication Technologies, JeJu Island, 2013, pp. 776-781

- [10] C. Kaushal and H. Singh, "Comparative study of recent sequential pattern mining algorithms on web clickstream data," 2015 IEEE Power, Communication and Information Technology Conference (PCITC), Bhubaneswar, 2015, pp.652-656
- [11] Qiang Su, Lu Chen, "A method for discovering clusters of e-commerce interest patterns using click-stream data" Published in: Electronic Commerce Research and Applications Volume 14, Issue 1, January–February 2015, Pages 1-13
- [12] VAPNIK, V., GOLOWICH, S. E. & SMOLA, A., 1996. Support vector method for function approximation, regression estimation, and signal processing. Advances in Neural Information Processing Systems, NIPS, 281-287