International Journal of Computational Intelligence Research
ISSN 0973-1873 Volume 13, Number 5 (2017), pp. 1261-1269
© Research India Publications
http://www.ripublication.com

EFFICIENT WEB USAGE MINING BASED ON K-NEAREST NEIGHBORS

Manisha Kumari[1] and Sarita Soni[2]

*[1]Department of Computer Science & Engineering, Research Scholar,
B.B.A.U. University, Lucknow, India*

*[2]Department of Computer Science & Engineering, Assistant Professor,
B.B.A.U. University, Lucknow, India*

ABSTRACT

The aim of Web usage mining is to discover patterns of user activities in order to better serve the needs of the users for example by dynamic link handling, by page recommendation etc. The aim of a Web site or Web portal is to supply the user the information which is useful for him. There is a great competition between the different commercial portals and Web sites because every user means eventually money in this work, we present a study of automatic web usage data mining based on The K-Nearest-Neighbor (KNN) classification method has been trained to be used on-line and in Real-Time to identify the class of the data. The K-Nearest Neighbor (K-NN) algorithm is one of the best methods for solving classification problems; it often yields competitive results and has significant advantages over several other data mining methods. Our work is therefore based on the need to establish a flexible, transparent, consistent straightforward, simple to understand and easy to implement approach. This is achieved through the application of K-Nearest Neighbor technique, which we have tested and proved to be able to overcome some of the problems associated with other available algorithms. The result shows that the K-Nearest Neighbor classifier is transparent, consistent, and straightforward. The work is implemented on WEKA which presents collection of machine learning algorithms for data mining tasks including the K-Nearest Neighbor (K-NN).

Keywords: Automated, Data mining, K-Nearest Neighbor, Weka, Real-Time

1. INTRODUCTION

The World Advanced Web (WWW) is continuously growing with the advice transaction aggregate from Web servers and the bulk of requests from Web users. Accoutrements Web an administrator with allusive advice about users' admission behavior and acceptance patterns has become a call to advance the superior of Web advice account performances. As such, the hidden ability acquired from mining Web server cartage and user admission patterns could be activated anon for business and administration of E-business, E-services, E-searching, E-education and so on. The change of the Internet has advance to an astronomic admeasurements of the accessible advice and the personalization of this advice amplitude has become a necessity. The ability acquired by acquirements web users' preferences can be acclimated to advance the capability of their web sites by adapting the web advice anatomy to the user's behaviour. Automated ability abstraction from web log Files can be advantageous for anecdotic such account patterns and infer user profiles.

Web acceptance mining, Web agreeable mining and Web anatomy mining [1] Web acceptance mining, the arrangement assay consists of several accomplish including statistical analysis, clustering, and allocation and so on. A lot of the accepted assay is absorption on award patterns but with little accomplishment on the abundant pattern/trend assay that varies with the Web environments and the able paradigms advised [5].

This apriorism is mainly accompanying to web acceptance mining, which is an important annex of web mining. Web acceptance mining can be authentic as the appliance of abstracts mining techniques to web log abstracts in adjustment to ascertain user admission patterns. Web acceptance mining has assorted appliance areas such as web pre-fetching, hotlink prediction, and website about-face and web personalization.

2. PROPOSED METHODOLOGY

The proposed work decomposes the library requirements based on the common characteristics shared by the requirements using clustering technique. Thus the requirements that are grouped in each clusters exhibit certain properties that can be used for requirements re- use. Fuzzy c means and k means algorithm are the techniques applied for requirement clustering. The work has been carried out in WEKA. The Waikato Environment for Knowledge Analysis (WEKA) 3.8 serves as an intelligent tool for data analysis and predictive modelling. WEKA was chosen for its wide collection of free analytical tools and data mining algorithms.

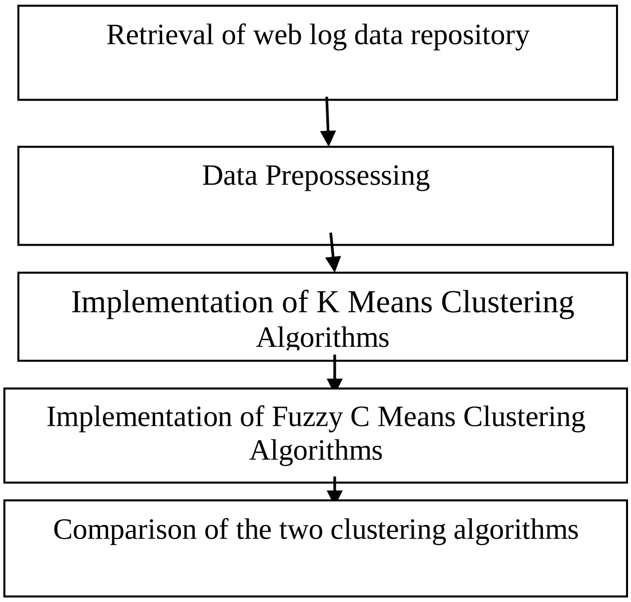
Retrieval of web log data repository
Data Prepossessing
Implementation of K Means Clustering Algorithms
Implementation of Fuzzy C Means Clustering Algorithms
Comparison of the two clustering algorithms

Fig 1: Proposed methodology for comparative analysis of clustering algorithm

3. K-MEANS CLUSTERING

Suppose that the accustomed set of N samples in an n-dimensional amplitude has somehow been abstracted into K-clusters $\{C_1, C_2, C_3... C_K\}$. Each C_K has n_K samples and each sample is in exactly one cluster, so that $\sum n_K = N$, where k=1… n. The mean vector M_k of cluster C_K is defined as the centroid of the cluster.

$$M_K = (1/n_k) \sum \sum_{i=1}^{n_k} x_{ik}$$

Where x_{ik} is the i^{th} sample belonging to cluster C_K. The square-error for cluster C_K is the sum of the squared Euclidean distances between each sample in C_K and its centroid. This error is also called the within-cluster variation:

$$e_k^2 = \sum_{i=1}^{n_k} (x_{ik} - M_k)^2$$

The square-error for the entire clustering space containing K cluster is the sum of the within-cluster variations [28]. The following expression is describe the sum of square-error

$$E_k^2 = \sum_{k=1}^{K} e_k^2$$

Where e is the sum of Cluster.

The basal accomplish of the K-means algorithm are:

- select an antecedent allotment with K clusters absolute about called sample, and compute the centroids of the clusters,
- Generate a new allotment by allotment anniversary sample to the abutting array centre,
- Compute new array centre as the centroids of the clusters,
- Repeats accomplish 2 and 3 until optimum amount of the archetype action is begin or until the array associates stabilizes.

Advantages

KNN has many main advantages : simplicity, effectiveness, intuitiveness and competitive classification performance in many domains. It is robust to noisy training data and is effective if the training data is large.

4. FUZZY C-MEANS CLUSTERING

Fuzzy c-means (FCM) is a adjustment of absorption which allows one section of abstracts to accord to two or added clusters. This adjustment is frequently acclimated in arrangement recognition. It is based on abuse of the afterward cold function:

$$J_n = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m}\left\|x_i - c_j\right\|^2, 1 \le m < \infty$$

Where m is any absolute amount greater than 1, u_{ij} is the amount of associates of xi in the array j, xi is the ith of d-dimensional abstinent data, c_j is the d-dimension centermost of the cluster, and ||*|| is any barometer cogent the affinity amid any abstinent abstracts and the center.

Fuzzy administration is agitated out through an accepted enhancement of the cold action apparent above, with the amend of associates u_{ij} and the array centers c_j by:[31,32]

$$u_{ij} = \frac{1}{\sum_{k=1}^{C}\left(\frac{\left\|x_i - c_j\right\|^2}{\left\|x_i - c_\lambda\right\|^2}\right)^{\frac{2}{m-1}}}, c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} \cdot x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$$

This iteration will stop when $max_{ij}\left\{\left|u_{ij}^{(k+1)} - u_{ij}^{(k)}\right|\right\} < \varepsilon$, where ε is a termination criterion between 0 and 1, whereas k are the iteration steps. This procedure converges to a local minimum or a saddle point of J_m.

The FCM absorption is acquired by aspersing an cold action apparent in equation (1).

$$J = \sum_{i=1}^{n} \sum_{k=1}^{c} \mu_{ik}^{m} |p_i - v_k|^2 \qquad (1)$$

Where:

- J is the cold function
- c is the amount of clusters
- m is a fuzziness agency (a amount > 1)
- vk is the centroid of the kth cluster
- |pi – vk| is the Euclidean ambit amid pi and vk authentic by blueprint (2):

$$|p_i - v_k| = \sqrt{\sum_{i=1}^{n} (p_i - v_k)^2} \qquad (2)$$

The adding of the centroid of the kth array is accomplished application blueprint (3):

$$v_k = \frac{\sum_{i=1}^{n} \mu_{ik}^{m} p_i}{\sum_{i=1}^{n} \mu_{ik}^{m}} \qquad (3)$$

The down-covered associates table is affected application the aboriginal equation (4):

$$\mu_{ik} = \frac{1}{\sum_{l=1}^{c} \left(\frac{|p_i - v_k|}{|p_i - v_l|} \right)^{\frac{2}{m-1}}} \qquad (4)$$

$$|p_i - v_k| = \sqrt{\sum_{i=1}^{n} (p_{iR} - v_{kR})^2 + (p_{iG} - v_{kG})^2 + (p_{iB} - v_{kB})^2}$$

Steps of Fuzzy C-Mean Algorithm

The algorithm is composed of the afterward steps:

This algorithm determines the afterward accomplish.

Step1. Randomly initialize the associates cast (U) that has constraints

Step2. Calculate centroids (c_i)

Step3. Compute contrast amid centroids and abstracts credibility Stop if its advance over antecedent abundance is beneath a threshold.

Step4. Compute a new U Go to Step 2.

By iteratively after light the array centres and the associates grades for anniversary abstracts point, FCM iteratively moves the array centers to the "right" area aural a abstracts set.

5. RESULT AND ANALYSIS

The algorithms are developed in WEKA for analysis and comparison.

Table 1: Comparison table of K-Mean and Fuzzy C-Mean algorithm for time efficiency

Cluster	K-Mean Algorithm Time (ms)	Fuzzy C-Means Algorithm Time(ms)
20	393	260
40	581	338
60	806	425
80	997	586
100	1038	697

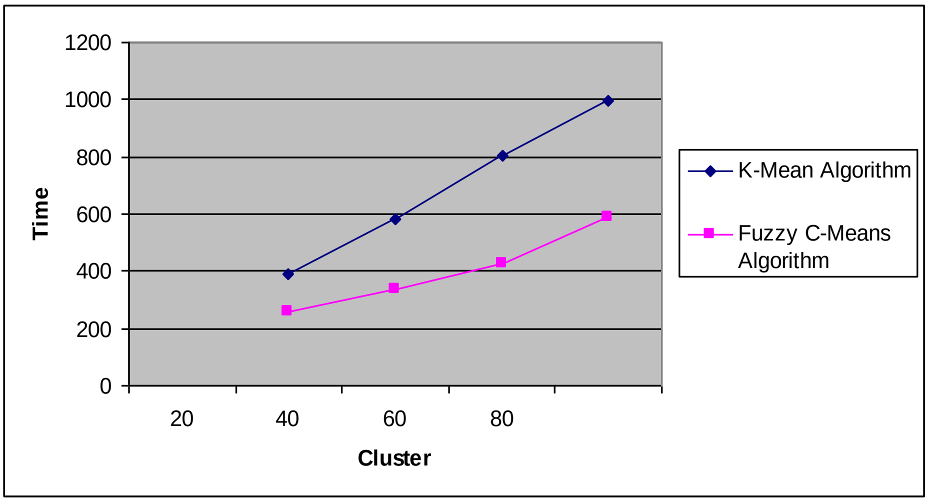

Figure 2: Comparison Graph of K-Mean and FCM Algorithm for Timing Efficiency

In aloft we accept accustomed Comparison Blueprint of K Mean and Fuzzy C Mean. As it is apparent from the graph, as the amount of array increase, again the constant of time (ms) in k-mean algorithm will be top but Fuzzy C Mean algorithm yield everyman amount of time (ms) as compared to k-mean algorithm. Fuzzy C Mean algorithm performs abundant bigger than the K-Mean algorithm in advertent user sessions for all kinds of parameters.

Now we calculate the accuracy of cluster in term of time

Table No 2: Comparison table of K-Mean and FCM Algorithm in term of accuracy

Cluster	K-Mean Algorithm	FCM
	Accuracy (%)	Accuracy (%)
20	46.07	47.40
40	44.19	46.62
60	41.96	45.75
80	40.03	44.14

We construct the graph in term of accuracy

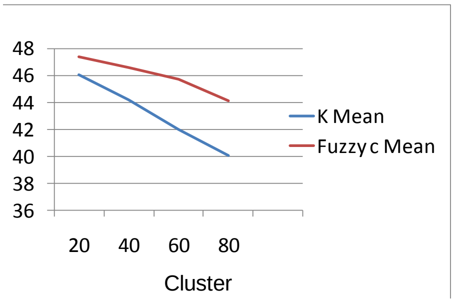

Figure 3: Comparison Graph of K-Mean and FCM Algorithm for accuracy Efficiency

6. CONCLUSION AND FUTURE WORK

Web acceptance mining is the above appliance of abstracts mining admission to apprentice acceptance patterns from Web data, with the ambition of added compassionate and serve the requirements of Web-based applications. The K-Means and Fuzzy C-Means algorithms are one of the important absorption algorithms in abstracts mining domain. We presented a absorption web acceptance abstracts which is advantageous in award the user admission patterns and the adjustment of visits of the hyperlinks of the anniversary user. The appropriate admission was acclimated for ability independent a harder absorption of the web log abstracts set and as the assay adumbrated anniversary of the clusters seems to accommodate observations with specific accepted chacterstics and advance the algorithm ability with advice of FCM algorithm. Experiments prove the bigger algorithm has able to analyze the antecedent array centres. The numbers of data points as well as the number of clusters are the factors upon which the behaviour patterns of both the algorithms are analyzed. FCM produces close results to K-Means clustering but it still requires more computation time than K-Means clustering because of the fuzzy measures calculations involved in the algorithm. Thus for the data points generated using statistical distributions, the K-Means algorithm seems to be superior to Fuzzy C-Means.

In the future, this analysis programme will abide to investigate both K-means and FCM absorption algorithms in appellation ambit similarity. In particular, we are investigating methods to accredit the optimal amount of clusters to be automatically and consistently identified. Further tissue sections will be calm and acclimated to appraise our allegation in this apriorism and approaching research.

REFERENCES

[1]. J. Srivastava, R. Cooley, M. Deshpande and P-N. Tan (2000). “Web Usage Mining: Discovery and Applications of usage patterns from Web Data”, SIGKDD Explorations, Vol 1, Issue 2.

[2]. Cooley, R. Mobasher, B. and Srivastave, J. (1997) “Web Mining: Information and Pattern Discovery on the World Wide Web” In Proceedings of the 9th IEEE International Conference on Tool with Artificial Intelligence, 1082-340919,1997

[3]. Murat Ali Bayir, Ismail H. Toroslu, Ahmet Cosar "A New Approach for Reactive Web Usage Data Processing"Proceedings of the 22nd International Conference on Data Engineering Workshops (ICDEW'06), 0-7695-2571, 7/06

[4]. M. Baglioni, U. Ferrara, A. Romei, S. Ruggieri, and F. Turini," Preprocessing and Mining Web Log Data for Web Personalization

[5]. Raymond Kosala, Hendrik Blockeel," Web Mining Research: A Survey" SIGKDD Explorations, Vol 2, Issue 1, july 2000.

[6]. S. K. Pani, L. Panigrahy, V.H.Sankar, Bikram Keshari Ratha, A.K.Mandal, S.K.Padhi"Web Usage Mining: A Survey on Pattern Extraction from Web Logs" *International Journal of Instrumentation, Control & Automation (IJICA), Volume 1, Issue 1, 2011*

[7]. Sita Gupta, Vinod Todwal "Web Data Mining & Applications" International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-1, Issue-3, February 2012

[8]. Vijayashri Losarwar, Dr. Madhuri Joshi"Data Preprocessing in Web Usage Mining" International Conference on Artificial Intelligence and Embedded Systems (ICAIES'2012) July 15-16, 2012 Singapore

[9]. Renáta Iváncsy, István Vajk"Frequent Pattern Mining in Web Log Data" Acta Polytechnica Hungarica Vol. 3, No. 1, 2006

[10]. Ravi Bhushan, Dr. Rajender Nath"Automatic Recommendation of Web Pages for Online Users Using Web Usage Mining" International Conference on Computing Sciences, 2012.