Improving Cluster Quality by using Ripley's K-Function

Mandeep Kaur
CSE Student
Sri Guru Granth Sahib World University
Fatehgarh Sahib

Usvir Kaur
Assistant Professor (CSE)
Sri Guru Granth Sahib World University
Fatehgarh Sahib

Roop Kamal Kaur
CSE Student
Sri Guru Granth SahibWorld University
Fatehgarh Sahib

ABSTRACT
As the data on the web is growing rapidly, more and more people rely on the search engines to explore the web .Due to heterogeneous and unstructured nature of the web data, Web mining uses various data mining techniques to extract hidden useful knowledge from Web hyperlinks, page content and web usage logs. Web Usage Mining is one of the applications of data mining techniques that are used to discover interesting usage patterns from Web data, in order to understand and better serve the needs of Web-based applications. Web usage mining consists of three phases: preprocessing, pattern discovery, and pattern analysis. In this paper Ripley's k-function is used to refine the original clusters obtained by k-mean and weighted k-mean clustering algorithms.

Keywords
Web mining, Web Usage mining, k-means, weighted k-means, Ripley's k-function, entropy, accuracy, precision, recall, f-measure.

1. INTRODUCTION
Clustering is the process of grouping a set of objects in such a way that objects in the same group (called a **cluster**) are more similar to each other than to those in other groups (clusters).

Clustering of users or sessions into meaningful clusters is now most challenging branch of research. k-means clustering algorithm partition the set of data objects into meaningful clusters by minimizing the sum of squared error distances between objects and centroids of the clusters. Weighted k-means algorithm is the extension of k-means algorithm that includes some more information i.e. a set of weights associated with each of the data points. The weighted K-Means algorithm utilizes a weight vector to minimize the effects of irrelevant attributes and reflects only the semantic information of objects. Ripley's K(t) function is basically a tool that is used for analyzing completely mapped spatial point process data, i.e., data

Based on the locations of events. The Ripley's K- function is given as:

K (t) = 1/ λ E[no. of extra events within distance 't' of a randomly chosen event];

Where λ denotes the density (number per unit area) of events.

K (t) describes the characteristics features of the point processes at many distance scales.

1.1 Web Mining
Web mining is an application of data mining that is used to discover useful information from web documents and web services.web mining is categorized into mainly 3 types:-

A.web content mining.

B.web structure mining.

C.web usage mining.

1.1.1 Web content mining
It provides the knowledge found by going through the web pages contents like image, videos, text etc.

1.1.2 Web structure mining
It basically gives the idea about the structural layout of the web. It also uses the connectivity among websites that are called as **"Hyperlinks".**

1.1.3 Web usage mining
It deals with analyzing the user's navigational behavior by accessing patterns from web log files.

1.2 Web Log Files
Web Log Files are those files that contain information about each and every website visitor's activity. Log files are automatically created by web servers. Each time when a visitor requests any file (pages, images, etc.) from the website, the information of user's request is added to the current log file. log files are mostly in text format and each log entry /hit is saved as a one line of text. Log file ranges from 1KB to 100MB.Typically a web log file has following fields: IP Address, date and time of request, session time, page requested, referrer, server domain, server response, size, agent log.

Types of Web log files: access log file, referrer log file, agent log file, error log file.

1.3 K-Means Algorithm
One of the most popularly used clustering methods is k-means clustering algorithm. The k-means algorithm is very effective in producing clusters for many practical applications.

Algorithm 1: The k-means clustering algorithm

Input: *D = {d1, d2, d3... di, dn }* // Set of 'n' data points.

k =Number of required clusters.

Output: A set of *k* clusters.

Steps:

1. Randomly choose *k* data points from '*D' dataset as* initial centroids;

2. **Repeat**

Assign each of point '*di*' to the cluster from which it has minimum distance from centriod.

Calculate the mean for each new formed cluster;

Until convergence criteria met.

1.4 Weighted K-Mean Clustering Algorithm

Weighted k-mean algorithm is the extension of traditional k-means algorithm. As in the k-means clustering algorithm there is one limitation that it does not tell anything about which attribute contribute more to the clustering process. Weighted k-mean algorithm uses some additional information such as set of weights associated with each data point.

The procedure for **Weighted K-Means clustering algorithm** is given below.

Input: A set of 'n' data points and the number of clusters to be formed 'K'.
Output: Centroids of the K clusters.
(i) Ist initializes the number of clusters k.
(ii) Then randomly selecting the centroids (1, 2, …, K) in the data set.
(iii) Choosing the Static weights.
(iv) Find the distance between the centroids using the Euclidean Distance equation. dij = w*(xi-ck)^2
(v) Update the centroids using given equation.
(vi) Stop when the new centroids are found nearer to the old one. Otherwise, go to step (iv).

1.5 Ripley's K-Function

Ripley's K(t) function is basically a tool for analyzing the completely mapped spatial point process data, i.e., data based on the locations of events'(t) describes characteristic features of the point processes at many distance scales .Alternative summaries for example mean nearest neighbor or cumulative distribution function(cdf) do not possesses this property.
The main drawback of the G^-1- and F^-1-functions, and any of the nearest-neighbor method is that they heavily rely on local information to observe departures from CSR. In other words, we can say that, they do not use spatial information over a broad range of scales. Quadrant analysis is also suffered from the opposite problem i.e it only consider spatial information at broad range of scales and ignored the finer-scale distances information used by all nearest neighbor methods. Ripley's K-function is a compromise between these two types of analysis which also addresses the edge and overlap effects.

Estimating the value of k (t)

We are given the locations of all events within a defined study area , k(t) is defined as ratio of a numerator and density of events λ.The density of events can be estimated as λ=N/A where 'N' is the observed no. of points and 'A' is the area of the study region. if we ignore the edge effects then the numerator can be estimated by N-1 I(dij<t) where 'dij' is the distance between the ith and jth points and I(x) is the indicator function which has the value 1 if x is true and 0 otherwise.
Edge effects mainly arises because of the points that lie outside the boundary and are not counted in the numerator even if they are within the distance't' of a point in the study area. Ignoring edge effects distorts the k (t) estimate, especially at larger't' values.

A wide variety of edge-corrected estimators have been proposed. The most commonly used is given as follows:

$$K(t) = 1/\lambda \sum N\, w(li, lj) I(dij<t)$$

Where 'dij' is the distance between ith and jth points and, I(x) is the indicator function .The weight function w(li , lj) provides the edge correction. Weight function has the value of 1 when circle centered at 'li' and passes through a point 'lj' (i.e. within a radius 'dij')is completely inside the specified study area. If part of the circle falls outside the study area (i.e if dij is larger than the distance from 'li' to at least one boundary) then w (li,lj) is defined as the proportion of the circumference of the circle that falls within study area. The effects of edge correction are more important for large values of't' because large circles are more likely found to be outside the study area. Although we can determine value of k(t) for any 't', but it is a common practice to consider values of 't' less than one half of the shortest dimension of the study area. If the study area is approximately rectangular or t< (A/2) ^1/2, where 'A' is the area of the study region.

Advantages of ripley's k-function:

1. k (t) is easy to compute.

2. By considering more than just the nearest neighbor distances, Ripley's K-function combines distance measurement with the quadrate counting, and so it contains more information than the nearest-neighbor distances and thus provides a more sensitive analysis.

3. The Ripley's K-function sometimes also referred to as the second moment cumulative function, as it is related to the second-order intensity function.

4. Edge corrected estimator is used for refining boundaries of clusters.

2. PROPOSED METHODOLOGY

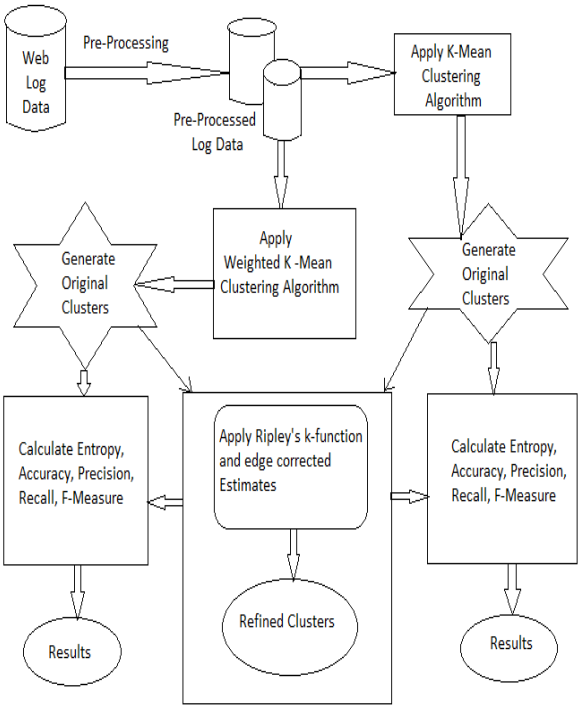

Fig 1: Architecture of Proposed Work

The raw web log files used in this experiment are of size 1MB Containing information like IP address, date and time of request, page requested, referrer, server domain, server response, size, agent log. Before using the web log data for mining process it needs to be preprocessed and converted into the format that can be fed into the clustering algorithm. A session timeout interval of 30 minutes is considered for generating final sessions. By applying the two algorithms we can see that clusters with similarity among sessions have been obtained and can be used for the prediction of pages to a user of similar interests. Then we apply Ripley's k-function to refine the original clusters. The quality of obtained clusters is evaluated using evaluation metrics such as Entropy and Accuracy along with External quality measures such as Precision, Recall, and F-Measure.

2.1 Entropy

Entropy is a most commonly used clustering performance evaluation metric i.e. it measure amount of disorder in a vector. There exist several variations of entropy. The most commonly used is called Shannon's entropy. Shannon's entropy is given as:

$$H(X) = -\sum_{i=0}^{n-1} P(x_i) * log_2(P(x_i))$$

Where 'H' is the symbol used for entropy. 'X' represents a vector of zero-indexed symbols, and 'P' refers to "probability of." The log2 function (log to base 2) assumes that log2 (0) = 0.0 rather than considering the true value of negative infinity.

2.2 F-Measure, Precision And Recall

F-Measure is a measure that combines the precision and recall and is also called as balanced F-score. we can calculate the recall and precision values of that cluster for each class as follows:

Recall (i,j) = xij/xi

And

Precision (i,j) = xij/xj

Where 'xij' is the no. of objects of class i that are in cluster j, 'xj' is the no. of objects in cluster j, and 'xi' is the no. of objects in class i.

Precision and **recall** are two most widely used metrics for evaluating the pattern recognition algorithms for correctness. Precision and recall are the measures that are used to evaluate the quality of set of retrieved documents. The F-Measure of cluster j and class i is given by:

F(i,j)=2*Recall(i,j)Precision(i,j)/Precision(I,j)+Recall(i,j)

The F-Measure values lies in the interval [0, 1] and the larger values indicates higher clustering quality.

3. OUTCOME

3.1 Results for Entropy, Precision, Recall, F-Measure for dataset 1.

The below table compares the results obtained by applying k-mean clustering algorithm on dataset1 in terms of Entropy ,Accuracy, Precision ,Recall and F-Measure. It is clear from the table that the combinational k-mean and Ripley's K-Function has improved values for Entropy, Accuracy, Precision, Recall and F-Measure.

3.1.1 Comparison of k-mean and k- mean +Ripley's k-function

Table 1.comparison of k-mean and k- mean +Ripley's k-function

Parameters	K-Mean clustering algorithm	K-mean +Ripley's K-Function (Refined Clusters)
Entropy	0.3	0.04
Accuracy	69.3	78.4
Precision	0.006	0.7
Recall	0.078	0.13
F-Measure	0.0088	0.012

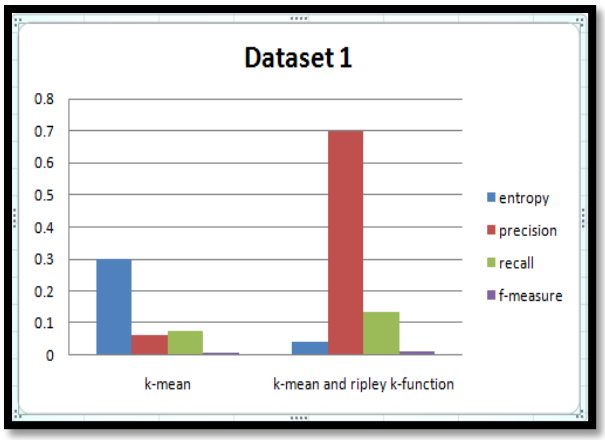

Fig 2: comparison between k-mean and k-mean +Ripley's-function

3.1.2 Comparison between weighted k-mean and weighted k-mean and Ripley's k-function:

The table given below compares the performance of weighted k-mean clustering algorithm and combinational weighted k-mean +Ripley's k-function by accessing the parameter values obtained by applying these algorithms on dataset 1.we observe from the parameters values in the table that combinational weighted k-mean and weighted k-mean +Ripley's k-function has reduced entropy value and high accuracy. The values of precision, recall and f-measure are also revised.

Table 2.Comparison between weighted k-mean and weighted k-mean+Ripley's k-function

Parameters	Weighted K-mean (Original Clusters)	Weighted k-Mean+Ripley's k-function(Refined Clusters)
Entropy	0.02	0.009
Accuracy	83.4	90.9
Precision	0.0066	0.76
Recall	0.088	0.138
F-Measure	0.010	0.0125

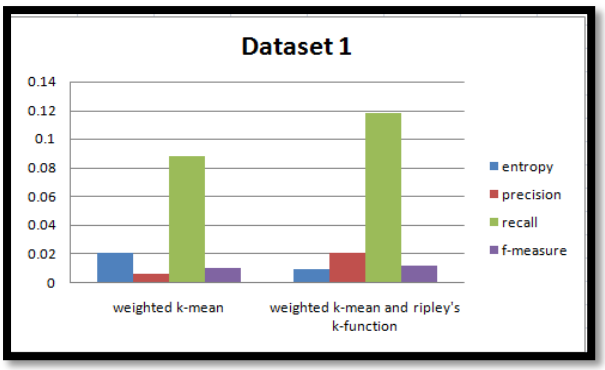

Fig 3.Comparison between weighted k-mean and weighted k-mean +Ripley's k-function

3.2 Accuracy

The accuracy of the clustering algorithms is computed by comparing the results of K-Means clustering algorithm and combinational K-mean+ Ripley's K-Function

3.2.1 Comparison of Accuracy of K-Means clustering algorithm and combinational K-mean+ Ripley's K-Function

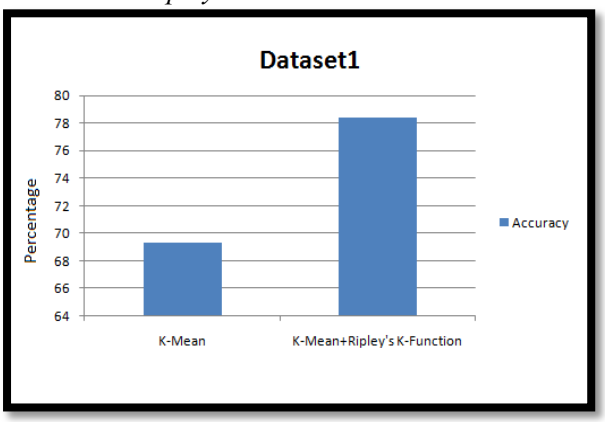

Fig 4. Comparison of Accuracy of K-Means clustering algorithm and combinational K-mean+Ripley's K-Function

3.2.2 Comparison of Accuracy of Weighted K-Mean clustering algorithm and combinational Weighted K-Mean +Ripley's K-Function:

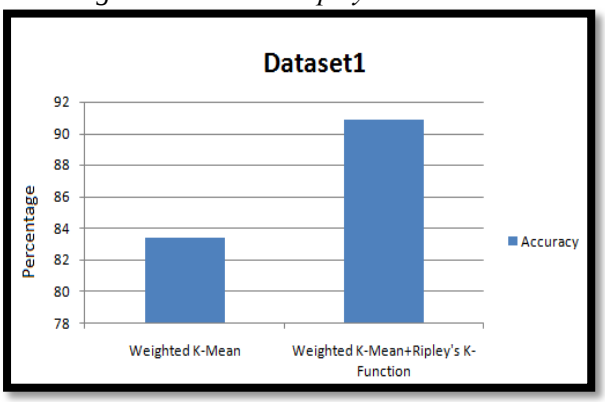

Fig 5. Comparison of Accuracy of Weighted K-Mean clustering algorithm and combinational Weighted K-Mean +Ripley's K-Function

3.2.3 Comparison of accuracy of combinational K-Mean+ Ripley's K-Function and combinational Weighted K-Mean+ Ripley's K-Function

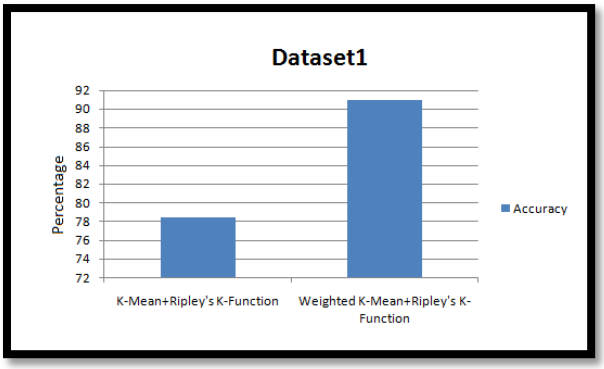

Fig 6. Comparison of accuracy of combinational K-Mean+Ripley's K-Function and combinational Weighted K-Mean+ Ripley's K-Function

4. CONCLUSION AND FUTURE SCOPE

This research work represents a comparative analysis of k-mean clustering algorithm combined with Ripley's k-function and weighted k-mean clustering algorithm combined with Ripley's k-function. The idea was to determine optimal clustering algorithm in terms of performing clustering of web log files. Weighted k-mean clustering algorithm converts the data into intermediate weight format which is more compatible to data cluster rather than reading the data into string, no and other formats. The evaluation results have been computed using parameters Entropy, Accuracy, Precision, Recall, and F-Measure. It is concluded that the combinational weighted k-mean and Ripley's k-function produces a reduced entropy value for web log clusters. Accuracy is found to be increased. Respectively the values of Precision, recall and F-Measure has been supervised and determined to be more efficient while comparing with k-mean and Ripley's k-function. The test has been performed on different datasets of log files and the results found to be consistent. The current research work have widened the future aspects for the same. The future research work might give it a try in optimizing the current result using some swarm intelligent technique like PSO or hierarchy of genetic algorithms.

5. REFERENCES

[1] Supinder Singh , Sukhpreet Kaur , " Web Log File Data Clustering Using K-Means and Decision Tree", International Journal of Advanced Research in Computer Science and Software Engineering Volume 3, Issue 8, August 2013.

[2] Dilpreet kaur, Sukhpreet kaur, " A Study on User Future Request Prediction Methods Using Web Usage Mining", International Journal of Computational Engineering Research, Vol 3, Issue 4,April 2013.

[3] Marjan Eshaghi, S.Z. Gawali, " Web Usage Mining Based on Complex Structure of XML for Web IDS", International Journal of Innovative Technology and Exploring Engineering (IJITEE) , Volume 2, Issue 5, April 2013.

[4] V.Shanmuga Priya, S.Sakthivel, "An Implementation of Web Personalization using Web Mining Techniques", International Journal of Computer Science and Mobile Computing Vol.2 Issue. 6, June 2013.

[5] Monika Yadav Mr. Pradeep Mittal, "Web Mining: An Introduction", International Journal of Advanced Research in Computer Science and Software Engineering,Volume 3, Issue 3, March 2013.

[6] Zhang et al., "An Improved K-means Clustering Algorithm",Journal of Information & Computational Science, Volume 10, Issue 1 ,2013.

[7] Dr. Mohammed Otair, "APPROXIMATE K-NEAREST NEIGHBOUR BASED SPATIAL CLUSTERING USING K-D TREE", International Journal of Database Management Systems (IJDMS) Vol.5, No.1, February 2013.

[8] Soumi Ghosh, Sanjay Kumar Dubey, "Comparative Analysis of K-Means and Fuzzy C-Means Algorithms", International Journal of Advanced Computer Science and Application((IJACSA), Vol. 4, No.4, 2013.

[9] Raed T. Aldahdooh, Wesam Ashour, "DIMK-means - Distance-based Initialization Method for K-means Clustering Algorithm", I.J. Intelligent Systems and Applications, Volume 02, Issue 41-51, January 2013.

[10] Fahim et al., "An efficient enhanced *k*-means clustering algorithm", Journal of Zhejiang University SCIENCE A, Volume 7, Issue 10, 2006.

[11] Rostyslav Kosarevych, Bohdan Rusyn, "Application Of The Ripley's K-Function For Image Segmentation" , TCSET, 2012.

[12] Thibault Lagache, Vannary Meas-Yedid, Jean-Christophe Olivo-Marin, "A Statistical Analysis Of Spatial Colonization Using Ripley's K-Function", IEEE 10th International Symposium On Biomedical Imaging:From Nano To Macro, 2013.

[13] Guohui Zhu, Ying Ge,Huachen Wang, "A Modified Ripley's K-Function To Detecting Spatial Pattern Of Urban System",

[14] Sampaio, Wener B, Diniz,Edgar M,Silva, Aristofanes C,Paiva, Anselmo C, de, "Detection Of Masses in Mammograms Using Cellular Neural Network, Hidden Markov Models And Ripley's K-Function", IEEE, 2009.

[15] Philip M.Dixon , "Ripley's K-Function",Volume 3, 2002.

[16] Mandeep kaur, Usvir Kaur, Dr.Dheerendra Singh, "Web log file clustering algorithms: A survey", International Journal of Computer Application and Technology (IJCAT), Volume 1, April 2014.