Australian Journal of Basic and Applied Sciences, 9(6) Special 2015, Pages: 51-55

ISSN:1991-8178

Australian Journal of Basic and Applied Sciences

Journal home page: www.ajbasweb.com

Data Clustering of Web Documents using Simple K-means Algorithm

[1]Prabaharan S, [2]Bhuvaneswari T. and [3]Subramaniyaswamy V

[1,2,3]Department of Computer Science and Engineering
[1,2]Vinayaka Missions Kirupananda Variyar Engg. College, Salem, India
[3]SASTRA University, Thanjavur, India.

ARTICLE INFO

Article history:
Received 12 November 2014
Received in revised form 26 December 2014
Accepted 29 January 2015
Available online 10 February 2015

Keywords:
Data mining, Web mining, Clustering, Simple K-Means, Web documents.

ABSTRACT

In data mining, clustering is the one of the well known techniques and K-means algorithm is the most frequently used clustering technique. The initialization of cluster centers depends on clustering technique used. In dynamic clustering based on the initial centroids chosen randomly several elements of the cluster changes over time. The number of clusters is equivalent to the number of Centroids which is taken as input parameter. In this paper, Simple K-Means algorithm is employed to cluster the web documents. Three types of Iris flower with 250 instances and five attributes is used as test and training data. The results obtained are compared with the existing IDC method. It shows that the response time is decreased when compared to existing method.

To Cite This Article: Prabaharan S, Bhuvaneswari T. and Subramaniyaswamy V., Data Clustering of Web Documents using Simple K-means Algorithm. ***Aust. J. Basic & Appl. Sci.,*** *9(6):51-55, 2015*

INTRODUCTION

The WWW is huge information and services pool which provides documents, contents, subjects, and powerful search engines. Web and text documents are not structured as these documents consist of sections, paragraphs, sentences, words, letters, punctuation marks, and HTML tags (Odukoya, O.H., 2010). Hence, it is desirable to develop enhanced techniques for machine learning in this semi structured, enormous amount of non tabular web data. Web and text documents are semi structured instead of the well organized tabular data upon which most of the machine learning technique is anticipated to function (Ramchandra Yenape, Sharvari Govilkar, 2012). Search engines mainly provide documents list that contained on the World Wide Web which matches the key terms and/or phrases. In order to reach full potential of World Wide Web, keyword matching is the only indicative of a document's relevance, and there is a need to improve keyword matching (Mark, P., 2002).

Clustering is the one of the most important methods used in data mining occurring from several fields like medical informatics, banking, information retrieval, bio-informatics and also useful in probing pattern-analysis, decision-making, grouping, machine-learning situations, image segmentation, and document retrieval (Jain, A.K., 1999; Amir Ahmad, Lipika Dey, 2007). Clustering is an unsupervised learning algorithm. The aim of clustering is to partition a specified set of data elements into identical groups, i.e. clusters, such that objects in the same group are similar to each other and different to items in other group or to discover different groups in a dataset.

Cluster analysis techniques can be classified as hierarchical clustering and non-hierarchical clustering. Single linkage, average linkage, complete linkage, median, and ward are the examples for hierarchical techniques. Some of the non-hierarchical techniques are k-means, adaptive k-means, k-medoids, and fuzzy clustering (Oyelade, O.J. 2010). The main difference between partitioned and hierarchical clustering is that, in partitioned clustering algorithm data is partitioned into more than two subgroups in one step and in hierarchical clustering algorithm data is divided into two subgroups in each step (Azhar Rauf, 2012). In this paper, Simple K-Means algorithm is employed to cluster the web documents.

Related work:

Various algorithms have been proposed to facilitate cluster the data, which take into account the input parameters and nature of the data. Fixed number of clusters is taken as an input in most of the algorithms. It is not easy to predict the number of clusters for the unknown domain data set in the real-world application (Ahamed Shafeeq, B.M., K.S. Hareesha, 2012). Hierarchical techniques make a nested series of partitions with singleton clusters of

Corresponding Author: Prabaharan S., Department of Computer Science and Engineering

individual items at the bottom and a single inclusive cluster at the top (Zamir, O., O. Etzioni, 1998). In divisive hierarchical clustering, initially the entire data set is considered as a single cluster, and the cluster is divided in each step until only singleton clusters of individual items remain. Groups with different non-overlapping boundaries should preferably be created by a superior clustering algorithm, although an ideal separation cannot typically be attained in practice. According to linear time clustering algorithm, the most common class is the K-means and its variants. It works by recognizing possible clusters as updating the clusters iteratively (Zamir, O., O. Etzioni, 1998).

Web Documents Clustering algorithms begin with the set of items as own clusters. Two most similar clusters are combined at each step. Until a minimum number of clusters are reached this process continues. If entire hierarchy is needed then the procedure continues until only one cluster is omitted [13]. Self-Organizing Maps (SOM) is an elementary data analysis method that uses an unsupervised learning algorithm with no previous information of how input and output are linked. SOM is used to cluster multi-dimensional data into lower-dimensional space, and to disclose unseen formation of data. It holds local similarity and adjacent relationships amidst data items (Lavneet Singh, 2010; Toomas Kirt, 2007).

A number of clustering systems have been proposed that uses dynamic elements to cluster fixed data where groups of items adjust over time and develop into interesting clusters (Diday, E., 1973). Cluster Initialization methods have received interest in several fields including medicine, engineering, and biology. The grouping of data points which are close to one another is the objective of clustering. The K-means algorithm is the broadly accepted method in clustering (Belal, M., A. Daoud, 2005). According to the basic k-mean clustering algorithm, clusters are fully dependent on the selection of the initial clusters centroids. K data elements are selected as initial centers; then distances of all data elements are calculated by Euclidean distance. Data elements having less distance to centroids are moved to the appropriate cluster. The process is continued until no more changes occur in clusters (Azhar Rauf, 2012). Odukoya *et al* formulated an Improved Data Clustering (IDC) algorithm with a novel initialization scheme in which a set of medians extracted from a dimension with highest variances is found. The algorithm was simulated using fuzzy logic (Odukoya, O.H., 2010).

Several variants of K-mean algorithm have been proposed such as K-median. A recent partitioning algorithm called K-mode algorithm uses simple matching coefficient measure to deal with categorical elements (Huang, Z., 1998). Simple K-means is the one of the well-known partition clustering algorithms. It has been elected and listed amongst the top ten most important data mining algorithms in recent time due to its simplicity and scalability as it has linear asymptotic running time with respect to any variable of the problem (Raed, T., 2013).

1. Proposed work:

In this paper, Simple K-Means algorithm is employed to cluster web documents. This model consists of the following steps: initialization of clusters, centroids clustering, assigning records, distance measure, distance calculation. Fig.1 shows the block diagram of the model. In this work, three types of Iris flower with 250 instances and five attributes is used as test and training data.

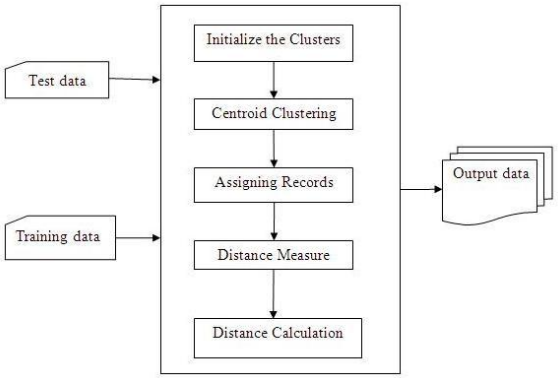

Fig. 1: Block diagram of the model.

1.1 *Simple K-means algorithm:*

Simple *K*-means algorithm is a partition clustering algorithm which divides the data points into K partitions where every partition is a cluster. Based on certain objective function the partitioning is done. One of these functions is the minimizing square error which is computed by given equation (1).

$$m = \sum \sum | q - \mu_i |^2 \quad (1)$$

Where m is the sum of square, q is the point in a cluster, and μ_i is the centroid of the cluster. Each cluster must have at least one point and every point have to be in one and only one cluster. The variance of data in each dimension is computed for a given

data set with dimensionality *d*, the column with maximum variance *dvmax* is found and sorted in descending order. The data points of *dvmax* are divided into *k* subsets, where *k* is the preferred number of clusters, and median of each subset is found. The resultant data points (vectors) for each median is used to initialize the cluster centers. The number of clusters *k* is assumed to be unchanged. Let the *k* prototypes ($z_1 \ldots z_k$) be initialized to one of the *m* input patterns ($j_1, \ldots, j_n$). Therefore,

$$Z^i = j^i \text{ such that } {}_{i \in \{1,\ldots,k\},\, p \in \{1,\ldots,n\}} \quad (2)$$

Where, Z^i is the j^{th} cluster whose value is the disjoint subset of input pattern. The error function is used to determine the quality of the clustering.

Initialization of clusters:

On the initial instant (t = 0), the execution can generate the initial values of mean (μ_j) and covariance matrix (Q_j) arbitrarily. Gaussian probability distribution is used to describe the observed and unobserved entities of the data set *x*. Each class *j* of *N* classes (or clusters) is comprised of mean (μ_j), parameter vector (Θ), and covariance matrix (Q_j). This is represented in Equation (3).

$$\theta(t) = \mu_i(t), Q_j(t),\ j = 1 \ldots N \quad (3)$$

1.2 ***Centroids of clusters:***

Let y_1, y_2, and y_3 be data points. Define a function *g* as

$$g(X) = ((y_1 - X)^2 + (y_2 - X)^2_+ (y_3 - X)^2) \quad (4)$$

Now, it turns out that the minimum of the function g occurs at the centroid of the three points. This interpretation can be easily being generalized across dimensions and centroids.

1.3 ***Assigning records in data set:***

Partition the observations into k clusters such that the sum of squares of the observations to their allotted cluster centers is a minimal i.e. assign data points to their closest centroids. Each observation is allotted to the cluster with the minimal value.

$$m(k) = \sum_{j=1}^{n} \sum_{i=1}^{n} (X_{ji} - \bar{X}_{ki})^2 \quad (5)$$

Where *k* is the cluster, x_{ji} is the value of the i^{th} variable for the j^{th} observation, and $\bar{X}_{ki}$ is the mean of the i^{th} variable for the k^{th} cluster.

3.5 Distance Measure:

Equation (6) represents the Euclidean distance between the points *A* and *B*, which is the length of the line segment connecting them ($\overrightarrow{AB}$). If *A* and *B* are n-dimensional vectors where *A*= (a_1, a_2,.,a_n) and *B* = (b_1, b_2,.., b_n), then the Euclidean distance from *A* to *B*, or from *B* to *A* is given by

$$\left.\begin{matrix} d(AB) \\ d(BA) \end{matrix}\right\} = \sqrt{\sum_{i=1}^{n} (A_i - B_i)^2} \quad (6)$$

3.6 Distance calculation:

The Manhattan distance between two points calculated along axes at right angles where distance that would be traveled to get from one data point to the other if a grid-like path is followed. In a plane with *A* at (x_1, x_2) and *B* at (y_1, y_2), it is $|x_1 - y_1| + | x_2 - y_2|$. The Manhattan distance between two *n*-dimensional vectors is the sum of the differences of their corresponding components.

$$d(A, B) = \sum_{i=1}^{n} | X_i - y_i \ | \quad (7)$$

4. Experimental setup and Results:

The Iris flower data set collected from Wikipedia is used for experiment. It consists of 3 classes of 250 instances each class. One class corresponds to one species of Iris flower named Iris Setosa, Versicolor, and Virginica. Each class has 5 attributes. It represents Sepal Length, Sepal Width, Petal Length, Petal Width, and Species. The metrics used for evaluating the experiments results are accuracy, adjusted rand index, entropy, and time complexity.

4.1 Accuracy:

The accuracy of clustering is calculated using confusion matrix.

Overall Accuracy = $\frac{TN + TP}{TP + FP + FN + TN}$ (8)

Table I and II shows the confusion matrix for the existing IDC method and Simple K-means method at *k*=3 respectively. Table III shows the comparison of the accuracy of IDC and Simple K-means methods for *k*=3 to *k*=8.

Table I: Confusion Matrix for *k* =3 (IDC method).

	setosa	versicolor	virginica
setosa	50	0	0
versicolor	0	48	14
virginica	0	2	36

Table II: Confusion Matrix for *k*=3 (Simple K-means).

	setosa	versicolor	virginica
Setosa	15	0	0
Versicolor	0	19	0
Virginica	0	2	15

Table III: Comparison of Accuracy.

No. of clustering (*k*)	IDC	Simple K–means
3	0.893	0.961
4	0.693	0.963
5	0.666	0.965
6	0.666	0.961
7	0.66	0.961
8	0.66	0.977

4.2 Entropy:

Entropy is calculated as

$$E(A) = \sum_{i=1}^{n} \frac{|si|}{|s|} I(si) \quad (9)$$

The total entropy for the existing IDC method and Simple k-means is shown in Table IV. The entropy of both the models was found stable except for *k*=3 to *k*=8.

Table IV: Entropy for IDC and Simple K-means.

No. of clusters (*k*)	IDC (entropy)	Simple K-means (entropy)
3	0.248	0.248
4	0.2154	0.2154
5	0.2154	0.2154
6	0.1715	0.1715
7	0.1715	0.1715
8	0.1897	0.1897

4.3 Adjusted rand index:

The adjusted rand index is calculated using Euclidean distance which is shown in Fig. 2 against a number of clusters using The rand index *R* is defined by,

$$R = \frac{a+b}{a+b+c+d} = \frac{a+b}{n/2} \quad (10)$$

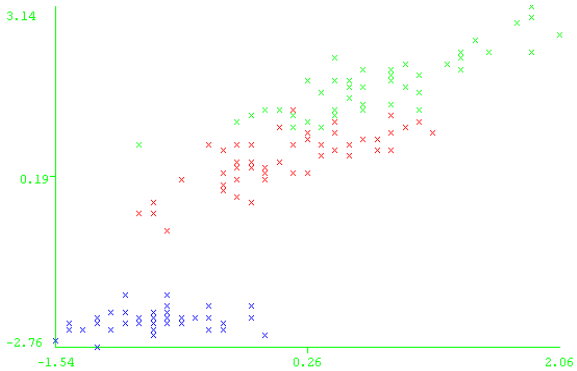

Fig. 2: Adjusted rand index.

4.4 Time Complexity:

If *k* and *d* (the dimension) are fixed, the time complexity of the simple k-means is $O(n^{dk+1} \log n)$, where *n* is the number of entities to be clustered. Table V shows the time taken to cluster Iris dataset using existing and proposed methods. The simulation for 10 iterations was done and the time taken were recorded at the end of each iteration i.e. at iteration = 1, 3, 5, 7, 9, 10. One can found that Simple K-means is faster at iterations 2, 4, 6, 8, and 10.

Table VI shows the average time taken by Simple K-means and IDC which are 0.009s and 0.00439s respectively. Multivariate normal distribution datasets at fixed number of clusters are taken in this experiment.

Table V: Time complexity of IDC and Simple K-Means at *k*=3 for 10 iterations

Trial No	IDC (sec)	Simple K- means
1	0.362	0.278
2	0.011	0.028
3	0.007	0.027
4	0.005	0.020
5	0.008	0.028
6	0.006	0.001
7	0.007	0.015
8	0.009	0.015
9	0.023	0.002
10	0.013	0.010

5. Conclusion:

In this paper, Simple K-means algorithm is employed to cluster web documents. Three types of Iris flower with 250 instances and five attributes is used as test and training data. The results obtained are compared with the existing method. The response time of the proposed method is decreased as compared to existing IDC method and performed better in clustering web documents. The accuracy of the Simple k-means is also found better when compared to existing IDC method.

Table VI: Average Time of IDC and Simple K-means at *k*=3.

Method	Average (seconds)	Range (Low-High) (seconds)	Mean (seconds)
IDC	0.0439	0.010 -0.0320	0.0439
Simple K-Means	0.009	0.543-0.205	0.828

REFERENCES

Odukoya, O.H., G.A. Aderounmu, E.R. Adagunodo, 2010. An Improved Data Clustering Algorithm for Mining Web Documents, *Proceedings of the International Conference on Computational Intelligence and* Software Engineering (CiSE), 1-8.

Ramchandra Yenape, Sharvari Govilkar, 2012. New Data Clustering Algorithm for Mining Web Documents, *International Journal on Advanced Computer Theory and Engineering* (IJACTE) , 1.1.

Mark, P., Sinka David, W. Corne, 2002. A Large Benchmark Dataset for Web Document Clustering, *Soft Computing Systems: Design, Management and Applications*, 87: 881-890.

Jain, A.K., M.N. Murty, P.J. Flynn, 1999. Data Clustering: A Review, *ACM Computing Surveys* (CSUR), 31 (3): 264-323.

Oyelade, O.J. O. Oladipupo, I.C. Obagbuwa, 2010. Application of k- Means Clustering algorithm for prediction of Students' Academic Performance, *International Journal of Computer Science and Information Security*, 7-1.

Azhar Rauf, Sheeba, Saeed Mahfooz, Shah Khusro, Huma Jave, 2012. Enhanced K-Mean Clustering Algorithm to Reduce Number of Iterations and Time Complexity, *Middle-East Journal of Scientific Research*, 12(7): 959-963.

Amir Ahmad, Lipika Dey, 2007. A k-mean clustering algorithm for mixed numeric and categorical data, *Data & Knowledge Engineering*, 63: 503-527.

Ahamed Shafeeq, B.M., K.S. Hareesha, 2012. Dynamic Clustering of Data with Modified K-Means Algorithm, *Proceedings of the International Conference on Information and Computer Networks*, 27: 221-225.

Raed, T., Aldahdooh, Wesam Ashour, 2013. DIMK-means ―Distance-based Initialization Method for K-means Clustering Algorithm, *I.J. Intelligent Systems and Applications*, 2: 41-51.

Belal, M., A. Daoud, 2005. A new algorithm for cluster initialization, *World Academy of Science, Engineering and Technology*, 4: 74-76.

Wang Jun, OuYang Zheng-Zheng, 2010. The Research of K-means Clustering Algorithm Based on Association Rules, *Proceeding of the International Conference on Challenges in Environmental Science and Computer Engineering*, 285-286.

Zamir, O., O. Etzioni, 1998. Web document clustering: a feasibility demonstration, *Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval*, 46-54.

Oren, E.Z., 1999. Clustering Web Documents: *A Phrase-Based Method for Grouping Search Engine Results, Ph.D. Thesis, University of Washington*.

Toomas Kirt, Ene Vainik, Leo Vohandu, 2007. A Method for Comparing Self Organizing Maps: *Case Studies of Banking and Linguistic Data, Local proceedings of ADBIS*, 107-115.

Lavneet Singh, Savleen Singh, Parminder Kumar Dubey, 2010. Applications of Clustering Algorithms and Self Organizing Maps as Data Mining and Business Intelligence Tools on Real World Data Sets, *Proceedings of the International Conference on Methods and Models in Computer Science* (ICM2CS-2010), 27-33.

Diday, E., 1973. The dynamic cluster method in non-hierarchical clustering, *International Journal of Computer & Information Sciences*, 2(1): 61-88.

Huang, Z., 1998. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values, *Data Mining and Knowledge Discovery*, 2: 283-304.