International Journal of Current Advanced Research
ISSN: O: 2319-6475, ISSN: P: 2319-6505, Impact Factor: 6.614
Available Online at www.journalijcar.org
Volume 7; Issue 7(F); July 2018; Page No. 14273-14277
DOI: http://dx.doi.org/10.24327/ijcar.2018.14277.2581

Research Article

ANALYZING AND CLUSTERING FOR SOCIAL NETWORK DATA USING SELF-ORGANIZING MAPS

Anu Sharma[1]., MK Sharma[2] and RK Dwivedi[3]

[1]Uttrakhand Technical University Dehradaun,
[2]Amrapali Institute Haldwani,
[3]Teerthanker Mahaveer University Moradabad

A R T I C L E I N F O

Article History:

Received 5th April, 2018
Received in revised form 24th May, 2018 Accepted 20th June, 2018
Published online 28th July, 2018

Key words:

Self-Organizing Map, Data Mining, Neural Networks, Clustering, K-Means, Matlab, Social Network

A B S T R A C T

With an onset of social network amount of data is increasing day by day. In order to analyze the data and extract useful information, the data mining technology can be used. Clustering is one of the popular method of data mining. Clustering can be used for visualizing and analyzing of data. We are discussing Kohonen SOM. We are using neural networks, as a data mining tool which provides statistical observation and layout from big data-sets. We determine how Self-Organizing Kohonen Maps, can be efficiently used for data mining purposes. The Self Organizing Map (SOM) unsupervised learning is an effective computational tool in data mining processes. Self-Organizing Maps (SOMs) used to visualize social network dataset. We used Self-Organizing Map for clustering and analyzing high-dimensional and complex social network datasets. This paper also visualizes SOM neighbor connection, SOM neighbor weight distance, SOM weight position. We perform self organizing map algorithm for social network dataset in matlab.

INTRODUCTION

Social Network is a structure consisting of individuals or organizations. It is considered as social network data has mapping of all edges between vertices. Network illustrated in the form of social network diagram having points indicate vertices and lines indicate edges that are relationship between the vertices [19]. Clustering is an important task .Clustering is the process of grouping similar data into identifiable clusters and dissimilar data into different clusters. We use Neural Network in data mining for data classification and clustering. classification can be done using supervised or unsupervised learning system. We have two types of variables in supervised learning i.e. input variables and output variables and to learn the mapping function we use an algorithm. Approximation of the mapping function is a goal so that when we have new input data so that we can anticipate the output variables for that data [24]. Example: Regression, Classification. We have only input data and no corresponding output variables in unsupervised learning technique [24]. Example clustering, association. We use Kohonen's Self Organizing Map which is an unsupervised learning technique. In this, we can minimize the proportionality from a very high proportion data into 2 or 3 dimensional space. This reduction facilitates us to describe the results easily and intuitively. The advantage of using SOM is that the method can automatically clusters nodes [19].

Corresponding author:* **Anu Sharma
Uttrakhand Technical University Dehradaun,

Partitioning Algorithms

This is used to bifurcate the data into disjoint clusters. The most famous partitioning based algorithms are k -mean, k-medoid, k-mode and k –prototype.

According to (Berkhin, 2006) there are two approaches to partition the data. [7][8]

- Conceptual Approach
- Objective Function.

In conceptual point of view, clusters are identified with the help of predefined model and in Objective function based partitioning approach either the pair wise computation of cluster or similarity-based relation between the clusters of dataset is considered.

Main advantages and disadvantages of partitioning methods are discussed below:

Advantages

- It is suitable for the dataset that includes the well separated compressed spherical
- clusters.
- It is a simple method.

Disadvantages

- User has to define number of clusters in advance.
- It is unable to deal with non-convex clusters with different sizes and density. It is
- very sensitive to noise and outliner.

K-Means Clustering

K-means clustering is an unsupervised clustering algorithm which is used to find groups within the data (Rousseeuw and Kaufman, 2005, Burkardt, 2009). K-mean was proposed by (Macqueen. 1967). It is the simplest unsupervised clustering algorithm [9]. In k-mean, we assume the number of clusters (K) prior before partitioning the data [7]. It requires the user defined parameters: Number of clusters (K), Distance metrics and cluster initialization. Basic algorithm has some simple steps. First, we have to choose number of clusters as initial centroid, afterword it generates the number of clusters as a cluster center. In next step, it allocates each point to its nearby cluster center and again recomputed the center of each new cluster. This process will continue until some convergence criteria are met, in other words, until the centroids do not change. Fuzzy c mean (Dunn 1973), X-mean (Pelleg *et al*., 2000) and Kernel K-means (Schölkopf *et al*., 1998), K-prototype (HUANG, 1998) are some extension of k mean.

The basic steps of K-mean clustering algorithm are

a. Initially take any k objects as centroids.
b. Find distance of all objects from those k centroids, less the distance the object is in that centre of centroids.
c. Now find the centroids from the objects which are in that clusters.
d. Repeat step 2 and step 3 until the value of centroids is same. [14]

Advantages

- It is easy to understand.
- It gives good result when data is well separated.
- it is very easy to implement.
- It is suitable for very large data sets.[8]

Disadvantages

- We have to define number of clusters prior.
- It chooses center of the cluster randomly, which might not give positive results.
- It is applicable when mean value is defined.
- It is not a good choice for noisy data.
- It is Sensitive to the outliner.
- Final result always depends on the initial partition.
- the algorithm is data-dependent. [8][9].

K-medoid

It is another important clustering algorithm based on partitioning. It was introduced by (Kaufman *et al*., 1987). In k-medoid algorithm, each cluster is represented by the most centric object (medoid) in the cluster. Medoids are more inflexible to noise and outliers as compared to centroids.
The K-medoids algorithm as follow:

- It starts with a random selection of objects as medoids for every k clusters then it assign each point to a cluster that is associated with cluster medoids.
- Afterward, it recalculates the k-medoids position.
- This process will continue until medoid becomes fixed.

Advantages

- Easy to implement and understand.
- It is less sensitive to the outliner as compared to k-mean.

Disadvantages

- It needs a prior knowledge about the number of cluster parameter.
- Final result and run time always depend on the initial partition

FCM - Fuzzy CMEANS algorithm

For the study of data and structure of models fuzzy clustering is a powerful unsupervised method. Fuzzy c-means algorithm is most widely used. [22]

HIERARCHICAL METHOD

Primary aim of the hierarchical method is to demonstrate the cluster similarity into tree pattern that is also called dendrogram. The nested clusters in the dendrogram represent the clusters that are related to each other in dataset [8]. There are mainly two types of algorithm of the hierarchical method:

i. Agglomerative Method
ii. Divisive method. [11]

Dendrogram can demonstrate both methods. Hierarchical clustering approach uses different restraint to decide locally which cluster should be merged at every step. Hierarchical cluster formed the document group into a tree like structure (dendrogram) where parent/child relationships can be viewed as a topic/subtopic relationship [8].

An Agglomerative Method

It is a bottom up approach by one by one connecting nearest pairs of clusters together until the total objects form one large cluster. The nearest cluster can be resolved by calculating the distance between the objects of n dimensional space [6]. It can generally classified on the basis of inter-cluster similarity measurements.[6] The most popular inter-cluster similarity measures are single-link, complete-link, and average-link [6]. Agglomerative algorithms According to (Jain *et al*., 1988), it is also known as bottom-up method. Agglomerative method considers each point as cluster, and it merges the point until we do not get the final desired cluster. Rock, BIRCH, Cure, CFT, Chameleon are main extension of agglomerative algorithm.

Divisive algorithms (top to down)

According to (Kaufman *et al*., 1990), it is opposite to agglomerative algorithm. In this method, all the points or objects are considered as part of only one cluster but further points are subdivided into a small cluster until we get the final desired result.[8]

Self-Organizing Maps (SOM) as a Data Mining Tool

Kohonen proposed Self-Organizing Map (SOM) [20]. This is the most popular artificial neural algorithm. SOM is used in unsupervised learning, clustering, classification and data visualization [16]. SOM is widely used in pattern recognition, biological modeling, data compression, signal processing and data mining [20]. It provides an approach for cluster analysis and achieves a mapping of high dimensional input vectors into a two dimensional output space [17].The resulting map is accomplished of performing the clustering task in a completely unsupervised fashion [17]. Design of the data patterns onto n-dimensional grid of neurons or units [18] is the basic idea of SOM. That grid known as the output space. Moments in time are called as epochs. Learning rate is 0 if no learning happens.

Clustering data is an excellent application for neural networks. Grouping data by similarity is involved in this process. All SOM visualizations in this study have been made with the MATLAB program using the SOM Toolbox program package.

Analyzing Social Media Data

In this research we use facebook ego network data set and Slashdot social network dataset. We perform self organizing map algorithm for both the dataset in matlab.

Facebook Ego Network Dataset

In this research, SNAP Facebook Dataset is used. This dataset contains personal networks of connections between friends of survey participants. Such personal networks represent friendships of a focal node, known as "ego" node, and such networks are therefore called "ego" networks. We will just reload the preprocessed data. [14] This dataset consists of 'circles' from Face book. This dataset consist of 4039 nodes and 88234 edges. [14]

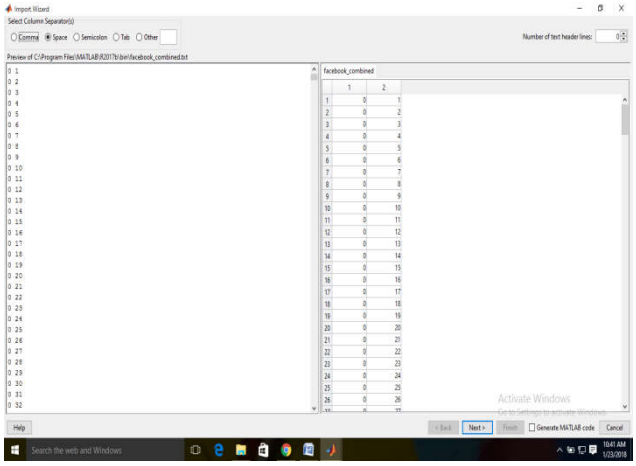

Fig 4.1 Facebook_combined Dataset in Matlab

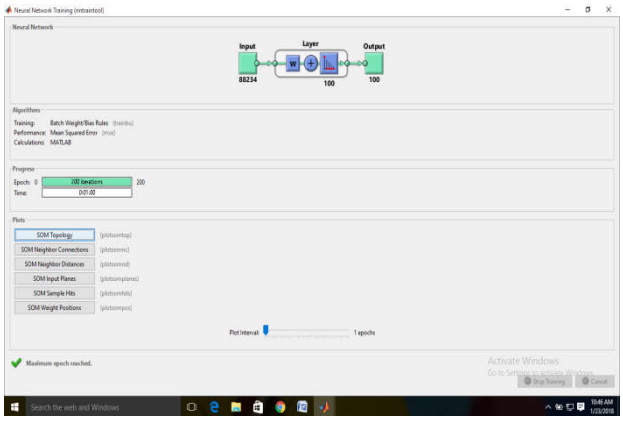

Fig 4.2 Facebook_combined training

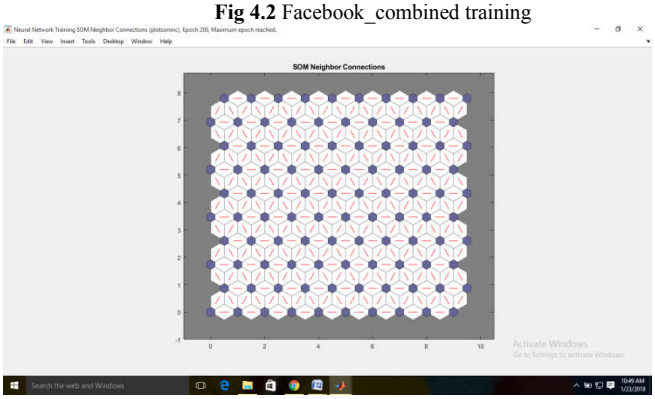

Fig 4.3 Facebook_combined SOM Neighbor Connections

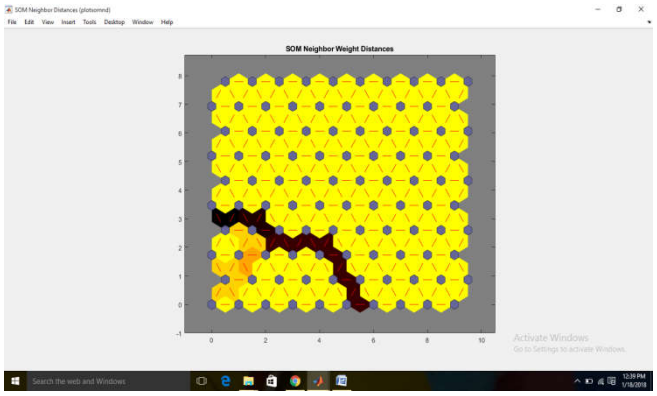

Fig 4.4 Facebook_combined SOM Neighbor Weight Distances

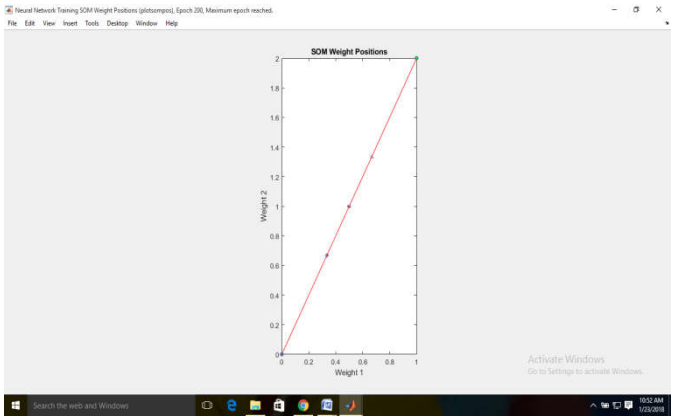

Fig 4.5 Facebook_combined SOM Weight Positions

Slashdot Social Network Dataset information

Slashdot introduced the Slashdot Zoo feature which allows users to tag each other as friends or foes. The network contains friend/foe links between the users of Slashdot [13]. We download Slashdot 0902.txt.gz. This dataset consists of 77360 nodes and 905468 edges.

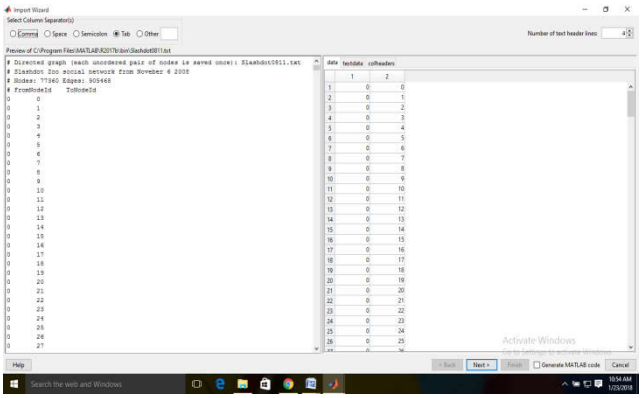

Fig 4.6 Slashdot Dataset in Matlab

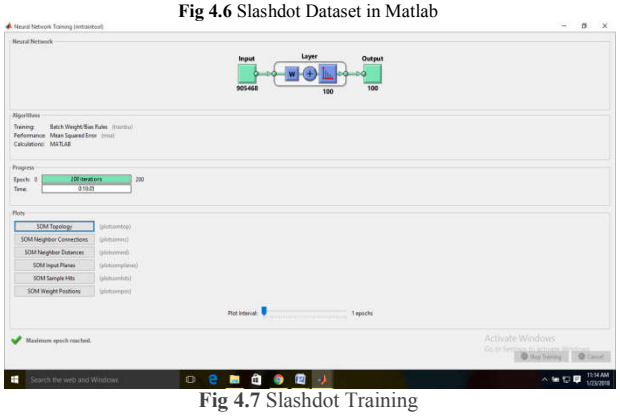

Fig 4.7 Slashdot Training

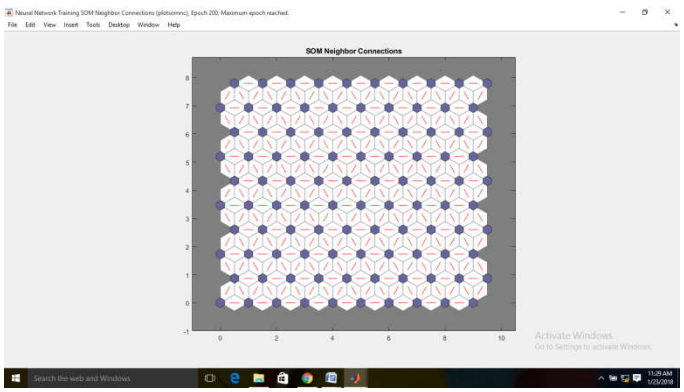

Fig 4.8 Slashdot Neighbor Connections

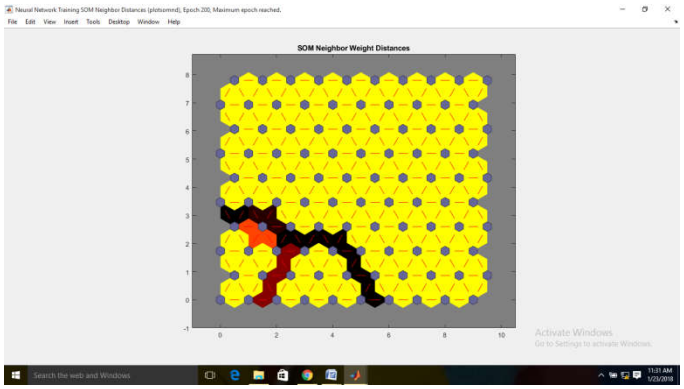

Fig 4.9 Slashdot Neighbor Weight Distances

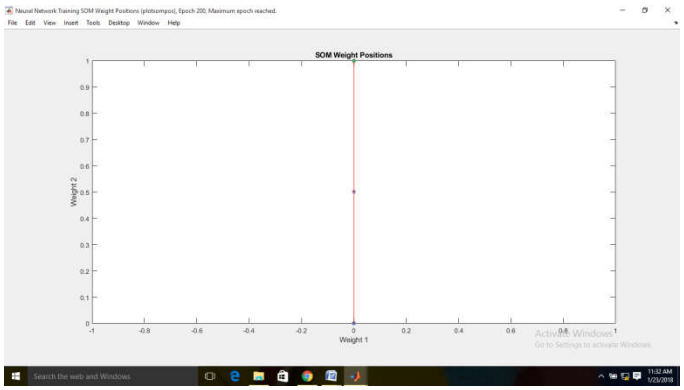

Fig 4.10 Slashdot Neighbor Weight Positions

RESULT & DISCUSSION

We have applied SOM on face book and Slashdot social network datasets. Here we have illustrated that for data mining SOM can be used as an effective tool. We have trained facebook _combined and Slashdot data set with epoch 200 iteration. Time taken for facebook _combined was 0:01:00 and for Slashdot was 0:18:03.We also visualizes SOM neighbor connection, SOM neighbor weight distance, SOM weight position on both social networking sites. SOM can be used for large network dataset. We have use 10 x10 size map in this study. Every neuron depicts the number of input vectors which it classifies. The size of a colors patch shows the relative number of vectors for each neuron. SOM layer denotes neurons as gray-blue patches and its direct neighbor relations with red lines in SOM neighbor weight distances. Black to yellow color patches shows how close each neuron's weight vector to its neighbors. In SOM weight positions, green dots denotes input vectors and shows how SOM classifies the input space by showing blue-gray dots for each neuron's weight vector and connecting neighboring neurons with red lines. It represents the data on various clusters. Each cluster is shown in different colors. Gray-blue patches denote SOM layer neurons and red lines as their direct neighbor relations.

References

1. https://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R/Clustering/Self-Organizing_Maps_(SOM)
2. https://en.wikipedia.org/wiki/Self-organizing_map
3. Shigeru Obayashi and Daisuke Sasaki, “Visualization and Data Mining of Pareto Solutions Using Self-Organizing Map”.
4. Minji Maria Lee, Enrico Steiger, Alexander Zipf, “Clustering and Analyzing Air Pollution Data using Self-Organizing Maps”, AGILE 2016 – Helsinki, June 14-17, 2016
5. Er. Gurpreet Singh, “Review on Kohonen -SOM and K-Means data mining clustering algorithm based on academic data set”, *An International Journal of Computer & IT.*
6. Jason Ong, Syed Sibte Raza Abidi, “Data Mining Using Self-Organizing Kohonen maps: A Technique for Effective Data Clustering & Visualization”, In International Conference on Artificial Intelligence (IC-AI'99), June 28-July 1 1999, Las Vegas.
7. S.Amudha. “An Overview of Clustering Algorithm in Data Mining”, *International Research Journal of Engineering and Technology (IRJET)*
8. Meenu Sharma, Mr. Kamal Borana, “Clustering In Data Mining: A Brief Review”, *International Journal of Core Engineering & Management (IJCEM)* Volume 1, Issue 5, August 2014.
9. Yujie Zheng+, “Clustering Methods in Data Mining with its Applications in High Education” , 2012 International Conference on Education Technology and Computer (ICETC2012) IPCSIT vol.43 (2012) © (2012) IACSIT Press, Singapore
10. Mythili S1, Madhiya E2, “An Analysis on Clustering Algorithms in Data Mining”, *International Journal of Computer Science and Mobile Computing*, Vol.3 Issue.1, January- 2014, pg. 334-340
11. Amandeep Kaur Mann, Navneet Kaur, “Survey Paper on Clustering Techniques”, *International Journal of Science, Engineering and Technology Research* (IJSETR) Volume 2, Issue 4, April 2013, and ISSN: 2278 – 7798
12. https://snap.stanford.edu/data/egonets-Facebook.html
13. https://snap.stanford.edu/data/soc-Slashdot0902.html
14. Anand M. Baswade1, Prakash S. Nalwade2, “Selection of Initial Centroids for k-Means Algorithm”, IJCSMC, Vol. 2, Issue. 7, July 2013, pg.161 – 164.
15. http://www.math.le.ac.uk/people/ag153/homepage/KmeansKmedoids/Kmeans_Kmedoids.html
16. https://arxiv.org/ftp/cs/papers/0611/0611058.pdf
17. Yi Ding∗ ,Xian Fu, “The Research of Text Mining Based on Self-Organizing Maps”, 1877-7058 © 2011 Published by Elsevier Ltd. doi:10.1016/j.proeng.2011.12.757, Procedia Engineering 29 (2012) 537 – 541
18. Fernando Bação1, Victor Lobo1, 2, and Marco Painho1, “Self-organizing Maps as Substitutes for K-Means Clustering”, V.S. Sunderam *et al*. (Eds.): ICCS 2005, LNCS 3516, pp. 476 – 483, 2005. © Springer-Verlag Berlin Heidelberg 2005

19. Fatemeh Ghaemmaghami, Reza Manouchehri Sarhadi, "SOMSN: An Effective Self Organizing Map for Clustering of Social Networks", *International Journal of Computer Applications* (0975 – 8887) Volume 84 – No 5, December 2013.
20. M.N.M. Sap1, Ehsan Mohebi2, "Hybrid Self Organizing Map for Overlapping Clusters", *International Journal of Signal Processing, Image Processing and Pattern Recognition.*
21. http://www.idconline.com/technical_references/pdfs/data_communications/Data_Mining_Cluster_Analysis.pdf
22. Raulji Jitendrasinh G, "A Review on Fuzzy C-Mean Clustering Algorithm", *International Journal of Modern Trends in Engineering and Research*, e-ISSN: 2349-9745 p-ISSN: 2393-8161
23. Nidhi Grover, "A study of various Fuzzy Clustering Algorithms", *International Journal of Engineering Research* ISSN: 2319-6890) (online), 2347-5013(print) Volume No.3, Issue No.3, pp: 177-181, 01 March 2014.
24. https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms
25. Dr. Gurpreet Singh, Amandeep Kaur, "Comparative Analysis of K-Means and Kohonen-SOM data mining algorithms based on student behaviors in sharing information on facebook", *International Journal Of Engineering And Computer Science* ISSN:2319-7242 Volume 6 Issue 4 April 2017, Page No. 20990-20993
26. Ivica Mar_i, Slobodan Ribari_, "Comparison Of A Back Propagation And A Self Organizing Map Neural Networks In Classification Of Tm Images".
27. http://www.pitt.edu/~is2470pb/Spring05/FinalProjects/Group1a/tutorial/som.html

How to cite this article:

Anu Sharma., MK Sharma and RK Dwivedi (2018) 'Analyzing and Clustering for Social Network Data Using Self-Organizing Maps', *International Journal of Current Advanced Research,* 07(7), pp. 14273-14277.
DOI: http://dx.doi.org/10.24327/ijcar.2018.14277.2581

****