International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-4, February 2019

Probability Density Based Fuzzy C Means Clustering for Web Usage Mining

Jayanti Mehra, RS Thakur

Abstract: ***The World Wide Web is huge repository and it is growing exponentially. It contains vast amount of information which is growing and updating rapidly. Various organizations, institutes, government agencies and service centers update their information regularly. The World Wide Web provides its services to the varieties of web users. Web users may have different interests, needs and backgrounds. Clustering is one of the most important tasks in the active areas of Web Usage Knowledge Discovery. It assures to handle the difficulty of information overload on the Internet while many users are connected on the social media. Clustering is utilized for grouping information into comparative access design for discovering client interest. There are two drawbacks of FCM algorithm, firstly the requirements of no. of clusters c and secondly assigning the primary relationship matrix. Due to these two drawbacks the FCM algorithm is hard to decide about the suitable no. of cluster and this algorithm is insecure. The determination of desirable preliminary cluster is an important problem, therefore a new technique called PDFCM algorithm is described.***

Index Terms: ***Clustering, FCM, Probability Based Fuzzy c means Clustering (PDFCM), Web Log Mining.***

I. INTRODUCTION

Nowadays internet has become a convenient foundation and source of information in everyone's daily activity. The World Wide Web had gone through enormous development in last two decades but its amount of swap and extent increased the trouble for different websites. To fulfill the demands of their users, the e-commerce website is quickly progressing hence them importance is obvious. Because of several tremendous benefits of web research, it is pretty interesting thing for organizations. It has helped to improve the profitability of the market and also for the benefit of the market intelligence; this also helps in marketing and comparative analysis for finding the customer relationships [4-5]. The web data were organized and assembled, and structured thorough the client's profiles. This advantage helps organizations to save current clients by giving more customized administrations; however, it additionally contributes in finding for potential clients. The two drawbacks of FCM algorithm which ensure its insecurity in completing any task are, firstly the requirements of 'c' i.e. no. of clusters and secondly assignment of initial value for membership matrix. In this chapter the probability density based fuzzy c-means clustering algorithm (PDFCM) is proposed keeping in mind these two shortcomings of the FCM algorithm, and furthermore, it is exceptionally delicate to the assurance of the two parameters.

Manuscript published on 28 February 2019.
*****Correspondence Author(s)
Jayanti Mehra, PhD. , Department of Computer Applications, Maulana Azad National Institute of Technology Bhopal, (M. P.), India.
Dr. Ramjeevan Singh Thakur, Associate Professor, Department of Computer Applications at Maulana Azad National Institute of Technology, Bhopal, India.

If FCM have these drawbacks the algorithm is hard to take the suitable no. of cluster and this algorithm is insecure [15-10].The

determination of desirable preliminary cluster is an important problem for that procedure, therefore PDFCM algorithm i.e. Probability density based fuzzy c-means clustering algorithm (PDFCM) is proposed here. The complete process of proposed PDFCM is shown in Fig. 1.

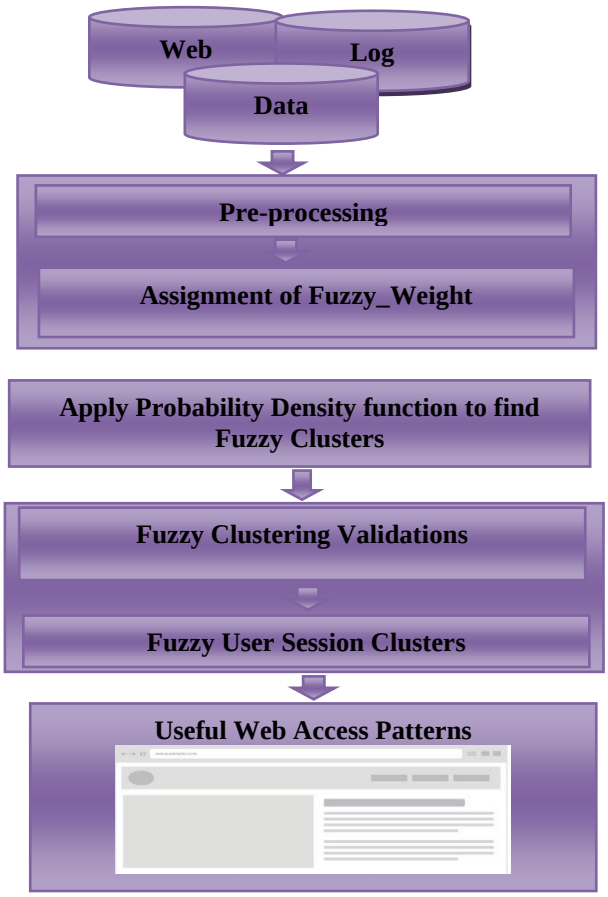

Figure 1: Framework of Probability Density function based fuzzy session clustering

II. LITERATURE REVIEW

S. K. Dwivedi et al. [17] have worked on different types of data preprocessing methods; it is to change raw data into appropriate format. When data is taken from server side it is not acceptable for our mining process. It is very important to pre-process the data. This paper discusses about different data pre-processing techniques. H.X. Pei et al. [14] proposed effective D-FCM algorithm to solve the problem of selection of suitable clusters and gave experimental result using different databases. Z. Ansari et al. [21] solved the problem of selection of suitable cluster centres.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

They proposed Mountain Density based-fuzzy C means clustering and fuzzy c median algorithms and compare different validity index with FCM and FCMed algorithm. A. Gupta and A. Khandekar [2] presented the idea of clustering and also discussed the use of fuzzy technique with different data mining process. Lastly, this paper gives a relative investigation of dual fuzzy algorithm, further described FCM algorithm and adaptive fuzzy clustering technique. V. Anitha and P. Isakki Devi [18] gave a consideration on web usage mining to anticipate the web user's behaviour from log files in web server. Users use website pages with a continuous way and access pages with links are stored in log file of web server and also taking about with the respect to behaviour from investigation of various algorithm and distinctive techniques. D Koutsoukos et al. [19] explained about the session identification algorithm for weblog data and fuzzy c means clustering has also been explained, the studied the impact of the cluster of distance framework that have on the clustering process, the proposed procedure use subtracting clustering for the partition for demonstration of the session information. The preliminary comes going to show that the proposed approach is incredible in the change of client sessions. M. Sampath and Prabhavathy et al. [16] proposed FLAME clustering algorithms for web page access prediction and compare FCM and FLAME algorithms and discovered patterns from the weblog data. Z. Ansari et al. proposed a Fuzzy set theoretic approach based fuzzy c means a framework for deciding the client session cluster in weblog data. At last distinguish this approach with regular approach. K. Suresh et al. [15] presented clustering for weblog data for finding the useful web access patterns in order to visit of hyperlinks and explained the improved fuzzy c means clustering for www.msbn.com data sets. The defined algorithm is able to identify the initial cluster and this is proved by experiment result. V. Chitraa and A.S. Thanamani [19] proposed a fuzzy c-means based novel approach to cluster the web user transactions. This approach is grouping the similar user navigation patterns. The algorithm enhances the FCM and Penalized FCM clustering algorithm by adding Posterior Probability to find highest membership for a member to add in a cluster. Classification is carried out by SVM and RVM for classifying a new user to a particular group. The method is evaluated based on different data and shows the better performance compared with other existing clustering techniques

III. PRINCIPLE OF FCM CLUSTERING

This method is first introduced by Dunn in 1973 and improved by Bezdek et al (1984) [20-8]. Clustering algorithms divide a data set A into C clusters which are normally disjoint and reproduce X when we unite them [13]. These clustering algorithms are termed as Hard (non fuzzy) clustering algorithms. Each sample in a data set would have membership to every cluster [1-9]. It is discussed that membership value for a data point could be determined with the help of a function called membership function. Membership value of a data point is always between 0 and 1 [7]. If the similarity between the sample and the cluster is of high degree, then the membership value is close to 1; otherwise it will be close to 0 [6-11]. Therefore, FCM uses fuzzy dividing in such a way that all members can divide to numerous clusters with a degree of membership which is specified by membership values [2-3].

For calculating membership using Eq. (3.1):

$$uij = 1/\sum_{i-1}^{k} [\|ai - cj/ai - ck\|]^{2/m-1} \quad \text{.....(3.1)}$$

For calculating cluster cancroids using Eq. (5.2)

$$cj = \sum_{i=1}^{N} u_{ij}^{m} a_i / \sum_{i=1}^{N} u_{ij}^{m} \quad \text{.....(3.2)}$$

To findout minimum distance (using Eucludean distance) [38] [42] using Eq. (3.3)

$$|ai - cj|^2 \text{.....(3.3)}$$

Fuzzy algorithm allows a data points belong to more than one clusters [12]. The minimization of the objective function is used, which is given in Eq. (3.4)

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{c} u_{ij} \, \|ai - cj\|^2 \quad \text{...(3.4)}$$

Where,
m is the weighting component; $1 \leq m < \infty$
A= $\{a_1, a_2 \ldots a_n\}$ datasets
C = number of clusters in A, $2 \leq c < n$
U = Fuzzy c-partition of A,
$u_{ij=}$ degree of membership
a_i =is the jth d-dimensional data
C_j=is the center of cluster j.

Steps involved in FCM algorithm are:

1. Define the Initial membership matrix is create using $U^{(0)}$.
2. Next determine the centers vectors $C^{(k)}=[C_j]$ with $U^{(k)}$.
3. Using the formula (5.3) we calculate the centroid of a cluster.
4. Objective function J_m is calculated using the formula (5.4). If value of J_m is less than a threshold, the process can be finished.
5. update $U^{(k)}$, $U^{(k+1)}$
6. New Membership of data points is calculated using the formula (5.1)
7. Stop

This centroid value serves as the reference value. Similarly, centroid value for all the clusters is generated accordingly. The reference value concludes the information of a set data points in the original dataset. These reference values are used for the successive processes of the approach.

IV. PROPOSED METHODOLOGY OF PDFCM

$S = \{s_1, s_2, \ldots, s_m\} \subseteq m$ user sessions.
In sequence to recognize the primary cluster center, from all user session s_i is deal with as a probable candidate and the starting PDF value for user session s_i, it is indicating like$P_1(s_i)$, is calculated using Eq.(4.1).

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

$P1(s_i)=\sum_{k=1}^{m} \exp(-E^2(s_i,s_k)/R^2$ for all i=1..m ……..…(4.1)

Neighborhood radius= R was set to $\sqrt{n}$ (here n is number of URLs)

$E_2(s_i,s_k)$ =Euclidean distance(s_i,s_k)

Where R is a +fixed that characterizes a neighbour-hood for client session s_i. The PDF estimation of the client session s_i is an estimate density of every client sessions in the neighborhood of s_i. Client sessions external the circular distance have small effect on its PDF value. The client session with the topmost PDF value is select as the main cluster center point v_1 as takes after.

$$i_1 \leftarrow \underset{i=1}{\overset{M}{\operatorname{argmax}}}\{P1(s_i)\};\ v_1 \leftarrow {}^{s}i_1 \qquad \ldots\ldots\ldots 4.2$$

the second PDF value calculating using Eq. (4.3).

$$P_2(s_i \leftarrow P_1(s_i) - P_1(v_1)\exp\left(-\frac{E2(si,v1)}{R2}\right) \text{for all } i=1..m \quad \ldots\ldots\ldots(4.3)$$

Continue work on PDF value for all user sessions, and next cluster center is selected with highest PDF value using Eq.(4.4).

$$i_2 \leftarrow \underset{i=1}{\overset{m}{\operatorname{argmax}}}\{P1(s_i)\};\ v_1 \leftarrow {}^{s}i_2 \qquad \ldots..4.4$$

Also, to select the *j*th cluster center, the PDF value calculates using Eq. (4.5).

$$P_j(s_i \leftarrow P_{j-1}(s_i) - P_{j-1}(v_{j-1})\exp\left(-\frac{E2(si,vj-1)}{R2}\right) \text{ for all } i=1..m \ldots\ldots\ldots..(4.5)$$

And the jth cluster centerv_j is selected is using Eq. (4.6).

$$i_j \leftarrow \underset{i=1}{\overset{m}{\operatorname{argmax}}}\{P_j(s_i)\}; v_j \leftarrow s_{ij} \qquad \ldots\ldots\ldots..4.6$$

4.1 Algorithm for PDFCM

Input: *neighborhood radius $R(\sqrt{n})$, c, maximum iterations η (100), error threshold (0.01), and set of user's sessions $S= \{s_1, \cdots, s_m\}$*

Output: *Set of c cluster centers $V= \{v_1, \cdots, v_c\}$and partition matrix P*

define the fixed of cluster centers V(0)

for *i←1, m* ***do***

Calculate the Probability Density values $P_1(s_i)$ using (4.1)

end for

Calculate the I^{st} cluster center $v_1(0)$ equation (4.2)

for *j←2, c* ***do***

for *i←1, m* ***do***

Calculate the revised probability values $P_j(s_i)$ using (4.5)

end for

Calculate the jth cluster center$v_j(0)$ using (4.6)

end for

t ←1

repeat

Calculate the partition matrix P(m) entries:

for *i←1, m* ***do***

for *j←1, c* ***do***

Calculate μij(m)

Stop for

Stop for

After that new cluster center calculate through V(m):

for *j←1, c* ***do***

Compute $v_j(m)$

end for

Calculate the objective function $J_{FCM}(m)$

m ← m +1

until $J_{FCM}(m) - J_{FCM}(m-1)| <$ *t =η*

V. EXPERIMENTALRESULTS

The probabilty density based fuzzy c means algorithm is implemented in MATLAB for session clustering in weblog data. The experimental work is done on processor-intel(R), RAM-4GB, system type-64-bit operating system in windows 8.1 environment.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

5.1 Analysis in Terms of Accuracy

The comparison of the traditional FCM algorithm with proposed PDFCM shows that the PDFCM algorithm perform well over the FCM, which states that the modification in the FCM has improved the performance of the normal FCM algorithm. It is decided to compare the accuracy of both algorithms over the weblog data sets. The comparison is shown diagrammatically.

It clearly shows that accuracy of PDFCM is better than FCM for different number of datasets. The accuracy of both FCM and PDFCM on various sizes of weblog datasets is shown in Fig. 2.

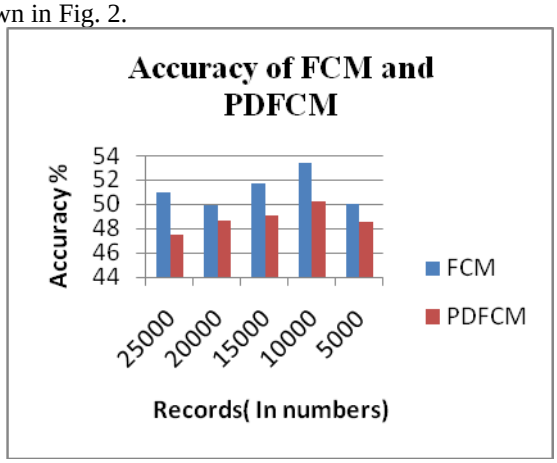

Figure 2: Accuracy of FCM and PDFCM

5.2 Analysis in Term of Execution Time

The execution time required by both FCM and PDFCM algorithms are compared. Which is shown in Table 1 The results are shown diagrammatically in fig.3 also from the diagram, result shows that computational time of PDFCM is lesser than compared to the FCM for each dataset.

Table 1 Running time (In Seconds) of FCM & PDFCM

Record Sets	Running Time (in Seconds)	
	FCM	PDFCM
25000	2.818464	1.537
20000	2.618464	1.437
15000	2.418464	1.337
10000	2.008464	1.237
5000	1.818464	0.537

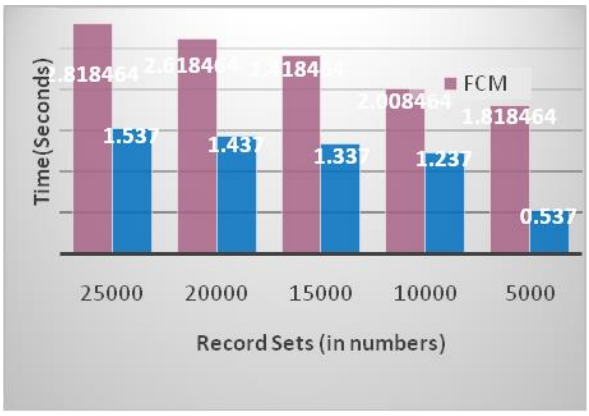

Figure 3 Graphical representation of Execution Time required by FCM and PDFCM

5.3 Analysis in Terms of Memory Requirement

Finally, the memory requirement in terms of bytes is compared between FCM and PDFCM in Table 2 The results are shown diagrammatically as in the Fig.4 From this diagram, it is clear that PDFCM requires less memory than the FCM for all size data set records.

Table 2: Memory requirements for FCM and PDFCM

Record Sets	FCM	PDFCM
25000	5236	4560
20000	4620	3652
15000	3254	3000
10000	2080	1900
5000	1500	1200

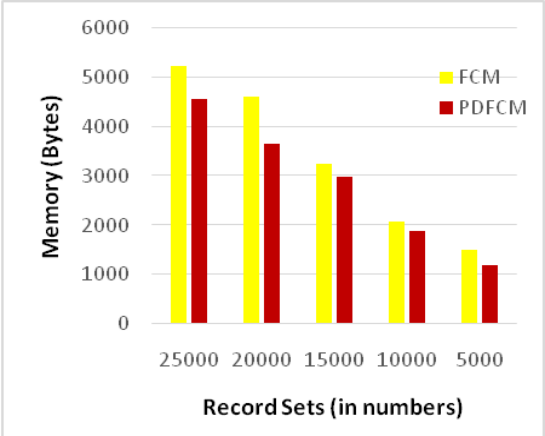

Figure 4 Memory requirements for FCM and PDFCM

VI. CONCLUSION

In this chapter, an enhanced approach for initialization of membership and cluster number for FCM algorithm is presented. And usual random assignment of initial parameter to the FCM algorithm is altered in this approach. Based on experimental results it is clear that PDFCM approach provides better accuracy, reduce running time and take less iteration to complete the experiments. PDFCM algorithm is used to discover web user session cluster. It is better than the FCM because in it prior section of suitable cluster centre is done which was FCM the drawback of algorithm In PDFCM no of iteration reduces from 46-49 to 3. After this it is identify by using index function. That's why PDFCM algorithm works better than FCM by looking at different measurement.

REFERENCES

1. A. Gupta and A. Khandekar, "Development of Weblog Mining Based on Improved Fuzzy C-Means Clustering Algorithm", International Journal of Science, Engineering and Technology Research, Vol.5 (3), pp.688-693, March 2016.
2. A. Kapoor and A. Singhal, "A comparative study of K-Means, K-Means++ and Fuzzy C-Means clustering algorithms", In Proc. of 3rd International Conference on Computational Intelligence & Communication Technology, IEEE, pp. 1-6, 2017.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

3. A. Zahid, A. V. Babuy, W Ahmed and M F Azeemz, "A fuzzy set theoretic approach to discover user sessions from web navigational data", In Proc. of Recent Advances in Intelligent Computational Systems, IEEE, pp. 879-884, 2011.
4. B. Chandra, M. Gupta, and M.P. Gupta, "A multivariate time series clustering approach for crime trends prediction", In Proc of International Conference on Systems, Man and Cybernetics, IEEE, pp. 892-896, 2008.
5. B. Maheswari and P. Sumathi, "A New Clustering and Preprocessing for weblog mining" In Proc. of World Congress on Computing and Communication Technologies, IEEE, pp. 25-29, 2014.
6. B. S. Shedthi, Shetty and M. Siddappa, "Implementation and comparison of K-means and fuzzy C-means algorithms for agricultural data", In Proc. of International Conference on Inventive Communication and Computational Technologies, IEEE, pp. 105-108, 2017.
7. C. Baviskar and S. Patil, "Improvement of data object's membership by using Fuzzy K-Means clustering approach", In Proc. of International Conference on In Computation of Power, Energy Information and Communication, IEEE, pp. 139-147, 2016.
8. C. T. Baviskar and S. S. Patil, "Improvement of data object's membership by using Fuzzy K-Means clustering approach", In Proc. of International Conference on Computation of Power, Energy Information and Communication (ICCPEIC)IEEE, pp. 139-147, 2016.
9. C. Yanyun, Q. Jianlin, G. Xiang, C. Jianping, J. Dan and C. Li, "Advances in research of Fuzzy c-means clustering algorithm", In Proc. of International Conference on Network Computing and Information Security, IEEE, vol. 2, pp. 28-31, 2011.
10. Chen, Y.L. and Huang, C.K., "Discovering fuzzy time-interval sequential patterns in sequence databases", IEEE Transactions on Systems, Man and Cybernetics, Vol. 35(5), pp. 959-972, 2005.
11. D. Koutsoukos, G. Alexandridis, G. Siolas, and A. Stafylopatis, "A new approach to session identification by applying fuzzy c-means clustering on weblogs", In Proc. of Symposium Series on Computational Intelligence, IEEE, pp. 1-8, 2016.
12. G. S. Chandel, K. Patidar and M. S. Mali, "A Result Evolution Approach for Web usage mining using Fuzzy C-Mean Clustering Algorithm", In Proc. of International Journal of Computer Science and Network Security, Vol.16(1). pp.135-140, 2016.
13. H. Gulat, and P. K. Singh, "Clustering techniques in data mining: A comparison", In Proc. of 2nd International Conference on Computing for Sustainable Global Development, IEEE, pp.410-415, 2015.
14. H. X. Pei, Z. R. Zheng, C. Wang, C. Li, and Y. H. Shao, "D-FCM: Density based fuzzy c-means clustering algorithm with application in medical image segmentation", Procedia Computer Science, Vol.122(1), pp. 407-414, 2017.
15. K. Suresh, R. M. Mohana, A. Rama Mohan Reddy, and A. Subramanyam, "Improved FCM algorithm for clustering on web usage mining." In Proc. of International Conference on Computer and Management, pp. 1-4. 2011.
16. P. Sampath and M. Prabhavathy, "Web Page Access Prediction Using Fuzzy Clustering by Local Approximation Memberships (Flame) Algorithm", Vol.10 (7), pp.3217-3220, 2006.
17. S. K. Dwivedi and B. Rawat, "A review paper on data preprocessing: A critical phase in web usage mining process", In Proc. of International Conference on Green Computing and Internet of Things, IEEE, pp. 506-510, 2015.
18. V. Anitha and P. Isakki, "A survey on predicting user behavior based on web server log files in a web usage mining", In Proc. of International Conference on Computing Technologies and Intelligent Data Engineering, IEEE, pp. 1-4, 2016.
19. V. Chitraa, and A. S. Thanamani, "Weblog Data Analysis by Enhanced Fuzzy C Means Clustering", International Journal on Computational Sciences & Applications, Vol.4 (2), pp. 81-95, 2014.
20. Y. Hu, Chuncheng Y. Y. Zuo, and F. Qu, "A cluster validity index for fuzzy c-means clustering", In System Science, In Proc. of International Conference on Engineering Design and Manufacturing Informatization (ICSEM) IEEE, vol. 2, pp. 263-266, 2011.
21. Z. Ansari, S. A. Sattar, A.V. Babu, and M. F. Azeem, "Mountain density-based fuzzy approach for discovering web usage clusters from weblog data, Fuzzy Sets and Systems", Vol.279 (1), pp.40-63, 2015.

AUTHORS PROFILE

Jayanti Mehra received Graduation degree From Barkatullah University Bhopal MP in 2003 and Post Graduation Degree in Computer Science from Makhanlal Chaturvedi National University of Journalism and Communication University Bhopal in year 2008. She is currently pursuing the PhD. Degree in the Department of Computer Applications, Maulana Azad National Institute of Technology Bhopal. M. P. His Research interests include Web Mining, Fuzzy c means and Clustering.

Dr. Ramjeevan Singh Thakur is Associate Professor in the Department of Computer Applications at Maulana Azad National Institute of Technology, Bhopal, India. He is a Teacher, Researcher and Consultant in the field of Computer Science and Information Technology. He earned his Master Degree from Samrat Ashok Technology Institute, Vidisha (M.P.) in 1999 and Ph.D. Degree (Computer Science) From Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal (M.P.) in 2008. His research area are Data Mining, Bioinformatics and Soft Computing

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)