Australian Journal of Basic and Applied Sciences, 9(27) August 2015, Pages: 793-799

ISSN:1991-8178

Australian Journal of Basic and Applied Sciences

Journal home page: www.ajbasweb.com

An Ensemble Link Clusterin`G Model With Swarm Intelligence Based Centroid Selection For Categorical Dataset

[1]N. Yuvaraj and [2]Dr.C.Suresh Gnana Dhas

[1]*Research Scholar, Department of Computer Science and Engineering, St. Peter's Institute of Higher Education and Research, St.Peter's University, Avadi, Chennai.*
[2]*Professor & Head, Vivekanandha College of Engineering for Women, Tiruchencode.*

ARTICLE INFO

Article history:
Received 12 July 2015
Accepted 28 August 2015
Available online 15 September 2015

Keywords:
Data Mining, Classification, Categorical Data, Cluster Ensembles, Jensen-Shannon Divergence (JSD), Firefly algorithm (FA), Link based Clustering Ensemble (LCE), University of California Irvine (UCI), Division Reminder Hashing (DRH).

ABSTRACT

This paper focuses on formulating the cluster ensemble framework with categorical data clustering. In previous approaches, initial Centroid values selection has been a tough issue to handle. In order to overcome the issue of initial Centroid selection, a swarm intelligence algorithm has been used in this research work for optimal solution. Firefly Algorithm (FA) has been used in this work due to its significance in attaining the optimal results. Thus, Link Clustering Ensemble Model with Firefly Algorithm (LCEMF) has been proposed in this work. LCEMF is employed to determine respective link based similarity measure in line with the categorical data. Then, a graph partitioning approach has been deployed to respective weighted bipartite graph which is formulated based on refined matrix. Finally, Support Vector Machine (SVM) classification technique has been used and applied for all relevant clustering techniques. The simulation results have been carried out on UCI data. It is observed that proposed LCEMF technique generates comparatively improvised results when compared with the existing techniques on evaluating in the context of cluster validity indexes.

To Cite This Article: N. Yuvaraj and Dr.C.Suresh Gnana Dhas., An Ensemble Link Clusterin`G Model With Swarm Intelligence Based Centroid Selection For Categorical Dataset. ***Aust. J. Basic & Appl. Sci.,*** *9(27): 793-799, 2015*

INTRODUCTION

Data clustering may be considered as a device based learning method that inherently has several integral practical applications, like sales data grouping that enables disclosure of consumer-buying behavior, or network data grouping network that provides insights into several communication prototypes. There are several popularly used clustering algorithms, which include k-means (Junjie Wu, 2012) and Partitioning Around Medoids (PAM) with two major steps that involve building and swapping (Kaufman, L. and P.J. Rousseeuw, 2009) for the facilitating numerical data by measuring a Euclidean distance between feature vectors (Huang-Cheng Kuo, *et al*., 2007; Liviu Octavian Mafteiu-Scai, 2013). Though, direct application of these is not feasible for categorical data clustering, wherein domain values are distinct and possess no definitive ordering as such.

Resultant from this is that there has been an introduction of several categorical data clustering algorithms in last couple of years, with respective application to relevant domains like in the case of handling uncertainty in process of clustering (Darshit Parmar, *et al*., 2007), Context-Based Distance learning (Dino Ienco, Ruggero G. Pensa, Rosa Meo, 2009) employed for real world as well as synthetic datasets, Concept-Drifting Categorical Data (Hung-Leng Chen *et al*., 2009). Every algorithm as has as such possesses its respective positive outcomes and drawbacks. Consequently, for users it becomes rather tough to determine which particular algorithm essentially would be the precise alternative taking in to consideration a specified data set. Though the above mentioned issues are contributing factors to the motivations at the back of cluster ensembles (Iam-On, N., *et al.*, 2001), there still exists an additional interesting application that comprises of the possibility of reusing existing knowledge regarding the data.

Lately, cluster ensembles (Vega-Pons, S., J. Ruiz-Shulcloper, 2011; Li, T.Y., Y. Chen, 2010) have gained popularity as a method used to overcome issues associated with clustering algorithms. Though, it is also apparent that clustering techniques possibly can find varying patterns for any particular data set. Essentially this happens as each clustering algorithm possesses it respective bias which is an outcome from optimization of several criterion. Additionally, there is no evidence against which results from clustering can thus be validated.

Corresponding Author: N. Yuvaraj, Research Scholar, Department of Computer Science and Engineering, St. Peter's Institute of Higher Education and Research, St.Peter's University, Avadi, Chennai.
E-mail: yraj1989@gmail.com

Resultantly existing cluster ensemble method performance possibly can be degraded as several matrix entries are not known, first Centroid selection of the k means clustering then poses several related issues and duplicate entries inherent in the cluster cannot thus be eliminated.

The following paper is aimed at introducing an innovative Firefly algorithm along with a link-based clustering ensemble framework approach that is aimed to refine matrix mentioned before, thus providing significantly fewer unknown entries. The initial Centroid values that are present in the clustering ensemble framework are chosen on the basis of the Firefly Algorithm (FA). Duplicate entries are eliminated from the cluster by employing the DRH techniques. In the presented study link-based similarity measure (Boongoen, T., Q. Shen and C. Price, 2010) has been exploited in order to ascertain unknown values on the basis of link network of clusters.

Proposed Methodology:

The following paper what has been suggested the LCEMF which has proven to quite effective, and differs from prevalent link based clustering ensemble framework. As suggested work chooses cluster Centroid values on the basis of fireflies behavior with respect to Fixed-k, Random-k clustering technique to formulate cluster ensemble framework. Current work complete generates cluster ensemble framework through the process of randomly choosing cluster Centroid values for every clustering techniques. Also reduction has to be feasible for the removal of duplicate data from inherent clusters, through the deployment of DRH technique which eliminates duplicate data present in Link-Based Similarity function. Duplicate data is eliminated so that the Consensus Function may then be applicable with respect to the Refined Matrix (RM). Checking of the results obtained from clustered data SVM classification techniques may be applied which basically segregate clustered data as positive and negative classes. Figure 1 illustrates proposed LCEMF methodology.

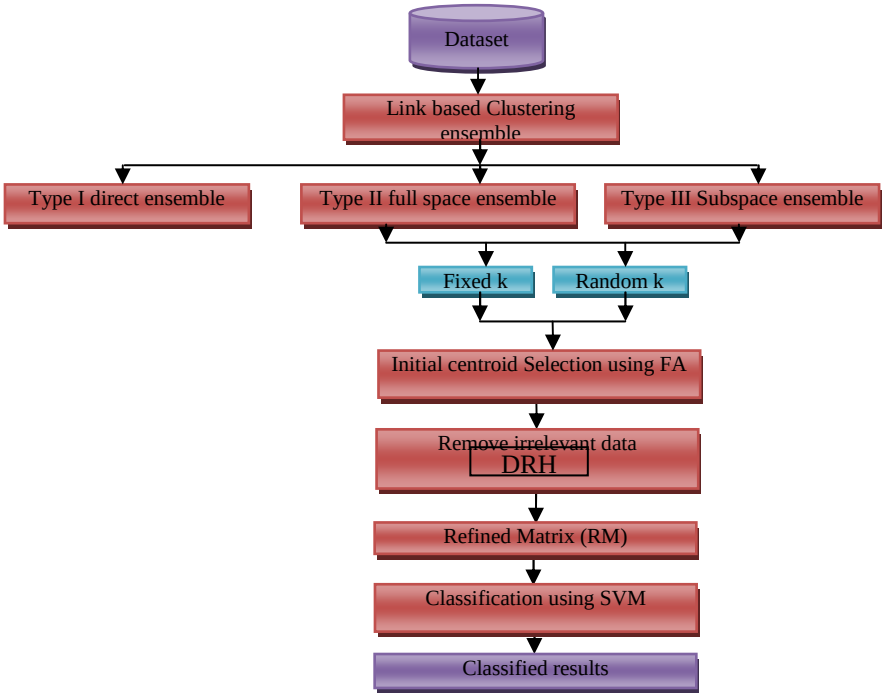

Fig. 1: Architecture representation of the proposed work

There are basically five main steps: 1) Generating base clustering's in order to formulate a cluster ensemble (π),2) Creation of base clusters for choosing initial Centroid values based on FA 3) Creation of a well refined cluster-association matrix or RM by deploying a link-based similarity algorithm, and 4) Eliminating duplicate data from existing cluster on the basis of DRH , thereafter which generating final data partition (π^*) through the exploitation process of spectral graph partitioning method considering the same as a consensus function. 5) In conclusion classify clustered data by employing the SVM.

Ensemble framework:

Type I, As such direct ensemble pursues same procedure as the study in (Can Gao, Witold Pedrycz, Duoqian Miao, 2013), wherein the first cluster ensemble type transforms issues related to categorical data .Let $D= \{d_1,\ldots d\}$ be a set of data points, $A= \{a_1,\ldots a\}$ be a set of categorical attributes, and $\pi= \{\pi_1,\ldots \pi\}$ be a set of partitions. Each partition is

Australian Journal of Basic and Applied Sciences, 9(27) August 2015, Pages: 793-799

generated for specific categorical attributes values of the attribute $\pi i = \{C_i^1$, (where $U^k i_{j=1}\ C^i_j = a_i$ and k_i is the number of values of attribute a_i. Using the formalism, categorical data thus can be transformed directly as a cluster ensemble, without implementation of any base clustering as such. Type II Full-space ensemble that pursues after the two ensemble types is engendered on the basis of base clustering results, wherein each of these is ascertained through the application of the clustering algorithm to the categorical data set. In the study here, the k-modes method (Liang Bai, *et al*., 2013) has been employed to create base clustering's, wherein each has random initialization of cluster centers. Two schemes mentioned have been deployed for selecting a definite number of clusters for each base clustering's: 1) Fixed-k, $K = \sqrt{N}$ (where N is the number of data points and 2) Random-k, $k \in \{2,| [\sqrt{N}]\}$. However for these methods the first Centroid selection is still selected in a random manner, it is overcome by FA for every point in the dataset.

Firefly algorithm (FA) for initial Centroid selection:

Fireflies utilize these generated flashes in order to communicate and attract prey. The following work here shows each firefly's selection process of the best Centroid values on the basis of flashing behavior using rhythmic flashes, the least distance that exists amid two, and then in that case one among the fireflies is then categorized as the best Centroid point for the cluster data points. YANG employed this particular firefly behavior and thus introduced in the year 2010 the Firefly Algorithm (Yang, X.S., 2009).

1) Fireflies are unisexual hence one definite Centroid point (firefly) has the tendency to attract other data point irrespective of sex;

2) Attractiveness and brightness are both proportional to each other. Hence in the case of two random flashing data points one which is less bright automatically will move closer to the one that is brighter as both attractiveness and brightness are proportional, also they decrease with increase in Jensen-Shannon divergence distance.

3) Cluster Centroid point brightness is influenced by cluster distance and is considered as the objective function. In the case of maximization, brightness and objective function value can be proportional (Yang, X.S., 2009). Now defining firefly attractiveness:

$$\beta(r)=\beta_0 e^{-\gamma r^2} \quad (1)$$

where, β_0 is the attractiveness at r=0 and γ is the light absorption coefficient at the source. It should be noted that the $r_{i,j}$ which is described by equation 2 for distance between two data points in the cluster and respectively. JSD is a measure of the "distance" between two probability distributions of the cluster Centroid and the cluster data point which can also be generalized to measure the distance (similarity) between a finite numbers of distributions. JSD is a natural extension of the Kullback-Leibler divergence (KLD) to a set of distributions compute the JSD of a set of size n as

$$JS(p_1,p_2,\ldots,p_n) = H(\sum^n_{i=1} w_i p_i) - \sum^n_{i=1} w_i H(p_i) \quad (2)$$

where w_i is the vote weight of the i-th cluster in the set; and H(P) is the Shannon entropy of the distribution $P = \{p_j, j= 1,\ldots K\}$, , defined as

$$H(P) = -\sum^k_{j=1} p_j \log p_j \quad (3)$$

The estimative of p_i and p_j can be calculated by any method, in this paper use Parzen windows (Nakariyakul, S. and D. Casasent, 2008) to obtain $P_i(D)$ and $P_j(D)$ in equation (4).

$$P(D_j) = 1/n \sum^n_{i=1}\ 1/V_n \varphi(D_i - D_m/h_n) \quad (4)$$

Where $\varphi(x)$ is the window function and n is the number of data points. V_n and h_n are the volume and edge length of a hypercube where the function will be evaluated. The movement of a Centroid value Firefly i, which is attracted another cluster Firefly is determined by:

$$D_i = Di + \beta_0 e^{-\gamma r^2} (D_j - D_i) + \alpha(rand - 1/2) \quad (5)$$

Where, the second term is the attraction while the third term is randomization including randomization parameter α and the random number generator rand which its numbers are uniformly distributed in interval [0, 1]. For the most cases of implementations β_0 and $\alpha \in [0, 1]$. In the most applications, it typically varies from 0.01 to 100. After the Centroid values selected then clustering is formed for both fixed k and random k clustering methods. Type III -Subspace ensemble another alternative to generate diversity within an ensemble is to exploit a number of different data subsets. Similar to the study in (Yu, Z., *et al*., 2007), for a given N> data set of N data points and d attributes, an N) data subspace (where q <d) is generated by,

$$q = q_{min} + [\delta (q_{max} - q_{min})] \quad (6)$$

Where δ is a uniform random variable, q and q_1 are the lower and upper bounds of the generated subspace, respectively. In particular, q and q_1 are set to 0. An attribute is selected one by one from the pool of d attributes, until the collection of is obtained. The index of each randomly selected attribute is determined as h= [1+ ω(d)], given that h denotes the hth attribute in the pool of d attributes and $\omega \in (0, 1)$ is a uniform random variable.

Generating a Refined Matrix (RM):

The refined cluster-association matrix is put forward as the enhanced variation of the original Binary Matrix (BM). Its aim is to approximate the

value of unknown associations ("0") from known ones ("1"), whose association degrees are preserved within the Refined Matrix (RM). Given a cluster ensemble of a set of data points , a weighted graph G= (V; W) can be constructed, where V is the set of vertices each representing a cluster and W is a set of weighted edges between clusters. Formally, the weight assigned to the edge wde∈W, that connects clusters C_d, $C_e \in V$, is estimated by the proportion of their overlapping members.

$$w_{de} = | L_d \cap L_e | / | L_d \cup L_e | \quad (7)$$

It is a number between 0 and 1; it is 0 when the two cluster data points are disjoint, 1 when they are equal, and strictly between 0 and 1 otherwise. It is a commonly used indicator of the similarity between two cluster in the data set are more similar when their W_{xy} is closer to 1, and more dissimilar when their W_{de} is closer to 0 duplicates are removed based on the calculation of the hash value, here the hash value of the two cluster data point sets such as L_d, L_e are calculated based on the Division Reminder Hashing (DRH). In division reminder method, key L_d, is divided by m larger than the number n of keys in L_d, and the remainder of this division is taken as index into the hash table i.e.

$$H(L_d) = L_d \bmod m \quad (8)$$

The number m is usually chosen to be a prime number, since this frequently minimizes the number of collisions. The above hash function for each one of the cluster will map the keys in the range 0 to m-1; But if want the hash addresses to range from 1 to m rather than from 0 to m-1use the formula,

$$H(L_d) = L_d \bmod m + 1 \quad (9)$$

Let be a hash function that maps the members of x and y to distinct clusters, and for any cluster in the set S define $h_{min}(S)$ to be the member x of S with the minimum value of h(x). Then $h_{min}(L_d) = h_{min}(L_e)$ exactly when the minimum hash value of the union $L_d \cup L_e$ lies in the intersection $L_d \cap L_e$. Moreover, our proposed algorithms are fully automated and robust without requiring many parameters. In the above steps performs the similarity between the two clusters, in order to measure the similarity between the pages in the cluster extend this work to weighted triple quality matrix (WTQ). WTQ aims to differentiate the significance of triples and hence their contributions toward the underlying similarity measure. WTQ is inspired by the initial measure which evaluates the association between home pages. In particular, features of the compared pages p_a and p_b are used to estimate their similarity s (p_a and p_b) as follows,

$$s(p_a \text{ and } p_b) = \sum_{\forall zc \in z} 1/\log(\text{frequency}(z_c)) \quad (10)$$

where Z denotes the set of features shared by home pages p_a and p_b and frequency (z_c) represents the number of times z_d appearing in the studied set of pages. The quality of each cluster is determined by the rarity of links connecting to other clusters in a network. WTQ measure of clusters C_d,$C_e \in$ V with respect to each triple $C_k \in$ V is estimated by,

$$WT\ Q^k_{de} = 1/W_k \quad (11)$$

Here, W_k is defined as $Wk = \sum_{\forall t \in Nk} w_{tk}$, that is directly linked to the cluster C_k, such that $\forall C_t \in N_k$, $w_{tk} \in$ W. The accumulative WTQ score from all triples (1…q) between clusters C_d and C_e can be found as follows:

$$WT\ Q_{de} = \sum_{k=1}^{q} WT\ Q^k_{de} \quad (12)$$

Following that, the similarity between clusters and can be estimated by

$$Sim(C_d, C_e) = WT\ Q_{de} / WT\ Q_{max} \text{ x } DC \quad (13)$$

Where $WT\ Q_{max}$ is the maximum $WT\ Q_{de}$ value of any two clusters C_d, $C_e \in V$ and DC ∈ (0, 1) is a constant decay factor (i.e., confidence level of accepting two non identical clusters as being similar). With this link-based similarity metric, $Sim(C_d, C_e) \in [0,1]$ with $Sim(C_d, C_e) = 1$, $C_x, C_y \in V$. It is also reflexive such that $Sim(C_x, Cy)$ is equivalent to $Sim(C_y, C_x)$.

Applying a Consensus Function to RM:

Having obtained an RM, a graph-based partitioning method is exploited to obtain the final clustering. This consensus function requires the underlying matrix to be initially transformed into a weighted bipartite graph. Given an RM representing associations between N data points and P clusters in an ensemble π, a weighted graph G= (V, W) can be constructed, where $V = V^d \cup V^e$ is a set of vertices representing both data points V^d and clusters V^{ϵ}, and W denotes a set of weighted edges that can be defined as follows, $w_{ij} \notin W$ when vertices $v_i, v_j \in V^d$, $w_{ij} \notin W$ when vertices $v_i, v_j \in v^e$ and Otherwise $w_{ij} = RM(v_i, v_j)$ when vertices $v^i \in v^d$ and $v^j \in v^e$. Note that the graph G is bidirectional such that w_{ij} is equivalent to w_{ji} . Given such a graph, a spectral graph partitioning method (Luxburg, U., 2007) is applied to generate a final data partition.

Support Vector Machine (SVM) for classification:

In the study here during final stage executing classification technique to clustered Conesus function matrix, the Support Vector Machine or SVM is deployed. SVM algorithm selects hyperplane which can optimally split clustered Conesus function clustered matrix data. Categorical clustered data

points are classified thus as under positive or negative classes. Hyperplane decision function is as follows:

$$f(x)=sgn((w,\phi(x))+b \quad (14)$$

Where **w** is the weight vector for clustered data, orthogonal to the hyper plane, "b" is a scalar that represents the margin of the hyper plane, "x" is the current clustered sample tested, $\phi(x)$ is a kernel function that transforms the input data into a higher dimensional feature space and ","representing the dot product. Sgn is the signum function. If w has unit length, then $(w,\phi(x)$ is the length of $\phi(x)$ along the direction of w. Kernel defines a Mercer Kernel according to Mercer theorem given in (XIA Guo-en and SHAO Pei-ji, 2009). This gives the mapping in to clustered data as,

$$\phi(x) = (\sqrt{\lambda_1}\Psi_1(x)\sqrt{\lambda_2}\Psi_2(y),\ldots.)^T \quad (15)$$

Experimentation Results:

Following section here evaluates suggested Firefly Algorithm with link based cluster ensemble framework or LCEMF, by employing a several validity indices as well as real data sets. Experimental evaluation has been executed for more than five data sets here. The "20Newsgroup" data set (http://people.csail.mit.edu/jrennie/20Newsgroups/.) is a subset of the well-known text data 1,000 documents collection from the UCI Machine Learning Repository (Asuncion, A. and D.J. Newman, 2007). Their details are summarized in Table 1, Number of Data Points (N), Attributes (d), Attribute Values (AV), and Classes (K). Missing values (denoted as "?") in these data sets are simply treated as a new categorical value. In particular, the frequency ($f \in \{0,1\ldots\infty\}$) that a key word appears in each document is transformed into a nominal value: "Yes" if $f > 0$, "No" otherwise.

There five cluster ensembles types investigated for the following evaluation: Type-I, Type-II (Fixed-k), Type-II (Random-k), Type-III (Fixed-k), and Type-III (Random- k). The k-modes clustering algorithm is specifically used to generate the base clustering's with clustering methods such as LCEMF, Link-Based Cluster Ensemble (LCE), Similarity matrix (CO) with single linkage (CO+SL), CO with average linkage (CO+AL), Cluster-based Similarity Partitioning Algorithm (CSPA), Hyper-Graph Partitioning Algorithm (HGPA) and proposed optimization based clustering methods. Table 2 shows the classification accuracy of the SVM methods for above mentioned method for all the dataset in the Table 1.

Table 1: Data Sets

Dataset	N	D	AV	K
Zoo	101	16	36	7
Lymphography	148	18	59	4
20 Newsgroup	1000	6084	12168	2

Table 2: Classification Accuracy of Different Clustering Methods

Dataset	Ensemble type	LCEMF	LCE	CO+SL	CO+AL	CSPA	HGPA
Zoo	I	0.932	0.894	0.883	0.889	0.801	0.594
	II-Fixed k	0.941	0.921	0.882	0.899	0.824	0.836
	II-Random k	0.942	0.931	0.861	0.873	0.824	0.826
	III-Fixed k	0.954	0.941	0.891	0.916	0.823	0.831
	III-Random k	0.946	0.931	0.862	0.873	0.825	0.835
Breast Cancer	I	0.956	0.94	0.652	0.653	0.849	0.673
	II-Fixed k	0.9856	0.972	0.653	0.954	0.836	0.873
	II-Random k	0.9863	0.973	0.653	0.943	0.812	0.870
	III-Fixed k	0.9835	0.969	0.649	0.952	0.831	0.849
	III-Random k	0.9756	0.964	0.653	0.932	0.812	0.844
20 Newsgroup	I	0.789	0.61	0.6	0.6	0.6	0.565
	II-Fixed k	0.831	0.782	0.6	0.6	0.6	0.6
	II-Random k	0.783	0.72	0.6	0.6	0.6	0.6
	III-Fixed k	0.823	0.726	0.6	0.6	0.6	0.6
	III-Random k	0.818	0.6	0.6	0.6	0.6	0.6

Table 3: Pairwise Performance Comparison among Examined Clustering Methods

Ensemble type	Methods	CA		NMI		AR	
		B	W	B	W	B	W
I	LCEMF	187	28	146	53	164	53
	LCE	170	35	137	65	149	61
	CO+SL	34	180	81	136	47	163
	CO+AL	72	143	114	93	84	120
	CSPA	105	92	132	82	109	95
	HGPA	21	193	19	208	24	201
II – Fixed K	LCEMF	225	3	228	5	212	09
	LCE	208	4	204	9	201	13
	CO+SL	26	171	37	149	28	163
	CO+AL	131	46	134	42	134	39

	CSPA	86	91	68	101	82	103
	HGPA	64	80	93	92	93	91
II –Random K	LCEMF	235	2	214	8	219	08
	LCE	209	4	203	12	201	11
	CO+SL	17	173	38	159	28	170
	CO+AL	97	80	94	63	117	53
	CSPA	76	113	47	131	51	125
	HGPA	67	121	41	139	49	132
III –Fixed K	LCEMF	219	5	203	08	197	09
	LCE	197	7	191	13	182	16
	CO+SL	17	167	37	142	32	162
	CO+AL	115	43	119	43	139	29
	CSPA	73	87	53	94	61	106
	HGPA	71	82	56	101	65	101
III –Random K	LCEMF	227	3	206	09	196	09
	LCE	203	6	191	13	181	11
	CO+SL	15	178	34	146	32	163
	CO+AL	81	67	81	59	103	43
	CSPA	52	112	36	121	45	123
	HGPA	51	117	38	134	63	131

Table 3 shows every technique's frequencies of significant better (B) performance, categorized as per evaluation indices Normalized Mutual Information (NMI), Adjusted Rand (AR) and Classification Accuracy (CA). "B" and "W" indicate number of times that a specific technique ascertains which is "better" and "worse" than the others. Effectively, resulting technique performance depends on decay factor (i.e., DC $\in$ [0, 1]) varied from 0.1 through 0.9, in steps of 0.1 with ensemble size (M) of 10. The following Figure 2 shows LCEMF results for ensemble type I across varying ensemble types, and are dependent on DC value. The performance of the models whose values are presented in X-axis and Y-axis, it indicative that proposed LCEMF generates higher results than current techniques. Figure 3 shows LCEMF results for ensemble type II with fixed K parameter and LCEMF delivers higher results than current techniques as shown.

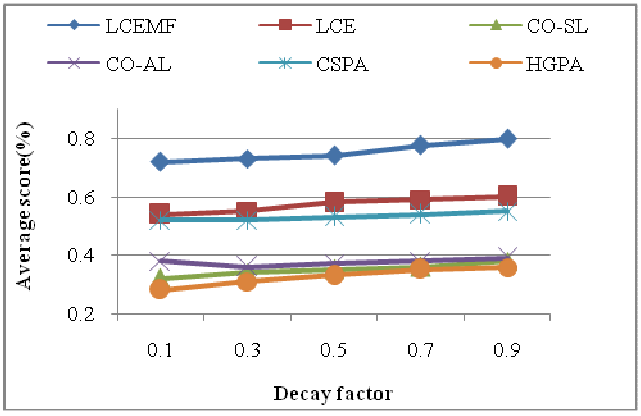

Fig. 2: Ensemble Type I vs. decay factor

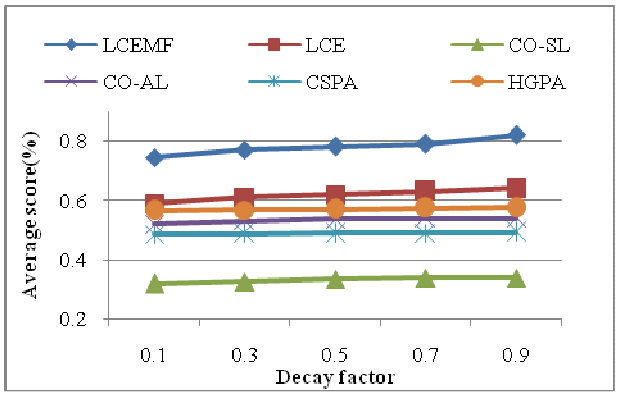

Fig. 3: Ensemble Type II-Fixed-k vs. decay factor

Conclusion And Future Work:

A novel LCEMF has been proposed in this work for categorical data**.** During the cluster ensemble phase, an efficient swarm intelligence algorithm called Firefly algorithm which is based on the biological behavior of the fireflies has been used to calculate the centroid values. Then, it further carries out transformation of original categorical data matrix into an information-preserving numerical variation RM, duplicate entries are eliminated using DRH wherein application is direct of effective graph partitioning method. Issue of constructing RM is thus effectively resolved using similarity amid categorical labels, by employing Weighted Triple-Quality similarity algorithm. SVM classification is carried out on clustered data to analyze LCEMF clustering technique results; also what has been suggested is a new LCEMF for categorical data that almost is reflective of the clustering algorithm's performance. The performance accuracy of the proposed LCEMF methods is 0.954 % for zoo dataset , 0.9863 % for breast cancer dataset and 0.823 % for 20 Newsgroup which is high when compare to all methods with different dataset samples. The proposed LCEMF approach attains almost 5 % higher than the existing LCE methods.

REFERENCES

Junjie Wu, 2012. " Advances in K-means Clustering: A Data Mining Thinking", Springer, 190.

Kaufman, L. and P.J. Rousseeuw, 2009. Finding Groups in Data: An Introduction to Cluster Analysis, vol. 344, John Wiley & Sons, New York, NY, USA.

Huang-Cheng Kuo, Nat. Chiayi Univ., Chiayi, Pei-Yuan Jou, Jen-Peng Huang, 2007." Adaptive Weighting Distance for Feature Vectors of Biological Sequences", 2007 International Conference on Machine Learning and Cybernetics, 4: 2269-2273.

Liviu Octavian Mafteiu-Scai, 2013. " A New Dissimilarity Measure between Feature-Vectors", International Journal of Computer Applications (0975 – 8887), 64(17): 39-44.

Darshit Parmar, Teresa Wu, Jennifer Blackhurst, 2007. " MMR: An algorithm for clustering categorical data using Rough Set Theory", 25th International Conference on Conceptual Modeling Data & Knowledge Engineering, 63(3): 879-893.

Dino Ienco, Ruggero G. Pensa, Rosa Meo, 2009. " Context-Based Distance Learning for Categorical Data Clustering", Advances in Intelligent Data Analysis VIII Lecture Notes in Computer Science, 5772: 83-94.

Hung-Leng Chen, Ming-Syan Chen, and Su-Chen Lin, 2009. " Catching the Trend: A Framework for Clustering Concept-Drifting Categorical Data", IEEE transactions on knowledge and data engineering, 21(5): 652-665.

Iam-On, N., T. Boongoen, S. Garrett and C. Price, 2011. A link-based approach to the cluster ensemble problem. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 33(12): 2396-2409.

Vega-Pons, S., J. Ruiz-Shulcloper, 2011. A survey of clustering ensemble algorithms. International journal patterns Recognition Artificial Intelligence, 25(3): 337-372.

Li, T.Y., Y. Chen, 2010. Fuzzy clustering ensemble with selection of number of clusters. J Comput, 5(7): 1112-1118.

Boongoen, T., Q. Shen and C. Price, 2010. "Disclosing False Identity through Hybrid Link Analysis," Artificial Intelligence and Law, 18(1): 77-102.

Can Gao, Witold Pedrycz, Duoqian Miao, 2013. "Rough subspace-based clustering ensemble for categorical data" , Soft Computing, 17(9): 1643-1658.

Liang Bai, Jiye Liang, Chuangyin Dang and Fuyuan Cao, 2013. "The Impact of Cluster Representatives on the Convergence of the K-Modes Type Clustering", IEEE transactions on pattern analysis and machine intelligence, 35(6): 1509-1522.

Yang, X.S., 2009. Firefly algorithm for multimodal optimization. in: stochastic Algorithms: foundations and applications SAGA lecture notes in computer sciences, pp: 169-178.

Nakariyakul, S. and D. Casasent, 2008. "Improved forward floating selection algorithm for feature subset selection," in IEEE Int. Conf. Wavelet Analysis and Pattern Recognition, ICWAPR'2: 793-798.

Yu, Z., H.-S. Wong and H. Wang, 2007. "Graph-Based Consensus Clustering for Class Discovery from Gene Expression Data," Bioinformatics, 23(21): 2888-2896.

Luxburg, U., 2007. "A Tutorial on Spectral Clustering," Statistics and Computing, 17(4): 395-416.

XIA Guo-en and SHAO Pei-ji, 2009. "Factor Analysis Algorithm with Mercer Kernel", IEEE Second International Symposium on Intelligent Information Technology and Security Informatics.

http://people.csail.mit.edu/jrennie/20Newsgroups/.

Asuncion, A. and D.J. Newman, 2007. "UCI Machine Learning Repository," School of Information and Computer Science, Univ. of California,
http://www.ics.uci.edu/~mlearn/MLRepository. html.