International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 15 (2018) pp. 11793-11801

Ensemble Distributed Noun Attribute Selection based on its First Appearance for Text Document Clustering

Dr. S.Vijayalakshmi

Assistant Professor, Department of Computer Applications, NMS S.Vellaichamy Nadar College, Madurai, India.

Abstract

It has been proved that the First appeared words are more important. In order to improve the word limit, first appeared noun is considered in each part of the document in a corpus and a novel ensemble attribute selection methods called Ensemble Distributed Noun Attribute Selection using its First Appearance (named EDNAFA). EDNAFA is proposed and HCLK-Means clustering algorithm and Clustering with Flocking Algorithm are used. Both theoretical certifications on some schematic examples and numerical results on a suit of computer-generated and it shows that EDNAFA can be used to generate efficient clusters, which provides meaningful evidence that EDNAFA can improve the performance of the clustering on 20 Newsgroups and Specific Crime Judgment Corpus.

Keywords: ensemble distributed attribute selection methods, RiTa WordNet, HCLK-Means algorithm, Flocking Algorithm

Biographical Notes: Dr. S.Vijayalakshmi is Assistant professor in Department of Computer Applications, NMS S.Vellaichamy Nadar College, and Madurai. Her current research interests include data mining, text mining, clustering and classification, artificial neural network and statistical learning theory. She is author of many research papers published at both national and international journals, conference proceedings.

INTRODUCTION

In recent years, the use of attribute selection for knowledge discovery has become increasingly important in many domains that are characterized by a large number of features, but a small number of samples. Classic examples of such areas include text mining, bioinformatics and biomedical field, where the number of attributes (problem with high dimensional space) often beats the number of samples by orders of magnitude [1]. Domain specialists would like a stable attribute selection algorithm over an unstable one when solely small changes are made to the dataset. Robust attribute selection techniques would permit domain experts to have more confidence in the selected features, as in most cases these attributes are subsequently analyzed further, requiring much time and effort, especially in text mining applications.

Ensemble Multi-label Feature selection algorithm based on data entropy [2] has become a growing research field that is freelance on any existing classifiers. Its elementary idea consists of exploitation the knowledge gain to assess the correlation between the feature and also the label set, and filtering out appropriate attributes additional with efficiency. They [2] calculated the information gain in an ensemble framework and separate valuable attributes in keeping with the threshold value determined by the effective factor. Attribute dimension reduction can be characterized as attribute extraction and attribute selection [3]. Presently, numerous approaches of attribute dimension reduction have been developed in multi-label classification, but the mainstream attention on attribute extraction. Feature extraction maps attribute variables from high dimensional area to low dimensional area.

Attribute selection removes inapplicable and redundant attribute out of original ones. Generally attribute selection has a triple purpose [4], refining the performance of classifier, creation classifiers faster, and providing a better understanding of the process that generated data. Recently a number of algorithms of feature selection are planned [5][6]. However, the computation quality of them depends on the concrete classifiers and therefore the attribute set varies with the classifiers. The ensemble method will work better than a traditional method. Since the ensemble method can overcome the errors some of the existing classifiers introduce embedded methods use internal information of the classification model to perform feature selection (e.g. use of the weight vector in support vector machines). They often provide a good trade-off between performance and computational cost [7].

Current work in this area mostly focuses on the stability document representation to be used for attribute selection, introducing measures based on Hamming distance[8], correlation coefficients[9], consistency and information theory[10]. Kalousis and coworkers also present an extensive comparative evaluation of feature selection stability over a number of high-dimensional datasets[9], However, most of this work only focuses on the stability of single attribute selection techniques, an exception being the work of [8] which describes an example combining multiple feature selection runs.

In this analysis, we tend to investigate whether or not the utilization of ensemble feature selection techniques is used to yield additional robust feature selection techniques, and whether or not combining multiple strategies has any impact on the clustering performance. The explanation for this concept stems from the sphere of ensemble learning, wherever multiple (unstable) attribute selection strategies are combined to yield a more stable, and higher performing ensemble attribute selection.

The rest of the paper is organized as follows. In Section 2, we tend to propose the definition of the distribution of Noun with first Appearances. In Section 3, the schematic examples in are going to be used to prove the development of ensemble feature selection strategies. In Section 4, HCLK-Means Clustering and Clustering with Flocking algorithm is selected to our proposed ensemble attributes. In Section 5, we compare the performance of our ensemble attribute selection methods by applying in the clustering with a real and synthetic datasets. Finally, conclusions and future work are mentioned in Section 6.

DISTRIBUTED NOUN ATTRIBUTE BASED ON ITS FIRST APPEARANCE

The nouns are extracted as shown in the following figure 1 from each and every document in the training corpus. In this research, we have considered fifty percentages of the documents as training corpus and the remaining fifty percentage of the document as testing corpus.

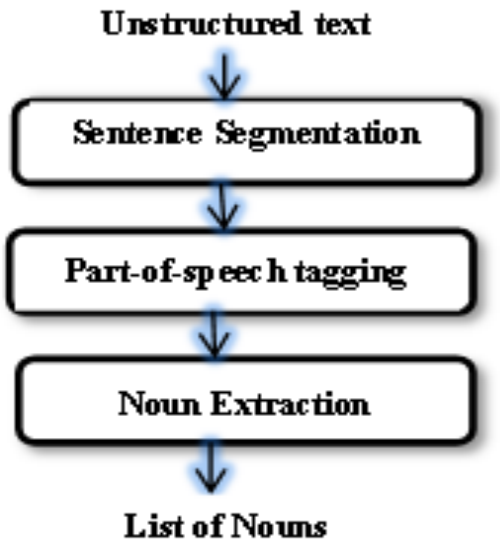

Figure 1. Shows initial level of Nouns Extraction

Each document in training document corpus is ten fixed partition or paragraph. Each partition of document's size may vary. Two dimensional matrixes are constructed for noun document matrix as follows.

$$\begin{array}{c} \\ N0 \\ N1 \\ \vdots \\ \vdots \\ Nm \end{array} \begin{array}{c} \begin{array}{cccccc} P0 & P1 & P2 & \cdots & \cdots & Pn \end{array} \\ \begin{bmatrix} v00 & v01 & v02 & \cdots & \cdots & v0n \\ v10 & v11 & v12 & \cdots & \cdots & v1n \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ vm0 & vm1 & vm2 & \cdots & \cdots & vmn \end{bmatrix} \end{array} \quad (1)$$

N0, N1, N2... Nm is noun attributes and extracted from the training corpus based on its occurrence (see Figure 1). P0, P1, P2...Pn is part number of the each document in the training corpus.

If the Noun (N0) is present in the Part of the document (P0), then the value (V00) is concern part number, otherwise the default value specified by the user. The above matrix will be constructed separately for each and every document in the training corpus. The distribution of first appeared noun is calculated using the following formula (see Equation 2) from the above matrix (see Equation 1).

$$FA(N0) = \min\{v00, v01, v02, \cdots v0n\} \quad (2)$$

Example:

$$FA(N0) = \min\{5,1,3,10,6\} = 1$$

We have considered the noun (N0) is present in the position 5(v00) of part 0(P0), N0 is present in position 1 (v01) of part1 (P1), N0 is present in position 3 (v02) of part2 (P2), N0 is present in position 10 (v03) of part3 (P3) and N0 is present in position 6 (v05) of part4 (P4). Minimum value is considered the first appearance of that noun. The same procedure is applied to find its first appearance for N1, N2, N3, etc. for document 0. The same process is repeated for remaining documents (d0, d1, d2, etc.) in the training corpus. Finally the first appearance of Noun- document vector space model is represented as follows (see Equation 3).

$$\begin{array}{c} \\ N0 \\ N1 \\ \vdots \\ Nm \end{array} \begin{array}{c} \begin{array}{cccccc} D0 & D1 & D2 & \cdots & \cdots & Dn \end{array} \\ \begin{bmatrix} FA(N0,D0) & FA(N0,D1) & FA(N0,D2) & \cdots & \cdots & FA(N0,Dn) \\ FA(N1,D0) & FA(N1,D1) & FA(N1,D2) & \cdots & \cdots & FA(N1,Dn) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ FA(Nm,D0) & FA(Nm,D1) & FA(Nm,D2) & \cdots & \cdots & FA(Nm,Dn) \end{bmatrix} \end{array} \quad (3)$$

We have implemented and used four different weighting functions to evaluate the importance of First appeared Noun (S.Vijayalakshmi, 2014) in our previous work. They are First Appeared Noun with Global Inverse ($DNAFA_{GI}$), First Appeared Noun with Global Log Inverse ($DNAFA_{GLI}$),, First Appeared Noun with Local Linear($DNAFA_{LL}$), and First Appeared Noun with Local V Linear($DNAFA_{LVL}$).

Inverse Document Frequency of Noun is calculated separately and we multiply with the value of first appeared Noun. Now we list first appeared noun with its weight. Quick sort is applied and top 'K' attributes are selected for clustering. The clustering results are published in our previous work. We investigate that by doing ensemble distributed noun attribute based on its first appearance to improve the clustering results.

IMPROVING CLUSTERING RESULTS BY ENSEMBLE DISTRIBUTED NOUN ATTRIBUTE SELECTION WITH FIRST APPEARANCE

Our proposed Ensemble Distributed Noun Attribute based on its First Appearance (EDNAFA), we have exposed the documentations on schematic examples (see Figure 2).

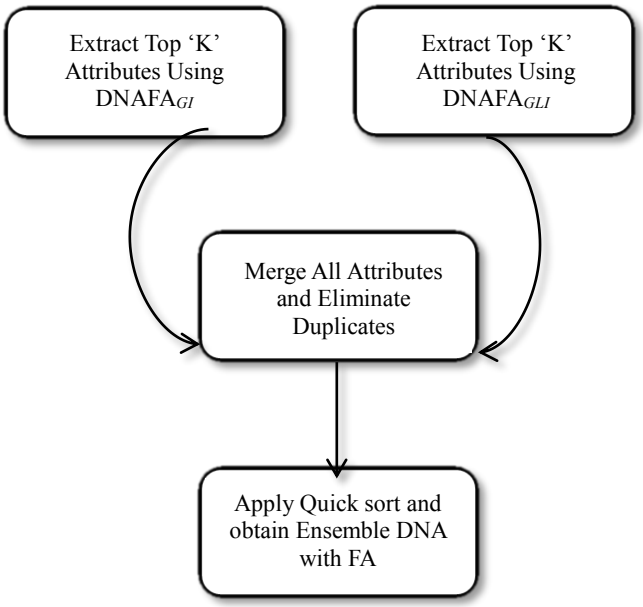

Figure 2. Schematic process of Ensemble DNA based on its First Appearance

In this section, we have implemented six different EDNAFA selection methods obtained from four different distributed nouns attribute based on its first appearance (DNAFA) by combining any two methods. They are

1. E(DNAFA$_{GI}$+ DNAFA$_{GLI}$)
2. E(DNAFA$_{GI}$+ DNAFA$_{LL}$)
3. E(DNAFA$_{GI}$+ DNAFA$_{LVL}$)
4. E(DNAFA$_{GLI}$+ DNAFA$_{LL}$)
5. E(DNAFA$_{GLI}$+ DNAFA$_{LVL}$)
6. E(DNAFA$_{LL}$+ DNAFA$_{LVL}$)

We are also called the above six different method as follows.

1. EDNAFA$_{(GI +GLI)}$
2. EDNAFA$_{(GI+ LL)}$
3. EDNAFA$_{(GI+ LVL)}$
4. EDNAFA$_{(GLI+ LL)}$
5. EDNAFA$_{(GLI+ LVL)}$
6. EDNAFA$_{(LL+ LVL)}$

Our previous work, we have applied distributed noun attribute based on its first appearance with Global Inverse (DNAFA$_{GI}$), distributed noun attribute based on its first appearance with Global Log Inverse (DNAFA$_{GLI}$), distributed noun attribute based on its first appearance with Local Linear (DNAFA$_{LL}$), distributed noun attribute based on its first appearance with Global Inverse (DNAFA$_{LVL}$).

$$\mathrm{DNAFA}_{GI} = \mathrm{f}(\mathrm{p}, len(d)) = \frac{1}{p+1} \tag{4}$$

$$\mathrm{DNAFA}_{\mathrm{GLI}} = \mathrm{f}(\mathrm{p}, len(d)) = \frac{1}{\log(p+2)} \tag{5}$$

$$\mathrm{DNAFA}_{\mathrm{LL}} = \mathrm{f}(\mathrm{p}, len(d)) = \frac{len(d) - P}{len(d)} \tag{6}$$

$$\mathrm{DNAFA}_{\mathrm{LVL}} = \mathrm{f}(\mathrm{p}, len(d)) = \frac{\left| p - \frac{[len(d) - 1]}{2} + 1 \right|}{len(d)} \tag{7}$$

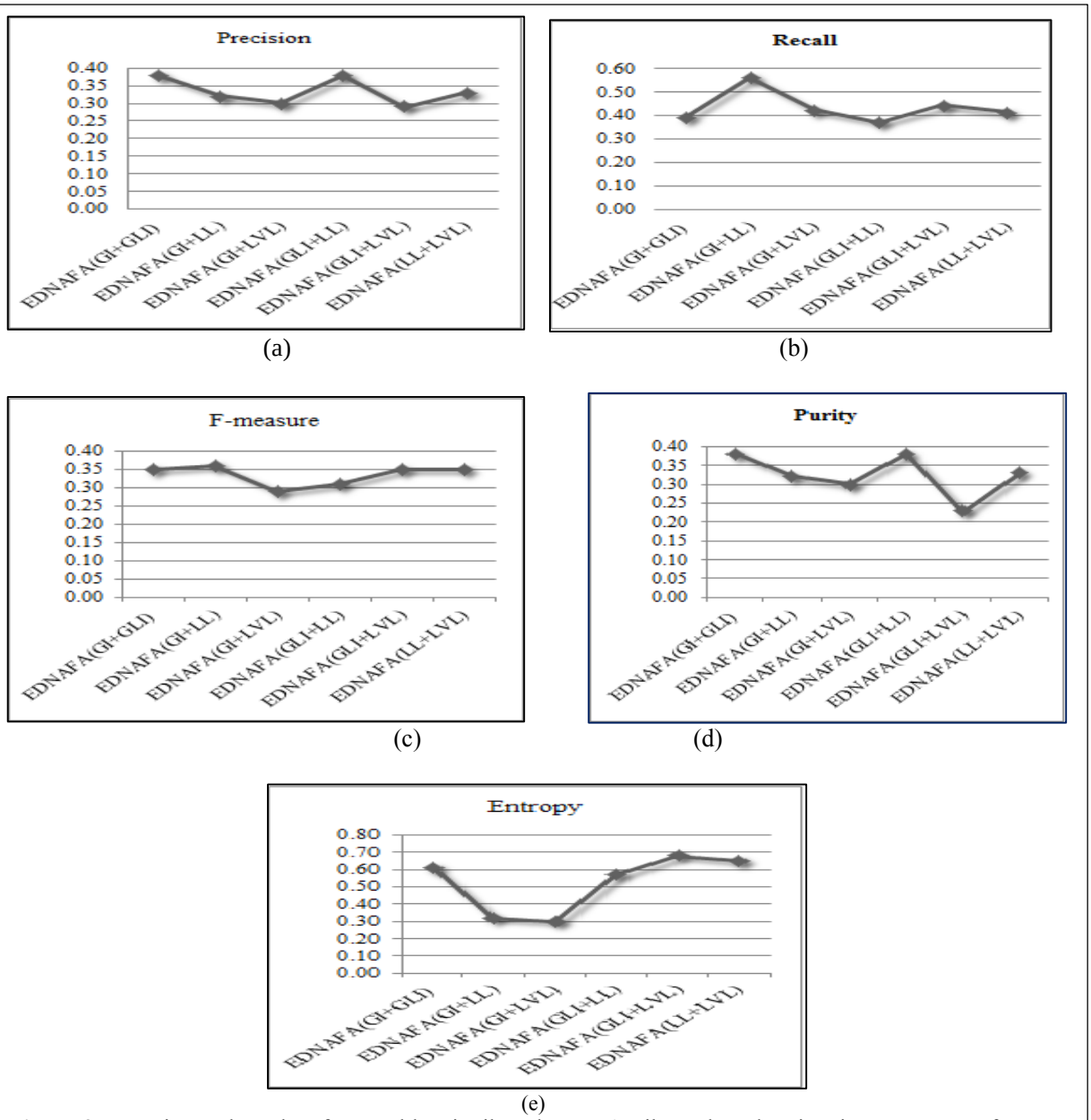

Figure 3. Experimental results of Ensemble Distributed Noun Attributes based on its First appearance for Text Document Clustering with flocking algorithm (20 Newsgroup dataset) (a). Precision (b). Recall (c). F-Measure (d). Purity (e). Entropy.

TEXT DOCUMENT CLUSTERING

For example, we consider two methods such as $DNAFA_{GI}$, and $DNAFA_{GLI}$. We collect the attributes with its weight, and also we check the attributes. If the attributes of $DNAFA_{GI}$ matches with the attribute of $DNAFA_{GLI,}$ We add the weight of the attribute weight. This attribute with updated weight is added to the new list, otherwise we directly add attributes with concern weight to new list. The same process applied for all attributes and merged together, and verified if any duplicate attribute exist. If exist, eliminate duplicates. Sort the nouns by applying quick sort and select Top 'K' attributes for clustering. The same process is applied the remaining methods also.

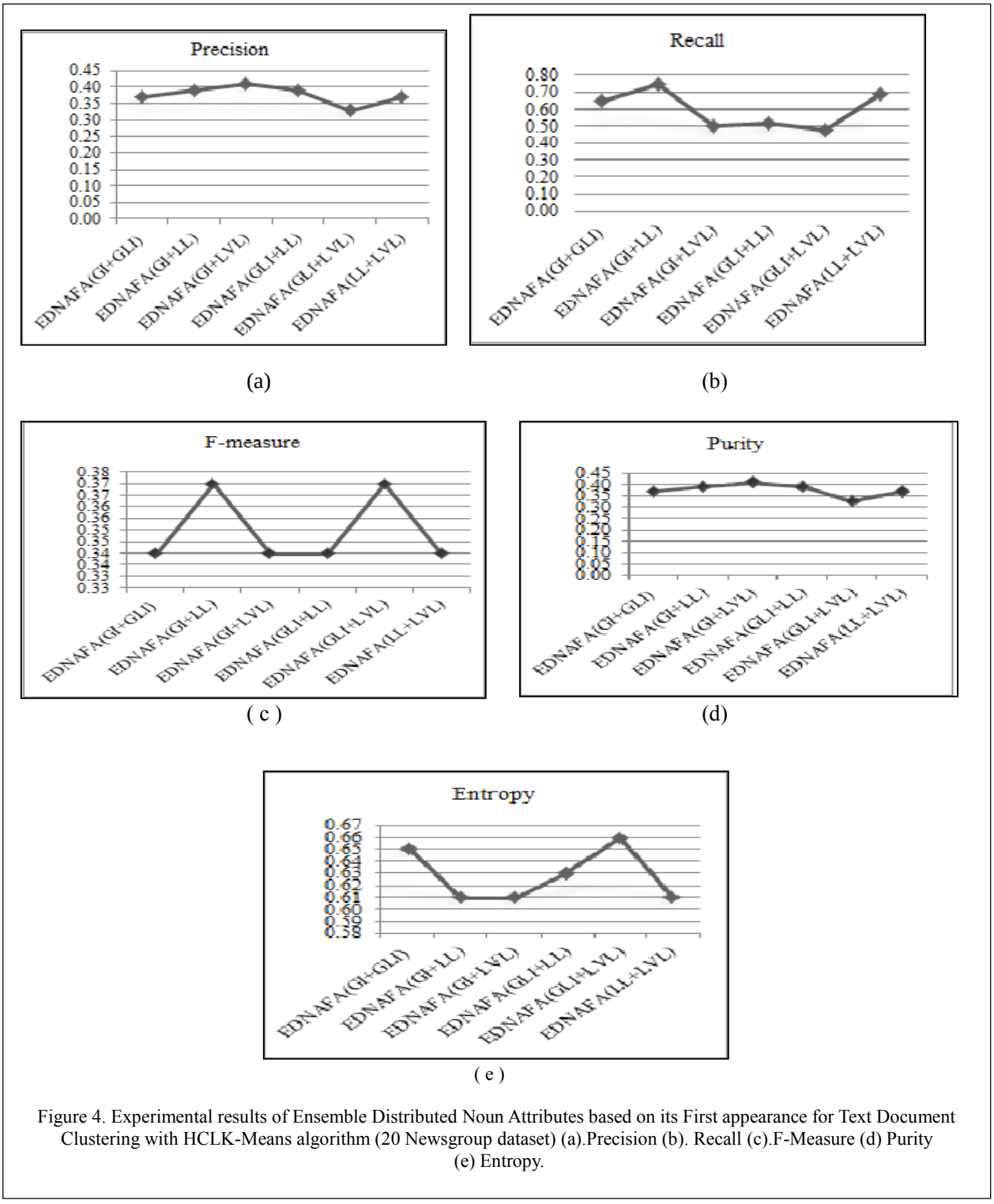

Figure 4. Experimental results of Ensemble Distributed Noun Attributes based on its First appearance for Text Document Clustering with HCLK-Means algorithm (20 Newsgroup dataset) (a).Precision (b). Recall (c).F-Measure (d) Purity (e) Entropy.

In this text document clustering algorithm, text documents are represented by vector space model. In this model, each document d is considered as a vector in the ensemble distributed noun - space and represented by the ensemble distributed noun frequency (NF) vector:

$$d_{EDNAFA} = [EDNAFA_1, nEDNAFA_2, \dots, EDNAFA_u] \quad (8)$$

Where $EDNAFA_i$ frequency of the ith ensemble distributed noun attribute with first appearance in the document and u is the total number of unique nouns. There are several pre-processing steps, including stop word removal, finding the pre-processed word is noun with the help of RiTa WordNet and then stemming on the documents. The standard model is used to calculate the weight by using Inverse document Frequency (IDF) in the test document set.

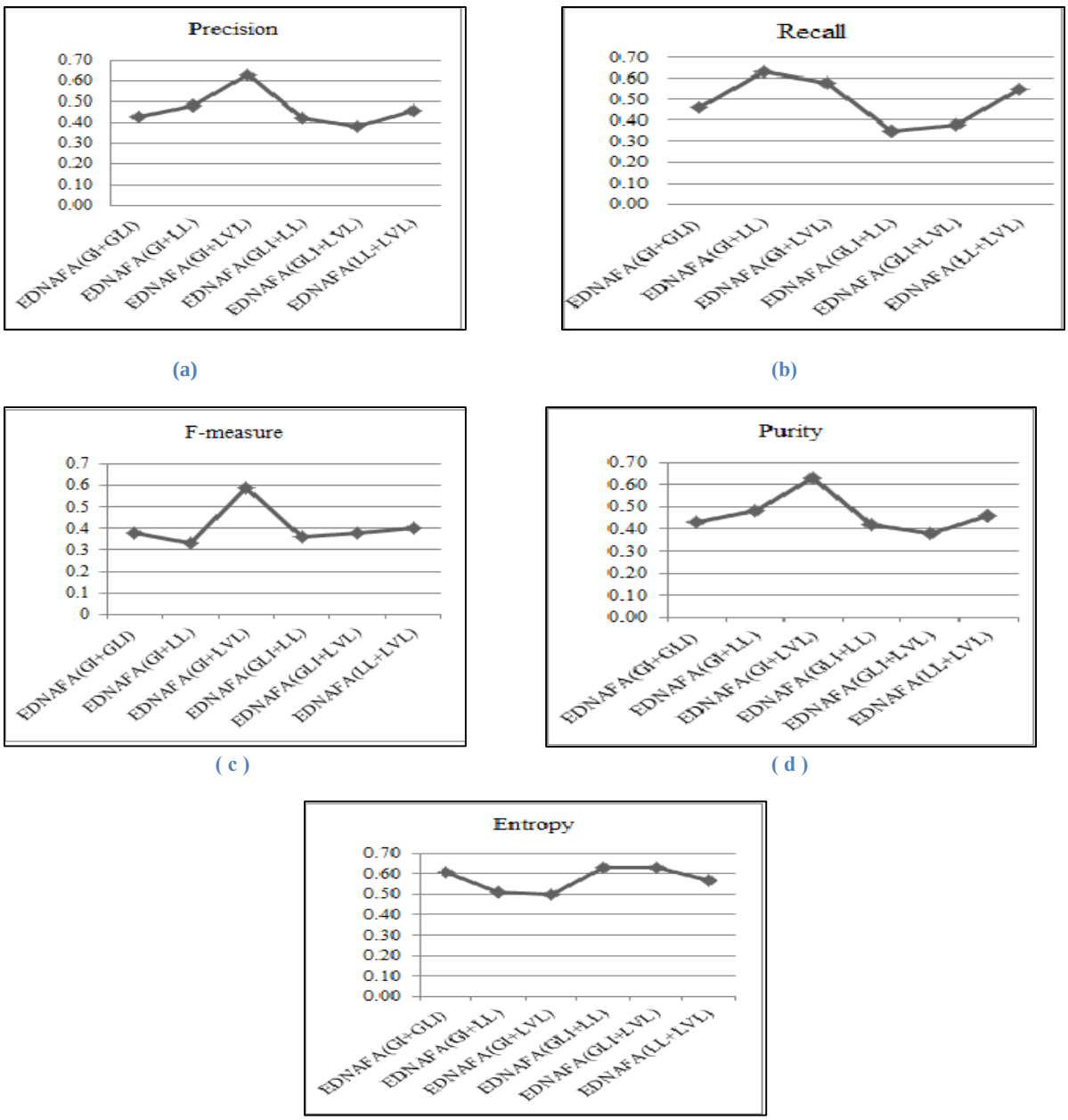

Figure 5. Experimental results of Ensemble Distributed Noun Attributes based on its First appearance for Text Document Clustering with flocking algorithm (SCJC dataset) (a).Precision (b). Recall (c).F-Measure (d) Purity (e) Entropy.

HCLK-Means clustering [13] requires learning rate as parameter. The learning rate is decreased during the process time to get improved optimal solution. When the cosine function is is employed to assign to the cluster with the foremost similar centroid and the international criterion function is maximized as a result. All two algorithm required prior knowledge about how many clusters are expected in the dataset.

The flocking algorithm [14], HCLK-Means algorithm are applied to the real document collection respectively. The cosine distance measure is used as similarity metric in each algorithm. The cosine value is one when the two documents are identical and zero if there is nothing common. The larger cosine worth indicates that these 2 documents shares a lot of nouns and area unit a lot of similar. every document is described as boid in Flocking rule. Each boid can only sense the flock mates located with sense range. Each boid will

solely sense the flock mates settled with sense vary. In this implementation, we have used the boids ranges from 100 to 2000.

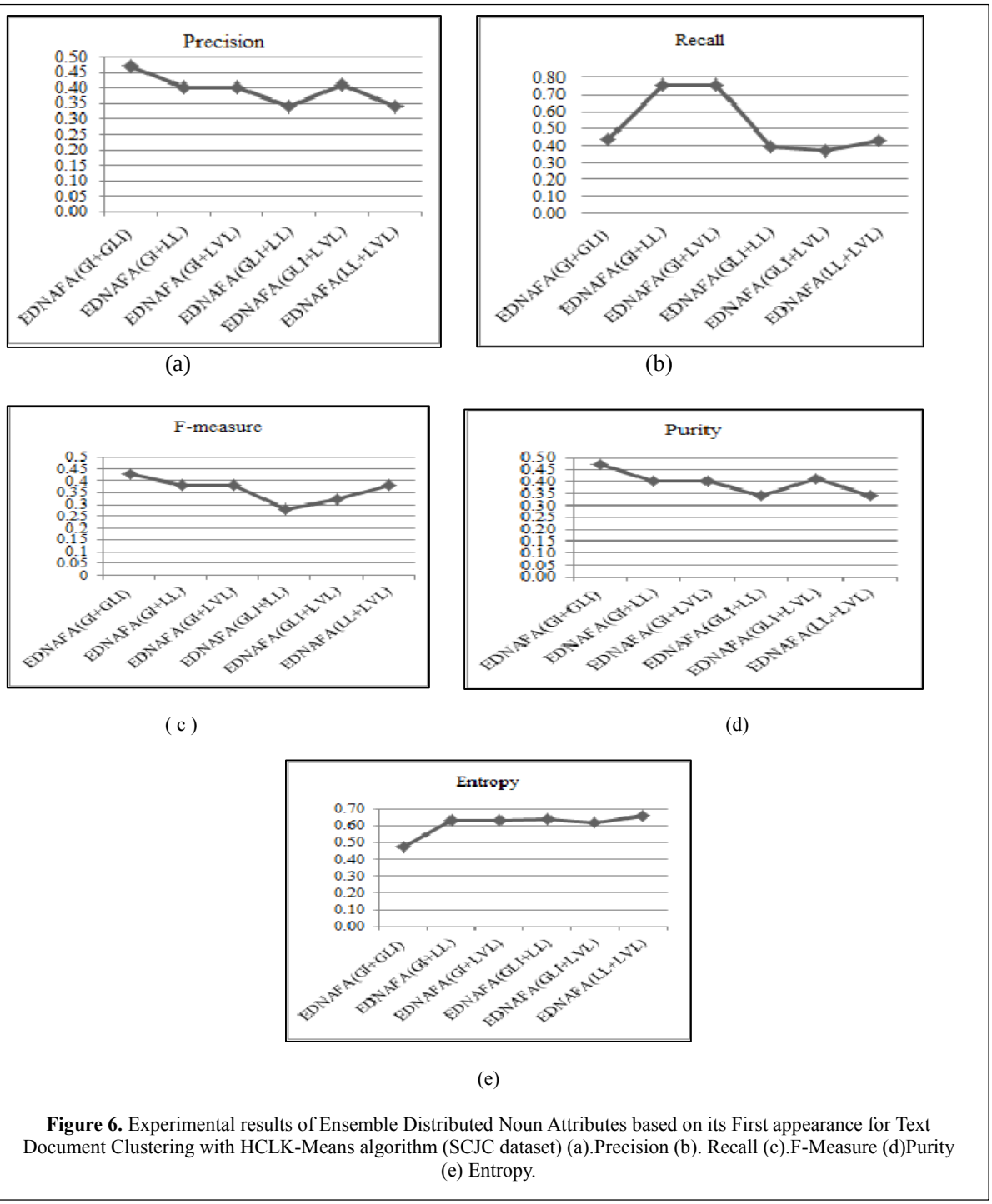

Figure 6. Experimental results of Ensemble Distributed Noun Attributes based on its First appearance for Text Document Clustering with HCLK-Means algorithm (SCJC dataset) (a).Precision (b). Recall (c).F-Measure (d)Purity (e) Entropy.

EXPERIMENTAL RESULTS AND DISCUSSION

In this section, we present some numerical results and one real dataset called Specific Crime Judgement Corpus (SCJC) and also used standard dataset 20 Newsgroups to show the efficiency in clustering technique when using ensemble distributed attribute selection based on its first appearance. Each corpus is separated into two, the first half (50%) is considered as training corpus and used for attribute selection, and the remaining half for testing corpus used to evaluate selected attributes(EDNAFA) and the quality of the cluster. Attribute selection technique can give more valid results.

In general, all methods of EDNAFA techniques provides healthier attributes than our previous work, which was implemented using DNAFA selection methods.

Figure 3,4,5,6 shows that the evaluation of the clustering results using our proposed ensemble attributes. The cluster is assessed based on Precision, Recall, F-Measure, Purity and Entropy. From the above, the first four attributes is considered as efficient, if it gained the result with maximum value that must be closer to 1(ranges from 0.0-1.0). Entropy Measure is considered as efficient, it has to be minimum value that must be nearby to 0 (ranges from 0.0-1.0).

The above measures (Yanjun Li, 2008)) are evaluated using the following formula: F-Measure is obtained by using precision and recall value. Thus we can calculate precision $P(i,j)$ and recall $R(i,j)$ of each cluster j for each class i. If n_i is the number of the member of the class i, n_j is the number of the member of the cluster j and n_{ij} is the number of the member of the class i and the cluster j.

$$P(i,j) = \frac{n_{ij}}{n_j}, \qquad n_j \text{ is the total number of documents obtained} \quad (8)$$

$$R(i,j) = \frac{n_{ij}}{n_i} \quad n_i \text{is the total number of documents obtained} \quad (9)$$

$$F(i,j) = \frac{2 * P(i,j) * R(i,j)}{P(i,j) * R(i,j)} \quad (10)$$

$$\text{Purity}(j) = \frac{1}{n_j} \max_i \left(n_{ij}\right) \quad (11)$$

Entropy of cluster j is calculated as:

$$E_j = -\sum_j \text{Precision}_{ij} * \log\left(\text{Precision}_{ij}\right) \quad (12)$$

Figure 3 illustrates that EDNAFA(GI+LVL) method of attribute selection gives better clustering results using clustering with flocking algorithm based on its Precision, Recall, F-Measure, Purity and Entropy on 20 Newsgroups testing corpus. For HCLK-Means clustering generates efficient cluster by using the attribute selection technique EDNAFA (GI+LVL) on 20 Newsgroups testing corpus in Figure 4.

On SCJC testing corpus, EDNAFA(GI+GLI) attribute selection method obtained better clustering results with both HCLK Mean Clustering and clustering with flocking algorithm. The combination of Global Inverse attributes gave efficient results. Clustering results may differ based on its data set.

At the outset, the clustering results based on its precision, recall, F measure, purity, and entropy shows that HCLK-Means clustering works efficiently than clustering with flocking algorithm on 20 Newsgroups.

CONCLUSIONS AND FUTURE WORK

In this paper, we introduced the use of ensemble methods for attribute selection. We showed that by constructing ensemble attribute selection techniques, robustness of attribute ranking and attribute subset selection could be improved, using similar techniques as in ensemble methods for supervised learning. Ensemble methods show great promise for high dimensional attribute space. It tries out that the best trade-off, clustering performance depends on the dataset. Attribute selection techniques will gain the importance in the future.

REFERENCES

[1]. Saeys.Y, Inza. I., Larranaga.P. "A review of feature selection techniques in bioinformatics", Bioinformatics 23(19), pp2507–2517 (2007).

[2] Shining Li, Zhenhai Zhang, and Jiaqi Duan,"An Ensemble Multi-Label Feature Selection Algorithm Based on Information Entropy", Vol. 11, No. 4,pp 379-385, The International Arab Journal of Information Technology, July 2014.

[3] Grigorios T., Ioannis K., and Ioannis V., "Mining Multi-Label Data," in Proceedings of Data Mining and Knowledge Discovery Handbook, Springer, USA, pp. 667-685, 2010.

[4] Isabelle G., and Andre E., "An Introduction to Variable and Feature Selection," Journal of Machine Learning Research, vol. 3, no. 1, pp.1157-1182, 2003.

[5] Li G., You M., and Ge L., "Feature Selection for Semi-Supervised Multi-Label Learning with Application to Gene Function Analysis," in Proceedings of the 1st ACM International Conference on Bioinformatics and Computational Biology, USA, pp. 354-357, 2010.

[6] Zhang Y., You L., and Chen J., "Feature Selection for Multi-Label Data by Using Simulated Annealing," Computer Engineering and Design, vol. 32, no. 7, pp. 2494-2500, 2011.

[7] Guyon, I., Elisseeff, A.: An Introduction to Variable and Feature Selection. Journal of Machine Learning Research 3, 1157–1182 (2003)

[8] Dunne, K., Cunningham, P., Azuaje, F.: Solutions to instability problems with sequential wrapper-based approaches to feature selection. Technical report TCD-2002-28. Dept. of Computer Science, Trinity College, Dublin, Ireland (2002)

[9] Kalousis, A., Prados, J., Hilario, M.: Stability of feature selection algorithms: a study on high-dimensional spaces, Knowledge and information systems,12(1), 95–116 (2007).

[10] Kuncheva.L. A stability index for feature selection. In: Proceedings of the 25th International Multi-Conference on Artificial Intelligence and Applications, pp. 390–395 (2007).

[11] S.Vijayalakshmi and Dr.D.Manimegalai, "Distributed Noun Attribute based on its First Appearance for Text

Document Clustering" conducted by Park College of Engineering and Technology(18 & 19 Dec, 2014), Page No. 780-784, 2014 IEEE International Conference on Computational Intelligence and Computing Research(2014 IEEE ICCIC), ISBN: 978-1-4799-3974-9.

[12] Yanjun Li, Congnan Luo and Soon M.Chung, Member IEEE," Text Clustering with Feature Selection by using Statistical Data", IEEE transaction of knowledge and Data Engineering (2008).

[13] S.Vijayalakshmi, Dr.D.Manimegalai,"Integrating Ontology to Enhance HCL-Based Text Document Clustering", Research Journal of Applied Sciences, 8(7):358-368, 2013 ISSN: 1815-932X, DOI:10.3923/rjasci.2013.358.368.

[14] S.Vijayalakshmi, Dr.D.Manimegalai, " Text Document Clustering with Flocking Algorithm using Specific Crimes Judgment Corpus", Asian Journal of Information Technology, 13(1):21-28, 2014, ISSN: 1682-3915, DOI: 10.3923/ajit.2014.21.28.