Evaluation of Clustering around Weighted Prototype and Genetic Algorithm for Document Categorization

Garima Jain
Samrat Ashok
Technological Institute
Department of Information
Technology, Vidisha, India

Shailendra Kumar Shrivastava
Samrat Ashok
Technological Institute
Department of Information
Technology, Vidisha, India

ABSTRACT
Document clustering is very important in the field of text categorization. Genetic algorithm, which is an optimization based technique which can be applied for finding out the best cluster centres easily by computing fitness values of data points. While clustering around weighted prototype technique is especially helpful when proper pairwise similarities are available. This technique does not find global solution of the objective function. Experimental result shows that F-measure and Normalized mutual information of genetic algorithm is better than clustering around weighted prototype for 20 Newsgroup dataset. F-measure and accuracy of genetic algorithm is better than clustering around weighted prototype for the Reuter-21578 dataset.

Keywords
Clustering, Similarity Based, Genetic Algorithm, Document Categorization, Text mining.

1. INTRODUCTION
Text categorization is the task of associating documents with predefined categories that are associated with their content. Text Categorization is a crucial and active analysis field for the reason that the large number of documents available and the resulting require organizing them. Text Categorization issue has been drawn with the pattern classification procedure, where documents are represented as numerical vectors and normal classifiers (e.g., naive Bayes and support vector machines) are applied [1]. This kind of representation is well known as the vector space model [2]. Underneath the Vector Space Model single assumes a document may be a point in an N-dimensional space and documents that are nearer in that space are related to each other [3]. Among the various instances of Vector Space Model, possibly the majority used form is bag-of-words representation. Within the bag of word it is assumed that the content of a document may be determined by the set of terms it contains. Documents can be represented as points within the vocabulary area, i.e., a document is represented by a numerical vector of length equal to the number of different terms within the vocabulary (the set of all different terms within the document collection). The basics of vector specify how essential the corresponding terms are describing for the semantics or the content of the document. Bag of word is mainly used for document representation in each text categorization and information retrieval. An important part of the text categorization systems using the bag of word representation therefore known as term-weighting scheme that is responsible for deciding while relevant term is for describing the content of a document [28, 29, 30, 31]. Term- weighting schemes be term frequency, where the value of a word in a document is given by its frequency of occurrence in the document.Even though, as a result of capturing statistical data from the original document provides simplicity of vector space model. It is not easy to perform clustering straightforwardly in the space, which is extremely sparse and high dimensional. Dimension reduction techniques are capable of reducing the dimensionality so to facilitate the data can be handled by existing approaches more simply [4]. To solve the problems of text mining there are various techniques presented for retrieval of relevant information. In text mining included information extraction, text categorization, text document analyzed on the basis of term [5], phrase [6], concept [7] and pattern.

During the earlier period, various clustering methods have been applied for document categorization. A clustering technique generally classified as hierarchical and partitional clustering based on the properties of generating clusters. Partitional clustering directly divides information into some predefined range of clusters without the hierarchical structure, while hierarchical clustering groups information with an order of nested partitions [11]. A wide collection of document clustering methods such as k-means [8], Clustering around weighted prototype [9] have been used vector space model to represent documents, where all document takes as “a bag of terms”. Numerous types of measures have been planned for calculating the relationship between two vectors. The well-known similarity measure is Euclidean distance [10] taken from the Euclidean geometry field. K-mean is the popular clustering method suitable to its good balance between simplicity and effectiveness, but disadvantage is final result is based on the initial selection of cluster centers. Its objective function has local minimum. Similarity-based clustering approach referred to as Clustering around weighted prototype for document analysis and categorization. In this approach, more than one object is assigned to every cluster with numerous weights. The larger weight of an object is the more representative within the corresponding cluster. It is often observed that several objects together characterize a cluster with different weights [9]. Generally, the more central an object is placed in a cluster; the larger weight is assigned to that object. This version enhances the ability to capture a lot of information of the cluster structure. This technique is especially useful when proper pairwise similarities are available. This algorithm is not finding global solutions of the objective function. In this research present is a new dynamic based on genetic method. Genetic technique is a global algorithm, which can locate the best centers easily.

Genetic Algorithms are applied to a variety of fields to search out approximate best solution in an efficient way. Genetic Algorithm usually first transforms initial solutions of the original problem into sequences of genes in chromosomes. Given a population of chromosomes, genetic operations are applied iteratively to produce a new generation of population.

For every generation, each chromosome is evaluated by a given fitness function. The fitness value is to calculate approximately whether a chromosome can produce major performance. [12] Each chromosome is a sequence of integers representing the category labels. In the initial population, take the total set of input clusterings that an ensemble is to be generated. The selection procedure select chromosomes for the later breeding directed by the survival of the "fittest" conception of natural genetic systems Crossover is a probabilistic procedure that exchanges information between two parent chromosomes for generating two child chromosomes [13].

2. RELATED WORK

Clustering based on similarity has been examined for a long moment. Hierarchical clustering collected the data with a sequence of nested partitions. It creates a dendrogram comprising of sequence of clusters in an agglomerative or divisive way [11]. In hierarchical clustering, complete linkage, single linkage and group average linkage are the three exemplary approaches to compute the nearness of two clusters based on the similarities between objects within the two clusters. In these standard linkage based hierarchical clustering methodologies, every group is represented by every object in that group. Experimental learning on document clustering by demonstrate that partitioning approaches are better to hierarchical ones for lower computational cost and better class of clusters [14]. Another usual proximity-based approach is k-medoids approach. It produces k partitions of the dataset, where every cluster is represented by objects belonging to that cluster. A well famous k-medoids algorithm is the Partitioning around Medoid technique [15]. As a consequence of the variety of learning structure in world, the conventional ''one medoid for one cluster'' approach used is k-medoids method might not be sufficient. To be capable to capture the cluster structure enhanced, specialist creates clustering approaches based on multiple representative objects. In [16] hierarchical clustering approach Clustering Using Representative, where a particular number of representative objects are choose to be well separated in order to capture wealthy cluster shapes. Those representative objects are uniformly weighted and aren't necessary to be real objects. An improved version of Clustering Using Representative is proposed in [17] to pick different numbers of representatives for a variety of clusters based on cluster density. A multi-representative approach based on density is planned in [18]. In the fuzzy clustering approach called fuzzy clustering with weighted prototype [19], every cluster is characterized by a variety of weighted medoids. Unlike from other multi representatives based methods [16] [18]. Where representatives of every cluster are choose to be a pre-specified variety with equal weights, in Partitional Fuzzy Clustering, the weights as well as the variety of representative objects in every cluster are determined based on the nature of the dataset. Efficiency and scalability are two vital factors to applications with huge scaled data. Genetic Algorithm is more capable and scalable compared to clustering around weighted prototype.

3. TEXT DOCUMENTS CLUSTERING PREPROCESSING

In this research, 20 newsgroups and reuter-21578 datasets are used for implementation. Select some documents and apply preprocessing.

3.1 Preprocessing

3.1.1 Stopword Removal Phase

Stopwords (a, about, an, are, as at be, from, when, what…….) occur frequently, but do not represent any content of documents. Articles, Prepositions, conjunctions and some pronouns are stopwords. Such words should be removed before documents are stored and indexed. Stopwords in the query are also removed before retrieval is performed.

3.1.2 Stemming Phase

In many languages, a word has various syntactical forms depending on a context that it is used. For example in English, noun has plural forms, verbs have gerund form and verbs used in past tense are different from the present tense. Stemming is a procedure that produces stems or roots. A stem is a left part after removing its prefixes and suffixes. For example "computer", "computing" and "compute" are reduced to "comput".

3.1.3 Other pre-processing tasks for Text

Digits: Numbers and terms that include digits are removed for example-dates, times.

Hyphens: Breaking hyphens are generally applied to deal with variation of usage.

Punctuation Marks: Punctuation is able to be dealt with similarly like hyphens.

Case of Letters: All the letters are typically changed to either the upper case otherwise lower case.

3.2 Feature Selection Techniques

Feature selection techniques term variance, document frequency and term frequency-inverse document frequency scheme are used to choose most important feature.

3.2.1 Document Frequency

Document frequency [26] assumes that frequent terms are extra informative than non frequent terms. Document frequency is the amount of documents during which f_j appears. Document frequency is sometimes used as a criterion to calculate efficiency. Document frequency of f_j is specified by the following formula:

$$\mathrm{DF}(f_j) = \mathrm{n}(f_{ij} > 0) \quad (1)$$

(i = 1, 2,.....,n; j = 1,2,.....,m)

3.2.2 Variance Score

Variance score [26] ranks feature by calculating variance of every feature f_j in document term matrix. Variance is specified by the following formula:

$$\mathrm{Var}\,(f_j) = \frac{1}{n}\sum_{i=1}^{n}(f_{ij} - \bar{f}_j)^2 \quad (2)$$

Where, $\bar{f}_j = \frac{1}{n}\left(\sum_i f_{ij}\right)$

A discriminative feature gets high variance score. Variance score are easy feature selection process used for selecting the feature [27].

3.2.3 Document Vectors

A document within the vector space model is represented as a weight vector, within which every component weight is computed based on some difference of word frequency otherwise term frequency-inverse document frequency scheme.

Term Frequency Scheme: during this methodology, the weight of term in document is the count of appearance in document.

TF-IDF Scheme-This is the generally well-known weight scheme wherever TF is term frequency and IDF is inverse document frequency [33].

Let *N* is the total amount of documents within the system or the collection and df_i be the number of documents within which term t_i found at least once. Let f_{ij} is raw frequency count of word t_i in document d_j. And $|v|$ be vocabulary size of set then the normalized term frequency denoted by tf_{ij}-

$$tf_{ij} = \frac{f_{ij}}{\max\{f_{1j}, f_{2j} \ldots\ldots\ldots\ldots f_{|v|j}\}} \quad (3)$$

Inverse document Frequency of term t_i is given by

$$idf_i = \log \frac{N}{df_i} \quad (4)$$

The Final TF-IDF term weight is given by-

$$w_{ij} = tf_{ij} \times idf_j \quad (5)$$

3.3 Similarity Measure

A lot of measures have been defined for computing the proximity between two vectors. Cosine similarity [20] is a measure that takes cosine angle between two vectors. The cosine similarity is unable to provide information on the magnitude of differences. The Jaccard coefficient [21] is a statistic used for comparing the relationship of two sample sets, and is defined as the range of the intersection separated by the range of the combination of the sample sets. Euclidean distance [10] is a recognized similarity metric taken from Euclidean geometry field. Euclidean distance is better than other measures because here continuous space are given where the entire dimensions are correctly scaled and related, then Euclidean distance is used. The distance measures can be generalized to n dimensions from the common 2 dimensional case. The formula for Euclidean distance in dimensions is given in Equation-

$$d(\bar{p}, \bar{q}) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \quad (6)$$

Where

$\bar{p} = (p_1, p_2, \ldots \ldots p_n)$

$\bar{q} = (q_1, q_2, \ldots \ldots q_n)$

4. CLUSTERING AROUND WEIGHTED PROTOTYPE

In this section present the details of the proximity based clustering approach clustering around weighted prototype.

In the dataset n objects are specified and the similarity matrix, in this method clusters n objects into non-overlapping clusters. As weight could be a continuous variable, the strategy of Lagrange multiplier is used to derive the solution of weight by locally maximizing objective function. Negative values of weight are also appeared throughout the iteration. To ensure that each one weight values be non-negative, require using the Karush-Kuhn-Tucker conditions in [19] to add the non-negative constrains into the Lagrangian. This might be time consuming for big datasets. To speed up the method, here this method used the subsequent heuristic as an estimated solution. For every cluster during this algorithm check weight, and set negative elements of weight to zero. Once the update weights have been derived, and then require deciding the way of cluster assignment. A typical method is to allocate every objects to the cluster that is the nearest. The alternating optimization is used to obtain local solutions of objective-based cluster. Following this procedure, in turn update the representative weights and cluster assignment based on one another. The representative objects are adjusted during the change of weights based on this partition, and therefore the updated representative objects of every cluster successively are used to generate new clusters. As been mentioned earlier, whenT_0, clustering around weighted prototype decreases to 1 object representation for every cluster; whereas T_1makes each objects equally represent each cluster.

In [9] present an increased version of the essential alternating optimization with ‘‘annealing’’. The deterministic annealing method has been planned to alleviate the local optimum drawback of cluster is the non-convex objective functions. The annealing method starts with a high-temperature that generates extremely fuzzy clusters means objects have close memberships altogether the clusters, and then the temperature is more and more decreased to provide less fuzzy clusters. The approach of incorporating ‘‘randomness’’ into the objective function produce additional opportunities for escaping from local solutions. Impressed by the deterministic annealing technique, that include an ‘‘annealing like’’ method into the Clustering around weighted prototype algorithm. With this procedure, clustering around weighted prototype is predicted to avoid additional local maximums and therefore tends to achieve better cluster results.

As shown in clustering around weighted prototype, rather than fixing the parameter T to a fixed value, Clustering around weighted prototype algorithm starts with a large T_0 and step by step decreases it once the method continues. This is analogy to the annealing procedure, where parameter T will be treated as the temperature. Once the temperature is high, i.e. T is large, the distribution of weight is close to random and therefore representative objects are searched in large space; once T is small, the distribution of weight becomes stable and therefore search the new representative objects only from the neighborhood of this one. Such a cooling method allows clustering around weighted prototype to avoid a number of the local optimums and is additional likely to produce better result [9].

Algorithm-

Input: Similarity matrix$S_{n\times n}$, the number of cluster k, parameterT_0, T_f,

M = Maximum iteration

Output: Weight matrix $W_{k\times n}$.

Method:

1. $T = T_0$, t = 0 generate an initial partition;
2. Repeat
3. $\{A_c^{t+1}\}_{c=1}^{k}$
4. $T = T_0 \times (T_f/T_0)^{t/M}$;
5. t=t+1.
6. Until t > M or T< T_f

5. GENETIC ALGORITHM

In this section present the details of genetic method. Genetic technique [34] is search techniques that related to the principle of natural selection. Clustering is a well-known unsupervised pattern categorization method that divides input space into K regions based on similarity/dissimilarity metric. The quantity of partitions/clusters might or might not be identified a priori. The parameters in the search space are represented in the form of chromosomes. A set of such chromosomes is known as population. An objective and fitness function related to each chromosome that represents the measure of fitness. Biologically inspired operators selection, crossover and mutation are applied to give up new child chromosomes. These operators continue a number of generations until the stop condition are fulfilled. The fittest chromosome seen up to the last generation gives the best answer to the clustering problem. In Genetic Algorithm each chromosome Chro_i in the population is initially encoded by a number of K centers, where K lies in the range $[K_{min}, K_{max}]$.

$$\text{Chro}_i = \{Center_{i,1}, Center_{i,2}, \dots\dots\dots, Center_{i,k}\}$$

For initializing $Center_i$, a row of elements are chosen randomly from the corpus matrix C in

$$Center_i = \{c_{i,1,}, c_{i,2}, c_{i,3}, \dots\dots\dots, c_{i,n}\}$$

n = number of total texts and the dimensions can be reduced from n to k .k < n.

$$Center_i = \{c_{i,1,}, c_{i,2}, c_{i,3}, \dots\dots\dots, c_{i,k}\}$$

Genetic Algorithm-

1. First initialize the population do

a) Randomly select k documents from n documents and consider them k centers, where k assumed to lie in the range$[k_{min}, k_{max}]$. These k centers consider as one chromosome of initial population.

$$Chro_i = \{center_{i,1}, center_{i,2}, \dots\dots center_{i,k}\}$$

b) for (n-k) documents do

i. find distances from m centers by n×n matrix

ii. cluster document with any center according to minimum distance

c) Find this chromosome fitness value by fitness formula.

2. Consider initial population as the old population

3. Find best chromosome from the final population.

This chromosome has best clusters of the documents

GA-Function-

1. Make new population by Crossover

a) Select two chromosomes from old population randomly.

b) If crossover probability genetic algorithm fulfilled then

i. Apply crossover between chromosomes and find out two offspring.

ii. Calculate two offspring fitness values by fitness formula.

iii. Choose best two among old chromosomes and two offspring.

iv. Add them into new population.

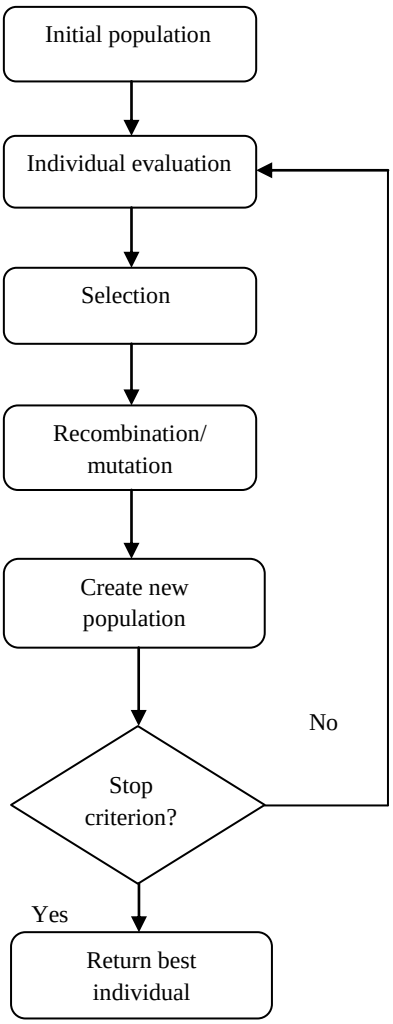

Fig 1: a generic diagram of Genetic Algorithm

6. EXPERIMENTAL RESULTS

In this section, simulate the task of document categorization to calculate the F-measure, Accuracy and Normalized Mutual Information. To comparison, run the genetic algorithm and similarity based technique clustering around weighted prototype. The simple vector model is used to represent the document. The similarity matrix is then generated as input data.

6.1 Datasets and Preprocessing

The datasets are extracted from several benchmarks. Multi5 is a subset extracted from the 20Newsgroups [22]. Multi5 contains around 100 documents from each of five categories comp.graphics, rec.motorcycles, rec.sports.baseball, sci.space, and talk.politi-cs.mideast. Reuters-21578 dataset contains the documents, which is present in Reuter's newswire [23]. Calculate each words weight with term frequency –inverse document frequency weight, and Euclidean distance similarity is used to calculate the similarity between documents.

6.2 Algorithms and Evaluation

For evaluation the following algorithms are run:

CAWP [9]: Clustering around weighted prototype algorithm

GA: Genetic Algorithm for text clustering

F-measure [24] as well as Normalized Mutual Information is used [25] to calculate the clusters quality produced by Clustering around weighted prototype and Genetic algorithm. Both F-measure as well as normalized mutual information compares the clusters created by clustering algorithm and taking values in range of [0, 1].

$$F - measure_M = \frac{(\beta^2+1) precision_M recall_M}{\beta precision_M recall_M} \quad (7)$$

Where, $precision_M = \frac{\sum_{i=1}^{l} \frac{tp_i}{tp_i + fp_i}}{l}$

$$recall_M = \frac{\sum_{i=1}^{l} \frac{tp_i}{tp_i + fn_i}}{l}$$

$$Average\ Accuracy = \frac{\sum_{i=1}^{l} \frac{tp_i + tn_i}{tp_i + fn_i + fp_i + tn_i}}{l} \quad (8)$$

Where, l = number of clusters, *tp* = true positive, *fp* = false positive, *fn* = false negative, *tn* = true negative

NORMALIZED MUTUAL INFORMATION

Another metric Normalized Mutual Information is used from the information theory which is defined as Let A and B is the random variables. Let I (A, B) indicate the mutual information between A and B, and H (A) stand for the entropy of A. One can illustrate that I (A, B) is a metric. There is no upper bound for I (A, B) as a result for easier explanation and comparisons, a normalized version of I (A, B) that lie in the range from 0 to 1. A number of normalizations are achievable based on the test that I (A, B) ≤ min (H (A), H (B)). These consist of normalizing using arithmetic otherwise geometric mean of H (A) and H (B). Since H (A) =I (A, A), choose the geometric mean for the reason that the analogy with normalized inner product within Hilbert space. Therefore normalized mutual information-

$$\text{NMI} = \frac{I(A,B)}{\sqrt{H(A)H(B)}} \quad (9)$$

One can see that NMI (A, A) = 1, as required. Equation second requires sampled quantities provided through the clustering. Let n_g is the number of objects within cluster g, n_h is the number of objects within cluster h.

$$\text{NMI} = \frac{\sum_{g=1}^{k} \sum_{h=1}^{m} n_{g,h} \log \frac{n.n_{g,h}}{n_g n_h}}{\sqrt{\left(\sum_{g=1}^{k} n_g \log \frac{n_g}{n}\right)\left(\sum_{n=1}^{m} n_h \log \frac{n_h}{n}\right)}} \quad (10)$$

RESULTS EVALUATION

For result evaluation perform statistical test and observe that genetic approach are significantly better than Clustering around weighted prototype. F-measure, normalized mutual information and accuracy to conclude difference. Results of F-measure, normalized mutual information and Accuracy are shown in Table-1and Table-2 ten number of features are taken.

Table 1. Results for 20Newsgroups dataset

DATASET SIZE	CAWP F-MEASURE	GA F-MEASURE	CAWP NMI	GA NMI
100	0.8204	0.8592	0.9037	0.978
200	0.8295	0.8589	0.7076	0.9069
300	0.6429	0.7076	0.9277	0.9621
400	0.8735	0.9037	0.8641	0.9211
500	0.5926	0.6957	0.6957	0.7059

Table 2. Results for reuter-21578 dataset

DATASET SIZE	CAWP ACCURACY	GA ACCURACY	CAWP F-MEASURE	GA F-MEASURE
1	0.894133	0.943686	0.888445	0.932012
2	0.959166	0.97799	0.966707	0.981362
3	0.86234	0.900305	0.838444	0.863966

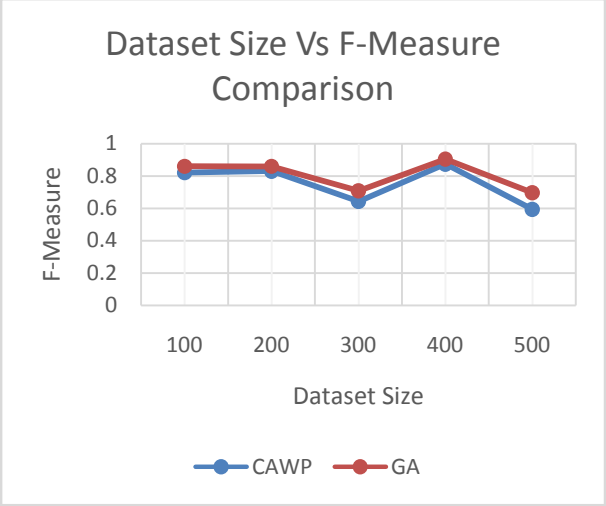

Fig 1: Dataset size and F-measure of CAWP and GA for 20 Newsgroups dataset

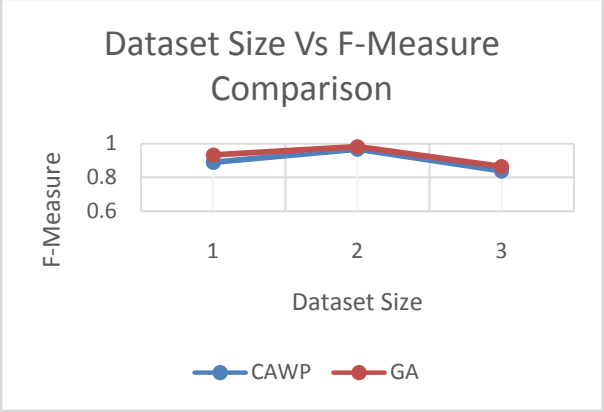

Fig 3: Dataset size and F-measure of CAWP and GA for Reuter-21578 dataset

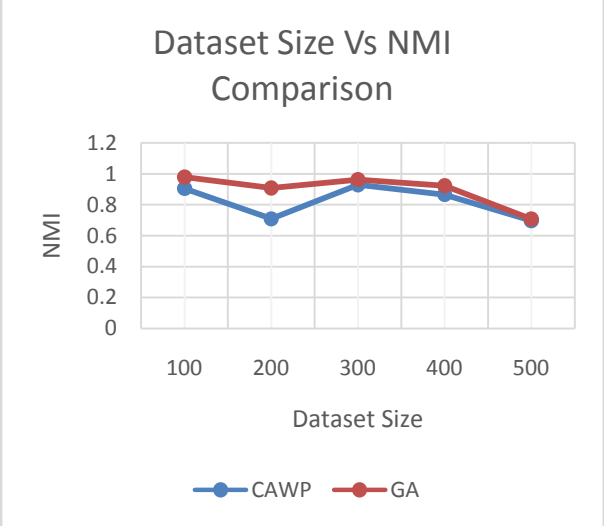

Fig 2: Dataset size and NMI of CAWP and GA for 20 Newsgroups dataset

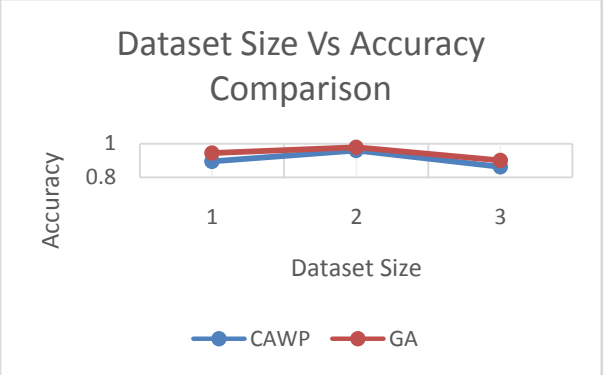

Fig 4: Dataset size and Accuracy of CAWP and GA for Reuter-21578 dataset

7. CONCLUSION

In this paper, evaluate Genetic algorithm and clustering around Weighted Prototype for document categorization. Genetic Algorithm calculates the similarities between documents and gives better clusters in comparatively less iteration than Clustering around Weighted Prototype. The techniques are tested on two datasets 20 newsgroups and Reuter-21578. It found that genetic algorithm provides efficient results as compared to clustering around weighted prototype.

There are several possible evolutionary techniques that may be used to improve results of genetic algorithm and enhance the work presented in this research. Newly developed evolutionary algorithms such as particle swarm optimization and cohort Intelligence [32] may be integrated to find better solutions of the objective function in future work.

8. REFERENCES

[1] F. Sebastiani, Machine learning in automated text categorization, ACM Comp. Surveys. 34 (1) (2008) 1–47

[2] G. Salton, C. Buckley, Term-weighting approaches in automatic text retrieval, Inf. Process. Manage. (1988) 513-523

[3] P. Turney, P. Pantel, from frequency to meaning: vector space models of semantics, J. Artif. Intell. 37 (2010)141- 188

[4] Jun, S., Park, S.-S., & Jang, D.-S. (2014). Document clustering method using dimension reduction and support vector clustering to overcome sparseness. Expert Systems with Applications, 41, 3204–3212.

[5] Yutaka Matsuo, Mitsuru Ishizuka "Keyword Extraction from a Single Document using Word Co-occurrence Statistical Information," FLAIRS 2003.

[6] M.F. Caropreso, S. Matwin, and F. Sebastiani, "Statistical Phrases in Automated Text Categorization," Technical Report IEI-B4-07-2000, Institution Elaborazione dell'Informazione.

[7] S.Shehata, F. Karray, and M. Kamel, "A Concept-Based Model for Enhancing Text Categorization," Proc. 13th Int'l Conf. Knowledge Discovery and Data Mining (KDD '07), pp. 629-637, 2007.

[8] Zhong, S. (2005). Efficient online spherical k-means clustering. In Proceedings of the IEEE international joint conference on neural networks (pp. 3180–3185).

[9] Jian-Ping Mei, Lihui Chen (2014). Proximity-based k-partitions clustering with ranking for document categorization and analysis. Expert System with Applications.

[10] T. W. Schoenharl and G. Madey, "Evaluation of measurement techniques for the validation of agent-based simulations against streaming data," in Proc. ICCS, Kraków, Poland, 2008.

[11] Rui Xu Donald C. Wunsch, II"Clustering" John Wiley & Sons, INC., Publication, 2009.

[12] Deng-Yiv Chiu, Ya-Chen Pan, Topic knowledge map and knowledge structure constructions with genetic algorithm, information retrieval, and multi-dimension scaling method, Knowledge-Based System, Vol. 67,

[13] Clustering Ensemble: A Multiobjective Genetic Algorithm based Approach, Science Direct, 2013.

[14] Zhao, Y., & Karypis, G. (2005). Hierarchical clustering algorithms for document datasets. Data Mining and Knowledge Discovery, 10, 141–168.

[15] Kaufman, L., & Rousseeuw, P. J. (1990). Finding groups in data: An introduction to cluster analysis. New York: Wiley.

[16] Guha, S., Rastogi, R., & Shim, K. (2001). CURE: An efficient clustering algorithm for large databases. Information Systems, 26, 35–58

[17] Bellec, J. -H., & Kechadi, M. -T. (2007). CUFRES: Clustering using fuzzy representative events selection for the fault recognition problem in telecommunication networks. In PIKM (pp. 55–62).

[18] Halkidi, M., & Vazirgiannis, M. (2008). A density-based cluster validity approach using multi-representatives. Pattern Recognition Letters, 29, 773–786.

[19] Mei, J.-P., & Chen, L. (2010). Fuzzy clustering with weighted medoids for relational data. Pattern Recognition, 43, 1964–1974.

[20] J. Han and M. Kamber, Data Mining: Concepts and Techniques, 2nd ed. San Francisco, CA, USA: Morgan Kaufmann; Boston, MA, USA: Elsevier, 2006.

[21] C. G. González, W. Bonventi, Jr., and A. L. V. Rodrigues, "Density of closed balls in real-valued and autometrized boolean spaces for clustering applications," in Proc. 19th Brazilian Symp. Artif. Intell., Savador, Brazil, 2008, pp. 8–22.

[22] Lang, K. (1995). NewsWeeder: Learning to filter netnews. In Proceedings of the 12th international conference on machine learning (pp. 331–339).

[23] http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html.

[24] Marina Sokolova, Guy Lapalme. A systematic analysis of performance measures for classification tasks. Information Processing and Management 45 (2009) 427–437

[25] Strehl, A., & Ghosh, J. (2002). Cluster ensembles – knowledge reuse framework for combining multiple partitions. Journal on Machine Learning Research, 3, 583–617

[26] L. Liu, J. Kang, J. Yu, and Z. Wang, "A comparative study on unsupervised feature selection methods for text clustering," in Proceedings of IEEE International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE '05), pp.597–601,November2005.

[27] D. Zhang, S. Chen, and Z.-H. Zhou, "Constraint score: a new filter method for feature selection with pairwise constraints," Pattern Recognition, vol.41, no.5, pp.1440–1451, 2008.

[28] R. Feldman, J. Sanger, The text mining handbook advanced approaches in analyzing unstructured data, ABS Vent. (2006)

[29] H. Altyncay, Z. Erenel, Analytical evaluation of term weighting schemes for text categorization, Patt. Recog. Lett. 31 (2010) 1310–1323.

[30] M. Lan, C.L. Tan, J. Su, Y. Lu, Supervised and traditional term weighting methods for automatic text categorization, Trans. PAMI 31 (4) (2009) 721– 735

[31] F. Debole, F. Sebastiani, Supervised term weighting for automated text categorization, in: Proceedings of the 2003 ACM Symposium on Applied Computing, SAC '03, ACM, New York, NY, USA, 2003, pp. 784–788.

[32] Krishnasamy, G., Kulkarni, A. J., & Paramesran, R. (2014). A hybrid approach for data clustering based on modified cohort intelligence and k-means. Expert Systems with Applications, 41, 6009–6016.

[33] Bing Liu. Web data mining. Second Edition, Springer, 2011.

[34] Wei Song, Soon Cheol Park, Genetic algorithm for text clustering based on latent semantic indexing, Computers and Mathematics with Applications 57 (2009) 1901_1907.