International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163
Volume 1 Issue 8 (September 2014) www.ijirae.com

HYBRIDIZATION OF EM AND SVM CLUSTERS FOR EFFICIENT TEXT CATEGORIZATION

S.Arul Murugan,
HOD, Assistant Professor,
Dept of Computer Applications,
Saradha Gangadharan College, Puducherry.
Research Scholar, Periyar University, Salem

Dr. P. Suresh,
Head of the Department,
Dept of Computer Science,
Salem Sowdeswari College (Govt. Aided), Salem.
Research Supervisor, Periyar University, Salem

Abstract— Text categorization is a dynamic research area in information retrieval and machine learning. Text classification is the task of mechanically transmitting semantic categories to natural language texts and has become one of the key methods for systematize online information. Fuzzy Self-Constructing Feature Clustering (FSFC) Algorithm condenses the dimensionality of feature vectors by membership function with statistical mean and deviation [1]. Yet, lexica are not included in order to make the preprocessing mechanism more effective and do not categorize the multi-label results. A wide range of Support Vector Machine (SVM) cluster and Expectation Maximization (EM) algorithm are introduced to solve the problem. Hybridization of EM algorithm and SVM cluster combines the classification power to produce the multi-label categorization results by removing noise effectively. Initially, EM algorithm extracts the potentially noisy article from the data set using the descending porthole technique. Descending porthole is a sliding window technique used from the top to bottom of the article for preprocessing. Subsequently, SVM cluster establish the content holdup method which generates a more efficient multi-label representation of the articles. Hybridization of EM algorithm and SVM cluster outperforms the Fuzzy Self-Constructing Feature Clustering Algorithm in terms of lexica inclusion and multi-label categorization of text results. The experimental performance of Hybridization of EM algorithm and SVM cluster is evaluated with Dexter Data Set from UCI repository against existing FSFC to attain lesser execution time, clustering efficiency, and increased net similarity score level in texts.

Keywords—Expectation Maximization, Support Vector Machine, Multi-label Representation, Feature Clustering, Fuzzy Self Constructing, Preprocessing, Membership Function

I. INTRODUCTION

Clustering is one of the traditional data mining techniques where clustering methods effort to distinguish intrinsic groupings of the text articles. A set of clusters is produced in which clusters exhibit high intra cluster resemblance and low inter cluster comparison. Clustering technique is used for verdict patterns in unlabelled data with numerous dimensions. Clustering has attracted interest from researchers in the field of data mining text categorization. The main advantage of clustering algorithm is the ability to learn from and detect similar data without explicit images.

Text categorization is a primary task in information retrieval with rich body of information that has been accumulated. The normal approach to text categorization has so far been using a document symbol in a word based input space. That is, as a vector in some high dimensional Euclidean space, and then has been relying on several classification algorithms, trained in a supervised learning manner. Since the early days of text categorization, the theory and follow of classifier design has considerably superior and numerous strong leaning algorithms have emerged.

In contrast, even though numerous attempts to initiate more sophisticated document representation techniques e.g. based on higher order word statistics the simple minded independent word-based representation, known as Dexter Data Set from UCI repository stay very popular. Indeed, to-date the best multi-class, multi-labeled categorization results are based on the Dexter Data Set.

A text classification assignment consists of the training phase and the text categorization phase. The former includes the feature extraction procedure and the indexing process. The vector space model has been used as a conservative method for text representation. The model represents a document as a vector of features using Term Frequency (TF) and Inverted Document Frequency (IDF). The model simply counts TF without considering where the term occurs. But each sentence in a article has different importance for identifying the content of the document. Thus, by assigning a different weight according to the importance of the sentence to each term, achieve better results.

For upcoming problem, weights are differently weighted by the location of a term, so that the structural information of a document is applied to term weights. But FSFC method supposes that only numerous sentences, which are located at the front or the rear of a article, have the significant meaning. Hence it can be applied to only documents with a fixed form such as articles. The next step uses the title of an article in order to choose the important terms. The terms in the title are handled importantly. But a drawback is that some titles, which do not properly contain the meaning of the article, rather increase the ambiguity of the meaning. The case often appears out in documents with a familiar style such as Newsgroup and Email.

Normally, text document clustering methods effort to separate the documents into groups where each group represents several themes that is different than that theme represented by the other groups. Text classification aims at assigning class labels to text records. Text classification is based on multi word with support vector machine investigates beneficial effects [7] which achieved only appropriate information. A multi-word extraction method based on the syntactical rules of multi-word does not integrate the learning method with the characteristics of the document vectors.

The noisy document is one of the foremost reasons of diminishing the performance for binary text classification. The classifiers need to professionally handle these noisy documents to attain the high performance. These noisy documents are one of the major causes of declining the performance for text classification. This work is related to develop an EM algorithm using the descending porthole technique which is used to efficiently handle these noisy documents to achieve the high performance. In contrast, the numerous greedy approaches for feature selection only consider each feature individually with single level label extraction. So to overcome that problem, SVM cluster achieved the content holdup method for providing a good solution to the statistical problem in data mining.

II. STATE OF ART

Clustering algorithm by extending affinity propagation with a novel asymmetric similarity quantity that captures the structural information of texts. A semi supervised learning approach; develop the knowledge from a minute quantity of labeled objects but the generic seeds construction strategy is not developed [2]. Concept-Based Mining Model professionally identifies the important matching concepts among articles, according to the semantics of their sentences. The similarity between documents is calculated based on a new concept based relationship [4].

Automatic Text Categorization (ATC) is studied under a Communication System perspective feature space dimensionality reduction. The feature space dimensionality reduction has been undertaken by a two-level supervised scheme. Communication theoretical modeling aspect, with special stress on the synthesis of prototype documents via the generative model are always depend on the document coding optimal design [6]. Multidimensional Scaling (MDS) self-possessed with Procrustes CCA (Canonical Correlation Analysis) and JOFC (Joint Optimization of Fidelity and Commensurability) developed for a particular text document classification application [8].

Subspace decision cluster classification (SDCC) model consists of a disjoint subspace decision clusters. Each one labeled with an overriding class to determine the class of new objects falling in the cluster. A cluster tree is at the start generating from a training data set by recursively calls a subspace clustering algorithm using Entropy Weighting k-Means algorithm [9]. Learning methods discovers the underlying relations between images and texts based on small training data sets [10]. Similarity based multilingual retrieval paradigm, using advanced similarity calculation methods fails to result in a better performance.

Sequence classification model is defined, based on a set of chronological patterns and two sets of weights. The two set of weights, one for the patterns and one for classes. The employment of different scoring functions and other machine learning approaches, such as linear mode or neural networks, for identifying the optimal weight values are not addressed. Finally, the extension of the methodology in order to handle time series, through the use of discretization techniques are not examined [3]. The above issues can be treated by employing a pattern reduction and selection algorithm,

Novel class detection problem becomes more difficult in the existence of concept drift, when the underlying data distributions evolve in streams. The classification model occasionally needs to remain and does not address the data stream classification problem under active feature sets [14]. An objective function is constructed by combining jointly the global loss of the local spline regressions and the squared errors of the class labels of the labeled data [11]. A transductive classification algorithm is initiated in which a globally optimal classification performed. Finally obtained but does not develop an algorithm for image segmentation and image matting.

Feature Relation Network (FRN) considers semantic information and also leverages the syntactic relationships between n-gram features but not appropriate for other text classification problems [15]. The core mechanism fails to add a more complex algorithm for the creation of the summaries [5]. In order not to make a too complex system that requires long execution times. Additionally, the fact that balancing factors were used, still, the greater in length sentences were gaining more weight than the shorter ones. Accordingly this implies that several short but comprehensive sentences may be omitted. ML-based methodology for building an application that is competent of identifying and disseminating healthcare information [12] fails to extend the experimental methodology. The focus is not in integrating the research discoveries in a framework to consumers.

Personalized ontology model represented over user profiles but fails to generate user limited instance repositories to go with the representation of a global knowledge base. The current system assumes that all user local instance repositories have contented based descriptors referring to the subjects, however, a large volume of documents existing on the web may not have sufficient such content-based descriptors [13]. Moreover, we now discuss to make the preprocessing mechanism more effective and categorize the multi-label results on text categorization. In summary, our contributions are:

(i) EM algorithm removes the potentially noisy articles from the dexter dataset using the descending porthole technique
(ii) After, EM algorithm based preprocessing; SVM cluster generates a more efficient multi-label representation of articles.

(iii) Multi label categorization of text results achieved on dexter datas.
(iv) Finally, Hybridization of EM algorithm and SVM cluster improve the similarity score level of text in articles.

III. METHODOLOGY

The noisy articles are usually located in a dataset and descending porthole technique is employed to effectively detect the noisy area. By estimate the entropy of mixed articles in a descending porthole, the noisy areas are found and all the articles in the area are regards as unlabeled data. EM algorithm efficiently holds unlabeled articles by providing the solution to extract and remove noisy articles from unlabeled data.

Subsequent process is clustering of a more sophisticated text categorization method using Content Holdup (CH) method. CH approach is used instead of articles in a feature cluster space, where each cluster is a distribution over article classes. Support Vector Machine (SVM) cluster allows for the best reported result for a multi level categorization of dexter dataset. SVM with content holdup method generates a more efficient multi-label representation of the articles.

Hybridization of EM algorithm and SVM cluster for efficient text categorization consists of two modules namely instruction module and text categorization module. The architecture diagram of the Hybridization of EM algorithm and SVM cluster for efficient text categorization is shown in Fig 1.

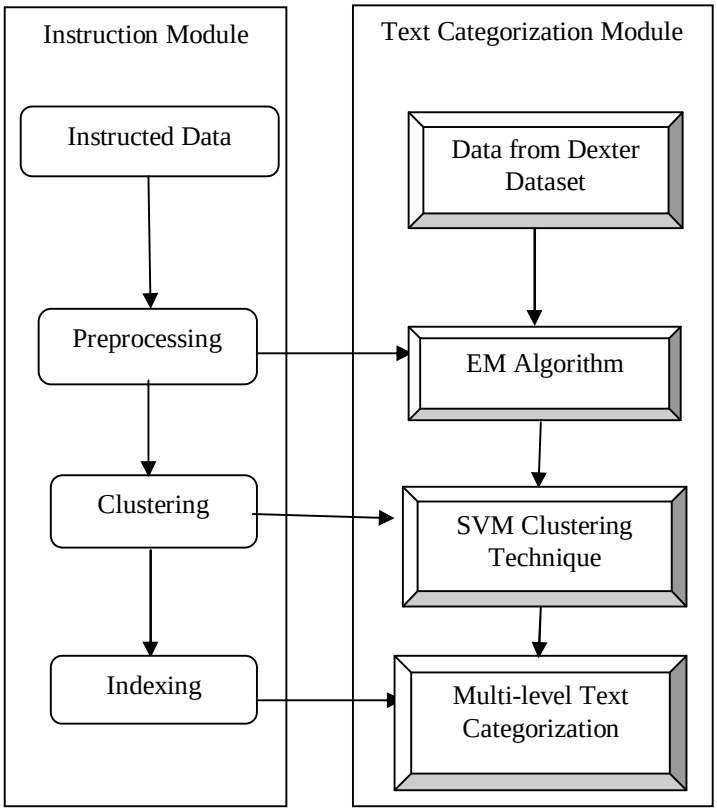

Fig 3.1 Hybridization of EM algorithm and SVM cluster Process

Fig 3.1 demonstrates the Hybridization of EM algorithm and SVM cluster on dexter dataset. Instructed module holds the list of process undergone, whereas the text categorization module holds the processing technique. EM algorithm and SVM cluster are combined together in order to make the preprocessing mechanism more effective and do not categorize the multi-label result.

3.1 Expectation Maximization for preprocessing

The proposed Expectation Maximization (EM) approach consists of the following four steps namely one adjacent to the rest method, calculating prediction scores to remove noise, calculating entropy using the descending porthole technique, and the EM algorithm. In the one adjacent to the rest method, the article of one group is regarded as positive examples and the documents of the other categories as negative examples. In order to set up training data into binary classification from dexter data Set, multi class setting is reformed into the binary setting using the one adjacent to the rest method.

The goal is to discover a edge area which denotes an area including many noisy articles. First of all, using a positive data set and a negative data set for each category from the one adjacent to the rest method, learn a Naive Bayes (NB) classifier and obtain a prediction score for each document by the following formula.

$$EM\ prediction\ Score\ (g_i, a_j) = \frac{P(positive|a_j)}{P(positive|a_j)+P(Negative|a_j)} \quad \text{..... Eqn (1)}$$

Where, g_i is a group and a_j means an article of g_i. $P(positive|a_j)$ means a probability of the article a_j to be positive in g_i, and $P(Negative|a_j)$ means a probability of the article a_j to be negative in g_i. According to these prediction scores, the entire articles of each group are sorted out in the descending order. Probabilities, $P(positive|a_j) + P(Negative|a_j)$ is generally calculated as follows

$$P(positive|a_j) = \frac{P(positive)\,P(d_j|positive)}{P(a_j)} \quad \text{..... Eqn (2)}$$

$$= P(positive) \prod_{i=1}^{V} P(v_i|positive)^{N(v_i|a_j)} \quad \text{..... Eqn (3)}$$

$$\frac{x \log(P(positive))}{n} + \prod_{i=1}^{|V|} P(v_i|a_j) \log\left(\frac{p(v_i|positive)}{P(v_i|a_j)}\right) \quad \text{..... Eqn (4)}$$

Where $v_i$the ith word in the vocabulary, V is is the size of the vocabulary, and $N(v_{i|a_j})$is the frequency of word v_i in article a_j.

In EM method, an edge detected in a block with the most mixed degree of positive and negative articles. The descending porthole technique is first used to detect the block. In EM technique, porthole of a certain size is descending from the top article to the last article in a list ordered by the prediction scores. An entropy value is calculated for estimating the mixed degree of each porthole to remove noise ratio for effective preprocessing as

$$\text{Entropy (P)} = -q_+ log_2 q_+ - q_- log_2 q_- \quad \text{..... Eqn (5)}$$

Where, given a porthole (P), q+ is the proportion of positive articles in P and $q_-$is the proportion of negative articles in P. For example, if a porthole of five articles has three positive articles and two negative articles, the proportions of positive articles and negative articles are 3/5 and 2/5 respectively. Thus the final predictable entropy value is calculated when using five articles for porthole operation.

Two portholes with the highest entropy value are picked up; one porthole is firstly detected from the top and the other is firstly detected from the bottom. If there is no porthole or only one porthole with the highest entropy value, porthole with the next highest entropy value becomes targets of the selected windows. Then maximum (max) and minimum (min) threshold values are searched from selected windows, respectively. The max threshold value is found as the highest prediction score of a negative articles in the former window and the min threshold value is as the lowest prediction score of a positive article in the latter porthole.

The articles between max and min threshold values has three classes for training articles namely absolutely positive articles , unlabeled articles , definitely negative articles. By applying the EM algorithm to these three data sets, extract actual noisy articles and remove them. EM algorithm is used to pick out noisy articles from unlabeled article for effective preprocessing operation. The universal EM algorithm consists of two steps originally trains a classifier using the obtainable labelled articles and labels the unlabeled articles by rigid classification.

//EM algorithm

Begin

Step 1: Each article with P (positive data) and N (negative data) are assigned

Step 2: a_+ is article of P, a_- is article of N, $a_?$ is unlabeled data.

Step 3: P' {}, N' {} uses the current NB classifier using adjacent to the rest method

Step 4: For each article, $a_? \in U$ *do*,

Step 5: If $P(g_+|a_?) \geq P(g_-|a_?)$ *then*,

Step 6: Continue $(P' = PU\{\})$

Step 7: Else $(P' = PU\{a_?\})$

End

EM instructs a new classifier using the labels of all the articles and iterates to convergence. E' step is reformed to effectively remove the noise articles located in the edge area and it does not assign an unlabeled article a_u to the positive data set, 'P', because it regards a_u as another noisy article. The positive articles 'P' are labeled by hand and have enough information for a group, additional positive articles decrease performance. As a final point, text classifiers with binary training dexter dataset produce effective preprocessing step.

3.2 SVM with content holdup method for Clustering

Support vector machine (SVM) is an inductive clustering scheme that newly proved to be successful along various application domains. In particular, there are numerous pieces of verification that designate that SVM is a good choice for text categorization. The data is linearly separable, linear SVM computes the maximum margin linear classifier. The SVM clustering case is an extension that allows effective clustering of similar text from article. A simple SVM approach deal with multi labeled text categorization with 'm' classes and it decompose the difficulty into 'm' binary problems. There exist recent decomposition methods that seem to be more dominant. However, for ease and for contrast with associated results choose simple decomposition for SVM clustering.

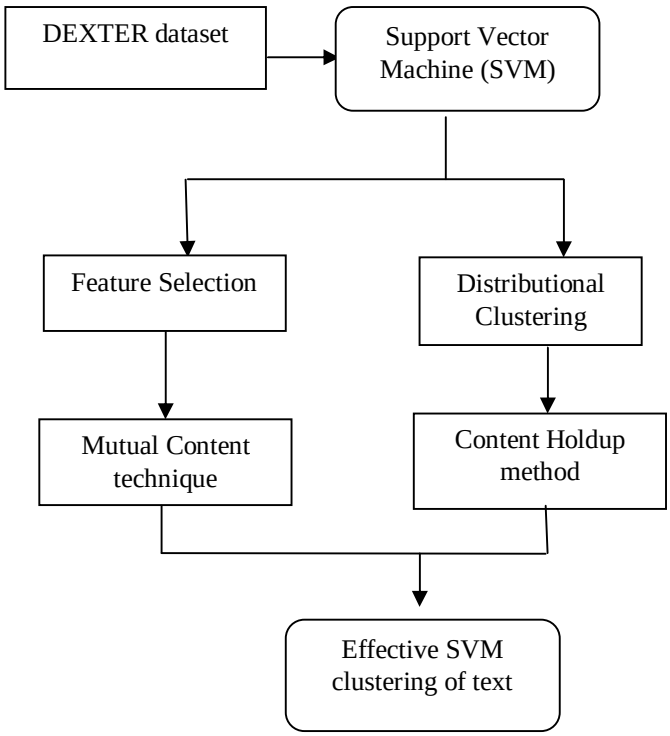

Fig 3.2 SVM Clustering Method

SVM consists of two processes as described in fig 3.2. The first process is based on feature selection using the mutual content technique. Where the k-most selective features (words) are selected, the articles are projected on them and then the SVM classifier is trained on the projections. The second process is based on Content holdup distributional clustering. Primarily, words of the training set are clustered into k-clusters using the deterministic annealing of the information (i.e.,) contents respectively.

3.3 Hybridization of EM and SVM

Hybridization of EM and SVM finds the similarity score level between the title and each sentence, and then assigns the higher importance to the sentences with the higher similarity. The title and each sentence of an article are represented as the vectors of content words. The similarity value of them is calculated by the inner product and the calculated values are normalized into values between 0 and 1 by a maximum value. The similarity value between the heading 'H' and the sentence S_i in an article 'a' are calculated by the following formula:

$$Similarity\ (S_i, H) = \frac{\vec{S}_i \vec{H}}{\max_{S_j \in a} (\vec{S}_j \vec{H})} \quad \text{..... Eqn (6)}$$

Where, $\vec{H}$ denotes a vector of the heading, and $\vec{S}_i$ denotes a vector of sentence. The importance value of a sentence is used for modifying the Word Frequency (WF) value of a term. That is, since a WF value of a term in a article is calculated by the sum of the WF values of terms in each sentence, the Modified WF (MWF) value (WTF, (a,t)) of the term 't' in the article 'a' is calculated as,

$$MWF(a,t) = \sum_{s_i \in a} wf(s_i,t) * score(S_i) \quad \text{…….. Eqn (7)}$$

Where, $wf(s_i,t)$ denotes WF of the term t in sentence s_i.Eqn (7) used to have effective indexing for multi label text categorization. The below describes the pseudo code of EM and SVM hybridization.

Input: Dexter Dataset, Article 'a', Cluster 'H'

Output: Set of multi-labels text categorization

For each Instructed Data
If Preprocess Document'd'
Uses EM descending porthole technique positive p' {} and negative n'{};
End if
End For Each

For each Preprocessed Data
Cluster h€'H' using Content Holdup method
Run h_i on (a) to obtain text
Effective distributional Clustering
End For Each

For Each Clustered Group
Perform Indexing operation based on 'H' and S_i
Multi-label text Categorization achieved
End For Each

The above pseudo code generally illustrates the steps followed for combining of EM and SVM cluster for producing the multi-label categorization results by removing noise effectively. Indexing method for text categorization uses the heading and the other uses the consequence of terms. For experiments, uses dexter dataset from UCI repository. Our system achieved a better performance than the basis system in all these classifiers and verified the effect of the proposed indexing method by measuring cohesion. The hybridization of EM and SVM indexing method reform the article vector space for a better performance in multi level text categorization.

IV. EXPERIMENTAL EVALUATION

Hybridization of EM algorithm and SVM cluster is evaluated with DEXTER dataset from UCI repository. Hybridization of EM algorithm and SVM cluster is implemented in JAVA. Dexter is a two-class text classification system with sparse continuous input variables and bag-of-word representation. Dexter dataset is single of five datasets feature selection challenge. The data are split into training, validation, and test set. Goal values are provided only for the 2 first sets. Test set performance results are obtained by submit prediction results.

The innovative data were 9947 features of which 2562 are always zeros for all the examples on behalf of frequencies of occurrence of word stems in text. Hybridization of EM algorithm and SVM cluster task are compared with Fuzzy Self-Constructing Feature Clustering Algorithm to learn which Reuter's articles are about corporate acquisitions with multi-label categorization of results. It is then added a number of distracter feature called probes having no predictive power. dataname.param, dataname.feat, dataname_valid.data, dataname_train.data are the data formats in Dexter dataset.

Clustering Efficiency is grouping of similar text in for the effective multi level text categorization. It is measured in terms of percentage (%). Execution time is defined as the amount of time taken to perform the preprocessing step using the EM. It is measured in terms of seconds (sec).

Execution time = Total no of time taken to perform features extraction - Amount of request time taken

Similarity score level is defined as the rate of matching accuracy between the original texts and the test text article. It is measured in terms of percentage (%).

V. RESULTS AND DISCUSSION

Hybridization of EM algorithm and SVM cluster evaluates the multi label text categorization using dexter dataset documents. The experimental performance of Hybridization of EM algorithm and SVM cluster is evaluated against Fuzzy Self-Constructing Feature Clustering (FSFC) Algorithm [1]. EM and SVM hybrid is measured in terms of execution time, clustering efficiency, and net similarity score level in texts with multi-label categorization results.

No. of extracted features	Execution Time (sec)	
	FSFC Algorithm	Hybridization of EM and SVM
50	199	175
100	203	187
150	234	211
200	247	220
250	262	233
300	276	255
350	288	253

Table 5.1 No. of extracted features vs. Execution Time

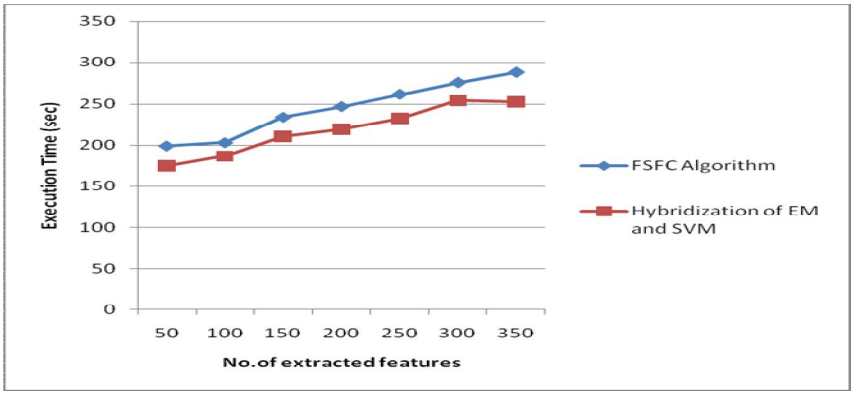

Fig 5.1 No. of extracted features vs. Execution Time

Fig 5.1 shows the performance evaluation of execution time using the EM algorithm and FSFC Algorithm. Execution time is measured based on the extracted features. The EM algorithm reduces the execution time to 7 – 15 % when compared with FSFC, because porthole is measured in descending order. The porthole measured from the top article to the last article in a list ordered by the prediction scores. EM algorithm removes the noise level in text categorization module using maximum and minimum threshold values calculation from selected porthole, correspondingly.

Text categorization Size (KB)	Clustering Efficiency (%)	
	FSFC Algorithm	Hybridization of EM and SVM
22	76	80
38	75	81
41	78	83
73	79	85
125	80	87
159	82	89
202	83	89

Table 5.2 Text categorization Size vs. Clustering Efficiency

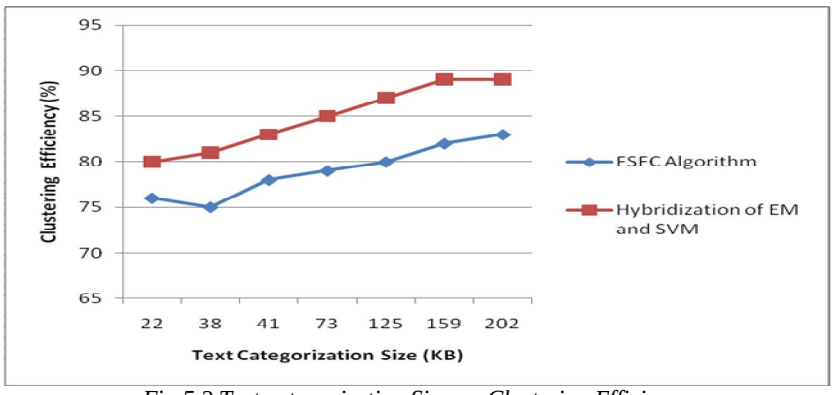

Fig 5.2 Text categorization Size vs. Clustering Efficiency

As the above parameter, clustering efficiency is measured based on the text categorization size. The clustering efficiency is 4 – 8 % is greater when compared with FSFC Algorithm [1]. FSFC algorithm condenses the dimensionality of feature vectors by membership function, so it is not effective in clustering. SVM with content hold up method improves the clustering efficiency using the resulting self consistent equations which essentially coincides with the preprocessed data. As the text size increases, clustering efficiency is also achieved in SVM clustering.

Test Text Article	Net Similarity Score Level (%)	
	FSFC Algorithm	Hybridization of EM and SVM
Test Set 1	87.55	91.12
Test Set 2	88.12	92.45
Test Set 3	88.79	92.32
Test Set 4	89.47	93.45
Test Set 5	89.95	94.52
Test Set 6	90.12	94.87
Test Set 7	91.88	95.13

Table 5.3 Test Text Article vs. Net Similarity Score Level

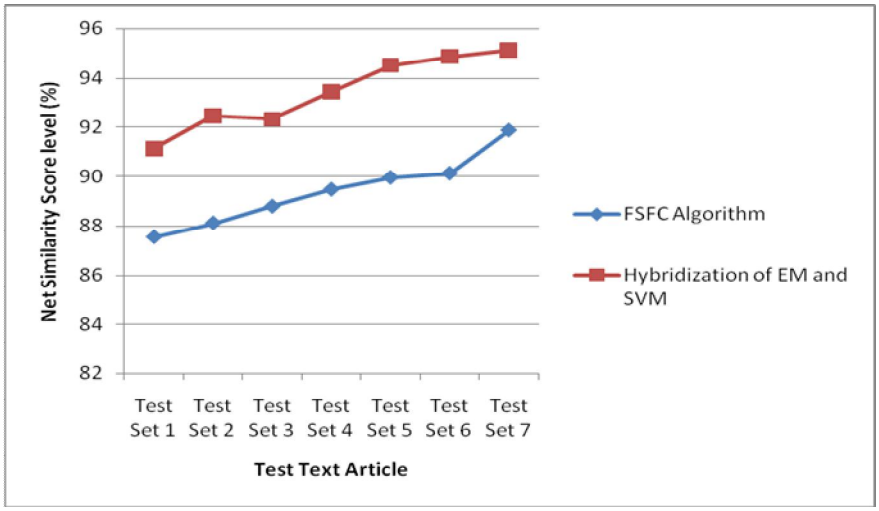

Fig 5.3 Test Text Article vs. Net Similarity Score Level

Similarity score level of the text article is measured to identify the score points. Different forms of articles set from dexter dataset are used for net similarity score level calculation in texts.Net similarity score level is 2 – 5 % higher in Hybridization of EM and SVM because of using Word Frequency (WF) value of a term. Modified word frequency initially uses the heading 'H' and sentence S_i similarity to improve the score level when compared with the FSFC algorithm.

As a final point, sorting the scores of all candidate groups, obtain a ranked list of group for each test set articles. The articles of dexter datasets have only one group, assign the only top ranking group to each article. Hybridization of EM and SVM algorithm improves the clustering effect with multi level text categorization. An alternative way is to count the decisions for all the group and compute the global result.

VI. CONCLUSIONS

Support Vector Machine (SVM) cluster and Expectation Maximization (EM) algorithm are combined together to make the preprocessing mechanism more effective and to categorize the multi-label results. Hybridization of EM algorithm and SVM cluster combines the classification power and produce the multi-label categorization results by removing noise effectively. Initially, EM algorithm extracts the potentially noisy documents from the data set using the descending porthole technique. Subsequently, SVM cluster establish the content holdup method which generates a more efficient multi-label representation of the documents. Experimental results are conducted using Dexter dataset, outperforms the proposed system with various statistical parameters. Hybridization of EM algorithm and SVM cluster with Dexter Data Set from UCI repository outperforms well against existing FSFC to attain lesser time taken for execution, 6.83 % effective in clustering, increased net similarity score level in texts with multi-label categorization results.

REFERENCES

[1] Jung-Yi Jiang., Ren-Jia Liou., and Shie-Jue Lee., "A Fuzzy Self-Constructing Feature Clustering Algorithm for Text Classification," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 23, NO. 3, MARCH 2011

[2] Renchu Guan., Xiaohu Shi., Maurizio Marchese., Chen Yang., and Yanchun Liang., "Text Clustering with Seeds Affinity Propagation," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 23, NO. 4, APRIL 2011

[3] Themis P. Exarchos., Markos G. Tsipouras., Costas Papaloukas., Dimitrios I. Fotiadis., "A two-stage methodology for sequence classification based on sequential pattern mining and optimization," Science Direct, Elsevier Journal, Data & Knowledge Engineering (2008)

[4] Shady Shehata, Fakhri Karray, and Mohamed S. Kamel, "An Efficient Concept-Based Mining Model for Enhancing Text Clustering," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 22, NO. 10, OCTOBER 2010

[5] Christos Bouras., Vassilis Poulopoulos., Vassilis Tsogkas., "PeRSSonal's core functionality evaluation: Enhancing text labeling through personalized summaries," Science Direct, Elsevier Journal, Data & Knowledge Engineering 64 (2008)

[6] Marta Capdevila., and Oscar W. Marquez Florez., "A Communication Perspective on Automatic Text Categorization," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 21, NO. 7, JULY 2009

[7] Wen Zhang., Taketoshi Yoshida., Xijin Tang., "Text classification based on multi-word with support vector machine," Science Direct, Elsevier Journal., Knowledge-Based Systems 21 (2008).

[8] Ming Sun., Carey E. Priebe., "Efficiency investigation of manifold matching for text document classification," Science Direct, Elsevier Journal., Pattern Recognition Letters 34 (2013)

[9] Yan Li., Edward Hung., Korris Chung., "A subspace decision cluster classifier for text classification," Science Direct, Elsevier Journal., Expert Systems with Applications., 2011

[10] Tao Jiang., and Ah-Hwee Tan., "Learning Image-Text Associations," IEEE TRANSACTION ON KNOWLEDGE AND DATA ENGINEERING., 2008

[11] Shiming Xiang., Feiping Nie., and Changshui Zhang., "Semi-Supervised Classification via Local Spline Regression," IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 32, NO. 11, NOVEMBER 2010

[12] Oana Frunza., Diana Inkpen., and Thomas Tran., "A Machine Learning Approach for Identifying Disease-Treatment Relations in Short Texts," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 23, NO. 6, JUNE 2011

[13]Xiaohui Tao., Yuefeng Li., and Ning Zhong., "A Personalized Ontology Model for Web Information Gathering," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 23, NO. 4, APRIL 2011

[14] Mohammad M. Masud., Jing Gao., Latifur Khan, Jiawei Han, "Classification and Novel Class Detection in Concept-Drifting Data Streams under Time Constraints," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 23, NO. 6, JUNE 2011

[15] Ahmed Abbasi., Stephen France., Zhu Zhang, and Hsinchun Chen., "Selecting Attributes for Sentiment Classification Using Feature Relation Networks," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 23, NO. 3, MARCH 2011

