International Journal of Computing Science and Information Technology, **2013, Vol.3,Iss.1 07-12**
ISSN: 2278-9669, January 2015 (http://ijcsit.org)

Aggregated Probabilistic Fuzzy Relational Sentence Level Expectation Maximization Clustering Algorithm using Subjective Method For Efficient Text classification

[1] Kartheek V.L, [2] Chandra Sekhar V

[1] Computer Science & Eng. Dept., S.R.K.R Engineering College, Bhimavaram, Andhrapradesh, India.
Email: kartheek.0531@gmail.com

[2] Computer Science & Eng. Dept., S.R.K.R Engineering College, Bhimavaram, Andhrapradesh, India.
Email: chandu.vasamsetty@gmail.com

Abstract

Now a days, Text clustering becomes an important application to organize the data and to extract relevant and accurate information from the available corpus. Many previous clustering techniques have difficulties in getting the relevant information belonging to a particular subject. In this paper we proposed an aggregated probabilistic Fuzzy relational sentence level expectation maximization clustering algorithm using subjective method. It will not only give the accurate and maximum similarities of sentences but also give the results in terms of accurate context. The practical results show that the proposed method obtains better and accurate results for getting best Subject wise sentence-level text classification when compared with the existing methods.

Keywords*:* Corpus, Fuzzy Clustering, Subjective method, probabilistic.

1. Introduction

Cluster analysis is useful tool for finding required information from available large sets of text with relevant subject or context. Cluster analysis is a technique for classifying data [1], i.e., to division of a given set of objects or things into a set of classes or clusters based on similarity. The goal is to divide the sentence into words and then know the probability of belonging to a particular class or cluster [2]. It is a method of finding the relevance of sentences to some class or cluster. The hard clustering methods restrict each point of the data set to exactly one cluster [1]. These methods yield exhaustive partitions of the example set into non-empty and pair wise disjoint subsets. Fuzzy cluster analysis, [3] allows accurate relevance of words to clusters in the range of [0, 1]. This tells the flexibility to express that data points belong to more than one cluster at the same Time. Furthermore, these relevance of words offer accurate estimation. Clustering is the process of grouping or aggregate of data items. Sentence level clustering used in different applications such as classify and categorization of documents and organizing the documents, etc [4]. In text processing, sentence clustering plays an important role and is used in various text mining activities [5]. Size of the clusters may change from one cluster to another [6]. The previous clustering algorithms have some problems in clustering the input dataset [3] and also not identifying the outliers. Against the drawbacks of these clustering algorithms, Later various clustering algorithms can be developed for the clustering of sentences [7]. In those, Contents present in text documents contain hierarchical structure and there are many of the terms present in the documents which are related to more than one point [8]. But in the previous algorithms, the accuracy of belonging to a particular class or cluster is very low.

Hence we proposed aggregated probabilistic Fuzzy relational sentence level expectation Maximization clustering algorithm. The various previous algorithms can be facilitates some poor performance. The fuzzy algorithms find all the possibilities of relevance. From this method a large variety of clustering techniques was derived with more complex prototypes, which are mainly interesting in data analysis applications [9]. However, the generalization of these techniques to clustering uncertain data or objects is not yet explored. The sentence can be accurately predicts the level of matching to a particular cluster. The Text classification plays major role today for all fields [10]. Recently, fuzzy set theory is more and more

frequently used because of its simplicity and usage in different applications [11]. The theory has been successfully applied to use in many fields. The Fuzzy concept is very popular to get accurate and efficient results.

1.1 Clusters of data

There are many algorithms already there for clustering of text or data [10]. Each algorithm will group the data objects based on some metrics or measure. The useful data can be classified for better way of getting things. Clustering is used in many different applications. The text mining [10] is the process of extraction or getting data efficiently. The retrieving of information is challenging today [13]. The similarity of words in a given sentence will decide the accuracy of belonging to a cluster. Sentence Clustering mainly used in variety of text mining applications. Clustering is one of the most [11] important concept for group of objects. When searching for a required data this technique is very useful.

2. Back Ground and Related Work

2.1 Efficient Subject wise Relevance Mining on Text Clusters

The subdivision of a given system always avoids the difficulty and problem. The sub division of a system concept simplifies the system easily and conveniently. The words can be taken from a particular related sentence and then solve it by getting aggregates. Based on this the estimation of nearest class can be identified. The text classification is very important for not only for the country but also for the entire world. The Clustering is one of the data mining techniques [13] that are used for classifying text. In this paper, we are going to present aggregation of probabilities of words in a given sentences with relevance to a subject or context.

2.2 The Relevance Similarity of Sentences

This proposal is a new approach for measuring the similarity between the collection of words and then the sentences. The relevance of words in a sentence can specify how much similar they are to the given classes of clusters. The finding of this important similarity gives the proof using this method [13]. The combining of sub systems gives the better results. The proposed approach outperforms the similarity between the words by find out the probabilities with the use of sub systems of a given system. Thus, we not only get the relevancy in terms of input text but also with accuracy of context. In this method the probabilities can be find out with respect to subject of context. So, it gives Better results when compared to previous algorithms.

2.3 Experiments on Probabilistic Sentence Level Clustering using Subjective

Identifying required data plays an important role in Text mining. The proposed method is based on the concept of the aggregated probability of sentences [15]. It used to find the relevant information or data sentences from a collection of documents. It uses the concept of sub division of a system.

2.4 Aggregated Expectation Maximization Clustering Algorithm using subjective method

Nowadays, large amount of data is available in the form of texts. It is very difficult for the people to find out useful and significant data [16]. For getting useful data we have many different algorithms [17]. The useful data can be taken from the large amount of data which is available and this data is in short and concise form [18]. This proposal describes a system, which consists of two steps. In first step, we are finding out the relevance of words in a given sentence [14]. In second step, we are implementing Expectation Maximization Clustering Algorithm [15] to find out sentence similarity between the sentences based on the value of aggregated relevance of sentences, we can easily estimate the matching of text data to a class.

3. Proposed Work

In this wok, the analysis of one can take advantage of the efficiency and stability of clusters, when the data to be clustered are available in the form of similarity relationships between pairs of words. More precisely, we propose a new aggregated probabilistic Fuzzy relational sentence level expectation Maximization clustering algorithm which does not require any restriction on the relation matrix. This APFRSEC algorithm is applied for the clustering of the text data which is present. APFRSEC will give the output as clusters which are grouped from text data which is present in a given documents. In this APFRSEC algorithm, Page Rank algorithm is used as similarity measure.

3.1 Page Score Value

In the proposed algorithm, the use of Page score algorithm is to calculate the number of page hits and traffic. Page score algorithm is used to determine the importance of a particular node or thing to visit for usage. This algorithm specifies the most occurrences of things in use. This score is known as Page rank Score. Sentence is rep-

resented by system in this assumption. Then find out the most possibility of nearer representing similarity between sentences. Sentence in a document is represented by a node in the directed graph and the probabilities specify the similarity to a class. The page score algorithm retrieves the data based on the number of hits.

3.2 Emax Algorithm

It is a method that tries to find aggregated probabilities of a sentence that has the maximum likelihood of getting nearer results. Its main role is to calculate nearest estimation. It is an important method, which is mainly used to finding the maximum aggregated probabilities of the model. The E-step consists the calculation of aggregated probabilities. The probabilities calculated from E-step are compared in Max-step for maximum values.

3.3 Subjective Method

This method is proposed to find out the required text with relevant subject or context. It finds out the accurate aggregated probabilities based on the context.

3.4 APFRSECS

Each of these subsystems of a given system can be calculated for relevance and then find out the matching. A System structure with n sub models fuzzy systems is depicted in Fig. 2. Each of this system structure has correspondent fuzzy system relevance Ri(x). In the following, the system is explained with the perspective of sentence and the words are taken as sub systems to reduce the burden and identify efficiently. Therefore, the output of the above model is the aggregation of the each sub system component of each fuzzy system. Let the classes be c1, c2, c3,…….cm. The sentence can be divided into appropriate words like w1, w2, w3,….wj. The probability of a word belonging to a class can be calculated from the following:

Relevance Ri(x) = P(cm / wi) = $\frac{\sum_{q=1}^{n} Yqi * Fj}{\sum_{q=1}^{n} yqi}$,------(1)

Yqi specifies the number of occurrences of word with in document yq.

Where Fj =1 ,if document yq belongs to class cm
0 ,otherwise .

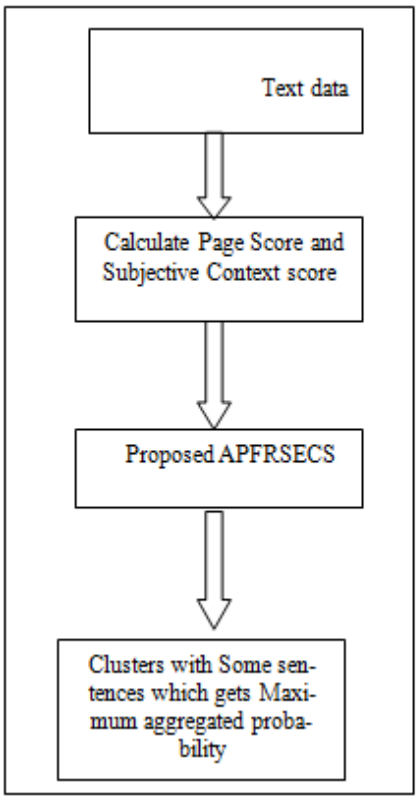

Figure1: APFRSECS Clustering Process

Aggregation means combining two or more attributes into a single attribute. The main purpose of using this aggregation method is to data reduction and also aggregated data tends to have less variability. For example, cities aggregated into regions, states, countries etc. Aggregation is also called as summation. The relevance of an aggregated system is calculated from the following equation.

3.5 Proposed APFRSECS Algorithm

Our proposed aggregated probabilistic Fuzzy Relational sentence level Expectation maximization clustering Algorithm using subjective method is developed using Sub system concept with the application of Probability. It is used in this work to implement the separation of information among the various subsystems, which are organized into a original System structures. Each of these subsystems may contain information related with particular aspects of the system.

Initialization:

# of original word patterns : m
# of sentences formed : s
# of classes : p

Initial #of cluster : k=0

Input:
$X_i = <x_{i1}, x_{i2}, x_{i3},\ldots\ldots x_{ip}>, 1<=i<=s>$

Output:
Clusters G1, G2; . . .; Gk

Procedure for APFRSECS Algorithm:

Based on the input text, find out the accurate context and subject.

For Each word pattern in a Sentence S_i,1<=i<=m
Probability Relevance R_i is finding out by (1);

Then aggregating the probability relevance of each word in a given sentence by (2).

The Sentence which gets maximum aggregated Probability of Relevance can belongs to a Class C_i .

Return with the Created K clusters;

End Procedure.

4. EXPERIMENTAL RESULT

DOC	Query (w1)	For (w2)	Max salary (w3)	Of (w4)	Emp (w5)	Class
d1	1	1	0	1	0	C1
d2	1	1	1	1	1	C1
d3	0	0	0	0	1	C2
d4	0	1	1	0	0	C2
d5	1	0	0	1	0	C2

Table 1: Document Set D1

For example,
Suppose there are two classes C_1 and C_2. The five documents d1, d2, d3, d4 and d5 belonging to c1, c1, c2, c2, c2 respectively. See the Table 1. For values of words appearing no. of times and belongs to a class. Let the occurrences of w1 in these documents be 1, 2, 3, 4 and 5 respectively. Then, the probability of word w1 belongs to a class c1 is calculated as

$$R_1(x) = P(c1/w1) = \frac{1*1+2*1+3*0+4*0+5*0}{1+2+3+4+5} = 0.2,$$

$$R_2(x) = P(c1/w2) = \frac{1*1+2*1+3*0+4*0+5*0}{1+2+3+4+5} = 0.2,$$

Similarly we have to find the $R_3(x)$ or P(c1/w3), $R_4(x)$ or p(c1/w4) and $R_5(x)$ or P(c1/w5). Then the finding of aggregating these all words probabilities belonging to a particular class C1 . Then getting a result of aggregated relevance of sentence S_1 belongs to a class C1 as follows. Then the values found are,

R1(x) = 0.2

$R_2(x) = 0.2$

$R_3(x) = 0.4$

$R_4(x) = 0.1$

$R_5(x) = 0.4$

$Agg_{R(x)}$ of S1 belongs to a class C_1 =
$R_1(x) + R_2(x) + R_3(x) + R_4(x) + R_5(x)$
= 0.2 + 0.2 + 0.4 + 0.1 + 0.4
= 1.3

After that take sentence S_1 and dividing it into possible words like w1, w2, w3, w4 by sub systems concept. Then calculate the following probabilities of relevance of words w_i in a given sentence Like $R_1(x)$ or P (c1/w1), $R_2(x)$ or P (c1/w2), $R_3(x)$ or P (c1/w3), $R_4(x)$ or P (c1/w4). Then the finding of aggregating these all words probabilities belonging to a particular class C_2. Then getting a result of aggregated relevance of sentence S_1 belongs to a class C_2 as follows.

$Agg_{R(x)}$ of S1 belongs to a class C_2 =
$R_1(x) + R_2(x) + R_3(x) + R_4(x)$
= 0.3 + 0.2 + 0.3 +0.4
= 1.2

If $Agg_{R(x)}$ of S_1 belongs to a class C_1 > $Agg_{R(x)}$ of S_1
Belongs to a class C_2. Then the sentence S1 is Belongs to Class C1.
Similarly, the aggregated relevance is find out for each and every sentence to get accurate outcome. The sentence gets the maximum relevance of belonging to a class can get the priority. When discuss about any algorithm, the specification of its importance and efficient performance is very important. Then getting a result of

aggregated relevance of sentence S_1 belongs to a class C1 as follows.
Then the values found are,

$R1(x) = 0.2$

$R_2(x) = 0.2$

$R_3(x) = 0.4$

$R_4(x) = 0.1$

$R_5(x) = 0.4$

$Agg_{R(x)}$ of S1 belongs to a class C_1 =

$R_1(x) + R_2(x) + R_3(x) + R_4(x) + R_5(x)$

$= 0.2 + 0.2 + 0.4 + 0.1 + 0.4$

$= 1.3$

After that take sentence S_1 and dividing it into possible words like w1, w2, w3, w4 by sub systems concept. Then calculate the following probabilities of relevance of words w_i in a given sentence Like $R_1(x)$ or P (c1/w1), $R_2(x)$ or P (c1/w2), $R_3(x)$ or P (c1/w3), $R_4(x)$ or P (c1/w4). Then the finding of aggregating these all words probabilities belonging to a particular class C_2. Then getting a result of aggregated relevance of sentence S_1 belongs to a class C_2 as follows.

$Agg_{R(x)}$ of S1 belongs to a class C_2 =

$R_1(x) + R_2(x) + R_3(x) + R_4(x)$

$= 0.3 + 0.2 + 0.3 + 0.4$

$= 1.2$

If $Agg_{R(x)}$ of S_1 belongs to a class $C_1 > Agg_{R(x)}$ of S_1
Belongs to a class C_2. Then the sentence S1 is Belongs to Class C1.

Similarly, the aggregated relevance is find out for each and every sentence to get accurate outcome. The sentence gets the maximum relevance of belonging to a class can get the priority. For, example when searching for the required data, this algorithm gives the low search results with highest accuracy.

The Proposed Subjective Method not only gives the accuracy just with the input keywords but also gives the accuracy in terms of accurate and correct context.

For Example, if we want search for "Query for Max salary of employee". Then First it want to check for the context and subject it is related to?. Then it can Find out the Relevant Aggregated probabilities and retrieve the information.

Table 3. Text sample vs. Average relevance % Level

Test	Average relevance percentage Level (%)	
	Fuzzy Self FC Algorithm	APFRSECS Algorithm
Text sample 1	87.69	92.30
Text sample 2	87.99	93.05
Text sample 3	88.73	92.68
Text sample 4	86.40	93.62
Text sample 5	88.49	95.01
Text sample 6	88.99	95.02
Text sample 7	91.00	96.00

Table 2. Text sample vs. Average relevance % Level

5. CONCLUSION & FUTURE WORK

In this Work, aggregated probabilistic Fuzzy Relational sentence level Expectation maximization clustering Algorithm using subjective method is proposed and applied to clustering fuzzy sets. . This algorithm gives the low search results with highest accuracy. It gives the More Accurate and efficient Results with relevant context when compared to Existing System. Thus we found maximum likelihood estimates of parameters. When the number of data sets increases, then the APFRSECS algorithm takes more time to perform clustering. This proposal can be useful in future for research work.

REFERENCES

[1] Jung-Yi Jiang, Ren-Jia Liou, and Shie-Jue Lee, Member A Fuzzy Self-Constructing Feature Clustering lgorithm for Text Classification, March 2011, IEEE, VOL. 23, NO. 3.

[2] N. Slonim and N. Tishby. "The power of word clusters for text classification," 23rd European Colloquium on information Retrieval Research (ECIR), 2001.

[3] Neha Mehta, Mamta Kathuria, Mahesh Singh, "Comparison of conventional and fuzzy Clustering Techniques: A survey", April 2014 IJARCCE, Vol. 2, Issue 1.

[4] G.Thilagavathi, J.Anitha, K.Nethra," Sentenc Similarity Based Document Clustering using Fuzzy algorithm", March 2014, IJAFRC, Vol1, Issue 3.

[5] K.Jeyalakshmi1, R.Deepa2, M.Manjula," An Efficient Clustering Sentence-Level Text Using A Novel Hierarchical Fuzzy Relational Cluster- Ing Algorithm", February 2014, IJARCCE, Vol.3, Issue.2

[6] M.S.Yang," A Survey of Fuzzy clustering", October 1993, Vol 18, No 11.

[7] F.Pereira, N. Tishby, and L.Lee."Distributional Clustering of English words," 31st Annual Meeting of ACL, 1993, pages 183– 190.

[8] Hathaway RJ, Bezdek JC Recent convergence results for the fuzzy C-means clustering algorithms, oct 1988. J Class 5:237-247.

[9] S.M. Jagatheesan1, V. Thiagarasu2," Development of Fuzzy based categorical Text Clustering Algorithm for Information Retrieval", January 2014, vol 2,issue1

[10] K. Nalini Dr. L. Jaba Sheela," Survey on Text Classification", IJIRAE, July 2014, Vol 1, Issue6.

[11] Roventa, E., Spircu, T." Averaging Procedures in Defuzzification Processes, Fuzzy Sets and Systems ", 2003,136, pp. 375-385.

[12] S.J.Lee and C. S. Ouyang. "A neuro-fuzzys system Modeling with self-constructing rule generation and Hybrid svd-based learning," IEEE Transaction on Fuzzy Systems, June, 2003, 11(3):341– 353.

[13] G. Salton and M. J. McGill. Introduction to Modern Retrieva,. McGraw-Hill Book Company, 1983.

[14] F. Sebastiani. "Machine learning in automated text Categorization," ACM Computing Surveys, 34(1):1 47 March 2002.

[15] L. X.Wang. A Couse in Fuzzy Systems and Control. Prentic-Hall International, Inc., 1997.

[16] Y. Yang and J. O. Pedersen. "A comparative Study On feature selection in text categorization.

[17] R. Bekkerman, R. El-Yaniv, N. Tishby, and Y. Winter ,"Distributional Word Clusters Versus Words for Text Categorization," J.Machine Learning Research, 2003 , vol. 3, pp.1183-1208.