International Journal of Pure and Applied Mathematics
Volume 118 No. 6 2018, 39-47
ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
Special Issue

ijpam.eu

Enhancing Accuracy in Clustering using Relevance Feature Selection

Senthil Selvi [1] and Dr. R.Parimala [2]

[1] *PG and Research Department of CS*
Periyar E.V.R. College,India
senthilselvikumar@yahoo.co.in
[2] *PG and Research Department of CS*
Periyar E.V.R. College,India
rajamohanparimala@gmail.com

Abstract

Nowadays most of the information is available in electronic format and a substantial portion of the information may contain text documents. Grouping of similar text documents is challenging because of the curse of dimensionality. The number of groups in text corpus is difficult to decide due to the implicit patterns and features. K-Means clustering algorithm group a set of text documents based on distance. This research work focus on grouping various text corpora such as news wires, email spam messages and medical abstracts using relevance features. Experimental results for the reduced features are reported with enhanced accuracy.

AMS Subject Classification: Data Mining

Key Words and Phrases: Relevance features, Inverse Document Frequency, Feature Selection

1 INTRODUCTION

Clustering has been studied widely in the database and statistics literature in the context of a wide variety of data mining tasks. The dimensionality of the text representation is very large, but the underlying data is sparse. In other words, the lexicon from which the documents are drawn may be of the order of 10^5[1]. The short sized documents also face the same problem. The text corpora collected undergoes pre-processing. The text corpus is converted to document term matrix whose rows are documents and columns are keywords, the elements are inverse

document frequency of all terms. In K-Means clustering various distance metrics are used to find clusters. The complexity of distance calculation depends on the dimensionality of document-term matrix. The distance between each data point and newly obtained cluster centers is recalculated. If no more data point to be recalculated the criterion for K-Means stops. Relevance feature selection reduces the dimensionality of text documents to be clustered. Feature selection is the process of selecting the most representative feature subset. In this paper document frequency threshold is used to select relevance features.

The paper is organized as follows: Section II outlines about Literature Review. Section III presents the Proposed Methodology. Section IV gives the detail about the Corpus used. Section V the paper outlines about used Environment and Libraries. Section VI discusses the Experimental Results and finally concludes.

2 LITERATURE REVIEW

The term for K-Means clustering algorithm was first developed by J.MacQueen[2] and then the idea was followed by J.A.Haritigan and M.A. Wong around 1975[3]. Staurt Lyold proposed the standard algorithm as a technique for pulse-code modulation in 1957 [4]. The K-Means algorithm was demonstrated as early as 1956 by Steinhaus [5]. Text mining applications include the automated clustering of documents to help organize a digital library [6, 7]. The extraction of key terms from text descriptions towards the development of biomedical ontology [8] and the identification of recurring topics from unstructured description of software customers technical questions [9]. Recently, Researchers addressed unsupervised feature selection in their study. Julia Handl and Joshua Knowles implemented unsupervised feature selection as a multi objective optimization problem in which a wrapper approach based on the Silhouette width showed better performance when compared to the DB-index and DB-Index/d_F [10]. Dimensionality reduction in learning tasks can be crucial for a number of reasons due to large feature sets, features may be redundant, noise or irrelevant and the number of terms is much larger than the number of documents. For the above reasons, feature selection is important in unsupervised data analysis [11].

Volker Roth and Tilman Lange proposed an algorithm with respect to Gaussian mixture model, which combines a clustering method with a Bayesian inference mechanism for automatically selecting relevant features [12] C.C. Aggarwal et.al., proposed Projected Clustering (ProClus) finds subsets of features using Manhattan distance defining for each cluster [13]. Jianbo Ye et. al., used D2-clustering (Discrete Distribution Clustering) for text corpus BBC Sports, BBC Abstract, 20 Newsgroups and Ohsumed's Medical abstract [14]. They implemented D2-clustering for entire subset of 20 Newsgroups and Ohsumed text collection with accuracy as in Table 1.

3 METHODOLOGY

The text corpus for this study is collected. The documents in corpus are in unstructured format transformed into vector format. The pre-processing of text documents

is done. This phase involves converting documents' to a plaintext document with white-spaces, numbers, punctuation and stop words removed. Finally stemming is performed. The document-term matrix created using Term Document – Inverse Document Frequency (TF-IDF). The vector is normalized in the interval [0, 1]. The relevant features are selected for clustering using Document Frequency Feature Selection (DFFS). DFFS selects the features whose frequency in the documents is greater than the threshold (T). The K-Means clustering is performed for the reduced DTM as

$$SSE = \sum_{k=1}^{k} \sum_{x_i \in C_k} \|(x_i - \mu_k)\|^2$$

$\boldsymbol{x}_i$is a data point in cluster C_i , $_i$ is the center for cluster $C_i$as the mean of all points in the cluster and ||.|| is the Euclidean distance.
The cluster accuracy is calculated as

$$\text{Accuracy} = \frac{\sum_{\text{i=1}}^{\text{K}} \text{Number of correctly classified cluster}}{\text{Total number of clusters}}$$

The proposed Relevance Unsupervised Feature Selection Algorithm (RUFSA) is depicted. The diagrammatic representation of the architecture is shown in figure 1.

RUFSA Algorithm

1. Collect text corpora
2. Perform preprocessing.
3. Create Document-Term-Matrix.
4. Normalize
5. Perform DFFS for various thresholds, T
6. Set minimum total-withiness as threshold T_1
7. Specify the number of clusters (K)
8. Perform K-Means Clustering for Reduced Features (RF).
9. Find total-withiness (Tw)

   If T1 > Tw then set T1<= Tw
10. Calculate accuracy(ACC) for the resultant cluster.

4 CORPUS USED

The corpora used in this work are different categories of text document collection.ata sets ranging from News wires, spam document to medical abstract documents.

The News Wire consists of BBC Sports, BBC News Abstract, 20 News Group ranging from Ng20-group1, Ng20-group2, Ng20-group, Ng20-group4, Reuters and C50. The spam documents consist of Enron1, Enron2, Enron3, Enron4, Enron5,

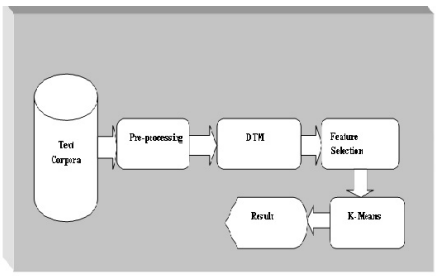

Figure 1: Relevance Unsupervised Feature Selection for Text Corpora

Enron6 and Lingspam documents. The medical Abstract the Ohsumed documents was considered for study.

The BBC Sports consists of documents from the BBC Sport website corresponding to sports news articles in five areas (athletics, cricket, football, rugby, tennis)[15]. The total number of documents is 737 with 5 class labels. The BBC News Abstract documents consists of documents from the BBC news website corresponding to news in five topical areas which are business, entertainment, politics, sport and tech[15]. The total number of documents is 2225 with 5 class labels. The 20 newsgroup corpus is organized into 20 different newsgroups, each corresponding to a different topic [16]. The documents are organized into four groups which are heterogeneous in nature. In Ng20-group1 contains documents of atheism, autos, base, crypt, electricity, in Ng20-group2 contains documents of forsale, graphics, hockey, ibmhardware, machinehardware, in Ng20-group3 contains documents of medical, motorcycles, os, religion, space, in Ng20-group4 contains documents of tp-guns, tp-mideast, tp-misc, tp-r and windows-x. All these groups account to 5000 documents with 5 class labels. A subset of Reuters 90 categories is taken [17]. This consists of 7316 documents with 15 classes and C50 documents is the subset of RCV1 [18]. This consists of 2500 documents with 15 class labels. The Enron1, Enron2, Enron3, Enron4, Enron5, Enron6 consists of 5172, 5857, 5512, 6000, 517 and 6000 documents respectively with 2 class labels. There are 6 legitimate users in Enron Email Documents. They are beck-s, farmer-d, kaminski-v, kitchen-l, lokay-m and williams-w and three spam users GP, BG and SH (SpamAssassin and Honeypot) [19]. Lingspam Corpus consists of 2894 documents with 10 class labels from part1-part10. These correspond to the 10 partitions of the corpus. Each one of the 10 subdirectories contains both spam and legitimate messages, one message in each file and finally the Medical abstract documents Ohsumed corpus which includes abstracts from the *MeSH* categories of the year 1991[20]. Some of the cases considered for study are Bacterial Infections and Mycoses, Virus Diseases, Parasitic Diseases, Neoplasm, Musculoskeletal Diseases. The total number of documents is 2092 with 5 class labels.

5 USED ENVIRONMENT AND LIBRARIES

R is a programming language is an environment for writing code used for statistical computing. R has many packages and includes R functions which are reusable. The package "tm"[21] is used to perform all manipulations regarding text. The

"broom" package converts statistical analysis objects from R into tidy data frames [22]. The"stats" package is used for the K-Means clustering process [23].

6 EXPERIMENTAL RESULTS AND DISCUSSION

RUFSA were applied to the text corpora mentioned in Section IV. The results of experiment showed the proposed accuracy is compared with existing K-Means algorithm accuracy. A sample of BBC Sports Corpus document is given in figure 2. Figure 3 represents a sample of results after preprocessing and the Document Term matrix created. A portion of DTM is shown in Figure 4. DFFS is applied for different threshold. A sample of DTM after DFFS is given in figure 5. K-Means clustering is performed and the results are tabulated using Algorthim 1. Figure 6 shows the K-means clustering and Figure 7 shows the accuracy calculated.

```
>trwrap(corp[[1]])
[1] "Claxton hunting first major medal"
[2] ""
[3] "British hurdler Sarah Claxton is confident she can win her first s
[4] "medal at next month's European Indoor Championships in Mads
[5] ""
[6] "The 25-year-old has already smashed the British record over 60
[7] "twice this season, setting a new mark of 7.96 seconds to win the
[8] "title. \"I am quite confident,\" said Claxton. \"But I take each rac
[9] "it comes. \"As long as I keep up my training but not do too muc
[10] "there is a chance of a medal.\" Claxton has won the national 60
[11] "title for the past three years but has struggled to translate her"
[12] "domestic success to the international stage. Now, the Scotland-
[13] "athlete owns the equal fifth-fastest time in the world this year. A
[14] "at last week's Birmingham Grand Prix, Claxton left European r
[15] "favourite Russian Irina Shevchenko trailing in sixth spot."
```

Figure 2: A sample of BBC Sports Corpus Document

```
[1] "claxton hunt first major medal british hurdler sarah claxton confid can"
[2] "win first major medal next month european indoor championship madrid"
[3] "yearold alreadi smash british record m hurdl twice season set new mark"
[4] "second win aaa titl quit confid said claxton take race come long keep"
[5] "train much think chanc medal claxton won nation m hurdl titl past three"
[6] "year struggl translat domest success intern stage now scotlandborn"
[7] "athlet own equal fifthfastest time world year last weekbirmingham"
[8] "grand prix claxton left european medal favourit russian irina"
[9] "shevchenko trail sixth spot
```

Figure 3: A sample of BBC Sports Corpus Documents after preprocessing

From the above table the total number of Initial terms is larger and hence using dimensionality reduction is performed and the reduced features are tabulated. It was found that handling large number of terms the complexity increases and hence reduced terms was observed and taken for accuracy and better results are shown than the existing accuracy result.

CONCLUSION

The proposed work used a filter feature selection DFFS decreases the time complexity of Clustering. The selected features for DFFS differ for various threshold values. In future a better method will be proposed to overcome the disadvantage and RUFSA algorthim applied for unlabelled corpus in future.

Docs \ Terms	England	first	game	last	play	player	said	two	will	win
103	9	5	0	0	4	0	0	5	1	0
181	11	8	0	0	3	0	0	5	1	0
243	1	1	0	3	0	3	4	0	14	1
249	0	0	0	1	0	1	0	0	12	0
252	1	1	0	3	0	3	4	0	14	1
432	1	2	4	3	1	0	0	1	0	8
578	1	14	18	0	4	1	2	5	0	14
69	0	1	0	3	0	0	3	4	3	4
693	0	5	0	6	5	3	2	3	4	2
708	0	10	19	0	3	3	0	7	1	5

Figure 4: A portion of DTM

Terms Docs	chelsea	club	cup	england	match	play
260	0.00000000	0.00000000	0.00000000	0	0.00000000	0.0000000
294	0.00000000	0.00000000	0.01925608	0	0.00000000	0.0000000
365	0.00000000	0.04491498	0.04186104	0	0.09241281	0.0000000
37	0.00000000	0.00000000	0.00000000	0	0.00000000	0.0000000
397	0.08382573	0.00000000	0.00000000	0	0.00000000	0.0000000
400	0.00000000	0.00000000	0.00000000	0	0.00000000	0.0000000
422	0.00000000	0.00000000	0.00000000	0	0.00000000	0.0160695
531	0.00000000	0.00000000	0.00000000	0	0.00000000	0.0000000
674	0.00000000	0.00000000	0.04279129	0	0.03148881	0.0349968
80	0.00000000	0.00000000	0.00000000	0	0.00000000	0.0000000

Figure 5: A portion of DTM after DFFS

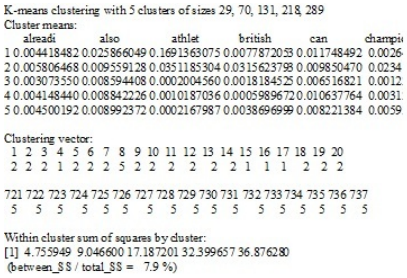

Figure 6: Results of k-Means Clustering

	i	acc	totss	tot.withinss	betweenss	iter
[1,]	1	1.0000000	1.0000	1.0000	1.000000	1
[2,]	1	0.7910448	108.8281	100.4770	8.351081	3
[3,]	2	0.7720488	108.8281	100.2616	8.566513	4
[4,]	3	0.8398915	108.8281	100.0892	8.738872	5
[5,]	4	0.7829037	108.8281	100.2608	8.567310	4
[6,]	5	0.8398915	108.8281	100.0892	8.738872	5
[7,]	6	0.8398915	108.8281	100.0892	8.738872	4
[8,]	7	0.8385346	108.8281	100.0937	8.734401	5
[9,]	8	0.7598372	108.8281	100.2657	8.562422	3
[10,]	9	0.9253731	108.8281	100.2686	8.559555	5

Figure 7: Accuracy Calculation for BBC Sports

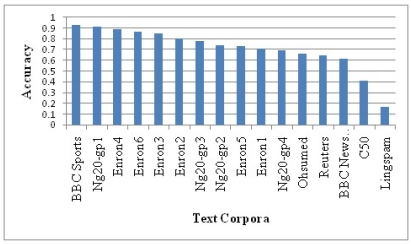

Figure 8: Clustering accuracy

Table 1: Results of Clustering for RUFSA Algorithm

Type	Text Corpora	Total No. of Documents	Features	K	RF	ACC %	Existing ACC %
News wires	BBC Sports	737	13057	5	403	**92.5**	81.2
	BBC News Abstract	2225	29069	5	1957	62.5	75.9
	Ng20-group1	5000	57467	5	1468	**92.0**	53.4
	Ng20-group2	5000	56459	5	1138	74.6	-
	Ng20-group3	5000	10020	5	1511	81.1	-
	Ng20-group4	5000	75731	5	2274	69.0	-
	Reuters	7316	23529	15	941	**64.5**	53.4
	c50	2500	28785	50	2055	39.6	-
Spam	Enron1	5172	49766	2	1642	70.9	-
	Enron2	5857	39542	2	1596	77.6	-
	Enron3	5512	53005	2	1564	84.8	-
	Enron4	6000	68654	2	2041	87.9	-
	Enron5	5175	41496	2	2115	75.5	-
	Enron6	6000	70413	2	1153	76.7	-
	Lingspam	2894	57057	10	2335	16.2	-
Medical Abstract	Ohsumed (5 cases)	2894	42266	23	900	**64.7**	26.0

References

[1] Mehdi Allahyari, Seyedamin Pouriyeh,Mehdi Assefi,Saied Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, Krys Kochut, A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques, *In Proceedings of KDD Bigdas.*, Halifax, Canada, August (2017).

[2] J.B.MacQueen. Some Methods for Classification and Analysis of Multivariate Observations, *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, (1967),281-297.

[3] Hartigan, J.A. *Clustering Algorithms (Porbability & Mathematical Statistics)* ,John Wiley & Sons Inc, (1975).

[4] *Reprinted in: IEEE Trans. Information Theory, IT-28* (Murray Hill, NJ.,Bell Telephone Labs Memorandum)**2** (1982): 129-137.

[5] Steinhaus, H., Sur la di vision des corps materiels en parties. Bulletin del'Academie Polonaise des ciences, **IV**, 12 ,(1956), 801-804.

[6] Efron, M., Marchionini, G., Elsas, J., & Zhang, J., Machine learning for information architecture in a large governmental website, *Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries*, (2004), 151-159.

[7] Krowne, A., & Halbert, M. ,An initial evaluation of automated organization for digital library browsing., *Proceedings of the 5th ACM/IEEE-CS joint conference on digital libraries*, (2005), 246-255.

[8] Inniss, T., Light, M., Thomas, G., Lee, J.R., Grassi, M.A., & Williams, A.B.,Towards applying text mining and natural language processing for biomedical ontology acquisition, *Proceedings of the 1st international workshop on Text mining in bioinformatics.* (2006), 14-17.

[9] S.Brindha, Dr.K.Praba,Dr.S.Sukumaran, The comparison of term based methods using text mining, *International Journal of Computer Science and Mobile Computing*, **5**(9), (2016): 112-116.

[10] Julia Handl, Joshua Knowles,Feature Subset selection in Unsupervised Learning via Multiobjective Optimization, *International Journal of Computational Intelligence Research*,**2**(3),217-238.

[11] Hartigan, J.A. *Clustering Algorithms (Porbability & Mathematical Statistics)*, John Wiley & Sons Inc, (1975)

[12] Volker Roth, Tilman Lange,Feature Selection in Clustering Problems, *Institute of Computational Science*

[13] C.C. Aggarwal C. Procopiuc, J. L.Wolf, P. S.Yu and J. S. Park,Fast algorithms for projected clustering, *In Proceedings of ACM SIGMOD Conference on Management of Data*,(1999), 61-72.

[14] Jianbo Ye, Yanran Liy, ZhaohuiWuz, James Z. Wang, Wenjie Liy and Jia Li,Determining gains acquired from word embedding Quantitively using Discrete Distribution clustering

[15] P.Cunningham, D. Greene, Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering, *Proc. ICML.*, (2006).

[16] `http://www.ai.mit.edu/people/jrennie/20Newsgroups/`

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

[18] ZhiLiu, C50 Dataset creator and donator, Hubei Wuhan,China: National Engineering research center for E- learning.

[19] `http://www2.aueb.gr/users/ion/data/enron-spam/`

[20] `ftp://medir.ohsu.edu/pub/ohsumed`

[21] Ingo Feinerer, Kurt Hornik, and David Meyer (2008), *Text Mining Infrastructure in R.* Journal of Statistical Software ,1-54, http://www.jstatsoft.org/v25/i05/.,

[22] Robinson, David, broom: Convert Statistical Analysis Objects into Tidy Data Frames. R package version0.4.2, `https://CRAN_R-project.org/package=broom`, (2017).

[23] R Core Team , R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, `https://www.R-project.org`,(2017).