ISSN (ONLINE): 2395-695X
ISSN (PRINT): 2395-695X

International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST)
Volume 2, Special Issue 19, October 2016

Estimating Number of Clusters in Sonnet Stanza Cluster

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

[1]Research Scholar and Assistant Professor, [2]Research Adviser and Assistant Professor
PG and Research Department of Computer Science
Periyar E.V.R. College (Autonomous),
Trichy-23, India

senthilselvikumar@yahoo.co.in[1], rajamohanparimala@gmail.com[2]

Abstract-**This paper considers clustering stanza of Shakespeare Sonnets (SSC) and its optimal size. K-Means clustering is a very effective clustering technique well known for its observed speed and its simplicity. The clusters are formed and distance are calculated using the Euclidean distance. In this work, Sonnet Stanza Clustering is evaluated and validated. This work concentrates on the entropy as performance measure. The optimal number of clusters is evaluated using elbow method. This method is used to evaluate optimal cluster size. The paper concludes with the findings of the results of the proposed algorithm for different feature subsets.**

Keywords— K-Means Clustering, Feature selection, elbow method, Entropy

I. INTRODUCTION

Clustering has a long and rich history in the different field of science. One of the most popular and simplest clustering algorithms is the K-Means. The clustering algorithm was proposed over 50 years ago and since then thousands of clustering algorithms have been published. Hartigan introduced the K-Means algorithm where the algorithm proposed repeatedly picks a point and determines its optimal cluster assignment [6]. It starts with a random initial point and keeps reassigning the patterns to clusters based on the similarity between the pattern and the cluster centers until a convergence criterion is met [8]. Hieatt found the frequency occurrence of rarely used words and also found the links between different groups of the Shakespearean sonnet [3].

Shakespearean sonnet has been the enchantment that has spellbound readers and critics of the Sonnets since their first publication in 1609. The Sonnets are slippery, ambiguous and challenging. Recently, Researchers has begun to find hidden patterns in Sonnets using computational methods. The goal of Information Extraction transforms machine text into structured format, a machine readable structure [2]. Today's organizations are moving towards the tremendous growth of unstructured data. The automated analysis performs a data mining technique which aims to discover patterns, and describe the relations among them. Throughout this work, a corpus of sonnets written by Shakespeare is considered for clustering. The Sonnet Corpus is 515.6 Kilo Bytes of data and includes 154 Stanza each with 14 lines, which equates to a total of 2614 lines of poetry. The Sonnet Stanza Clustering (SSC) is the process of grouping a set of lines in such a way that lines in the same group are more similar to each other than to those in other groups. The performance of algorithm is evaluated on Sonnet Corpus.

The feature set is created for the Sonnet Stanza and hence Sonnet Stanza Term Matrix is built (SSTM). The number of terms found in SSTM is large. Preprocessing is performed in order to remove unwanted terms before the clustering process is done. This reduces the corpus size greatly which is considered for clustering process. Accuracy and efficiency of clustering algorithms depends greatly on the input data. Removing unimportant features from the dataset can help us to form better clusters in lesser time. Therefore, it is essential to have a proper feature selection in order to reduce the sparseness. Sparse term removal at different threshold is proposed for feature selection. In addition, the proposed work randomly selects a feature subset. K-Means clustering is performed for the selected terms. K-Means algorithm is employed to generate different clusters for different runs.

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 Dataset used. Section V the paper outlines about used Environment and Libraries. Section VI discusses the experimental results and finally concludes.

II. LITERATURE REVIEW

K-Means algorithm is one of the partitioned clustering methods [8]. In 1967 Mac Queen developed the simplest and the easiest clustering algorithm – the K-Means clustering algorithm. Bhoomi proposed that before the K-Means converges, the centroids are computed and all points are assigned to their nearest centroids [2]. Cluster validity refers to formal procedures that evaluate the results of cluster analysis in a quantitative and objective fashion [20]. Cluster validity indices can be defined based on three different criteria: internal, relative, and external [20]. There has been significant work on characterizing rhyme, [4] poetry generation, case based reasoning to induce the best poetic structure, rhyme identification [10] and also visualization of poetry[5] .Okafor stated that Entropy is a good measure for determining the

quality of clustering [11] and problem of dimension reduction. Simonton analyzed the 154 sonnets written by William Shakespeare. Each sonnet was partitioned into four consecutive units (three quatrains and a couplet), and then a computer tracked down how the number of words, different words, unique words, primary process descriptions and secondary process descriptions is changed within each sonnet unit[21].

III. METHODOLOGY

Data representation is one of the most important factors that influence the performance of the clustering algorithm. The data available in Sonnet Corpus is homogeneous and unstructured. The pre-processing step converts unstructured text into structured form before preparation to clustering. They are first preprocessed for elimination of stop words, special characters, punctuation etc., after that stemming, Sparse Term Removal and tokenization steps are applied and TF-IDF score is calculated for all the words. A Sonnet Stanza Term Matrix (SSTM) is created which is fed into the K-means algorithm for clustering. Given M stanza $S=\{s_1, s_2, s_3,\ldots s_m\}$ containing all together V different words $W=\{w_1, w_2, w_3, \ldots w_v\}$ The term frequency of the term W_j in Stanza S_i is t_{ij}. If the representation (choice of terms) is good, the clusters are likely to be compact and isolated. The objective of K-Means is to minimize the total sum of the squared distance of every point to its corresponding cluster centroid. The SSE is defined as the sum of the squared distance between each member of the cluster and its centroid c_i . Mathematically:

$$SSE = \sum_{i=1}^{K} \sum_{x \in c_i} (x - c_i)^2 \qquad (1)$$

The K-means algorithm takes the centroid of a cluster as mean value of the points within cluster. It randomly selects K, number of objects in dataset, each of which initially represents a cluster mean. For each K of remaining objects, an object is assigned to cluster to which it is most similar, based on Euclidean distance. The algorithm iteratively improves within cluster variations. Automatically determining the number of clusters has been one of the most difficult problems in data clustering. Usually, clustering algorithms are run with different values of K; the best value of K is then chosen based on some criteria. The algorithm Sonnet Stanza Clustering (SSC) is given in Algorithm 1 and finding the optimal number of clusters is given in Algorithm 2.

Algorithm 1:

Step 1: Create a corpus from Shakespeare's Sonnet.
Step 2: Performing the Pre-processing on the Corpus.
- a. Transform characters to lower case.
- b. Converting to Plain Text Term
- c. Remove punctuation marks.
- d. Remove digits from the stanza.
- e. Remove extra whitespaces from the documents.
- f. Create the Sonnet Stanza Term Matrix (SSTM)
- g. Extract features using Sparse Term Removal

Step 3: Apply K-means algorithm and validate using entropy performance measure
Step 4: Display the results
Step 5: Repeat Steps 3 and 4 for K=2 to 5
Step 6: Estimate the Optimum number of Clusters using elbow method.

Algorithm 2:

1. Start with two clusters (K=2)
2. Perform K-means clustering on the SSTM
3. Generate K clusters for K=3, 4 etc.,.
4. Compute the sum of squared error for each cluster using equation (1). Store this value as SS_k.
5. Find differences of adjacent elements in vector SS_k. if all difference $(SS_K)>0$ is false, the optimum cluster size of K is found..
6. The value (K-1) is the optimal number of clusters.

A. Performance Measure

The proposed approach uses entropy as performance measure. The lower entropy means better clustering. Let *CS* be a clustering solution. For each cluster, the class distribution of the data is calculated first, i.e., for cluster *j,* the probability that a member of cluster *j* belongs to class *i* is computed (p_{ij}). The entropy of each cluster *j* is calculated using the formula

$$E_j = -\sum_{j=1}^{L} p_{ij} \log(p_{ij}) \qquad (2)$$

where the sum is taken over all L classes. The total entropy for a set of clusters is calculated as the sum of the entropies of each cluster weighted by the size of each cluster

$$E_{CS} = \sum_{j=1}^{K} \frac{K_j E_j}{M} \qquad (3)$$

where K_j is the size of cluster *j*, K is the number of clusters and M is the total number of data points [12].

IV. DATASET USED

A collection of "Sonnets of Shakespeare", a Sonnet Corpus containing collection of poems from the Shakespearean era, were collected for clustering at Gutenburg Website Shakespeare's copyright 1990-1993 by World Library, Inc., and is provided by project Gutenberg E-text of Illinois Benedictine College [14].

V. USED ENVIRONMENT AND LIBRARIES

R is a programming language and software environment used for statistical computing [15]. R Language was used for implementing the results. The package "tm" is a framework used for text mining applications [13] , package "cluster" is used for cluster analysis[17] and the package "fpc" portrays various methods to be used for cluster analysis and cluster validation[14]. The above packages were used to validate the result.

VI. EXPERIMENTAL RESULTS

The Sonnet Stanza Clustering is a high dimensional data, a standard benchmark dataset the Sonnet Corpus was taken for study. Table.1 presents the preprocessing results of the Sonnet Corpus with the initial Sonnet Corpus Size being 154 documents of 4232 terms and after preprocessing the size was reduced to 154 documents of 3038 terms.

K-Means Clustering given in Algorithm 1 and Algorithm 2 are implemented. The results are shown in Table 2. From Table 2, it is observed that the entropy value for K=2 with 344 features is low. This result also confirms with the results obtained in elbow method.

TABLE 1. PREPROCESSING METHODS AND CORPUS SIZE

Preprocessing Methods	Size
Without preprocessing of Sonnet Corpus	515.6 KB
Tokenization + Punctuation Removal	180.3 KB
Tokenization + Punctuation Removal + Stopword Removal	162.6 KB
Tokenization + Stopword Removal + Stemming + White Space Removal	151.2 KB

Using Algorithm 2 the graph is drawn between number of clusters and SSE. Fig 1. shows that when K=2 optimum number of clusters is obtained.

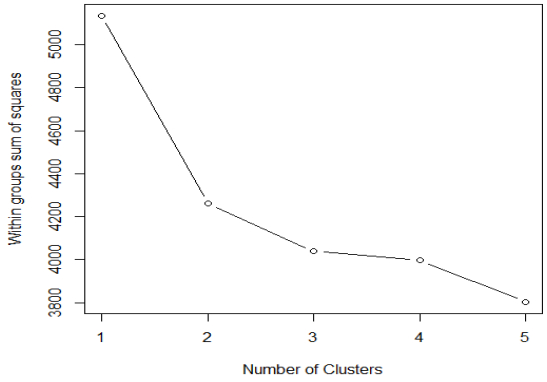

Fig 1: Elbow Method

TABLE 2: RESULTS OF K-MEANS CLUSTERING OF SSTM

Sonnet Sparse Term Threshold % (SSTP)	No. of Features (NF)	Corpus Size in KB (CSB)	K-Means Clustering for SSTM-Reduced Features(RF) and Entropy(E)							
			K=2		K=3		K=4		K=5	
			RF	E	RF	E	RF	E	RF	E
98.5	715	112.5	344	0.06937	339	0.40144	349	0.67066	347	0.92515
98.7	715	112.5	344	0.06937	339	0.40144	349	0.67066	347	0.92515
98.9	1211	146.2	631	0.12045	575	0.23342	625	0.73500	605	0.93416
99.1	1211	146.2	631	0.12045	575	0.23342	625	0.73500	605	0.93416
99.3	1211	146.2	631	0.12045	575	0.23342	625	0.73500	605	0.93416
99.5	3038	243.8	1527	0.12045	1567	0.41511	1545	0.61597	1562	0.90558
99.7	3038	243.8	1527	0.12045	1567	0.41511	1545	0.61597	1562	0.90558
99.9	3038	243.8	1527	0.12045	1567	0.41511	1545	0.61597	1562	0.90558

VII CONCLUSION

The optimum clustering is found using elbow method and the Sonnet Stanza Clustering studied. The issue of how to deal with the K initialization, and distance metric still remains open and in future this can be studied.

ACKNOWLEDGEMENT

I like to accord my gratitude to the R Community for providing an open source tool and providing the necessary packages for the successful implementation of this work.

REFERENCES

[1] Aljumily Refat, "Hierarchical and Non-Hierarchical Linear and Non-Linear Clusterng Methods to "Shakespeare Authorship Question"", Social Sciences, vol. 4, **(2015)**, pp. 758-799.

[2] M. Bhoomi Bangoria, "Enhanced K-Means Clustering Algorithm to ReduceTime Complexity for Numeric Values", International Journal of Computer Science and Information Technologies, vol. 5, no. 1, **(2014)**, pp. 876-879.

[3] S. Brian Everitt, Sabine Landau, Morven Lesse and Daniel Stahl, "Cluster Analysis", 5th Edition, Wiley, **(2011)** January.

[4] J. ByrdRoy and M.S.Chodorow, "Using an online dictionary to find rhyming words and pronunciations for unknown words", Proceedings of the 23rd Annual Meeting of ACL, **(1987)**, pp. 277-283.

[5] Alfie abdul Rahman,Julie Lein, Katharine Coles, Eeamonn Magurie, Miriah Meyer, Martin Wynne, Cris Johnson, Anne E. Trefethen and Min Chen, "Rule based Visual Mappings-with a case study on Poetry Visualization", In Compute Graphics Forum, **(2013)**, 32.

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

[7] Hieatt, A. Kent, W. Hieatt Charles and Prescott Anne Lake, "When Did Shakespeare Write "Sonnets 1609?"", Studies Philology 88.1", **(1991)**, pp. 69-109.

[8] 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)**, pp. 281-297.

[9] S. Julie, "A survey of the literature of clustev analysis",Comput. J., vol. 25, no. 1,**(1982)**, pp. 130-134.

[10] Manish Chaturvedi, Gerald Gannod, Laura Mandell, Helen Armstrong and Eric Hodgson, "Rhyme's Challenge: Hip Hop, Poetry, and Contemporary Rhyming Culture", Oxford University Press, Literary Criticism, **(2012)**.

[11] Okafor and Anthony, " Entropy based techniques with applications in data Mining", Florida, University of Florida, **(2005)**.

[12] E..Shannon Claude, "A mathematical theory of communication", Bell System Technical Journal, vol. 27, **(1948)** July and October, pp. 379-423 and 623-655.

[13] Ingo Feinerer, Kurt Hornik, and David Meyer. Text Mining Infrastructure in R, Journal of Statistical Software, vol. 25, no. 5, pp 1-54, **(2008)**, http://www.jstatsoft.org/v25/i05/.

[14] http://www.gutenburg.org/cache/epub/100/pg100.txt.

[15] http://www.R-project.org/.

[16] Christian Hennig,fpc: Flexible procedures for clustering. R package version 2.1-9. (2014), http://CRAN.R-project.org/package=fpc

[17] Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, Kcluster: Cluster Analysis Basics and Extensions. R package version 1.15.3..(2014).

[18] Robert L. Thorndike , "Who Belongs in the Family?", Psychometrika.,18,4,pp:267–276., doi:10.1007/BF0228926

[19] Colin Burrow, "William Shakespeare: The Complete Sonnets and Poems", Oxford University Press, **(2002)**.

[20] Anil K.Jain,R. Dubes,"Algorithms for data clustering",Prentice Hall,Englewood Cliffs,New Jeersey 07632,1988

[21] Simonton, Dean Keith 1990. Lexical choices and aesthetic success: A computer content analysis of 154 Shakespeare sonnets. Computers and the Humanities, Vol. 24, No. 4 ,Aug., 1990, pp. 251-264.