I J C T A, 9(15), 2016, pp. 7619-7626

Enhancement of Accuracy in K-means Clustering

Sharika Unnikrishnan*, Sreelakshmi S. and Deepa G.*****

ABSTRACT

Text Mining is remarkably a new and stimulating research area in this modern world of technological era. Text clustering is a text mining technique. It groups a set of objects in such a way that, objects in the same group (called a cluster) are more similar in one way or the other to each other than to those in other groups or clusters. There are several techniques to accomplish text clustering. Initially we did clustering using the K-means algorithm. It was found that text clustering requires text data that must be converted into numerical to get more accurate results.

The paper concentrates on K-means algorithm. It's the fastest algorithm and can deal with large data sets efficiently. However, k-means encounter certain problems while dealing with text data. The paper focuses on eliminating this drawback by directly converting the text data into a numeric value which results in more defined clusters and accurate running time.

Keywords: Text Mining, Text Clustering, K-Means Algorithm, WEKA

1. INTRODUCTION

Achievement of retrieving useful information from a large collection of unstructured data is challenging. But the introduction of text mining made access to these information possible. Text clustering is a text mining technique which determines the natural clusters of objects or texts and each of this clusters have text or objects of similar behavior. This has turn out to be an integral approach in many areas such as web, social networking sites, online review sites and other digital collections [1].Text clustering occurs in many kinds of text documents of various applications in large data sets with different attributes.

Different types of text clustering algorithms or techniques exist [2]. None was completely efficient for clustering the text documents as all have their own strength and weakness.

K-means algorithm is popular for cluster analysis in data mining. According to Joaquín Pérez Ortega [3] the definition of "means" limits the application only to numerical variables.

1.1. Motivation

When there is a data set with metadata having both text and numerical values, it is not possible to perform clustering using k-means algorithm. This is because k-means either rejects text field or will recommend to use a converter to change the text field into numeric data while running on a WEKA tool.

1.2. Objective

The paper proposes an algorithm which works as a support to K-means algorithm. With the help of this algorithm one can convert text data to integer. This data set when brought into WEKA gives more accurate clusters.

Department of CS & IT, Amrita School of Arts & Sciences, Kochi, Amrita Vishwa Vidyapeetham

* MCA Student, *Email: sharika.unnikrishnan4@gmail.com*

** MCA Student, *Email: lakshmiarvind88@gmail.com*

*** Assistant Professor, *Email: deepsgopi@gmail.com*

2. OVERVIEW OF TEXT MINING

Text mining attempts to find useful and previously unknown information from vast amount of data. It involves methods such as identification or extraction of representative features which is done as different phases and to find complex patterns[12]. Though text mining is used in many fields, some include E-discovery, National Security/Intelligence, search/information access.

Text mining consists of data analysis of texts as illustratred in Figure 1. Firstly text mining tool will do data analysis using a collection of texts. Through this phase, application of various algorithms takes place along with many sub process such as parsing, pattern recognition, syntactic and semantic analysis, clustering, and tokenization[13]. Subsequently evaluation of results and emergence of new data occurs.

2.1. Issues in Text Mining

The field of text mining has got high recognition due to large amount of text data. Text mining has many concerns. The major challenging issue in text mining arises due to the complexity of natural language [10]. One phrase or sentence can be interpreted in many ways. Thus, affecting the ambiguity and structuring of documents. In order to get rid of noisy data, before data processing we need to specify correct spellings, acronyms and abbreviations. Furthermore, text representation and structuring of documents is another major problem. To overcome these hindrances there are several Data/Text mining algorithms. Some of the familiar areas in data mining algorithms include classification, clustering, association analysis, statistical analysis.

3. CLUSTERING ALGORITHM

Clustering is the process of making a group of intangible objects into classes of similar objects. A cluster can be treated as one group. While doing cluster analysis, initially divide the set of data into groups based on data similarity and then allot labels to those groups. The main advantage of clustering over classification is that, it is flexible to changes and helps single out useful features that differentiate groups. Clustering is mainly used in fields like image processing, market search, WWW, sequence analysis and many more. The objective of clustering is to find the inherent grouping in a set of unlabelled data. Some of the clustering algorithms used are Hierarchical, Density Based and K-means. Here, K-means is found to be the most popular algorithm of text clustering [11]. Hence, this paper focuses on this prominent algorithm.

4. K-MEANS CLUSTERING

K-means Clustering algorithm is simple, best and efficient unsupervised algorithm for clustering large data sets. This algorithm was first developed by MacQueen in 1967. The aim of this algorithm is to separate or divide a set of objects into K clusters, where the count of K is predefined based on their features. So this algorithm is the most widely used partition algorithm. The idea is to define K centroids, one for each cluster. The centroid of a cluster is configured in such a way that it is closely related to all objects in that cluster. The closely relation can be measured using Euclidean distance.

Figure 1: Text Mining Process

Here explains how K-means works:

1. Select K count of clusters to be achieved.
2. Select K objects arbitrarily as the initial cluster centres.
3. Repeat
4. Assign each object to the closely related clusters.
5. Determine new clusters, i.e work out the mean value of the objects for each cluster.
6. Till no change.

4.1. Drawbacks of K-Means

K-means is a powerful clustering algorithm. Clustering producesnumber of disjoint non-hierarchical sets. It can be easily implemented and are well suited for generating clusters, even with large data set. With all these functionalities, the method still experiences several limitations as follows:

- If there are fixed number of clusters, it is difficult to predict the K value.
- It does not work with non-globular clusters and with clusters having different size and different density
- Initial partitioning is important as it can affect the final partitioning. So different initial partitioning might results in different final partitioning.
- Above all, the major difficulty encountered by the method is its limitation to numeric values; it defines only integers [3].

This paper presents an algorithm to make K-means more efficient by converting unused text field in the data set to integers.

5. WEKA

WEKA (Waikato Environment for Knowledge Analysis)is an open- source machine learning and data mining software [4].For normal users, it gives an interface for simple data analysis whereas for the researchers, a field to work on the concept of data mining. For example, if they want to know the performance of their own algorithm then, they can compare it with the existing algorithms using a WEKA tool [5].The main drawback of WEKA is the inability to perform multi-relational process and the lack of a merge tool for interconnected tables. The package can be downloaded from the website of Waikato University in New Zealand (http://www.cs.waikato.ac.nz/ml/weka/), the latest version is 3.7.

6. RELATED WORKS

K means is the most prominent and distinctive text clustering algorithm. In spite of that, it holds certain drawback and weaknesses. The paper [6] describes the depth view and discusses the appropriateness of the K-means algorithms, and also provides direction to text mining researchers, concerning about the selection of K Means algorithm for text clustering. It gives the overall view about the K Means algorithm and disadvantages associated with it.

To depose the limitations of k-means algorithm, three different modified k-means algorithm was introduced [7]. It makes the algorithm more effective, accurateincrease speed of clustering and reduces time complexity in order to get an optimal solution.

The researchers have showed a model that gives a way to extract features from products reviews [2]. They concluded that if there are large numbers of review for a product, this model inform future buyers what aspects of the product are good and bad according to the previous buyers.

In order to mine feature-opinions from product reviews, researchers have proposed many techniques using natural language processing and data mining methods [8]. Since customers shop and express their opinions online ,these techniques are important since it provides a feature- based summary of customer reviews of products that is sold online.

A multi-knowledge based approach is proposed for movie review mining and summarization[9]. The features and opinions in movie review domain are recorded using a key-word list. Feature-opinion pairs are mined using some grammatical rules and the keyword list. The experimental results automatically generate a feature-class based summary for random chosen online movie reviews.

7. METHODOLOGY

7.1. Experimental Results

To understand the efficiency of K-means in more practical situation, we performed a number of experiments on different data sets. Each test involves the use of K-means algorithm on data sets (from reviews) using WEKA. These data sets include both texts and integer values. First, we have performed clustering only on attributes containing integer values. The Figure 2 shows the WEKA result of this clustering.

As mentioned before k- means fail to accept the text fields while dealing with both text and integer value, the WEKA recommends the use of a converter. This is illustrated in Figure 3.

It is not always convenient to use a converter. A new method is introduced to support K-means algorithm to perform clustering that converts text field in data set.

8. SUPPORTIVE ALGORITHM FOR K-MEANS CLUSTERING

From all of the above experiments, it is clear that when data set contains both text and numeric values, K-means do not support clustering on text data. The proposed algorithm works to make the unused text data

```
=== Model and evaluation on training set ===

kMeans
======

Number of iterations: 9
Within cluster sum of squared errors: 89.54824091765983
Missing values globally replaced with mean/mode

Cluster centroids:
                                  Cluster#
Attribute       Full Data            0            1
                    (999)        (194)        (805)
===================================================
3.1                4.1394       2.7358       4.4777
4.3                 4.333       3.1807       4.6107
4.4                4.0974        2.579       4.4634
3.4                3.9636       2.5133       4.3131

Time taken to build model (full training data) : 0.05 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       194 ( 19%)
1       805 ( 81%)
```

Figure 2: Clustering of integer data sets.

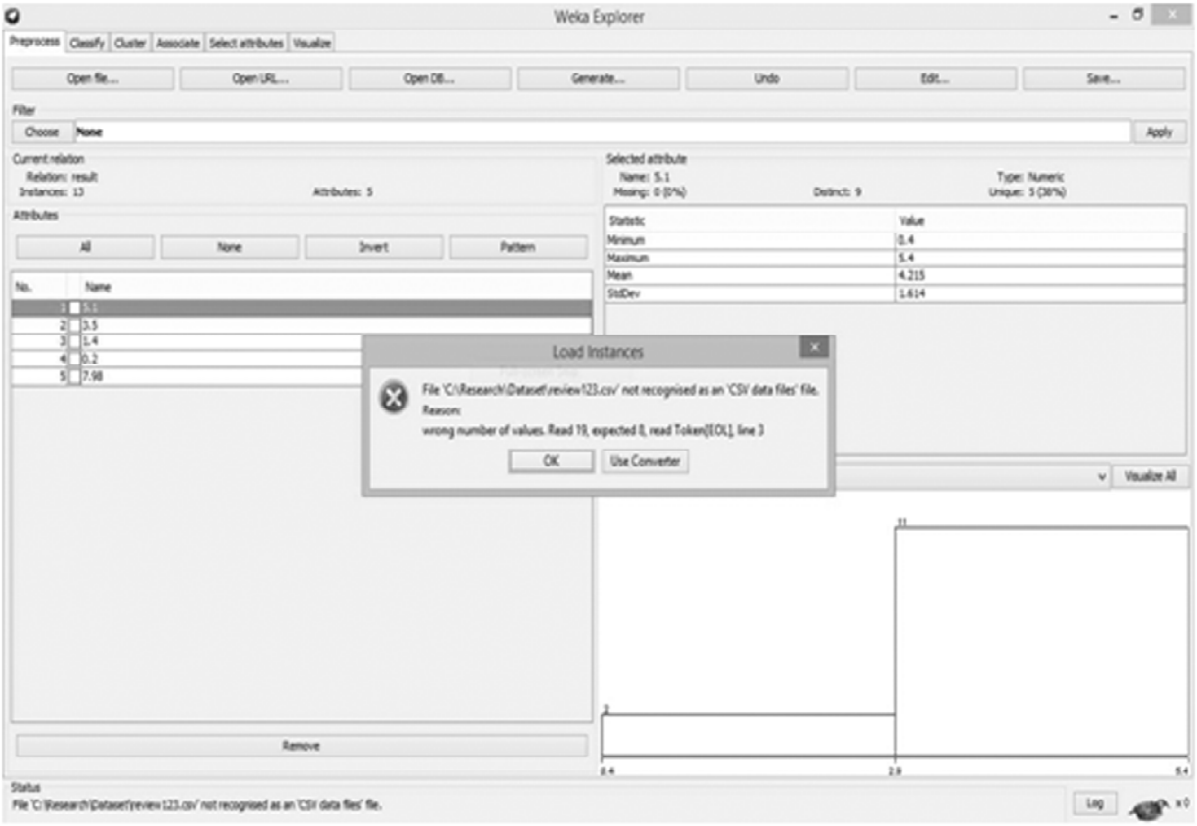

Figure 3: Attempting to load the data set which includes review (text) field

to useable integer data. The text data is converted into integer data based on the text content and the way the ordinary users expresses their reviews.

For making our experiment more stable, we consider the review dataset which contains many attributes. The last attribute in the dataset discussed is a long description about the personal review of a customer. For developing the proposed system, we develop three trained dataset namely, positive words dataset, negative words dataset and average words dataset. The following are the steps for the proposed algorithm:

1. Initially, take the last attribute of the data set, the customer reviews. Then replace any white gaps and special characters with spaces.
2. Split each review into words and store in an array.
3. Initialize positive count, negative count and average count as zero.
4. Take each word in sequence from the array and check their dependency by comparing it with previous words within the same sentence to determine whether it belongs to positive, negative or average.
5. Consequently increment the value of positive count, negative count or average count.
6. Continue step 4 and 5, until the end of current review.
7. Check which count is larger (positive count, negative count or average count).
8. Take the larger count and convert it into corresponding value based on their probability ratio.
9. Replace the last attribute in the dataset (i.e., the customer review) with the new numerical value.
10. Continue steps from 4 to 9 to convert the text into numerical value until the last review in the dataset.

The Figure 4 shows the flowchart of proposed algorithm.

Using the above algorithm, the text field get converted into numeric values. We use resultant data set to implement K-means algorithm in WEKA tool. Figure 5 shows the clustering results using proposed supportive algorithm.

Figure 6 shows the output (graphs) of k-means with proposed algorithm.

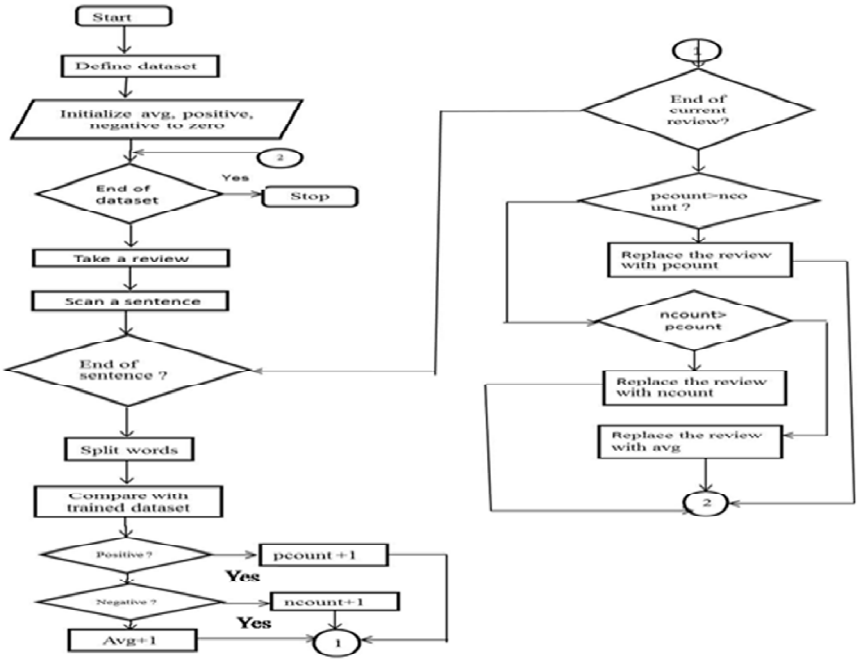

Figure 4: Flowchart of proposed algorithm

```
=== Model and evaluation on training set ===
kMeans
======
Number of iterations: 12
Within cluster sum of squared errors: 185.58470255787296
Missing values globally replaced with mean/mode

Cluster centroids:
                            Cluster#
Attribute     Full Data           0           1
                 (999)       (304)       (695)
=============================================
3.1                4.1394      3.6057      4.3729
4.3                 4.333       3.869       4.536
4.4                4.0974      3.5485      4.3376
3.4                3.9636       3.357       4.229
8.71               7.0167      3.5229      8.5449

Time taken to build model (full training data) : 0.03 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       304 ( 30%)
1       695 ( 70%)
```

Figure 5: Clustering using proposed algorithm

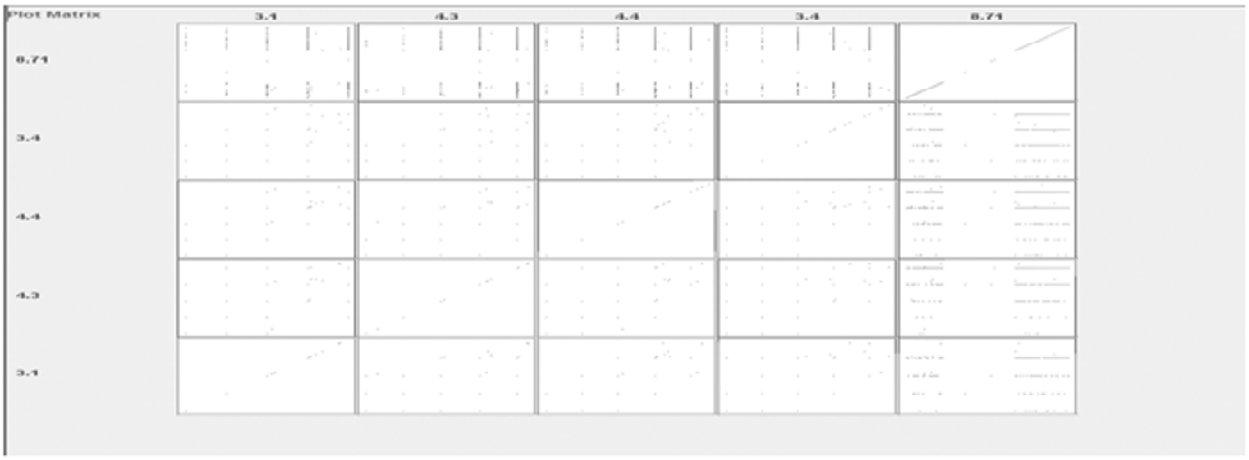

Figure 6: Output using k-means with proposed algorithm

Table 1
Comparison of Different Experimental Results

Name of Algorithm	*Number of Iterations*	*Time Taken*	*Clustered Instances*
K-means Clustering Algorithm	9	0.05	0 194(19%)1 1 805(81%)
K-means Clustering after using proposed algorithm	12	0.03	0 304(30%) 1 695(70%)

The below Table I shows the comparative results with and without using proposed algorithm.

These results represent overall accuracy rather than its conventional accuracy. This approach shows how k-means works to produce correct clusters by including texts field in the data set.

9. CONCLUSION

K-means is a well- known and best clustering algorithm for text mining. Despite of that, it resists all the text field character that comes across. WEKA being a prominent data analysis tool does support text fields. However WEKA doesn't implement K-means algorithm when the dataset contains texts data. To overcome this restriction an algorithm was implemented that converts texts data to integers to support k-means on a WEKA tool resulting in a more efficient and accurate clusters.

REFERENCES

[1] Charu C. Aggarwal,Yuchen Zhao, Philip S. Yu , "*On Text Clustering with Side Information*" IEEE.org

[2] Jingye Wang, HengRen,"*Feature-based Customer Review Mining*" On the Move to Meaningful Internet Systems: OTM 2014 Workshops.

[3] Joaquín Pérez Ortega1, Ma. Del Rocío Boone Rojas,María J. SomodevillaGarcía, "*Research issues on K-means Algorithm: An ExperimentalTrial Using Matlab*".

[4] Yinghua Lu, Tinghuai Ma,Changhong Yin, Xiaoyu Xie, Wei Tian,ShuiMingZhong "*Implementation of the Fuzzy C-Means Clustering Algorithm in Meteorological Data*" International Journal of Database Theory and Application Vol.6, No.6 (2013), pp.1-18.

[5] Y. Shen, J. Liu and J. Shen, Editors. IEEE Computer Society Washington, DC, USA. Proceedings of International Conference on Intelligent Computation Technology and Automation, Changsha, China, (2010) May 11-12.

[6] Francis Musembi Kwale, "*A Critical Review of K Means Text Clustering Algorithms*", International Journal of Advanced Research in Computer Science, Volume 4, No. 9, July-August 2013.

[7] JyotiYadav, Monika Sharma, "*A Review of K-mean Algorithm*", International Journal of Engineering Trends and Technology (IJETT) – Volume 4 Issue 7- July 2013.

[8] MinqingHu,Bing Liu, "*Mining Opinion Features in Customer Reviews*".

[9] Li Zhuang, Feng Jing , Xiao-Yan Zhu ,"*Movie Review Mining and Summarization*".

[10] Sayantani Ghosh, Mr. Sudipta Roy, Prof. Samir K. Bandyopadhyay "*A tutorial review on Text Mining Algorithms*".

[11] Kiri Wagstaf , Claire Cardie , Seth Rogers , Stefan Schroedl , "*Constrained K-means Clustering with Background Knowledge*" Proceedings of the Eighteenth International Conference on Machine Learning, 2001, pp. 577–584.

[12] Pavel Brazdil, "*Introduction to Text Mining*".

[13] Anna Stavrianou, Periklis Andritsos, Nicolas Nicoloyannis, "*Overview and Semantic Issues of Text Mining*".