New Text Clustering Method Based On Arithmetic Encoding Algorithm

Mr. Nikhil Pawar
Dr. P. K. Deshmukh

Abstract— in this paper, we proposed system that uses arithmetic encoding algorithm to encode data instances in to integer, clustering performed on integer instances is much more effective than clustering performed on string instances. This is very effective technique to improve clustering accuracy of text data; it has been observed that traditional clustering methods not perform well on string attributes. Compression ratio of Arithmetic encoding is better than Huffman encoding.

Index Terms — accuracy, cluster analysis, huffman encoding, arithmetic encoding machine learning, text mining.

1 INTRODUCTION

Text mining, also referred to as text data mining, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text. High-quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning. Text Mining has become an important research area. Text Mining is the discovery of unknown information from different data resources. Text mining is a process of extracting interesting and non-trivial information and knowledge from unstructured text. Text mining is a field which draws on information retrieval systems, data mining systems, computational linguistics, machine learning. As over 80% information is stored as text, text mining has a high commercial potential value. There are many sources of information from which knowledge may be discovered; yet, unstructured text is the source of knowledge. The problem of extract explicit and implicit concepts and semantic relations between concepts using Natural Language Processing (NLP) techniques is faced in Knowledge Discovery from Text (KDT). It increases efficiency in large quantities of text data. KDT, is rooted in NLP, draws on methods from statistics, information extraction, knowledge management, machine learning, reasoning, and others for its discovery process. KDT plays an important role in emerging applications, such as Text Understanding. Most of Text Mining operations are as follows: Feature Extraction, Text-based navigation, Search and Retrieval, Categorization (Supervised Classification), Clustering (Unsupervised Classification), Summarization. Like Data mining is about looking for patterns in data, In same way text mining is about finding patterns in text, it is the process of analyzing text to extract information that is useful for particular purposes. Aside from the raw analysis step, it involves interestingness metrics, data pre-processing, database and data management aspects, model and inference Considerations, complexity considerations, post-processing structures, implementing visualization, and online updating.

Arithmetic coding is a form of entropy encoding used in lossless data compression. Normally, a string of characters such as the words "hello there" is represented using a fixed number of bits per character, as in the ASCII code. When a string is converted to arithmetic encoding, frequently used characters will be stored with fewer bits and not-so-frequently occurring characters will be stored with more bits, resulting in fewer bits used in total. Arithmetic coding differs from other forms of entropy encoding such as Huffman coding in that rather than separating the input into component symbols and replacing each with a code, arithmetic coding encodes the entire message into a single number. Texts are always compressed with lossless compression algorithms. This is because a loss in a text will change its original concept. Repeated data is important in text compression. If a text has many repeated data, it can be compressed to a high ratio. This is due to the fact that compression algorithms generally eliminate repeated data. In order to evaluate the compression algorithms on the text data, a comparison between arithmetic and Huffman coding algorithms for different text files with different capacities has been performed in. compression ratio of the arithmetic coding for text files is better than Huffman coding.

This paper is composed further as: Section II talks about related work studied till now. Section III presents implementation details, algorithms used, mathematical model and experimental setup tended to by this paper. Section IV depicts results and discussion part. Section V draws conclusions and presents future work.

2 RELATED WORK

In previous method, Huffman encoding algorithm is used to encode text data. In computer science and information theory, Huffman coding is an entropy encoding algorithm used for lossless data compression. The term refers to the use of a variable length code table for encoding a source symbol (such as a

- *Nikhil Pawar is currently pursuing masters degree program in Computer Engineering in Pune University, India, PH-9503456776. E-mail :nikhilhpawa@gmail.com.*

character in a file) where the variable length code table has been derived in a particular way based on the estimated probability of occurrence for each possible value of the source symbol. Huffman coding is based on frequency of occurrence of a data item. The principle is to use a lower number of bits to encode the data that occurs more frequently. The average length of a Huffman code depends on the statistical frequency with which the source produces each symbol from its alphabet. A Huffman code dictionary, which associates each data symbol with a codeword, has the property that no codeword in the dictionary is a prefix of any other codeword in the dictionary. Huffman encoding is a way to assign binary codes to symbols that reduces the overall number of bits used to encode a typical string of those symbols. This Huffman encoding method is used in this method for encoding string data. This is new method for improving the clustering accuracy of text data. This method encodes the string values of a dataset using Huffman encoding algorithm, and declares these attributes as integer in the cluster evaluation phase. It is observed that when the dataset to be clustered has only string attributes, a traditional clustering method does not recognize, or recognize with a low accuracy, When we first convert it in to integer then clustering is perform well, the category of instances and it is Demonstrated that this method clusters with a higher accuracy the instances of such a dataset. The clustering methods have generally focused on the case of quantitative data, in which the attributes of the data are numeric. The problem has also been studied for the case of categorical data, in which the attributes may take on nominal values. There are also approaches on improving the clustering of text data streams. In the paper, we present a method for massive-domain clustering of data streams. The results obtained that a sketch-based clustering method can provide similar results to an infinite space clustering algorithm with high probability. We focus on view points and measures in hierarchical clustering. The research is particularly focused in studying and making use of cluster overlapping.

3 ISSUE

In previous method, Huffman encoding algorithm is used to encode text data, Compression ratio of Arithmetic encoding is better than Huffman encoding.

4 TEXT CLUSTERING TECHNIQUES

Clustering techniques apply when there is no class to be predicted but we want to divide instances in to natural groups. These clusters give signs of some mechanism that is at work in the domain from which instances are outlined, a mechanism that causes some instances to take a stronger likeness to each other than they do to the still in the same way examples. Clustering naturally has need of different techniques to the classification and association learning methods. With the popularity of Internet and great-scale getting better in the level of undertaking information, the bursting substance growth of useable things, the research of text mining, information filtering and information search. So, the cluster technology is becoming the core of text information mining technologies. The main objective of clustering is to partition unlabelled patterns into homogeneous clusters. Clustering algorithm can be divided into the following categories: hierarchical clustering partitioned clustering, density-based algorithm, self organizing maps algorithm. At the same time, the text clustering problem has its particularity. On one hand, the text vector is a high-dimensional vector, usually thousands or even ten thousands; On the other hand, the text vector is usually sparse vector, so it is difficult for the choice of cluster centre. As an unsupervised machine learning method, because of not need to train the process and manual label document at category in advance, clustering has certain flexibility and high automation handling ability. It is become an important mean which pays attention for more and more researchers. The purpose of text clustering is large-scale text data sets which can be grouped into several categories, and made between the text information in the same class which has high similarity, rather than the difference of text between the different types. There are many clustering techniques used for clustering text such as: Hierarchical clustering, Partitioned clustering, Density-based algorithm, and Organizing Maps algorithm. In this paper we focus on Hierarchical clustering to improve clustering efficiency. Text clustering is a typical problem of unsupervised machine learning. Hierarchical clustering algorithm by combining the appropriate similarity measure similarity such as cosine similarity, Dice coefficient, Jaccard similarity coefficient, has become the mainstream technology on the document clustering. Hierarchical clustering is commonly text clustering method, which can generate hierarchical nested class. Hierarchical clustering method takes category as hierarchical, in other words, with the change of category hierarchical, object also corresponding change. This method allows classifying data at different granularity. In accordance with generation methods of the category tree, hierarchical clustering method can be divided into two categories, one kind is integration method (bottom-up method), and the other kind is to split methods (top-down method).Hierarchical clustering accuracy is relatively high, but when each class merges, it needs to compare all classes' similarity in the global and selecting the most similar of two classes, so it's relatively slow. The defect of hierarchical clustering is that once a step (merge or split) completed, it cannot be revoked, so it can not correct the wrong decision. Hierarchical clustering methods are generally divided into bottom-up hierarchical clustering method and top-down hierarchical clustering method. Bottom-up (merge) hierarchical clustering method starts from a single object, first takes an object as a separate category, and then repeatedly merges two or more appropriate categories, until meeting stop conditions. Top-down (splitting) hierarchical clustering method starts from the objects complete

works, and gradually be divided into more categories. The typical approach is to construct a minimum spanning tree on similar graphs, and then at each step choosing a side which in the smallest similarity of the spanning tree (or in the farthest distance of the spanning tree) and removing it. If it deletes one side, it can create a new category. When the smallest similarity achieves some threshold value the cluster may stop. In general, the amount of computation of top-down method is greater than the bottom-up method, and the applications of top-down method is inferior widespread than the latter.

5 APPROACH

There should be a mechanism to improve clustering efficiency; to improve clustering accuracy of text data we can use a new text clustering method based on Arithmetic encoding algorithm. The main concept behind this study is, It is observed that when the dataset to be clustered has only string attributes, a traditional clustering method does not recognize, or recognize with a low accuracy, When we first convert it in to integer then clustering is perform well, the category of instances and it is Demonstrated that this method clusters with a higher accuracy the instances of such a dataset.

5.1 System Architecture:

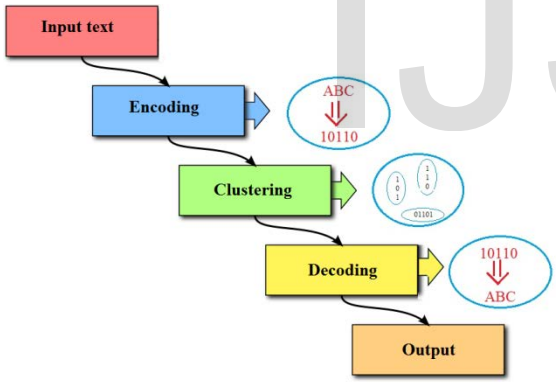

Fig 1: propose system Architecture

- Input : User Provide input as text instances
- Process Mechanism : Encoding text data in to integer then clustering is perform on integer attributes, to get text in original form decoding is done in next step
- Output: we have clusters to perform text mining.

5.2 Mathematical model

Let the system S is represented as: S = { T,E,C,D }

A. Learning Phase Consider, T is a set for learning dataset
T = {t1,t2,t3,. . . tn} Where t1,t2. . . tn are the learn data

B. Encoding Phase Let E be the set for Encoding phase
E = {e1,e2,e3,. . . ,en} Where, e1,e2,e3,. . . en are the Encoded data

C. Clustering Phase Let C is a set for Clustering
C = {c1,c2,c3,. . . .,cn} Where, c1,c2,c3. . . cn are the number of data form after Clustering
step.

D. Decoding Phase Let D is set for decoding Phase
D = {d1,d2,d3,. . . dn} Where, d1,d2,. . . dn are the Decoded data .

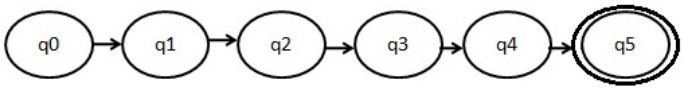

Fig 2: State Transition Diagram

The above model shows the representation of training the data set
q0 Initial state (Input)
q5 Final state (Output)
q1 Arithmetic Encoding
q2 Clustering
q3 Decoding
q4 text mining

5.3 Algorithm

Arithmetic Encoding Algorithm
For time=1:100
tic
K=0;
VECTOR-ARITH=V(1);
For 1=1:m;
a=0;
for q=1:k
if (V(1)==VECTOR-ARITH(q));
a= a+1;
end
end
if (a==0)
k=k+1;
VECTOR-ARITH(k)=V(1);
End
End
For u=1:k
a=0;
for 1=1:m
if(V(1)== VECTOR_ARITH(u))
a=a+1;
Varith(1)=u;
End
VECTOR-ARITH-NUM(u)=a;

```
End
End
Code= arithenco (Varith,VECTOR-ARITH_NUM);
[f1,f2]= size(code);
Compression ratio =b0/f2
toc
end
```

6 Results

6.1 Data set:

Here Text Mining is performed on two datasets named Physics dataset and the Biology dataset; these datasets are obtained from the online UCI Machine Learning Repository. UCI is the Centre for Machine Learning and Intelligent Systems. It holds a repository of datasets which are used by practitioners and researchers in the fields of Artificial Intelligence, Pattern Recognition, Machine Learning, Neural Networks, Data Mining, Bio-informatics and others these are referred to as the UCI datasets. Both dataset consist of string attributes.

6.2 Result

Table 1 presents the clustering results of the datasets with Physics and biology data sets. Our proposed method performs clustering with great efficiency, after observing these reading we can estimate clustering accuracy as 84.6%.

Table 1: THE CLUSTERING RESULTS

Dataset	Cluster 0	Cluster 1
Physics	61	331
Biology	509	99

We take input of 1000 instances of physics and biology datasets; we performed experiment on these 1000 instances, from experiment, *cluster 0* represents Biology instances and cluster 1 represent Physics instances. From experiment we get 509 instances in biology cluster and 331 instances in Physics cluster, so we get 84.6% accurate clustering with these datasets. We get 840 instances correctly clustered.

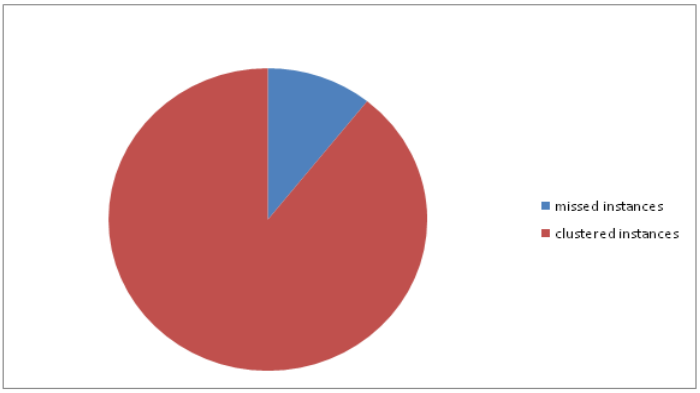

Fig 3: biology Cluster

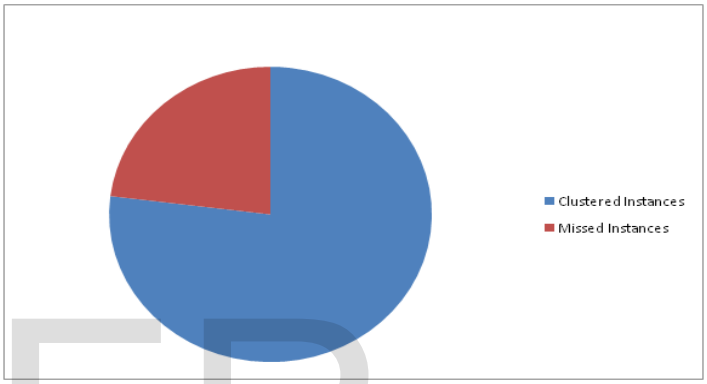

Fig 4: Physics Cluster

Figure 3 shows biology cluster in which we get 509 instances clustered correctly and figure 4 shows Physics Cluster in which we get 331 instances correctly, We observe a huge improvement comparing with the first experiment, where we used the initial dataset for the evaluation.

7 CONCLUSION

This is very effective technique to improve clustering accuracy of text data, it has been observed that traditional clustering methods not perform well on string attributes to improve clustering accuracy, Arithmetic encoding algorithm is used to encode data instances in to integer, clustering performed on integer instances is much more effective than clustering performed on string instances, here additional time require in encoding and decoding phase but these time is covered in clustering phase. Compression ratio of Arithmetic encoding is better than Huffman encoding.

REFERENCES

[1] B., Zheng, J., Chen, S., Xia, Y., Jin, "Data Analysis of Vessel Traffic Flow Using Clustering Algorithms", 2008 International Conference on Intelligent Computation Technology and Automation, Changsha, Hunan, China, pp. 243 – 246, 2008.

[2] M., Moslem, A., Hosein, and M.-B., Behrouz, "Neural Network nsembles using Clustering Ensemble and Genetic Algorithm", Third 2008 International Conference on Convergence and Hybrid Information Techology, Busan, South Korea, pp. 1924-1929, 2008.

[3] N., RaghavaRao, K., Sravankumar, P., Madhu, "A Survey On Document Clustering With Hierarchical Methods And Similarity Measures", International Journal of Engineering Research & Technology (IJERT), Vol. 1 Issue 7, ISSN: 2278-0181, pp. 1-7, 2012.

[4] C., C., Aggarwal, C., X., Zhai, Mining Text Data, chapter: A Survey of Text Clustering Algorithms, Springer US Publisher, Print ISBN 978-1- 4614-3222-7, Online ISBN 978-1-4614-3223-4, pp. 77-128, 2012.

[5] P., R., Suri, and M., Goel, "Ternary Tree and Clustering Based Huffman Coding Algorithm", IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 5, ISSN (Online): 1694-0814, 2010.

[6] http://www.cs.waikato.ac.nz/ml/weka/

[7] N., RaghavaRao, K., Sravankumar, P., Madhu, "A Survey On Document Clustering With Hierarchical Methods And Similarity Measures", International Journal of Engineering Research & Technology (IJERT), Vol. 1 Issue 7, ISSN: 2278-0181, pp. 1-7, 2012.

[8] Y., B., Liu, J., R., Cai, J., Yin, A., Wai-Chee Fu, "Clustering text data streams", Journal of Computer Science and Technology, 23(1), pp. 112- 128, 2008.

[9] P., R., Suri, and M., Goel, "Ternary Tree and Clustering Based Huffman Coding Algorithm", IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 5, ISSN (Online): 1694-0814, 2010.