Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.017

IJCSMC, Vol. 6, Issue. 10, October 2017, pg.46 – 50

Clustering and Classification in Sentimental Data Analysis

S.Maguda Gowreeswari, Dr. R.Manicka Chezian
Research Scholar, Dr. Mahalingam Centre for Research and Development, NGM College, Pollachi, India
Department of Computer Science, N G M College (Autonomous), Pollachi, Coimbatore - 642001, India
Email: gowrisivakrishnan@gmail.com, chezian_r@yahoo.co.in

Abstract: The essential process of determining the various types emotion behind a series of words, used to gain an understanding of the attitudes and feelings communicated on a post or remark. Sentimental data analysis is a greatly valuable media tool in social media monitoring as it allows us to gain an overview of the more extensive popular conclusion behind specific points. Also sentiment analysis are broad and powerful. The ability to extract insights from social data is a practice that is being widely adopted by various organisations across the world. At the same time there are challenges in implementing a perfect sentimental analysis model. The most important factor is unwanted data in the dataset. This unwanted data will totally affect the result. In order to overcome this problem, general clustering method has been implemented. This method will remove all the unwanted data from the data set and provided purified dataset for better analysis model. Here balanced clustering and k-Means classification has been used to implement the analysis part. These two methods combined together to form a best pre-processing method.

Keywords: Dataset, Data analysis, Clustering, Classification, Pre-processing.

I. INTRODUCTION

In clustering methods the idea is not to predict the target data as like classification methods, it is the method of trying to group the similar kind of things or removing unwanted data by considering the most satisfied condition all the items in the same group should be similar and no two different group items should not be similar. [1] For grouping the similar kind of items in clustering, different similarity measures could be used. In classification, the idea is to find the target class by analysis the training dataset. [2] This will be done by finding proper boundaries for each target class. In a general way the training dataset to

get better boundary conditions which could be used to determine each target class. [2] Once the boundary conditions determined, the next task is to find the target class as we have said earlier. The whole process is known as pre-processing.

While combining clustering with Classification techniques in data mining can capable to proceed with large number of data. [7] It can be used to find all the categorical classes, labels and classifies data based on the training set and class labels and it can be used for classifying newly available dataset. The term could cover on any context in which some decision or forecast is made on the basis of presently available of information. [9] Some Classification procedures will be a recognized method for repeatedly making such decisions in new situations. In that situation we can assume that problem is a concern with the construction of a procedure that will be applied to a continuing sequence of cases in which each new case must be assigned to one of a set of pre defined classes on the basis of observed features of data. Creation in the classification procedure from a set of data for which the classes are known in advance is termed as the pattern recognition or supervised learning.

II. BALANCED CLUSTERING

It is the method of clustering the unwanted data in equal groups. The model has two main restrictions, one on the total number of machines and another on the memory available on each machine. In particular, given an input of size N, and a sufficiently small $\gamma > 0$, in the model there are $N1-\gamma$ machines, each with $N1-\gamma$ memory available for the computation. As a result, the total amount of memory available to the entire system is $O(N2-2\gamma)$. In each round a computation is executed on each machine in parallel and then the outputs of the computation are shuffled between the machines. In this model the efficiency of an algorithm is measured by the number of the 'rounds' of cluster in the algorithm. A class of algorithms of particular interest are the ones that run in a constant number of rounds. This class of algorithms are denoted MRC0 . The high level idea is to use corset construction and a sequential space-efficient α-approximation algorithm (as outlined above). Unfortunately, this approach does not work as such in the cluster model because both the corset construction algorithm, and the space-efficient algorithm, require memory quadratic in the size of their input. Therefore we perform multiple 'levels' of our framework.

Given an instance (V, d), the balanced Clustering algorithm proceeds as follows:

1. Partition the points arbitrarily into $2n\ (1+\gamma)/2$ sets.
2. Compute the composable 2 p -mapping coreset on each of the machines (in parallel) to obtain f and the multisets S1, S2, . . . , S2n (1+γ)/2 , each with roughly Oe(k) distinct points.
3. Partition the computed coreset again into n 1/4 sets. 4. Compute composable 2 p -mapping coresets on each of the machines (in parallel) to obtain f 0 , and multisets S 0 1 , S0 2 , . . . , S0 n 1/4 , each with Oe(k) distinct points.
5. Merge all the S 0 1 , S0 2 , . . . , S0 n 1/4 on a single machine and compute a clustering using the sequential space-efficient α-approximation algorithm.
6. Map back the points in S 0 1 , S0 2 , . . . , S0 n 1/4 to the points in S1, S2, . . . , S2n (1+γ)/2 using the function f 0−1 and obtain a clustering of the

points in S1, S2, . . . , S2n (1+γ)/2 .

7. Map back the points in S1, S2, . . . , S2n (1+γ)/2 to the points in V using the function f −1 and thus obtain a clustering of the initial set of points.

Figure1.Balanced Clustering Algorithm

III. K-MEANS CLASSIFICATION

K-Means Classification rule distinguishes the classification of unknown data point on the basis of its closest neighbour whose class is already purpose K-Means in which most nearest neighbour is computed on the basis of estimation of k that indicates how many nearest neighbours are to be considered to characterize class of a sample data point. It makes utilization of the more than one closest neighbour to determine the class in which the given data point belongs to and consequently it is called as K-Means classification. These data samples are needed to be in the memory at the run time and hence they are referred to as memory-based technique. Enhanced K-Means is focused on weights. The training points are assigned weights according to their distances from sample data point. But at the same time the computational complexity and memory requirements remain the primary concern dependably. To overcome memory limitation size of data set is reduced. For this the repeated patterns which don't include additional data are also eliminated from training data set. To further enhance the information focuses which don't influence the result are additionally eliminated from training data set. The NN training data set can be organized utilizing different systems to enhance over memory limit of K-Means. The K-Means implementation can be done using ball tree, k-d tree, nearest feature line (NFL), principal axis search tree and orthogonal search tree. The tree structured training data is further divided into nodes and techniques like NFL and tuneable metric divide the training data set according to planes. Using these algorithms we can expand the speed of basic K-Means algorithm. Consider that an object is sampled with a set of different attributes. Assuming its group can be determined from its attributes different algorithms can be used to automate the classification process. In pseudo code K-Means classification algorithm can be expressed as follows

Step 1: K ← number of Mean neighbours

Step 2: For each object X in the test set do

Step 3: calculate the distance D(X,Y)
between X and every object Y in the training set

Step 4: neighbourhood ← the k neighbours in the training set closest to X

Step 5: $K(C| X) = D(X_1 | Y_1) + D(X_2 | Y_1)$ ..... $D(X_n | Y_1)$ - $X D(K_n)$

Step 6: K continues its process upto → $D(X_n | Y_1)$

Step 7: D → finds neighbourhood for X Class

Step 8: X.class ← Select Class (neighbourhood)

Step 9: End for

Figure 2.K-Means Classification Algorithm

IV. RESULT

The below mentioned Table1.1 shows the actual result after pre processing. The preprocessing has been done using clustering and classification methods as mentioned above. It removes all the repeated users and repeated comments. After the removal of data, the analysis will we executed again in the same execution process.Figure.3 represents the comparison between before and after removing redundancy.

Data	Data with redundancy	Data After Removing redundancy
Total data	199	140
Number of positive data	100	59
Number of negative data	39	31
Number of moderate data	60	50

Table 1.1 Comparison Method

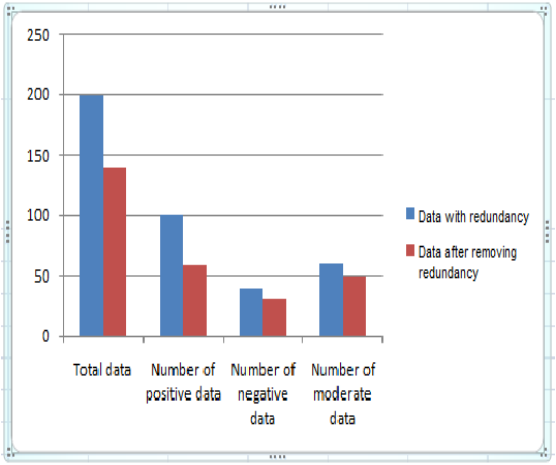

Fig 3. Comparison chart

V. CONCLUSION

A new active pre-processing algorithm is proposed which actively selects informative data by dealing with the clustering results of the dataset. Labelling these data and using them to label their k-Mean based on an adaptive threshold. The experimental results show that the proposed semi supervised clustering and classification can reach a stable state which only requires very small size of labelled dataset. However, the accuracy of the proposed semi supervised clustering is higher which clusters overlap each other than that in the dataset in which the boundaries between clusters are not very vague.

REFERENCES

[1] Delveen Luqman Abd Al.Nabi, Shereen Shukri Ahmed, "Survey on Classification Algorithms for Data Mining: (Comparison and Evaluation)" (ISSN 2222-2863) Vol.4, No.8, 2013

[2] Vidhya.K. G.Aghila. "Naïve Bayes Machine Learning approach in Text Document Classification",(IJIST) Vol. 7, No. 2, 2010.

[3] Nitin Bhatia, Vandana," Nearest Neighbor Techniques" (IJCSIS) Vol. 8, No. 2, 2010, ISSN 1947-5500.

[4] Riaan Smit" An Overview of Support Vector Machines, 30 March 2011.

[5] B. Kotsiantis · I. D. Zaharakis · P. E. Pintelas, "Machine learning: a review of classification and combining techniques", Springer Science10 November 2007

[6] Ashis Pradhan., "Support Vector Machines a survey, ISSN 2250-2459, Volume 2, Issue 8, August 2012

[7] Thair Nu Phyu," Survey of Classification Techniques in Data Mining", IMECS 2009, March 18 - 20, 2009.

[8] H. Bhavsar, A. Ganatra,"Variations of Support Vector Machine Classification: A survey", International Journal of Advanced Computer Research, Volume 2, Number 4, Issue 6 (2012) 230–236.

[9] Ms. Aparna Raj, Mrs. Bincy, Mrs. T.Mathu "Survey on Common Data Mining Classification Techniques", International Journal of Wisdom Based Computing, Vol. 2(1), April 2012

[10] Raj Kumar, Dr. Rajesh Verma," Classification Algorithms for Data Mining P: A Survey" IJIET Vol. 1 Issue August 2012, ISSN: 2319 – 1058.

