Journal of Engineering and Applied Sciences 13 (Special Issue 6): 5226-5229, 2018
ISSN: 1816-949X

Clustering Online News Comments Using Hadoop on Bigdata

Anu Sunil Kumar, Remya Anand and G. Deepa
Department of Computer Science and IT, School of Arts and Sciences,
Amrita University, Kochi, India

Abstract: Mining in news blog remarkably a new research area in this modern world of the technological era. Here, we propose a feature word selection of clustering online news comments using Hadoop on bigdata which realizes structurally superior clustering of online comments. Data is made to run on Hadoop platform, so as to convert the unstructured data from the news comments to a structured format for further classification. Here a Naive Bayesian classifier is included right before applying the k-means clustering algorithm. For clustering, the top most frequent nouns appearing across online comments are selected to construct an overall noun set. Local noun sets are constructed based on the frequently occurring nouns. The global noun set is the intersection of the local and overall noun set. The global noun set is reduced from the corresponding local noun set to construct the distinct noun set.

Key words: Mining, online news blog, Hadoop, classification, clustering, noun

INTRODUCTION

Feature selection is a relevant part of machine learning to minimize the inputs for processing, analysing and to detect the most meaningful inputs. In machine learning, "Feature selection" commonlycalled as variable selection, i.e., the technique of subset selection of apt attributes for constructing a model.

Nowaday's people are more dependent on social media like Facebook, Twitter etc., so one of the key way to get the public opinion regarding the online news comments is to make it available through social media like a blog. A blog is ananalysis area or informational website published on the web which consist of diverse informal posts.

People will go through the articles from the blog which have breaking news and information regarding political issues, society, science and health. People will provide their valuable opinions in the comment section based on the impact of particular news in their day to day life. Through these comments people will express their feelings about the decision taken by officials. As mentioned there will be large pool of online comments which often contains the similar opinions for news articles and which can be grouped into clusters. Clustering the similar opinions will make us easy to understand and analyse the online comments as a whole.

Motivation: The dataset is a collection of huge amount comments of each news article that may be structured or unstructured. If its is structured, then the k-means clustering algorithms can be easily applied but if unstructured then it is difficult to apply k-means clustering algorithms and also the efficiency will not be much consistent.

Objective: The study proposes an affective approach for feature word selection in online news comments. In order to cluster the vast number of online news comments, we are using clustering algorithms, i.e., k-means clustering algorithms. Before forming the clusters for each news articles a classifier is applied in order to enhance the cluster formation. Classifier used here is Naive Bayes classifier.

MATERIALS AND METHODS

The blog contains news articles, each of which contains number of comments. All these comments are stored to form the data sets. The data sets which we get from the online news comments may be structured or unstructured, so, if it is unstructured format then it must be converted to a structured format in order to easily implement classification using the Naive Bayesian classifier. For this, we run our data sets in Hadoop. Hadoop is the software for handling relatively huge data sets and is also an open source software. It contains commodity hardware built from computer clusters.

Hadoop: An open source project to implement MapReduce Model for scalable, reliable and distributed computing and it is written in Java. With the help of a simple programming model. It is a base for distributed processing of sizable datasets across computer clusters.

Corresponding Author: Anu Sunil Kumar, Department of Computer Science and IT, School of Arts and Sciences,
Amrita University, Kochi, India

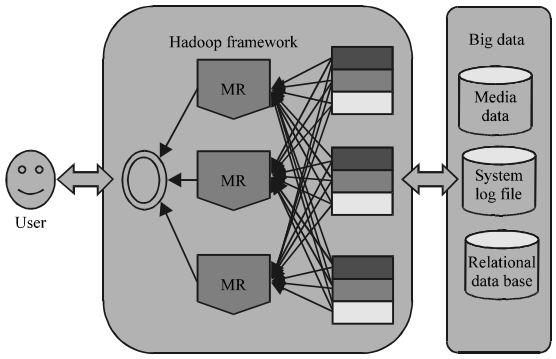


Fig. 1: The basic structure of Hadoop framework

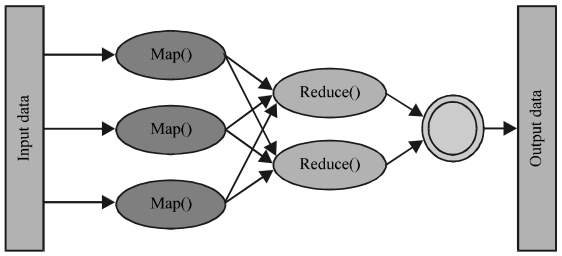


Fig. 2: The workflow of the MapReduce algorithm

Proposed by Doug Cutting, Mike Cafarella and team in 2005. Hadoop (Hamoud *et al*., 2014) framework is enough capable to develop applications which is capable of running on clusters and for a huge amounts of data it could perform complete statistical analysis (Fig. 1 and 2).

MapReduce: It is a programming model which is associated with the implementation of bigdata. Hadoop runs applications by using MapReduce algorithm. The algorithm contains two tasks, i.e., Map and Reduce. Map takes a set of data, then converts to another set of data and individual elements of set of data is broken down into tuples where as in reduce, output of map is taken as the input and also combines the tuples into smaller tuples.

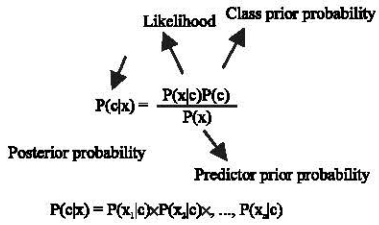
Algorithm; MapReduce:

i. First task is Map job, it takes the input data and processes it to produce a key/value pairs
ii. Then Reduce job takes these key/value pairs and then combines to form final result, i.e., Output data

Naive Bayesian classifier: Naive Bayes (Holmes *et al*., 1994) is a simple technique based on Bayes theorem for constructing classifiers: models that contains a class labels which is assigned to problem instances, represented as vector values. The class labels are drawn from some finite set.

They are highly scalable which require a number of parameters in a learning problem. Maximum probability can be evaluated in a closed-form expression which takes some less time when compared other types of classifiers. All Naive Bayes classifiers assume that, for the given class variable the value of a particular feature is independent of the value of any other feature.

Bayes theorem (Wettig *et al*., 2003) provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c). Look at the equation:



Where:

P(c|x) = The posterior probability of class (c, target) given predictor (x, attributes)
P(c) = The prior probability of class
P(x|c) = The likelihood which is the probability of predictor given class
P(x) = The prior probability of predictor

Algorithm 2; Naive Bayesian classifier probability:

- Create frequency table by converting dataset
- Find the probabilities and then create likelihood table
- Use the above equation to calculate the posterior probability for each class. The outcome is the class with highest posterior probability

Operation for feature word selection: Each news articles contain huge amount of comments which in turn contains a number of words which conveys an opinion about the news articles. The words which conveyed the opinion in the comment is taken into account and called as word tokens. For each news articles there will be a collection of word tokens. From the word tokens, commonly used nouns were used to create the overall noun set and it also other types of sets, i.e., local, global and distinct.

Each of the noun set were created as follows. Local noun set is the collection local word which convey the opinion of the article. Each news articles will contain its own local noun sets. For example, if there are 5 news articles, then for each of them there will be a local noun sets. The global noun set is intersection of local noun set and overall noun set. The distinct noun set is obtained by eliminating global noun set from local noun set (Cho and Lee, 2016). Based on these noun sets, the clusters are formed by applying k-means clustering algorithms (Tsai and Chui, 2008).

WEKA: Waikato Environment for Knowledge Analysis (WEKA) (Holmes *et al*., 1994) is Java based machine learning software developed in 1993 by the University of Waikato, New Zealand. WEKA supported applications which offers users, a tool to expose the masked information from database and file systems easily to use options and visual interfaces. The Weka is a collection of visualization tools and algorithms for solving real-world data mining problems and predictive.

RESULTS AND DISCUSSION

The online news comments may be structured or unstructured data. If it is structured, then data can be used more efficiently otherwise if the data is unstructured then it must converted into structured format, so as to efficiently apply the classification and clustering. For this all the dataset which is in executed on Hadoop. Then the WEKA libraries are imported to the Hadoop and then the classification and clustering is done. Steps involved in above process is as follows:

- Execute the Hadoop framework so as to convert the unstructured data into structured format
- Import the WEKA Libraries into the Hadoop and choose the main WEKA GUI
- Once the main GUI is opened, choose the dataset which is. arff file
- From the WEKA interface choose the classifier i.e.; Naive Bayesian classifier
- From the test options, choose the cross validation and enter the value of folds (Here, we entered the value as 3)
- Click on start

Figure 3 shows how the classification is being performed on the Hadoop framework and classifier used here is the Naive Bayesian classifier. Figure 4 and 5 shows the clustering result.

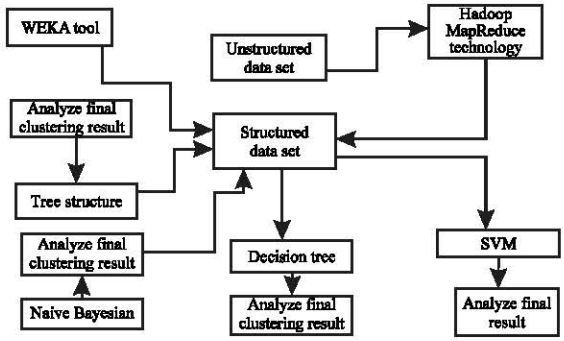


Fig. 3: The block diagram of the execution flow

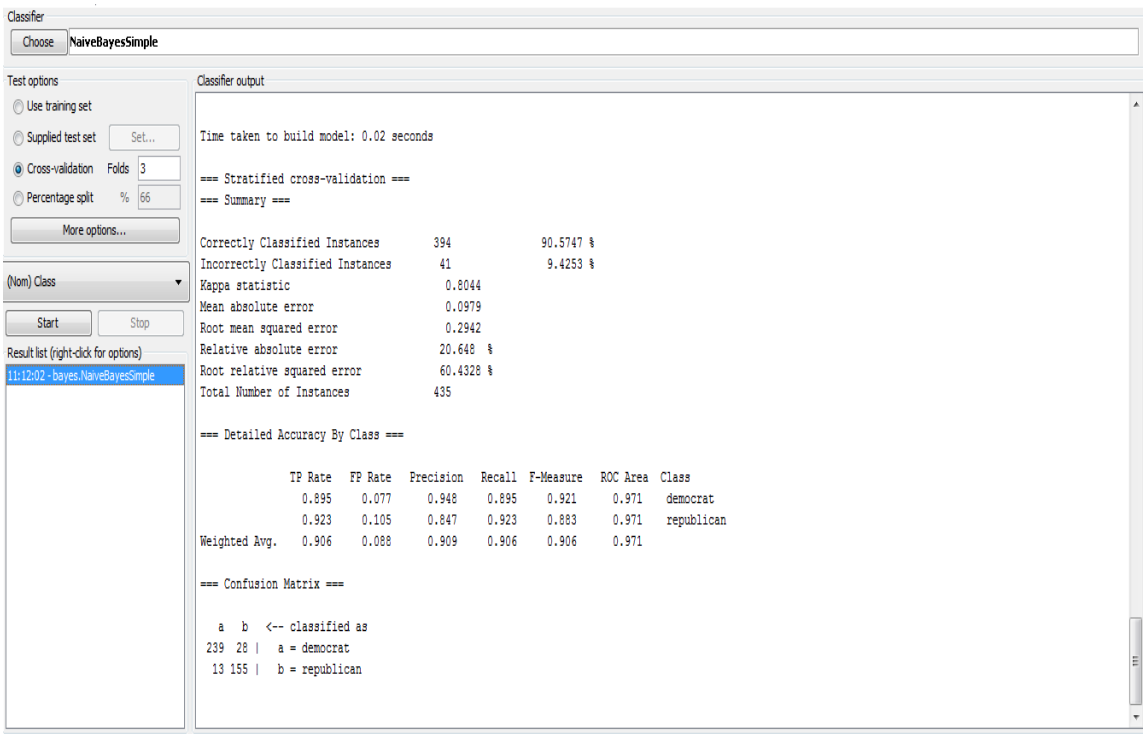


Fig. 4: The clustering results

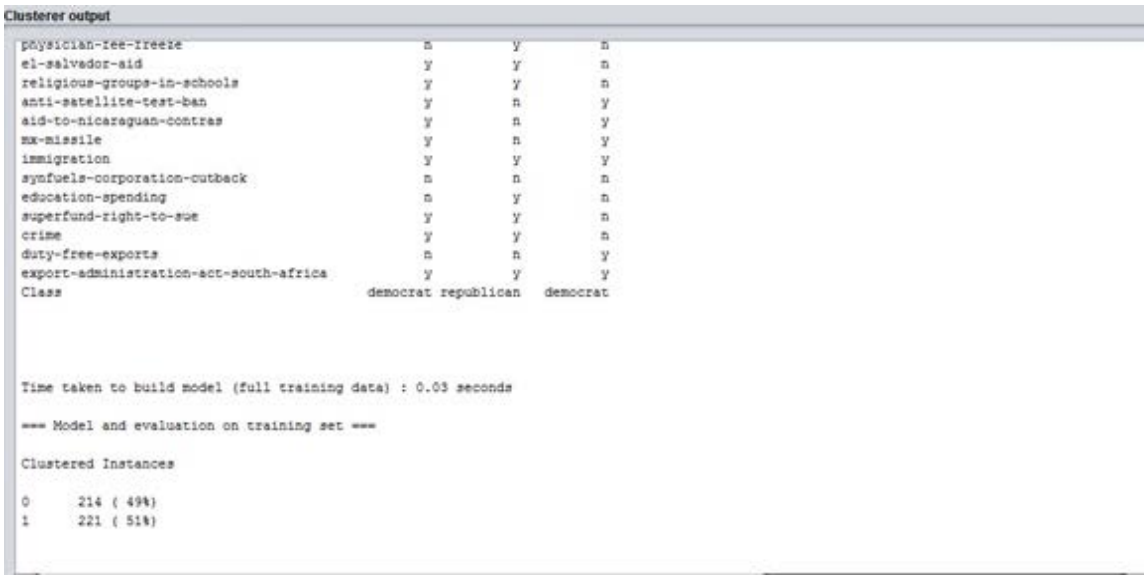


Fig. 5: The Hadoop framework

CONCLUSION

This study provides a study about an online news blog. By performing classification and clustering on Hadoop, we can decrease the time complexity, more time available to add value (rather than fix/maintain) and more stable operating environments.

RECOMMENDATIONS

In future algorithms can be improved and can be extended in further which would bring in a much more efficient output which would reduce the manual efforts related to comments which appear in a blog. The concept of Cyberbullying can be applied here, i.e., a form of bullying or harassment using electronic forms. Harmful bullying behaviour can include posting rumours about a person, threats and sexual remarks, disclose victim's personal information or pejorative labels in the comments of the news blog.

REFERENCES

Cho, H. and J.S. Lee, 2016. Data-driven feature word selection for clustering online news comments. Proceedings of the International Conference on Big Data and Smart Computing (BigComp'16), January 18-20, 2016, IEEE, Hong Kong, China, ISBN:978-1-4673-8796-5, pp: 494-497.

Hamoud, A., H. Bajwa and J. Lee, 2014. Hadoop based enhanced cloud architecture. American Society for Engineering Education, Salt Lake City, Utah.

Holmes, G., A. Donkin and I.H. Witten, 1994. WEKA: A machine learning workbench. Proceedings of the 2nd Australian and New Zealand Conference on Intelligent Information Systems, November 29-Dec. 2, 1994, IEEE Computer Society Press, pp: 357-361.

Tsai, C.Y. and C.C. Chui, 2008. Developing a feature weight self-adjustment mechanism for a k-means clustering algorithm. Comput. Stat. Data Anal., 52: 4658-4672.

Wettig, H., J. Lahtinen, T. Lepola, P. Myllymaki and H. Tirri, 2003. Bayesian analysis of online newspaper log data. Proceedings of the Symposium on Applications and the Internet Workshops, January 27-31, 2003, IEEE,Orlando,Florida,ISBN:0-7695-1873-7, pp: 282-287.