Volume 3, Issue 12, December-2016, pp. 659-666 ISSN (O): 2349-7084

International Journal of Computer Engineering In Research Trends

Available online at: www.ijcert.org

An Effective algorithm for Spam Filtering and Cluster Formation

Kavitha Guda

Associate Professor, Department of Computer Science and Engineering.
Vishwa Vishwani Institute of Technology, Hyderabad.
Telangana, India.

Abstract:- K-means clustering algorithm is one of the most widely used partitioning algorithms used for grouping the elements over spatiotemporal data. It is the fast, simple and can work with large datasets. It has some of the pitfalls regarding Number of iterations are more due to clusters details not known at an initial stage. It can detect only spherical clusters. Here we will propose a Hybrid K-Means clustering algorithm which will mostly work on the concept of splitting dataset and reducing the number of iterations. It will inherit the some of the features from two revised K-means algorithms. The advantage of separating more massive datasets is that handle easy, and the benefit of reducing iterations leads the easy cluster formation in this way the efficiency of the traditional K-means clustering algorithm is increased. Furthermore, we also proposed Naïve Bayes Algorithm for Email Spam Filtering on SPAMBASE Dataset.

Keywords: Data Mining, KDD, E-Mail, Spam, Naïve Bayes Algorithm, Spam Filter, K-Means Algorithm, Hybrid K-means Algorithm, SPAMBASE dataset.

1. INTRODUCTION

Tremendous growth of data since from few decades is unusually high. The cause behind the terrific increase in the size and the complexity of the data is due to various online commercial sites, work performed in the engineering field and other social media sites like Facebook, twitter, LinkedIn, and youtube etc. The internet contains a huge amount of raw data, to process the data several tools and techniques are used for the effective extraction of relevant data. Data Mining is a process established for the possible retrieval of unseen information for the sake of gaining knowledge. Facts can vary in dimension, difficulty to the formation. Data can be represented in the form of audio, video or simply a text data in the alphabetic or numeric form. Data mining is desirable, to tackle the large volume of data and to extract needed properties from the group of the data.

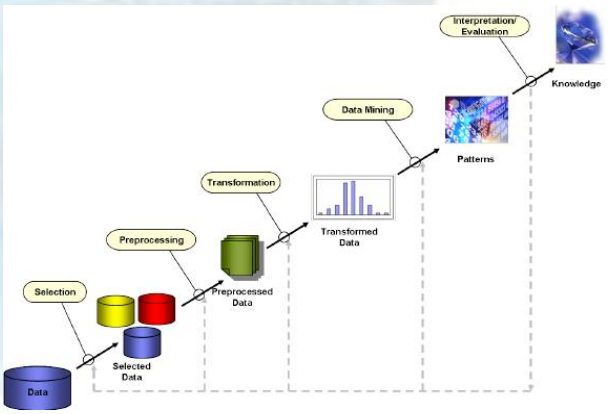

Figure 1. Steps in the KDD process

Knowledge or facts from the data can be acquired by undergoing many steps related to each other. Information mining is also categorized as Knowledge detection method, which means an action to extract valuable data from a collection of untreated data. Data mining is a concentric part of knowledge discovery [1]

Collection of Raw Data: Data-group can be gathered from various sources like online and offline, social media sources, public sector banks,

 DOI:10.22362/ijcert/2016/v3/i12/4321

retail sector, Insurance companies, Private sector banks, etc.

Data Selection: Data can vary in large volume, so it is necessary to extract relevant and essential data that is required for the further processing is selected.

Data Pre-Processing: Raw data can contain false information in the missing values or noise form. So, it is mandatory to pre-process the dataset, to remove any vague or incorrect data.

Transformation: The data is transformed into suitable shape so that mining job can be carried out.

Data Mining: Finding the relevance of the data is called as data mining. A variety of data mining approaches can be utilized to carry out the application in the data.

Evaluation: Gained information is evaluated for the exactness of the patterns and its compactness.

Knowledge: The final required information is called as Knowledge

Different Methods of Data Mining:

The diverse methods relevant in data mining are considered as mentioned underneath [2]. The following steps are performed on raw data to gain and access Knowledge.

- **Anomaly Detection:** Collected information that can be irrelevant or bogus is detected which is termed as an Anomaly or fake. Anomaly detection tracks the information that contributes to no fact or knowledge.
- **Association Rule Mining (ARM):** It is a procedure of establishing a relationship between the items in the dataset.
- **Clustering:** It is a procedure that labels the similar type of data in one group called as clusters without knowing any predefined model. The expressive process of grouping the data.
- Classification: It is a procedure that has a predefined known structure which groups the data into known predefined groups. Classification modeling is a predictive model for grouping the data. It helps to target data to different classes.
- **Summarization:** A process of labeling the data in a compact form so that we can visualize and represent it.
- **Electronic mail Spam Electronic mails are classified into two broad categories:** Spam emails and Ham emails. Spam emails are the unauthenticated emails received from the unknown sources that may contain the virus. Spam can originate from any external source like Web, Text messages, etc., depending upon the kind of broadcast; spam can be categorized into a variety of category similar to electronic mail spam, web spam, text spam, social networking spam [3].

The spam emails are scattering at the great pace due to the swift and offensive way of contribution data. It was noticed that account holders receive more spam emails than the ham emails. To avoid spam emails, spam filtration is important because spam can lead to time, energy, and bandwidth wastage, along with the misleading information [4]. Email can be labeled as a spam email only depending on these properties:

- **Uninvited Emails:** E-mails that are received from contacts that are not known to the user.
- **Bulk Mailing:** The kind of emails that are sent in mass or bulk to multiple account holders at the same time.
- **Unknown Mails:** In this type of mails in the identity and the details of the sender are not revealed or demonstrated.

For instance, when the user received a large amount of e-mail spam, the chance of the user forgot to read a non-spam message increase. As a result, many e-mail readers have to spend their time removing unwanted messages. E-mail spam also may cost money to users with dial-up connections, waste bandwidth, and may expose minors to unsuitable content. Over the past many years, many approaches have been provided to block e-mail spam [5]. For filtering, some email spam is not being labeled as spam because the e-mail filtering does not detect that email as spam. Some existing problems are regarding accuracy for email spam filtering that might introduce some error. Several machine learning algorithms have been used in spam e-mail filtering, but Naïve Bayes algorithm is particularly popular in commercial and open-source spam filters [6]. This is because of its simplicity, which makes them easy to implement and just need short training time or fast evaluation to filter email spam. The filter requires training that can be provided by a previous set of spam and non-spam

 DOI:10.22362/ijcert/2016/v3/i12/4321

messages. It keeps track of each word that occurs only in spam, in non-spam messages, and in both. Naïve Bayes can be used in different datasets where each of them has different features and characteristic.

2. RELATED WORK

Shi Na et al. has first explained the characteristics of k mean algorithm and then a newly enhanced k means algorithm is planned that reduces the measure of iterations. The improved algorithm avoids the calculation of the distance of each object to the cluster center again and again. First, it randomly chooses K data points and calculates the first cluster centers by smallest Euclidean distance. Two arrays are used to store the smallest distance of the clusters. The second one is used to store the cluster center of the object. This information is useful in reducing the number of times the loops are executed. In this way, it reduces the efficiency of the k-mean algorithm by increasing the execution speed. Two different types of datasets are used. Then both the k means and enhanced k mean algorithms are run on the dataset. The experiments show that the enhanced k mean algorithm provides superior performance as compared to traditional k mean algorithm [7].

Sourabh Shah et al. has taken three algorithms into consideration in this paper. K medoid, k mean and modified k mean algorithms are compared. In PAM algorithm initially, K objects are chosen as medoids. Then we calculate the distance of each object with the medoid and in this way we assign the object to the medoid with the smallest distance. In this way, every data item is allocated to the adjoining medoid. In next step, swapping is done. We swap a medoid m with nonmedoid o. Again the same procedure is followed. New cost is calculated. If this cost us lesser than the previous cost, then the newly chosen object becomes the medoid. After this iteration, we swap the non-medoid with the medoid, and the same procedure is repeated. The whole process continues until there is no change in the rate of medoids. The customized k means algorithm is as well described. It is approximately alike to the k mean algorithm. The only difference is that in modified k mean algorithm instead of implementing k mean on whole of the dataset; the dataset is split into smaller parts or subparts. Then k means is applied on these subparts. It is found experimentally that the customized algorithm k mean shows enhanced performance as compared to the traditional k mean algorithm and k medoid on the same dataset [8].

Kwai Han et al. clarifies that information concentrated shared (p2p) systems are finding expanding number of uses. Information mining in such P2P surroundings is a typical development. Be that as it may, common solid information mining configuration doesn't fit well in these sort of surroundings as they more often than not require bringing together the scattered information which is frequently not reasonable in a gigantic P2P arrange. Circulated information mining calculations that avoid huge scale synchronization or information centralization propose a distinctive decision. This paper considers the scattered k-implies grouping exertion where the information and figuring resources are spread over a vast P2P arrange. It offers two calculations which manufacture a gauge of the outcomes made by the standard concentrated k-mean bunching calculation. The essential is intended to work in a dynamic P2P arrange that can make grouping by limited synchronization as it were. The following calculation utilizes reliably inspected peers and gives intelligent certifications concerning the accuracy of bunching on a p2p arrange. Exploratory outcomes represent that both the calculations uncover excellent execution contrasted with their concentrated partners at the unobtrusive correspondence cost [10]

Konstantin Tretyakov et al., [11] have evaluated several most popular machine learning methods, i.e., Bayesian classification, k-NN, ANNs, SVMs and of their applicability to the problem of spam-filtering. In this work, the author proposed most trivial sample implementation of the named techniques and the comparison of their performance on the PU1 spam corpus dataset is presented. The author used extracting feature to convert all messages to vectors of numbers (feature vectors) and then classify these vectors. This is because most of the machine learning algorithms can only classify numerical objects like vector

3. METHODOLOGY

 DOI:10.22362/ijcert/2016/v3/i12/4321

3.1 Methodology for Hybrid K-Means Algorithm

K-Mean is the traditional partitioning algorithm. Till now various researchers have used it in many fields like biology, insurance, banking, marketing, etc. it has faced many modifications because it faces various drawbacks like we need to tell the number of clusters initially, how to choose initial points, the large number of iterations. Till date, many types of research have given their solutions for these problems. Some have used the hybrid approach. Some researchers have reduced the calculations by using their methods to increase the speed. Some researchers have used a different method to choose initial clusters. Others have used their methods to choose the no of cluster centers where as some researchers have used the median, mode or max-min distance to find the minimum distance. K-means deals with many problems like it are hard to assume the significance of K. For different values of K, clusters we get are different. It works only with numerical data. It is not capable of detecting the noise and outliers. It puts all the data into clusters. It cannot deal with irregular shapes. It cannot work with very large datasets. It does not work well with clusters of diverse thickness. With the analysis of k means algorithm, we have found that we can try to improve its speed or increase its efficiency by using our approach and moreover the algorithm can be enhanced to deal with the very large datasets. We can make it more robust comparatively. So what we have done is that we take a dataset first. Then that dataset is divided into the smaller dataset. Then we run an algorithm which is modified form of the k-mean clustering algorithm. In this algorithm, we have abridged the number of repetitions in k-mean clustering algorithm which increases its efficiency. But dividing the dataset into smaller datasets, we have made the traditional k-mean more robust in the way that now we can deal with comparatively larger datasets as compared to the traditional k-mean algorithm. In our study, we have merged two approaches basically, one is splitting the dataset into smaller datasets, and other is reducing the number of iterations. What we have done is that we take a dataset first.

Then that dataset is divided into the smaller dataset. Then we run an algorithm which is modified form of the k-mean clustering algorithm. In this algorithm, we have abridged the number of repetitions in k-mean clustering algorithm which increases its efficiency. By dividing the dataset into smaller datasets, we have made the traditional mean more robust in the way that now we can deal with comparatively larger datasets as compared to the conventional k-mean algorithm. While doing the research, the methodology we adopted is that first of all we collected the data on data mining which is known as literature survey. Then the second step was to choose the main topic in data mining on which we want to precede our research. Clustering was chosen as the main topic. Some research papers were studied to find the problem definition. Here we deal basically with the k-mean algorithm which is partitioning clustering algorithm. The data relevant to the k-mean algorithm was collected, and to deal with that problem, we present here an enhanced k-means clustering algorithm. Mainly means is an algorithm to select the early ideals to go after K-means clustering algorithm. If we choose wrong clusters initially, it leads to poor clustering. The k-means algorithm initialized with a random set of group centers. We introduced a different way of selecting the centres and then some method to reduce the number of iterations. Basically, we are splitting the data into smaller sets and then implementing am algorithm on these smaller datasets to reduce the number of iterations.

Steps for Proposed algorithms:

First of all, draw multiple sub-samples from the original data set.

- From every subsample arbitrarily choose k items from the dataset as initial cluster centres.
- Compute the area among every data items and all cluster mid-points as Euclidean area and allocate data items to the adjoining clusters.
- For every data item, locate the nearby centre and set the instance to cluster centre.
- Store the tag of cluster middle in which data item is and all the space of data item to the adjoining cluster and accumulate them.
- Recalculate the cluster centre for each cluster.

 DOI:10.22362/ijcert/2016/v3/i12/4321

• For every data item compute its space to the centre of the current adjoining cluster, if this space is fewer than previous distance, the data item stays in the first cluster else for all cluster centre calculate the space of each data item to all the centre, allocate data item to the adjoining middle.

• For every cluster, centre recalculates the centres until convergences criteria meet.

• Yield the clustering results.

3.2 Methodology for Email Spam Filtering

The methodology that is used for the filtering method is machine learning techniques that divide into three phases.

(i) Stage1: Pre-processing

(ii) Stage2: Feature Selection

(iii) Stage3: Naive Bayes Classifier

The following sections will explain the activities that involve in each phase to develop this project. Figure 2. Shows the process for e-mail spam filtering based on Naïve Bayes algorithm.

Stage: 1

Pre-processing

Today, most of the data in the real world are incomplete containing aggregate, noisy and missing values [12]. Pre-processing of e-mails in next step of training filter, some words like conjunction words, articles is removed from email body because those words are not useful in classification. As mentioned earlier, we are using WEKA tool to facilitate the experiments. For both experiments, the datasets are presented in Attribute-Relation File Format (ARFF) file

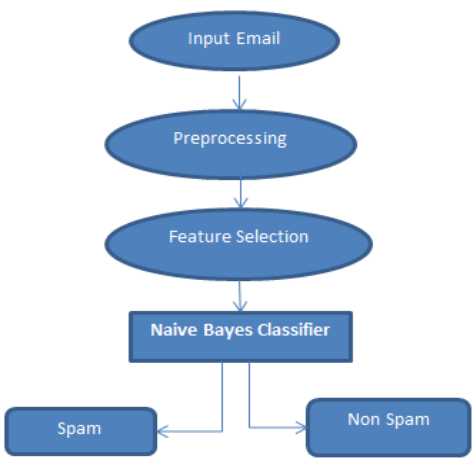

Figure 2. Practice of E-mail spam filtering based on Naive Bayes Algorithm

Stage2:

Feature Selection

After the pre-processing step, we apply the feature selection algorithm, the algorithm which deploys here is Best First Feature Selection algorithm [13].

Dataset 1: SPAMBASE

Data Set Characteristics:	Multivariate	Number of Instances:	4601	Area:	Computer
Attribute Characteristics:	Integer, Real	Number of Attributes:	57	Date Donated	1999-07-01
Associated Tasks:	Classification	Missing Values?	Yes	The number of Web Hits:	288415

Source: Mark Hopkins, Erik Reeber, George Forman, Jaap Suermondt Hewlett-Packard Labs, 1501 Page Mill Rd., Palo Alto, CA 94304

Stage3:

Naive Bayes Classifier

The methodology is used for the process of e-mail spam filtering based on Naive Bayes algorithm.

Naive Bayes classifier The Naive Bayes algorithm is a simple probabilistic classifier that calculates a set of probabilities by counting the frequency and combination of values in a given dataset [4]. In this research, Naive Bayes classifier use bag of words features to identify spam e-mail and a text is representing as the bag of its word. The bag of words is always used in methods of document classification, where the frequency of occurrence of each word is used as a feature for training classifier. This bag of words features are included in the chosen datasets. Naive Bayes technique used Bayes theorem to determine that probabilities spam e-mail. Some words have particular probabilities of occurring in spam e-mail or non-spam e-mail. Example, suppose that we know exactly, that the word Free could never occur in a non-spam e-mail. Then, when we saw a message containing this word, we could tell for sure that was spam email. Bayesian spam filters have learned a very high spam probability for the words such as Free and Viagra, but a very low spam probability for words seen in the non-spam e-mail, such as the names of friend and family member. So, to calculate the probability that e-mail is spam or

 DOI:10.22362/ijcert/2016/v3/i12/4321

non-spam Naive Bayes technique used Bayes theorem as shown in the formula below.

$$P(spam \mid word) = \frac{P(spam).P(word|spam)}{P(spam).P(word|spam) + P(non-spam).P(word|non-spam)}$$

Where:

(i) P(spamword) is the probability that an e-mail has particularly word given the e-mail is spam.

(ii) P(spam) is a probability that any given message is spam.

(iii) P(wordspam) is probability that the particular word appears in a spam message.

(iv) P(non − spam) is the probability that any particular word is not spam.

(v) P(word − spam) is the probability that the particular word appears in the non-spam message. To achieve the objective, the research and procedure are conducted in three phases. The phases involved are as follows:

The Evaluation Metric

Evaluation metrics are used to evaluate the performance of WEKA tool based on SPAMBASE dataset that had been chosen. The most simple measure is filtering accuracy namely percentage of messages classified correctly. Table 1 shows the evaluation measures for spam filters.

Table 1. Evaluation measure for spam filters

Evaluation Measure	Evaluation Function
Accuracy	$Acc = \frac{TN+TP}{TP+FN+FP+TN}$

4. RESULTS AND DISCUSSION

For this results we have been used an inbuilt dataset of weka Then we ran KMean clustering algorithm which is already defined in weka and then ran KMean updated on the same dataset, i.e., email dataset. By running both the algorithms on the same dataset, we came to know that our algorithm runs with more efficiency and robustness on the dataset. With efficiency, we mean that its processing speed is faster than the traditional KMean algorithm. With robustness, we mean that our proposed algorithm can work efficiently with large datasets as compared to traditional Kmean.

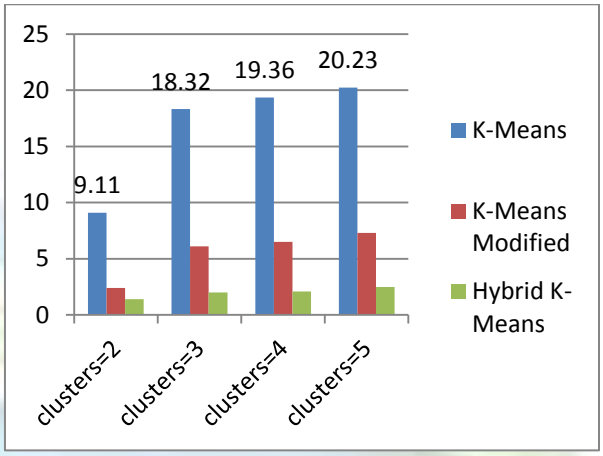

Fig 3. Time comparison

Email Spam with Naïve Bayes algorithm

The Accuracy, which refers to the proportion of emails classified as accurate type in the total emails. Accurate circumstances are True Positive (TP) and True Negative (TN), while false detected situations are False Positive (FP) and False Negative (FN). The accuracy of the system is calculated by the following equation:

Accuracy= ((TP+TN)/(TP+TN+FP+FN)) X 100

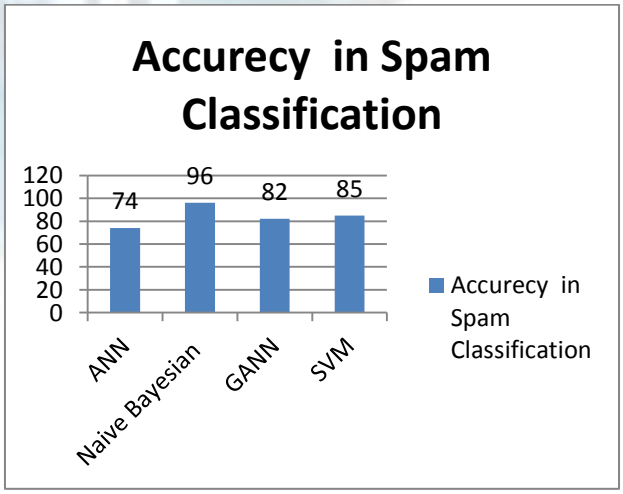

Fig 4. Accuracy in spam Classification

 DOI:10.22362/ijcert/2016/v3/i12/4321

Once the training completes, we observed that the classifier gets a training accuracy of about 96% and a test accuracy of about 97.23%.

5. CONCLUSION:

The proposed algorithm, i.e., K-Means Updated emphasizes the optimum utilization of resources while calculating KMeans. Processing speed increases, so processing time reduces. Comparatively, large dataset can be processed. With the analysis of the K-Mean algorithm, we have found that we can try to improve its speed or increase its efficiency by using our approach and moreover the algorithm can be enhanced to deal with very large datasets. E-mail spam filtering is an important issue in the network security and machine learning techniques; Naïve Bayes classifier that used has a very important role in this process of filtering e-mail spam. The quality of performance Naïve Bayes classifier is also based on datasets that used. Naïve Bayes classifier also can get the highest precision that gives most top percentage spam message manage to block if the dataset collects from single e-mail accounts.

REFERENCES:

[1] Marek Rychly, Pavlina Ticha, "A tool for clustering in data mining", International Federation for Information Processing, 2007.

[2]P.Verma, D.Kumar, "Association Rule Mining Algorithm's Variant Analysis", International Journal of Computer Application (IJCA), vol. 78, no. 14, September 2013, pp. 26–34.

[3]L.Firte, C.Lemnaru, R.Potolea, "Spam Detection Filter using KNN Algorithm and Resampling", 6th International Conference on Intelligent Computer Communication and Processing- IEEE, 2010, pp.27-33. [4] G.Kaur, R.K.Gurm, "A Survey on Classification Techniques in Internet Environment", International Journal of Advance Research in Computer and Communication Engineering, vol. 5, no. 3, March 2016, pp. 589–593.

[5] Rushdi, S. and Robet, M, "Classification spam emails using text and readability features", IEEE 13th International Conference on Data Mining, 2013.

[6] Androutsopoulos, I., Paliouras, G., and Michelakis, "E. Learning to filter unsolicited commercial e-mail", Technical report NCSR Demokritos, 2011.

[7]Na shi, "Research on k-means clustering algorithm", 3rd international symposium on intelligent information technology and security informatics, 2011.

[8] Shah Sourabh, Singh Manmohan, "comparison of a time efficient modified k-mean algorithm with k-mean and kmedoid algorithm" international conference on communication systems and network technologies, 2012.

[9] Boomjia M.D, "Comparison of partitioning based clustering algorithms".

[10] Han kwai, "Approximate distributed k-means clustering over a peer-to-peer network", IEEE transactions on knowledge and data engineering, 2009.

[11]Tariq, M., B., Jameel A. Tariq, Q., Jan, R. Nisar, A. S., "Detecting Threat E-mails using Bayesian Approach", IJSDIA International Journal of Secure Digital Information Age, Vol. 1. No. 2, December 2009.

[12]ML & KD- Machine Learning & Knowledge Discovery Group. http://mlkd.csd.auth.gr/concept drift.html.

[13] Rizky, W. M., Ristu, S., Afrizal, D. "The Effect of Best First and Spreadsubsample on Selection of a Feature Wrapper With Naïve Bayes Classifier for The Classification of the Ratio of Inpatients". Scientific Journal of Informatics, Vol. 3(2), p. 41-50, Nov. 2016.

[14]Feng, W., Sun, J., Zhang, L., Cao, C. and Yang,Q., "A support vector machine based naive Bayes algorithm for spam filtering," 2016 IEEE 35th International Performance Computing and Communications Conference (IPCCC), Las Vegas, NV, 2016, pp. 1-8.

[15] Lalchand G. Titare1, Prof. Riya Qureshi," Cloud Centric loT Based Farmer's Virtual Market place" International Journal of Computer Engineering In Research Trends., vol.3, no.12, pp. 654-658, 2016.

 DOI:10.22362/ijcert/2016/v3/i12/4321