ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

Outlier Detection and Analysis of Data Stream Classification Using Text Clustering

Neethu S[1], Sajni Nirmal[2]

M.Tech Student, Dept. of. CSE, Marian Engineering College, Trivandrum, Kerala, India[1]

Assistant Professor, Dept. of CSE, Marian Engineering College, Trivandrum, Kerala, India[2]

ABSTRACT: In fast growing technology, the emerging topic detection and classification is most important in social data streams. Here, we provide a new method for discovering emerging topics in social network stream to detect the anomalies in the social network based on links between the users that are generated dynamically. In this work Dynamic Threshold Optimization algorithm is used to detect outliers in streaming data. To find outliers dynamically by using various clustering methods. This paper also deals with efficient data mining procedure for medical records of patients. Here, we applied the data mining technique like text clustering methods to clustering the content present in the dataset. To provide privacy preservation LKC algorithms and compatibility checks are used. The performance analyses carried out by attributes present in the dataset with the number of clusters are generated.

KEYWORDS: Outlier detection, Social Network Stream, Text clustering methods, Privacy preservation.

I. INTRODUCTION

Social networking is the active online trend of the last few years. In particular, we are interested in the problem of identifying newly generated titles from social streams, which can be used to create automatically, exclusive news and other related details [1]. Detection of anomalies over streaming data is active research field from data mining that aims to detect patterns which have different behaviour, exceptional than normal behaviour [5].

Anomalies are also referred to as outliers that are does not exploit their actual behaviour. Clustering is the well-known procedure with successful applications on large area for finding patterns. Finding groups of objects will be same as one another or matched and differ from or unrelated to the objects in other groups are known as cluster analysis [4].

A hierarchical clustering method construct tree or taxonomy over the data. Agglomerative or bottom up approach and divisive or top down are two major types of hierarchical clustering methods. The problem of privacy-preserving data mining [3] has become more important issues in real life, because of upgrading the ability to save personal data about users, and their information.

II. RELATED WORK

In the area of topic detection and tracking [2], effectively detect and track the contents related to the various topics. The main task is to either classification or detection. The new documents are classified into one of the known topics. The detection process ensures that it belongs to none of the known categories. Data Mining – Clustering [5] discussing the motivation and aim of the clustering and classification. But there is not explained performance analysis based on argument factors.

Hierarchical agglomerative clustering [6] is our first example of a nonparametric, or instance-based, machine learning method. In many cases, association rule or classification rule mining are the results of data mining applications, which can compromise the secrecy of the data in theoretical challenges. In high dimensionality and Privacy-Preserving Data Publishing [3] are considered to perform privacy checking.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

III. PROPOSED METHOD

The proposed work consists of three different phases. They are training phase, working phase and text classification. In training phase, cluster management, class management, pseudo point management, edit features, search topics, and search title are generated. In working phase, existing class, new class and outliers are created. Partitional clustering method like K-means algorithm and dynamic threshold optimization are used to detecting the anomalies.In Text classification phase various text clustering methods like c-mean algorithm, Hierarchical agglomerative clustering and single- linkage algorithm are used to classify and clustering the dataset. Figure 1 shows the system architecture of the proposed work.

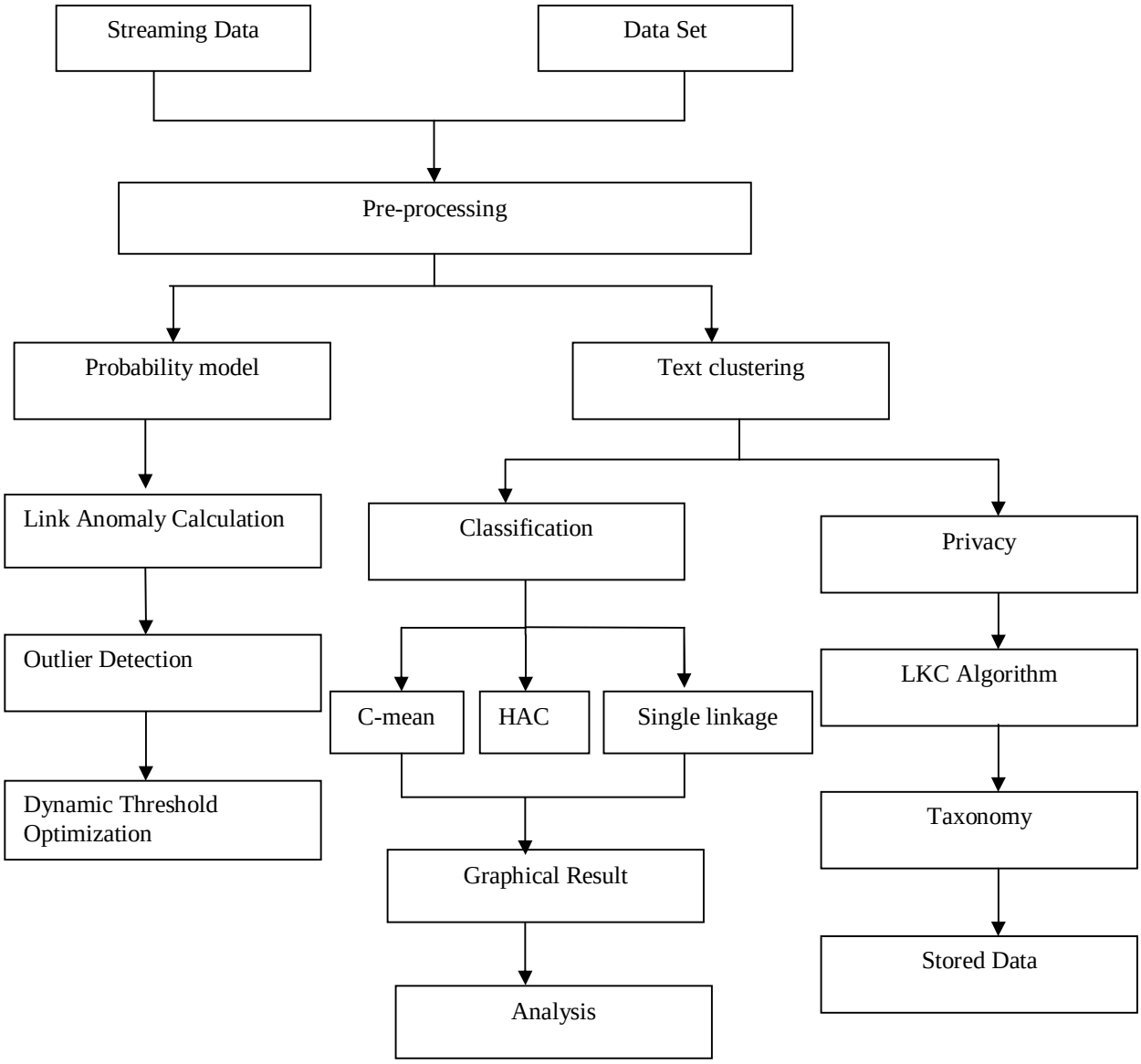

Figure 1: System Architecture

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

First, Pre-processing is done in streaming and dataset. The stream data are changing one. There is no exact structure, means any kinds of information are collected here. To assign the probability values and calculate the anomaly scores with help of k- means clustering method. Dynamic threshold optimization gives final resultant outlier by performing threshold comparison.
To create the graphical result of dataset by applying the various clustering algorithms like c-mean, HAC, and single linage. For providing privacy preservation of content in dataset uses a LKC algorithm which perform based on compactable taxonomy. The final result will be stored in a file.

A. Description Of Algorithms

For detecting outliers from the stream of data by the use of k-means clustering methods. In k-means, first calculate centroid or centre point associated with each cluster. C-mean allows centroid to be assigned into more than one cluster. Membership function of each cluster expresses the degree to which individual data points belong to the cluster. Hierarchical Agglomerative clustering [5] beginning with only one element sets. Merging them produce final taxonomy obtained.

HAC Pseudo code:

1) Start with all instances in their own cluster.
2) Until there is only one cluster:
3) Among these clusters, determine the two clusters, c_i and c_j, which are most similar.
4) Replace c_i and c_j with a single cluster $c_i \cup c_j$.

In single-link or single linkage hierarchical clustering, two clusters whose two closest members have the minimum distance merge together or the two clusters with the smallest minimum pair wise distance.

IV. PERFORMANCE ANALYSIS

The performance analysis can be made on medical records of patients in hospital dataset. Dataset consists of 18 attributes related to patient's medical treatment. This is a well known real life dataset, contains personal details of patients. For classification of dataset, we use three different algorithms like c-mean, HAC, and single-linkage algorithm. Based on number of clusters are formed with respect to each attribute, we can determines which is the very efficient algorithm for processing.

		NUMBER OF CLUSTERS GENERATED		
SL.NO:	ATTRIBUTES	C-MEAN	HAC	SINGLE LINKAGE
1	Hospital	2	3	3
2	Admission date	2	2	3
3	Admission diagnosis	2	17	3
4	History of present illness	2	2	3
5	Medication on admission	2	8	3
6	Hospital course	2	2	3
7	Past medical history	2	9	3
8	Allergies	2	9	3
9	Social history	2	12	3
10	Family history	2	4	3
11	Physical examination	2	6	3
12	Laboratory date	2	4	3
13	Discharge date	2	2	3

Table 1: Attributes Vs Number of clusters

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

The tabular representation gives the relationship between numbers of clusters generated and attributes present in dataset. Analysis performed based on this Table1 as shown in figure 2.

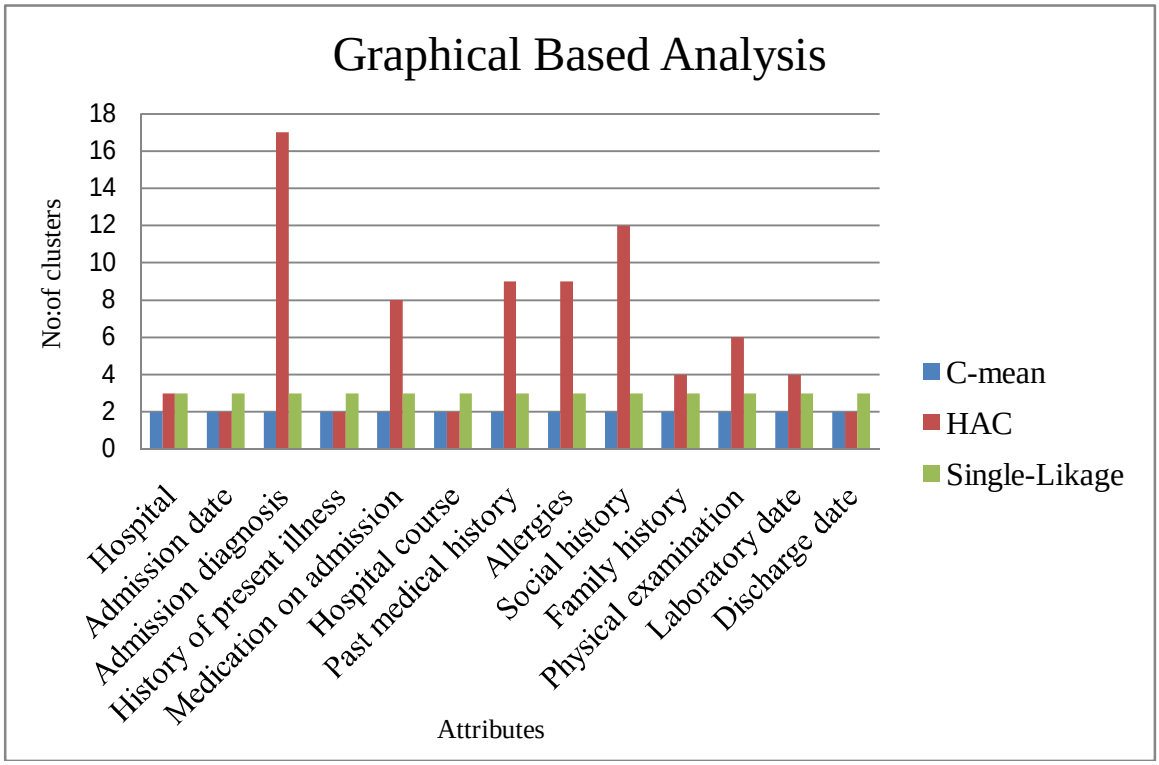

Figure 2: Graphical Based Analysis

V. CONCLUSION AND FUTURE WORK

From the experimental result, the graphical based analysis shows HAC is most efficient algorithm for processing the text classification of dataset. Here the C-mean clustering technique is not a deterministic one and also the single linkage algorithm only considered minimum shortest distance between two clusters. The single linkage method sensitive to detect noises and anomalies and it produces long, elongated clusters. As the performance of the proposed algorithm is analyzed between two parameters. In future with some modifications in design considerations the performance of the proposed algorithm can be compared with other attributes.

REFERENCES

[1] Toshimitsu Takahashi, Ryota Tomioka, and Kenji Yamanishi, "Discovering Emerging Topics In Social Streams Via Link Anomaly Detection", IEEE Transactions On Knowledge And Data Engineering, Vol.26, No.1, January 2014.
[2] J.Allan et al.,"Topic Detection and Tracking Pilot Study: Final Report", Proc.DARPA Broadcast news transcription and understanding Workshop, 1998.
[3] Agrawal R., Srikant.R,"Privacy-Preserving Data Mining", Proceedings of the ACM SIGMOD Conference, 2000.
[4]A. K. Jain and M. N. Murty and P. J. Flynn, "Data clustering: a review", ACM Computing Surveys, 31:3, pp. 264 - 323, 1999.
[5] Jerzy Stefanowski, "Data Mining – Clustering",IEEE Trans. Knowledge Discovery from Data, vol. 4, no.2, article 9, 2011.
[6]Ryan P. Adams, "Hierarchical Agglomerative Clustering", Proc. Ninth SIAM Data Mining Conf. (SDM '09), pp. 510-516, 2009.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

BIOGRAPHY

Neethu S received the B.Tech degree in Computer Science and Engineering from Marian Engineering College in 2014. Currently doing M.Tech degree in Computer Science and Engineering under Kerala University.

Sajni Nirmal received M.S in Computer Science Engineering from Technical University of Eindhoven, Netherlands in 2005 and working as assistant professor in Department of Computer Science and Engineering at Marian Engineering College, Kerala University.