ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

Time Based Analysis on Anomaly Detection and Classification of Data Stream

Neethu S[1], Sajni Nirmal[2]

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

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

ABSTRACT: In the rapid growth of social network, discovering emerging topics and classification of data is most important challenging factor. In social network stream, we can detect anomalies and emerging topics based on links between the users that are generated dynamically. For the emerging topic detection purpose to propose a new method in the area of streaming data. To find anomalies in social streams by using various clustering methods. This paper also implements the data mining technique like text clustering methods to clustering the dataset which contains medical records of patients. Here, we applied Hierarchical text clustering methods like C- Mean, Hierarchical Agglomerative clustering, and single linkage algorithms are used for clustering and classification of the dataset. LKC algorithm and compatibility is introduced to provide privacy preservation. The performance analysis carried out by time taken for the clustering and classification of various clustering algorithms based on attributes present in the dataset.

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

I. INTRODUCTION

Social networking is the effective online service trend of the last few years. Social networking sites allow users to share ideas, posts, activities and interests with people in their network. Mainly, people who interested in the problem of identifying freshly generated topics from social streams, which can be used to create automatically, exclusive news, discovering hidden market needs, detecting underground political movements and other political related details [1]. In the field of social data mining, current research works are concentrating more on achieving above listed challenges. All these challenges are comes under the category of detecting emerging topics from social media. The information exchanged over social networks is texts, URLs, images, and videos etc.

Anomaly detection over streaming data is active research area from data mining that aims to detect patterns or objects which have different behaviour, exceptional than normal behaviour [5]. Outliers are also known as anomalies that are exploit their abnormal behaviour. Clustering is the well-known technique with useful applications on huge area for finding patterns. To identifying set of objects will be similar to one another or related and differ from or unrelated to the objects in other sets are known as cluster analysis [4]. The probability model is used to mentioning the behaviour of a database user, and to detect the emergence of a new topic from the anomalies measured through the model.

A hierarchical clustering method create tree structure or taxonomy over the data. Major types of hierarchical text clustering technique are divisive or top down approach and agglomerative or bottom up approach. Different organizations like governmental and self financing agencies, hospitals, and financial institute collect and disseminate various specific data about multiple persons. In data mining, the privacy-preservation[3] has become more important issues in real life, because of increasing the ability to store personal data about users they are related to various governmental agencies, financial institutions or hospitals, and their information.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

II. RELATED WORK

For topic detection, a finite mixture model is used in early days. Finite mixture model [6] is a weighted average of a number of probabilistic approaches. This framework works well in dynamic topic trends tracked in a timely fashion. Mainly three methods are used for topic detection.

- Topic structure identification
- Topic emergence detection
- Topic characterization

In topic structure identification, identifying different kinds of topics exists and what are main one and importance and relevance of that topic. In topic emergence detection, emergence of a new topic are detected and recognizing how it grows and detecting the disappearance of an existing topic. In topic characterization**,** to identifying the characteristics for each of main topics. The main draw backs are:

- Context-based topic trend analysis: To analyse contexts, i.e., relations among words, in order to more deeply analyse the semantics of topics.
- Multi-topics analysis: one text comes from a single mixture component corresponding to a single topic and do not consider multi topics.

A unifying frame work model [8] is used for detection purpose. Detecting outliers and change points from time series is the main task. Detection of change point was used to reduce the issue of detecting outliers in that time series. Change point detection under the requirements of:

- The detection process should be online; that is, an outlier or change point should be detected immediately after it appeared.
- The detection should be adaptive to non-stationary data sources; that is an outlier or change point should be detected even if the nature of data source changes overtime.

The framework consists of two parts

- Data modelling: learn a probability density function from a data sequence.
- Scoring: give a score to each data or each time point.

The main advantage is change points from non stationary are much more efficient than conventional methods.

Temporal Text Mining (TTM) [7] task discovering and summarizing the evolutionary patterns of themes in a text stream. Here we use a probabilistic method for solving the problems through

- To discovering patterns of themes in text information together over time.
- Analysing the life cycle of each theme.
- Determine the globally attractive themes.
- Compute the strength of a theme in each time period.
- These methods are applicable to any text stream data.

The main drawback is not considered flat structure theme.

The Topic detection and tracking (TDT) [2] problem consists of three major tasks:

- Stream of data are segmented or divided, especially recognized speech, into distinct stories.
- Finding those news stories that are the initially discuss a new event occurring in the news.
- First, given a small number of sample news stories about an event, finding all following stories in the stream.

Topic Detection and Tracking (TDT) Pilot Study provide advance and accurately measure the state and to assess the technical challenges to be overcome.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

III. PROPOSED METHOD

Mainly three phases are proposed here. First phase is training phase. Second phase is working phase and last one is text classification of dataset. Cluster management, class management, pseudo point management, edit features, search topics, and search title are created in training phase. In working phase, existing class, new class and outliers are created. One of the clustering methods 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.

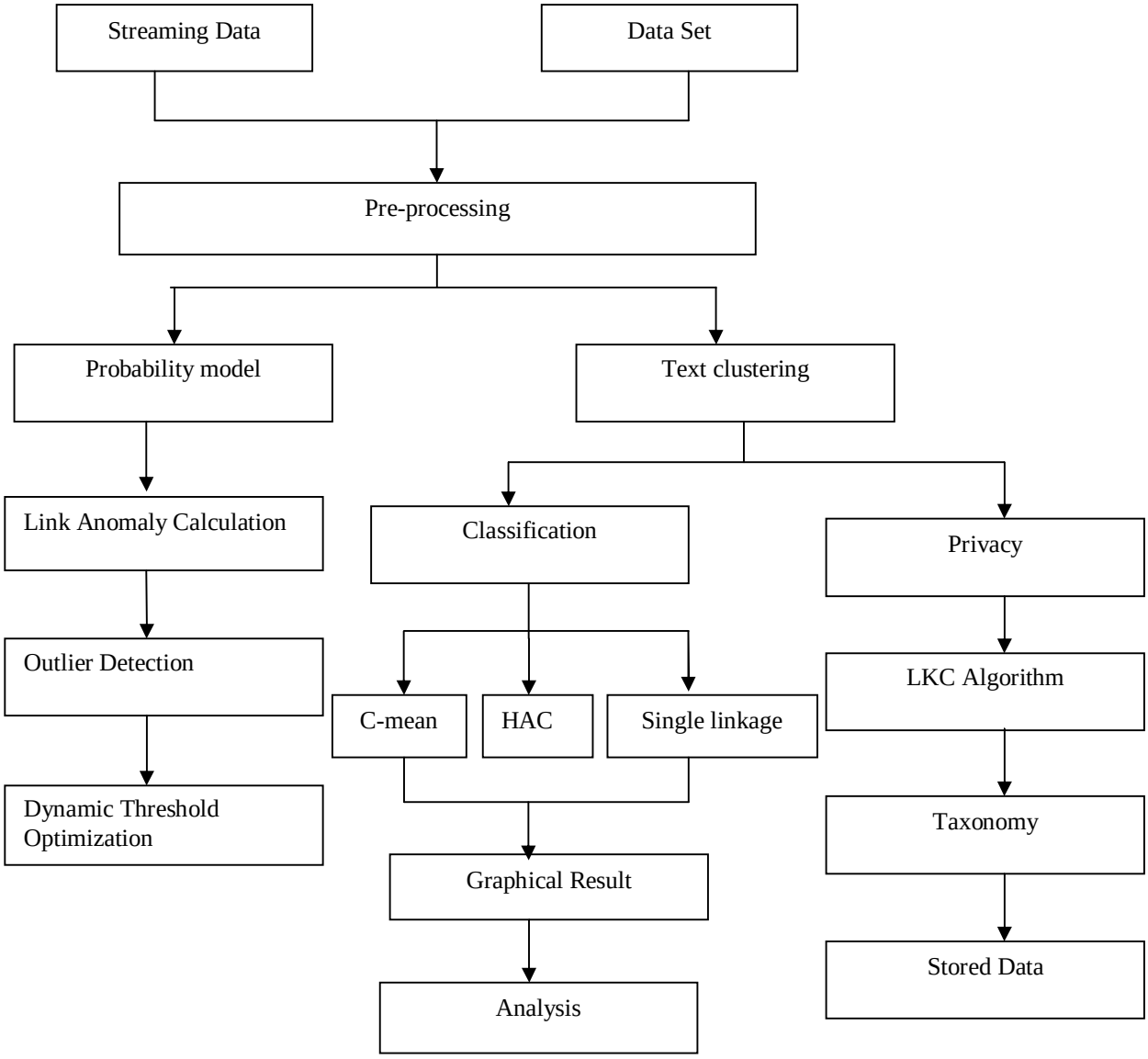

Figure-1 System Architecture

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

Pre-processing is done in streaming and dataset. The stream data has no exact structure, means any kinds of information are stored 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.

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.

Table 1- time Vs attributes

		TIME IN MILLISECONDS		
SL.NO:	ATTRIBUTES	C-MEAN	HAC	SINGLE LINKAGE
1	Hospital	109	264	118
2	Admission diagnosis	322	8946	12197
3	Admission date	225	1900	1130
4	History of present illness	1062	1894	587
5	Past medical history	1970	7993	12477
6	Allergies	614	9708	14075
7	Social history	501	1225	336
8	Family history	564	9035	13269
9	Medication on admission	598	444	1589
10	Physical examination	1471	37405	5790
11	Laboratory date	2740	18127	26480
12	Hospital course	3868	14718	22416

The tabular representation gives the relationship between Times generated for processing the various algorithms based on the attribute present in the data sets. The time in milliseconds used for generating the cluster levels. Analysis based on this Table 1 is shown in Figure 2.

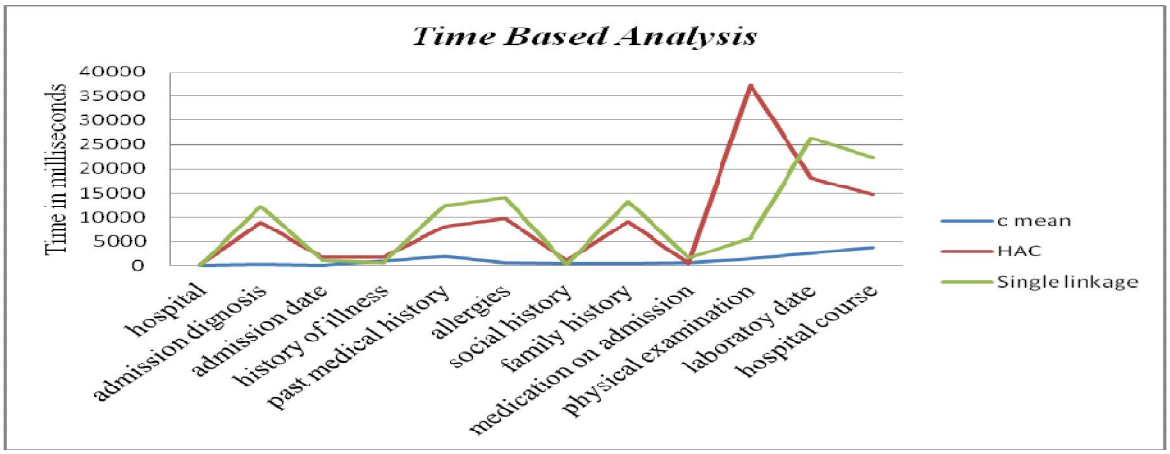

Figure 2: Time Based Analysis

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 6, June 2016

V. CONCLUSION AND FUTURE WORK

From the experimental result, the time based analysis shows C- mean is most efficient algorithm for taking less time for processing the text classification of dataset. The numbers of clusters are generated in less amount of time by c-mean algorithms. Single linkage and hierarchical agglomerative algorithms are takes more time compared to c-mean algorithm. As the performance of the proposed algorithm is analysed between two parameters. Here c- mean takes less than 5000 milliseconds for processing the datasets. In future with some modifications in design considerations the performance of the proposed algorithm can be compared with 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 understandingWorkshop, 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] S. Morinaga and K. Yamanishi, Tracking Dynamics of Topic Trends Using a Finite Mixture Model", Proc. 10th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 811-816, 2005.
[7]Q Mei and C. Zhai , "Discovering Evolutionary Theme Patterns from Text: An Exploration of Temporal Text Mining," Proc. 11th ACM SIGKDD Int'l Conf. Knowledge Discovery in Data Mining, pp. 198-207, 2010.
[8]Jun- ichi Takeuchi and Kenji Yamanishi , "A Unifying Framework for Detecting Outliers and Change Points from Time Series ", IEEE Transactions On Knowledge And Data Engineering, Vol. 18, No. 4, April 2006.

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.