Out Lier Detection and Clustering Analysis in Data Stream Classification

Neethu S[1], Sajni Nirmal[2]

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

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

Abstract: ***In social network stream, we can detect anomalies and emerging topics based on links between the users that are generated dynamically. In the rapid growth of social network, discovering emerging topics and classification of data is most important challenging issues. For the emerging topic detection purpose to propose a new method in the area of streaming data. Here, Dynamic Threshold Optimization algorithm is used to detect anomalies in streaming data. Anomaly detection refers to detecting clusters or objects in a given data set that conform to an established abnormal behavior. This classes and object are called anomalies or outliers that are critical information in several application domains. 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 different clustering algorithms posses processing of data in stream and dataset.***

Keywords: Dynamic threshold optimization, Outlier detection, Privacy preservation, Text clustering methods

1. 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 behavior, exceptional than normal behavior [5]. Outliers are also known as anomalies that are exploit their abnormal behavior.

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 behavior 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.

2. 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 analyze contexts, i.e., relations among words, in order to more deeply analyze 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.

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.

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.
- Analyzing 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.

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 modeling: 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.

3. Proposed Method

Training phase, Working phase and Text classification are three main phases in the proposed work Training phase consist of Cluster management, class management, pseudo point management, edit features, search topics, and search title. 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.

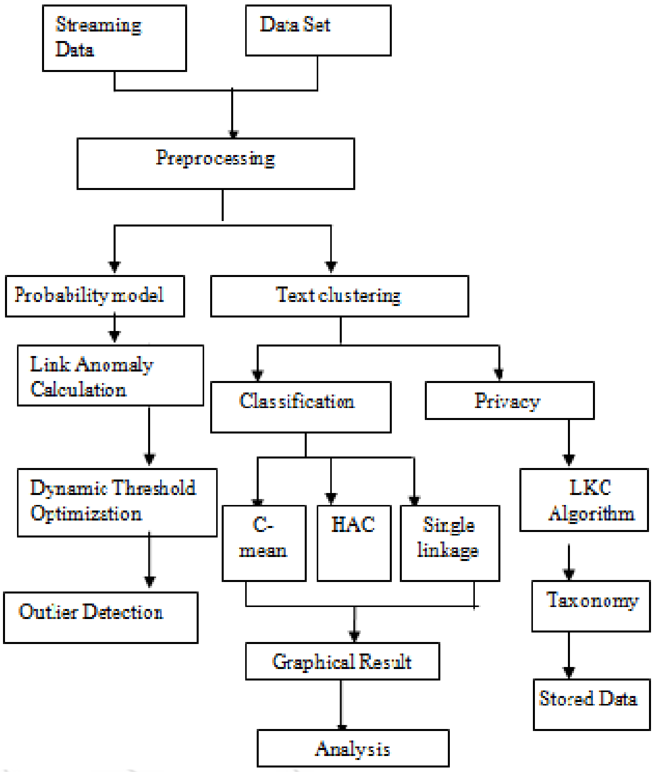


Figure 1: System Architecture

Preprocessing 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.

To create the graphical result of dataset by applying the various clustering algorithms like c-mean, HAC, and single linkage. 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.

C- Mean algorithm works by assigning membership to each data point corresponding to each cluster center on the basis of distance between the cluster center and the data point. More the data is near to the cluster center more is its membership towards the particular cluster center. The main advantages of this algorithms gives best result for overlapped data set and comparatively better then k-means algorithm. Unlike k-means where data point must exclusively belong to one cluster center here data point is assigned membership to each cluster center as a result of which data point may belong to more than one cluster center. The main disadvantage is Euclidean distance measures can unequally weight underlying factors.

In the beginning of the agglomerative clustering process, each element is in a cluster of its own. The clusters are then sequentially combined into larger clusters, until all elements end up being in the same cluster. At each step, the two clusters separated by the shortest distance are combined. The definition of 'shortest distance' is what differentiates between the different agglomerative clustering methods. Two types of hierarchical clustering technique are agglomerative and divisive. Agglomerative is a "bottom up" approach, here each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy. Divisive is a "top down" approach, here all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy..

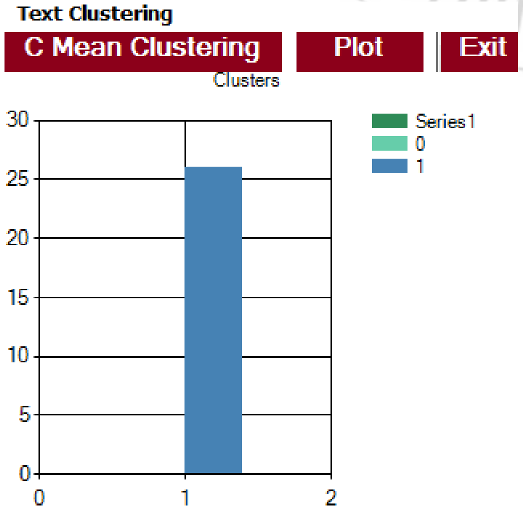
Single-linkage clustering is based on grouping clusters in bottom-up fashion (agglomerative clustering) , at each step combining two clusters that contain the closest pair of elements not yet belonging to the same cluster as each other. In single-linkage clustering, the distance between two clusters is determined by a single element pair, namely those two elements (one in each cluster) that are closest to each other. The shortest of these links that remains at any step causes the fusion of the two clusters whose elements are involved. The method is also known as nearest neighbor clustering. A drawback of this method is that it tends to produce long thin clusters in which nearby elements of the same cluster have small distances, but elements at opposite ends of a cluster may be much farther from each other than to elements of other clusters. This may lead to difficulties in defining classes that could usefully subdivide the data.

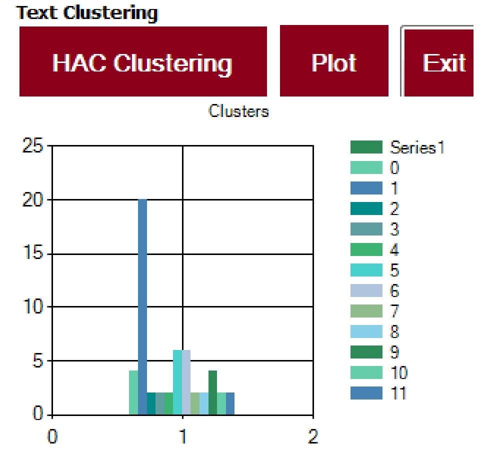
4. Performance Analysis

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.

Consider the various types of attributes and their performance in each of the clustering algorithms.

- **Admission Diagnosis**





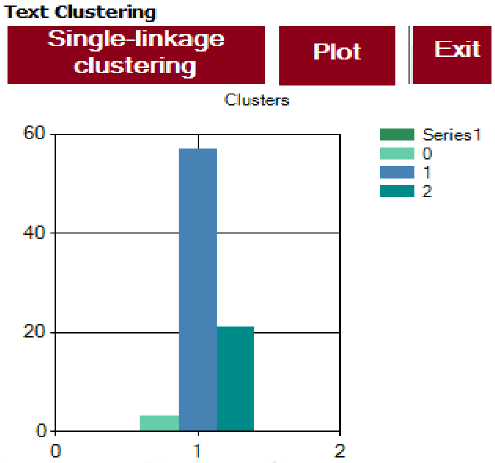
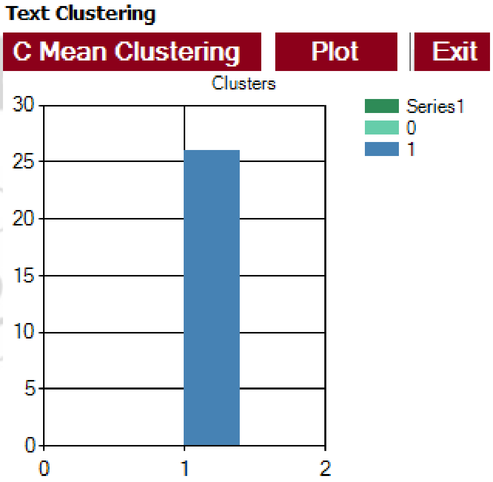
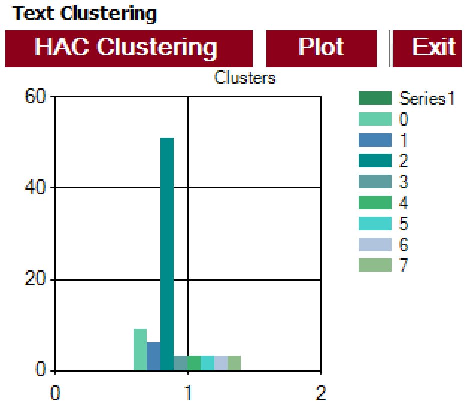


Figure 2: screenshots of admission diagnosis

- **Medication on Admission**





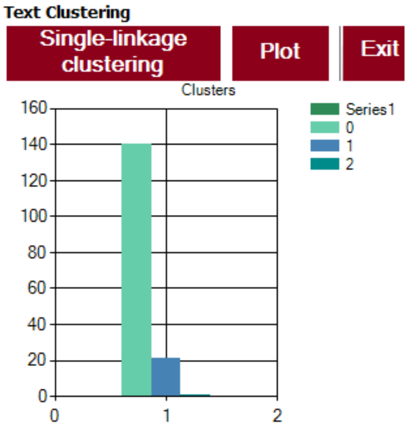


Figure 3: Screenshots of medication on admission

- **Physical Examination**

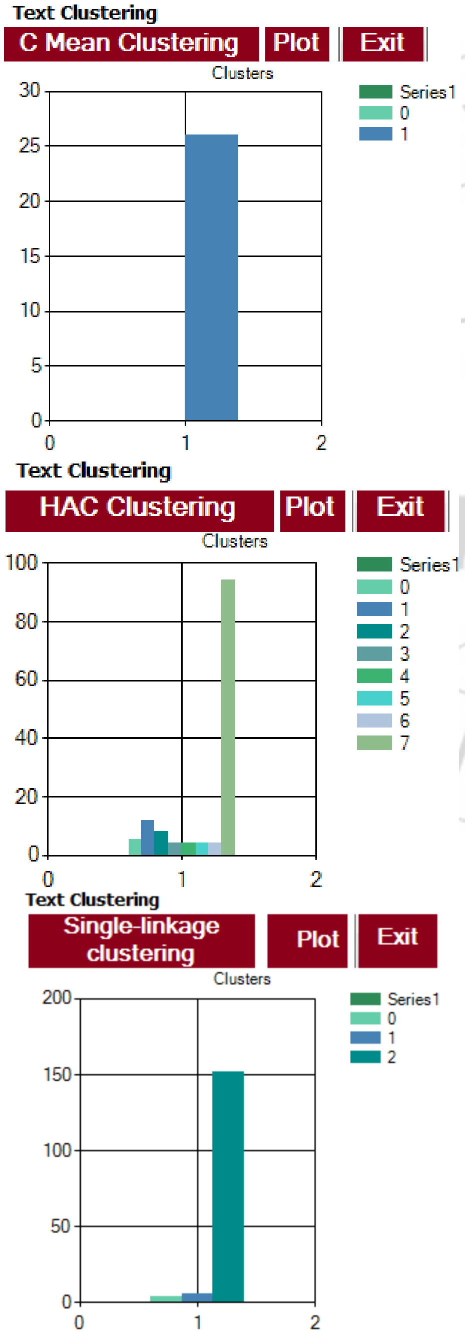


Figure 4: Screenshots of physical examination

5. Conclusion and Future Work

Analysis of various algorithms shows HAC is most efficient algorithm for processing the text classification of dataset. For analysis, we consider three attributes present in the dataset which contains medical records of patients like admission diagnosis, medication on admission and physical examination. Here the C-mean clustering techniques very less number of clusters are generated. But in HAC method huge numbers of clusters are formed. So more number of classifications and clusters are created in HAC. Clustering and classification process are very effective in hierarchical agglomerative clustering.

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] 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.

Author Profile

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.