Handling the varying classes of data in news feed

Ashwini Umarjikar[1], Janhavi Sonar[2], Prathyusha Rao Muthineni[3], Sheetal Shevate[4]

[1] *Ashwini Umarjikar Student, Computer Engineering, K.K.Wagh Institute of Engineering, Maharashtra, India*
[2] *Janhavi Sonar Student, Computer Engineering, K.K.Wagh Institute of Engineering, Maharashtra, India*
[3] *Prathyusha Rao Muthineni Student, Computer Engineering, K.K.Wagh Institute of Engineering, Maharashtra, India*
[4] *Sheetal Shevate Student, Computer Engineering, K.K.Wagh Institute of Engineering, Maharashtra, India*

ABSTRACT

Clustering of text data stream has a huge emphasis on data mining. Clustering of text stream used in news group filtering, document organization and category definition and clustering etc. Concept drift is the underlying distribution of the data that is varying. Previous systems have used live concept drifting detection methods which have used error rate classification. To overcome the drawbacks of those previous methods we have new efficient clustering method using three layer concept drift in text data streams. This new method adapt to real time changes rapidly and rigorously. In this, three layer indicate the layer of label space, the layer of feature space and the layer of mapping relationships between labels and features respectively.

Keyword : - *Clustering, concept drift, supervised learning, weight of evidence and information value*

1. Introduction

Text data stream such as news articles, customer reviews are in huge volume that are continuously growing and spreading in real time. In the new method we are grouping text data in clusters as per their classes using supervised learning. Clustering of text data stream is grouping similar documents and also adapt to the changes that are occurring in the real time data. It is not useful to store every news data, so it was necessity to design algorithm which only stores summary of just previous data. So, in this proposed method we have designed algorithm which stores summary of just previous data. And the proposed method is three layer concept drift detection method for clustering news data streams.

While clustering, we have used three layer concept drift method. Concept drift refers to, detecting the continuously changing cluster of text data streams. Changes include different features, different conditions and bidirectional changes. Clusters are created, updated and merged. It is applied to changes over real time. It have trending importance, as mostly dynamic data streams are taken into consideration for many application. It properly clusters the data by using knowledge and by detecting its unexpected class. In our three layer concept drift detection method, first layer is about clustering similar types of text data streams under same cluster. Second layer is about updating clusters centroid and the third layer is the most complex one which is about merging the two clusters which are being pointed by multiple documents of same type.

According to occurrence of concept drift detection, there are two types-

1) Sudden concept drift- It shows abrupt changes in concept drift and changes labels and feature space most frequently.

2) Gradual concept drift- It shows slow changes in concept drift than sudden concept drift.

Previous systems were based on change of error rate, since the change of error rate is a gradual process it is difficult to reflect all the changes in error rate. As text data stream are sparse and dimensions are large, it makes detection based on error much difficult because of the deviation of error rates is high.

In this new method it first detects concept drift then classifies and categorize text data stream according to triggers of concept drifts instead of their velocity. It is independent of classifiers and adapt to new data chunk quickly and has less missing detection. It works better, where concept drift occur frequently.

2. Existing methods

There were many existing approaches based on error rates of classification .Wang et al [2] detected concept drifts based on the error variance of concept similarity between the classifier and the data set. Gama et al [3, 4] proposed the DDM (drift detection method) concept drift detection method and used a threshold to determine whether the concept drift occurs or not. Baena-Gracia et al [5] put forward EDDM (Early Drift Detection Method) developed from DDM. It aggregates the ability to detect the gradual drift by ensuring the efficiency to detect the abrupt drift. Severo and Gama [6] proposed a detection system for regression problems, which is composed of three components: a regression predictor, a kalman filter and a Cumulative Sum of Recursive Residual (CUSUM) change detector. CUSUM is effective in detecting changes but it requires the availability of probability distributions which makes it inapplicable to online scenarios.

3. New method

In this section, we firstly divide the concept drifts in a text data stream into three categories, and then will apply a three-layer concept drifting detection approach based on the three categories. We will detect concept drifts according to the three-layer model.

3.1 Weight of Evidence and Information Value

WoE (Weight of Evidence) is a quantitative analysis method combining instances based on specific labels, and the WoE value is called ”contribution weight”. Given a feature F= { f 1 }and the label L= {0, 1}. WoE can only measure the contribution to the label for a feature value. Hence the cluster index is proposed to process a feature. The cluster index is an effective index for feature selection. Firstly, the computation mainly consists of counting which costs less time. Secondly, it contains the information of different granularity, such as cluster matrix, cluster vector and WoE data cube. Thus the drill-down and roll-up operations can be used to express the importance of feature or feature values in different granularity.

When the cluster value of a feature falls in [0,0.1], the feature has low contribution to the classification; when it falls in [0.1, 0.3], the feature’s contribution is medium; when falls in [0.3,], the contribution is high. In this method, we use the cluster index to detect concept drifts. Given a text stream $T = \{T_1, ..., T_m\}$, where T_i refers to the i-th data chunk in the text stream. In addition, we note the feature space of T_i as $F_i = \{ f_j \}$ and the label space as $T_i = \{l_k\}$.

3.2 Kinds of concept drift

Firstly, we consider concept drift detection from the label space. If L_i, which is the label space of T_i is different from L_{i+1} , we consider that there is a concept drift occurring which can be denoted as A-drift. An A-drift can be detected by a simple comparison of the label spaces of two data chunks. Moreover, we do not need to consider changes in the feature space and others for A-drift.

Secondly, we analyse concept drifts which is caused by the change of the feature space denoted as B-drift, which means that F_i is different from F_{i+1}. To measure the change of feature values, we need to select the informative features in each data chunk, and then compare the similarity between F_i and F_{i+1}.

Lastly, we concern concept drifts caused by the change of mapping relationship of labels and features, namely C-drift. In this case, the change of mapping relationship includes two aspects below. One is that feature f_i refers to l_0 in S_i, while $f_i \cup f_j$ refers to l_0 in S_{i+1} . The other is that feature f_i refers to l_0 in S_i but it refers to y_1 in S_{i+1}. To detect this drift, it needs to measure the change in mapping relation between labels and features in two data chunks. Considering the first aspect, we can detect this drift through computing the contribution of an feature to a label and comparing the similarity the two informative feature sets.

Let us take the online news feed as an example. The readers might be concerned from "Sports" to "political" over time, and this concept drift is denoted as A-drift due to the change of labels. If the focus of readers on political news changes from "Nation" to "state", it can be considered as the change of feature distribution, denoted as B-drift. In C-drift we can classify sub-categories of news under main category.

We will do further see causes of concept drifts. Generally, the C-drift happens less than A- and B-drifts in real world applications. In a real-world application, a concept drift may be triggered by more than one type.

In summary, we divide concept drifts into A-, B- and C-drifts based on their incentives. A-drift is the simplest and C-drift is the most complex.

4. Architectural Diagram:

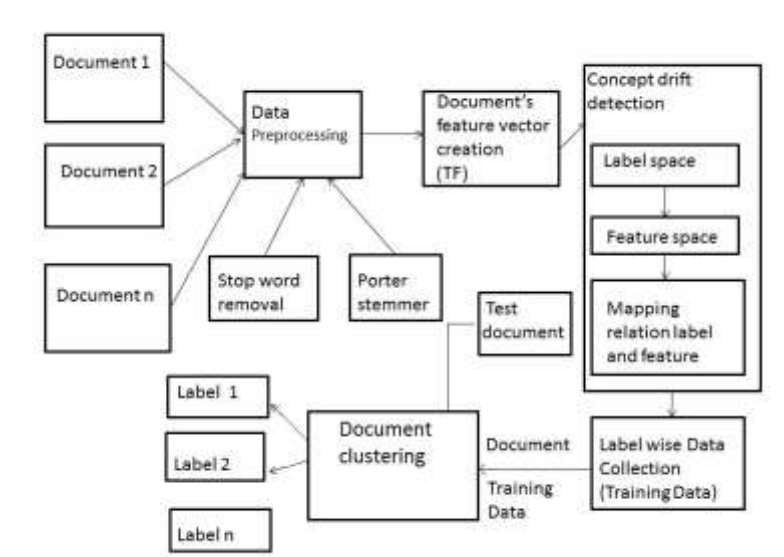

Fig: architectural diagram

4.1 Data pre-processing

4.1.1 Data cleaning

Data cleaning is the process of removing data in a database that is incorrect, incomplete, improperly formatted or duplicated. It also includes validation of the changes made and may require normalization. The goal of data cleaning is to achieve consistent, complete, accurate and uniform data.

4.1.2 Null value elimination

Stop words do not contribute to the context or content of textual documents. Due to their high frequency of occurrence, their presence in text mining presents an obstacle in understanding the content of the documents. So, we

remove the stop words from the documents. This process also reduces the text data and improves the system performance.

4.2 Data feature vector

4.2.1 TF/IDF calculations

In this process we create a document term matrix which gives us the information about unique words and the frequency of their occurrence in the document. It helps in calculating the frequency of the words that are present in each document in the given dataset. Tf-idf stands for term frequency-inverse document frequency and the tf-idf weight is a weight often used in information retrieval. The importance increases proportionally to the number of times a word appears in document but is offset by the frequency of the word in the dataset.

TF: Term Frequency, which measures how frequently a term occurs in a document. Thus, the term frequency is often divided by the document length (aka. the total number of terms in the document) as a way of normalization:

TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document).

IDF: Inverse Document Frequency, which measures how important a term is. Thus we need to weigh down the frequent terms while scale up the rare ones, by computing the following:

IDF(t) = log_e(Total number of documents / Number of documents with term t in it).

4.2.2 Document term matrix

A document-term matrix or term-document matrix is a mathematical matrix that describes the frequency of terms that occur in a collection of documents. In a document-term matrix, rows correspond to documents in the collection and columns correspond to terms. It becomes easy to evaluate the dataset with the help of document term matrix.

4.3 Data grouping

4.3.1 Data Clustering

Clustering of text data stream is grouping similar documents and also adapt to the changes that are occurring in the real time data. Clustering is one of the main techniques used for organizing documents to enhance retrieval and support browsing. The similarity is computed by using a similarity function. Text clustering can be in different levels of granularities where clusters can be documents, paragraphs, sentences or terms. Clusters are created, updated and merged.

4.4 Concept drift detection

4.4.1 Competitive learning

4.4.1.1 First Layer

In this layer, we aim to detect concept drifts caused by the change of the label space. Comparing two data chunks, if there are only a few instances with the same labels, we think that A-drift happens, otherwise there are no drifts. More specifically, for each label $l_k \in (L_i \cap L_{i+1})$, we count the numbers of instances with L_k in both chunks and sum up the smaller numbers of each l_k . All concept drifts caused by the changes of the label space will be found in this layer. Therefore we only need to consider the rest of data with the same labels in the next layer for detecting B-drift and C-drift.

$$\Phi 1(ni, ni+1) = \sum_{k=1}^{q} \min(nilk, ni+1lk)\,/n\,, \quad l_k \epsilon l_i/l_{i+1} \qquad \text{-------I}$$

where $n_i(l_k)$ indicates the number of instances with label l_k in Ti, a indicates the size of Li \ L_{i+1} and n is the number of instances in Ti.
According to Eq. I, if the value of 1 is less than the threshold t, we can determine that a concept drift has occurred. All concept drifts caused by the changes of the label space will be found in this layer.

4.4.1.2 Second layer

In this layer, we aim to detect concept drifts caused by the change of the feature space, namely B-drift. A concept drift will be detected by comparing the feature sets based on feature selection with other clusters. The informative features in two chunks are firstly selected, and then the similarity is calculated between the two feature sets. Especially, the cluster values of all features are computed and then we exclude those features whose cluster values are less than threshold. Then the similarity between the two feature sets is calculated according to

$\Phi 2(C_{i,}\ C_{i+1}) = Jac(F_i, F_{i+1}) / (F_i \cup F_{i+1})$ --------- II

When the value of Φ_2 is less than the threshold t, it can be considered that a concept drift has occurred, because features relevant to the classification in the two data chunks are not similar.

4.4.1.3 Third layer

We first select two feature sets by the cluster matrix which have the same label, and then compute the similarity of the two sets. Lastly, the average of each label can be used to detect drifts, and the definition is provided in Eq III,

$$\Phi 3(Ri, Ri+1) = (\sum_{k=1}^{a} \Phi 2(Ci|lk, Ci+1|lk))/l \quad , l_k \in l_i/l_{i+1} \quad \text{---------III}$$

where a indicates the size of $L_i \setminus L_{i+1}$.

When the value of 3 is less than the threshold t, it can be considered a C-drift occurs.

5. Conclusion

In this system we have classified news according to their types in real time system using three layer concept drift detection method. This method is independent of classifier. This new method improves efficiency by detecting the concept drift, layer by layer. In future work this model can be converted such that it will be able to operate on unlabeled data stream.

6. REFERENCES

[1]. Yuhong Zhang, Guang Chu, Peipei Li, Xuegang Hu, Xindong Wu, Three-layer Concept Drifting Detection in Text Data Streams, Neurocomputing (2017)
[2]. H.Wang,W. Fan, P. S. Yu, J. Han, Mining concept-drifting data streams using ensemble classifiers, in: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2003, pp. 226–235.
[3]. J. Gama, P. Medas, G. Castillo, P. Rodrigues, Learning with drift detection, in: Brazilian Symposium on Artificial Intelligence, Springer, 2004, pp. 286–295.
[4].J.Gama, G. Castilo, Learning with local drift detection, in: International Conference on Advanced Data Mining and Applications, Springer, 2006, pp. 42–55.
[5] M. Baena-Garcıa, J. del Campo-A´ vila, R. Fidalgo, A. Bifet, R. Gavalda, R. Morales-Bueno, Early drift detection method, in: Fourth international workshop on knowledge discovery from data streams, Vol. 6, 2006, pp. 77–86.
[6] M. Severo, J. Gama, Change detection with kalman filter and cusum, in: International Conference on Discovery Science, Springer, 2006, pp. 243–254.