International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

An Ensemble Classification Framework to Evolving Data Streams

Naga Chithra Devi. R

MCA, (M.Phil), Sri Jayendra Saraswathy Maha Vidyalaya, College of Arts and Science, Coimbatore, India

Abstract: ***Data stream classification poses many challenges to the data mining community. In this thesis, we address four such major challenges, namely, infinite length, concept-drift, concept-evolution, and feature-evolution. Since a data stream is theoretically infinite in length, it is impractical to store and use all the historical data for training. Concept-drift is a common phenomenon in data streams, which occurs as a result of changes in the underlying concepts. Concept-evolution occurs as a result of new classes evolving in the stream. Feature-evolution is a frequently occurring process in many streams, such as text streams, in which new features (i.e., words or phrases) appear as the stream progresses. Most existing data stream classification techniques address only the first two challenges, and ignore the latter two. In this thesis, we propose an ensemble classification framework, where each classifier is equipped with a novel class detector, to address concept-drift and concept-evolution. To address feature-evolution, we propose a feature set homogenization technique. We also enhance the novel class detection module using the Principle component analysis by making it more adaptive to the evolving Stream and enabling it to detect more than one novel class at a time with heterogeneous technique for novel data's. Comparison with state-of-the-art data stream classification techniques establishes the effectiveness of the proposed approach.***

Keywords: Information Retrieval, Data Classification, Outlier Detection, Novel Data extraction

1. Introduction

Technically, big data analysis is analysis of data mining and techniques .Novel mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Several types of analytical software are available: statistical, machine learning, and neural networks. As Novel contents keeps extending, the no. of pages crawled by the search engines is increases. With such large amount of data, estimating the relevant information satisfying the user query is a challenging task. Data prediction, Extraction and Alignment of big data from Novel databases is research area to obtain better mechanism and methodology to derive high precision and accuracy. Although many data extraction concepts such as [1], [2] and [3] have proposed in literature related to research area but they still lag in some measurement regarding the data mining properties like precision and recall measures etc. Therefore, it's a mandatory to ascertain the suitable solution for extraction and alignment of the big data. Another widespread application of Novel prediction is "personalization," in which users are categorized based on their interests and tastes [4]–[7]. In Novel prediction and Extraction, we face challenges in preprocessing, clustering, classification and prediction. In existing works, [8], [9], prediction model based on fusing several prediction models like Markov and SVM models has been utilized, even it fails to reduce the false positive rate. This exploitation has enabled us to considerably improve the prediction accuracy. In this paper, we introduce an efficient framework for Novel Data extraction and clustering mechanism to user query obfuscations to alleviate the issue of scalability, ambiguity and precision in the number of query suggestions (prediction) and Query Result Records (QRR)[10] as a clusters. In addition, the results indicate a dramatic improvement in prediction time for our objective. Moreover, the results demonstrate the positive effect of our proposed user specific clustering model in reducing the size of the prediction models through multi correlation factors estimations without compromising the prediction accuracy. Finally, we present experiments to study the effect of sparsity of pages, training partitioning, and ranking on the prediction accuracy. The advantages of this method novel structure of data results through options for aligning iterative and disjunctive data items to form Results sets of query. The rest of the paper is organized as follows: Sections 2 describes the related work of state of art methods about Novel data clustering and extraction with alignment technique, section 3 describes the overall framework with Methods and solution to achieve the Novel document clustering. Section 4 describes the experimental results of our method and performance measures with state-of-the-art methods. Section 5 concludes the paper and outlines possible future work.

2. Related Works

2.1 Data Collaboration based extraction and Content based Prediction

In Big data Analysis, Collaborative filtering approaches are the most popular prediction methods and are widely adopted in Data collaboration based extraction [11]. User-based approaches predict the ratings of active users based on the ratings of their similar users, and item-based approaches predict the ratings of active users based on the computed information of items similar to those chosen by the active user. However, on the Novel, in most of the cases, rating data are always unavailable since information on the Novel is less structured and more diverse. Query suggestion is closely related to query expansion or query substitution, which extends the original query with new search terms to narrow down the scope of the search. But different from query expansion, query suggestion aims to suggest full queries that have been formulated by previous users so that query integrity and coherence are preserved in the suggested queries [18]. Query refinement is another closely related notion, since the objective of query refinement is interactively recommending new queries related to a particular query.

2.2 Concept based mining and Click through Data Analysis

Concept based mining model [12][13] has also been utilized in big data community that analyzes terms on the sentence, document, function, dependency level and corpus levels is introduced. The concept Inclusion dependency clustering algorithm can effectively discriminate between non-important terms with respect to sentence semantics and terms which hold the concepts that represent the sentence meaning. The similarity between documents is calculated based on a new concept inclusion dependency measure. The proposed dependency measure takes full advantage of measures on the sentence, document, and corpus and function levels in calculating the dependency range between documents by the importance of dependency discovery, a method for discovering XML functional dependencies. Functional and inclusion dependency discovery is important to knowledge discovery, database semantics analysis and data quality assessment. In Click through data analysis, the most common usage is for optimizing Novel search results or rankings [10], Novel search logs are utilized to effectively organize the clusters of search results by learning “interesting aspects” of a topic and generating more meaningful cluster labels.. Besides ranking, click through data is also well studied in the query clustering problem [11]. Query clustering is a process used to discover frequently asked questions or most popular topics on a search engine.

3. Proposed Methodology

3.1 Establishing the Indexing of Data Warehouse for Evolution of Data from Different stream

A fundamental tool in construction of text classification is a list of 'stop' words (stop word list) that is used to identify frequent words that are unlikely to assist in classification and hence are deleted during pre-processing. Currently, we only remove English stop words (e.g., and, into, or will) as source code is almost exclusively written with English acronyms and comments. Till now, many stop word lists have been developed for English language. Then we use the cleaning filter to remove unnecessary punctuation characters like commas or semicolons at the start or end of the token that might have been inserted at formulas or (for example, name= or ’rech’ from an expression like int name=‘rech’ are changed to name and rech). Special characters that represent multiplications, equals, additions, subtractions, or divisions from formulas should have been eliminated in this process (e.g., from a =b * c +d only a, b, c, and d should get through).

3.2 Problem Formulation

The main objective of the proposed problem is to predict the user specific Query results state through an optimized clustering for the big data analysis. The linear clustering Suffix tree separates the data, but it maximizes the distance between the given data point to the nearest data point of each class.

The training data set is given by

$$D = \{(x_1, y_1)\Lambda, (x_i, y_i)\Lambda, (x_l, y_l)\},$$

$$x \in R^n, y \in \{-1,1\} \quad (1)$$

Where, l – number of training data,
X_i – Training data,
y_i – class label as 1 or -1 for x_i for large data with drifting
A nonlinear function is adopted to map the original input space R^n into N-dimensional feature space of the large dataset.

$$\psi(x) = \varphi_1(x), \varphi_2(x), \ldots, \varphi_N(x) \quad (2)$$

The separating hyper plane is developed in this N-dimensional feature space. Then the clustering function represented as,

$$y(x) = \mathrm{sgn}(\omega.\psi(x) + b) \quad (3)$$

Where ω - weight vector and b- scalar.
In order to obtain the optimal clustering through ensemble classifier $\|\omega\|$ should be minimized subject to the following constraints

$$y_i[\varphi(x_i).\omega + b] \geq 1 - \xi_i, \quad i = 1, 2\ldots l. \ (4)$$

The variable ξ_i is the positive slack variables, necessary for misclassification of data in different cluster.

3.3 Determining a feature Evolution and Feature selection for data classification using ensemble classification

One of the most assumptions of ancient data processing is that knowledge is generated from one, static and hidden perform from the data evolving in the data streams. However, it is hard to be true for data stream learning, where unpredictable changes are likely to eventually happen. Concept drift is said to occur once the underlying function that generates instances changes over time. The Suffix tree clustering is known to be efficient in clustering large datasets. This clustering is one in all the best and also the best far-famed unsupervised learning algorithms that solve the well-known clustering problem in terms large data through the steps of big data community.

The objective function is given in Eq. (5),

$$\min J(\omega, \xi) = \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{l}\xi_i \quad (5)$$ We have,

$$y_i[\varphi(x_i) * \omega + b] \geq 1 - \xi_i \quad (6)$$

where, C – margin parameter, ω - weight vector, x_i - training data, y_i - class label (1 or -1)
for x_i, ξ_i - positive slack variables; $\xi_i \geq 0, i = 1, \ldots, l$, b – scalar, l – number of training data.

Objective function obeys the principle of structural risk minimization in order to obtain the optimal solution with less false positive rate for the data clustered. The objective function in Eqn (5) can be re-modified by following Legrangian principle for the data segmentation and prediction as, L

$$(\omega, b, \xi, a, \gamma) = \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{l}\xi_i - \sum_{i=1}^{l} a_i(y_i[\varphi(x_i)*\omega + b] - 1 + \xi_i) - \sum_{i=1}^{l}\gamma_i\xi_i$$

(7)

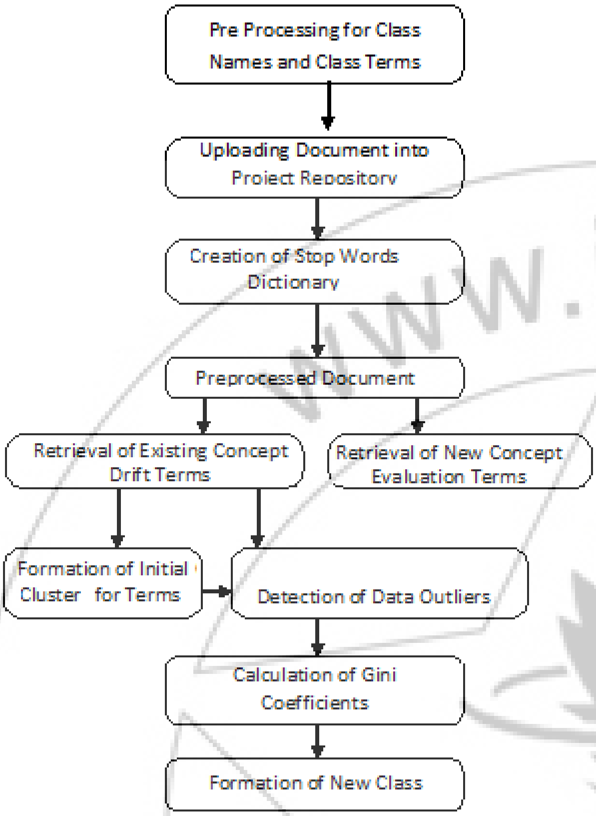

Figure 1: Block Diagram of Framework for Ensemble Classification

Below equation explains the similarity assignment as follows

Where, $a_i \geq 0, \gamma_i \geq 0 (i = 1,2,.....,l)$, a_i, γ_i -

On substituting Eq. (8) in Eq. (7), the dual problem becomes,

$$\max W(a) = -\frac{1}{2}\sum_{i,j=1}^{l} a_i a_j y_i y_j (\varphi(x_i), \varphi(x_j)) + \sum_{i=1}^{l} a_i$$

$$\max W(a) = -\frac{1}{2}\sum_{i,j=1}^{l} a_i a_j y_i y_j k(x_i, x_j) + \sum_{i=1}^{l} a_i$$

The suffix algorithm aims to partition a group of objects supported their attributes/features, into no. of feature clusters, wherever x may be a predefined or user-defined constant into x clusters

Prediction based on the query preferences and query frequency suggestions

The prediction of the query relevance is calculated based on the query preferences and query frequency of the user or community to the particular type of data. Frequency suggestion is employed through prediction and equivalence of the system in the data evolution and concept drifting in the data streaming in the network to the server.

Determining temporal probability and temporal pattern relevance of data to the query

Temporal probability is carried out the density based clustering technique and its cluster employed through the ranking of the document, temporal pattern relevance is also estimated from the cluster in terms of entropy and Euclidean calculation.

Ranking Based on the Integration Values through following process

Pair wise alignment through Similarity estimation.

Pair wise alignment is carried through the ranking based on the analysis and pair wise alignment is carried out through the algorithm is based on the observation that the data values belonging to the same attribute usually have the same data type and may contain similar strings, especially since results records of the query for the user query.

Holistic alignment based prediction methods.

Vertices from the same record are not allowed to be included in the same connected component as they are considered to come from two different attributes of the record. If two vertices from the same record breach this constraint, a path must exist between the two, which we call a breach path.

Nested structure Alignment through user specific clustering

Holistic data value alignment constrains a data value in a Result set to be aligned to at most one data value from another Result set. If a Result set contains a nested structure such that an attribute has multiple values, then some of the values may not be aligned to any other values. Therefore, nested structure processing identifies the data values of a Result set that are generated by nested structures.

4. Experimental Results

In this section, Experimental Results for query based prediction from big data with data evolution and feature evolution were carried out using Novel data and results were performed with performance system configurations to perform the data scaling and extracting into the proper clusters through suffix tree clustering. Initially extracting the framework has been utilized by training, validation and testing data for classification of results using historical prediction models identify the results set estimation efficiently and effectively in large dataset. The performances of the clustering and classification are experimented and presented in terms of relative speed, computational time as properties measure of performance using the large data set.

4.1Query frequency estimation and temporal probability estimation

The temporal prediction states observed from the large data set are as follows: supervised data , unsupervised data and semi-supervised data.

- **Feature extraction through user query modeling**

Feature Extraction is employed in large dataset with data drifting and information retrieval with estimating various factors in the query analysis to the large dataset

Feature extraction:

(1) The data in the big data is evolved with several feature classification with novel features estimation in each sample such as, y1, y2, y3, y4 and y5, are extracted by the equation as follows:

$$y_k = \frac{c^k}{\max\limits_{i=1}^{5}\left(c^i\right)} \quad (9)$$

where k=1, 2,. . ., 5,
c^k – Absolute feature data per one sample.
(2) The absolute information is calculated for different samples given by,

$$Y_6 = \log_{10}\left(\max_{m=1}^{5} c^m\right) \quad (11)$$

Table 1: Parameters of classification and Prediction of data classification

Parameters	*Notations used*	*Values*
Learning rate	Λ	0.01
Scaling factor	Σ	1

Table 2: Performance Parameters to compute Data Extraction mechanism

Parameters	*Notations used*	*Values*
Number of iteration	I	15000
Order of the polynomial	Order	3
Scaling factor	Σ	1

4.3 Result Analysis

The proposed framework is implemented and tested using different types of datasets using user specific cluster modeling and multi correlation estimation. An extensive experimental study was conducted to evaluate the efficiency and effectiveness of the proposed methodology on various parameters of benchmark instances and the prediction states are obtained in the graph

User Specific Clustering has been utilized by the training the data through the analysing the user behavior in the personalization methods in the literatures. Popular notions of clusters include groups with small distances among the cluster members, dense areas of the data space, intervals or particular statistical distributions. Clustering can therefore be formulated as a multi-objective optimization problem as it requires to cluster based on the different user perspective. The appropriate clustering algorithm and parameter settings (including values such as the distance function to use, a density threshold or the number of expected clusters) depend on the individual data set and intended use of the results. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery or interactive multi-objective optimization that involves trial and failure. It will often be necessary to modify data pre-processing and model parameters until the result achieves the desired properties.

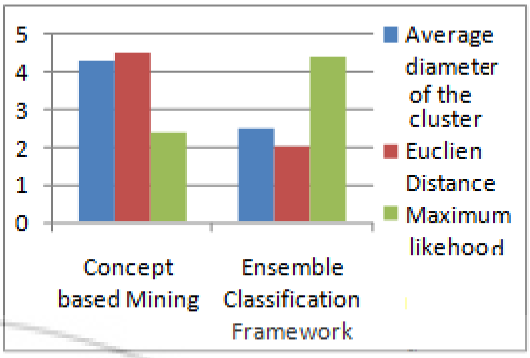

Figure 2: Estimation of the proposed framework against concept based mining

The following parameters are utilized to estimate the performance of the big data classification and prediction of data for user queries

Minimize average diameter of clusters
This factor estimates the performance of proposed framework in the classifying the data with concept drift .Proposed framework by suffix tree clustering proves the accuracy results set with precision and recall in the cluster achieved.

Maximum Likelihood:
It is a method of estimating the parameter of a statistical model. When applied to a data set and given a statistical model, maximum-likelihood estimation provides estimates for the model's parameters. We have proved the performance of system in clustering the query based on the several factors included in the framework and experiment to determine the performance factors with better results.

5. Conclusion

We have implemented classification and novel class detection technique for concept-drifting data streams that addresses four major challenges, namely, infinite length, concept-drift, concept-evolution, and feature-evolution with less false alarm rate and false detection rates in many scenarios. We have designed outlier detection, and novel class instances model, as the prime cause of high error rates. We also propose a better alternative approach for identifying novel class instances using discrete Gini Coefficient, and theoretically establish its usefulness. Finally, we propose a graph-based approach for distinguishing among multiple novel classes. However, we adopted some dynamic approach using drift detection technique utilizing the naïve baiyes which emphasizes mainly on concept-evolution. In performance evolution, results have been obtained with improved efficiency in terms of properties like probability determination, f measure. As a future work, we plan to enhance the classification solution for large data streams in terms optimization techniques to ensemble classifier to yield a more accurate results with less false positive rate. So we incorporate principle component analysis for classifier estimation as an optimization to the proposed solution.

References

[1] A. Bonnaccorsi, "On the Relationship between Firm Size and Export Intensity," Journal of International Business Studies, XXIII (4), pp. 605-635, 1992. (journal style)
[2] R. Caves, Multinational Enterprise and Economic Analysis, Cambridge University Press, Cambridge, 1982. (book style)
[3] M. Clerc, "The Swarm and the Queen: Towards a Deterministic and Adaptive Particle Swarm Optimization," In Proceedings of the IEEE Congress on Evolutionary Computation (CEC), pp. 1951-1957, 1999. (conference style)
[4] H.H. Crokell, "Specialization and International Competitiveness," in Managing the Multinational Subsidiary, H. Etemad and L. S, Sulude (eds.), Croom-Helm, London, 1986. (book chapter style)
[5] K. Deb, S. Agrawal, A. Pratab, T. Meyarivan, "A Fast Elitist Non-dominated Sorting Genetic Algorithms for Multiobjective Optimization: NSGA II," KanGAL report 200001, Indian Institute of Technology, Kanpur, India, 2000. (technical report style)
[6] J. Geralds, "Sega Ends Production of Dreamcast," vnunet.com, para. 2, Jan. 31, 2001. [Online]. Available: http://nl1.vnunet.com/news/1116995. [Accessed: Sept. 12, 2004]. (General Internet site)