RESEARCH ARTICLE OPEN ACCESS

A Distributed Ensemble Text Classification framework for Evolving Data Stream Classification

Sukanya.K *, Dr.N.Ranjith**

*(Department of Computer Applications, KSG College of Arts and Science, Coimbatore
Email: Ksukanya1375@gmail.com)
** (Department of Computer Applications, KSG College of Arts and Science, Coimbatore
Email: ranjith@ksgcollege.com)

--************************----------------------------------

Abstract:

Text Stream Classification is exploring research area in field of data mining. In last few decades, data volume and data velocity from various application has been growing enormously which has to be classified and clustered effectively for efficient data management. Much state of art approaches has been proposed in terms of supervised and unsupervised learning model to classify and cluster the distributed data from various data servers. Despite of various advantages, still some pitfall exhibited in terms of scalability and accuracy on classifying those large scale data. In order to overcome those challenges, a novel ensemble text classification framework has been employed to predict the novel classes from the evolving data streams. Proposed framework is considered as multi step classifier model utilizing principle component analysis for feature extraction and feature reduction and K nearest learning model is to generate the class decision boundaries to the extracted features. K-NN algorithm generates the regularized classes. Expectation maximization learning model is considered as optimization technique to classify the outlier features with concept drift and semantic drift. It generates the new classes to the outlier data. Experiment results on proposed dataset explain the efficiency of the proposed ensemble model against the state of art approaches in large stream data classification. Proposed classification model outperforms the existing state of art model in terms of Accuracy, precision, recall, and F measure.

Keywords — **Artificial Intelligence, Data Stream Classification, Ensemble Technique, Feature extraction, Latent Features.**

--************************----------------------------------

I. INTRODUCTION

Text classification is emerging research field in data mining towards handling enormous growth of data streams. Text Classification is employed for classifying the enormous data streams to achieve scalable and effective data management. Text classification utilizes the text mining in terms supervised and unsupervised learning towards extracting the relevant information from the data stream. Unsupervised learning model uses feature selection and feature Extraction techniques for extracting reliable and relevant features [2]. While irrelevant and noisy data can be eliminated using feature reduction technique [3] as it degrades the performance of classifiers in terms of accuracy. Further dimensionality of the data is also been reduced in pre-processing using singular value decomposition process in order to avoid curse of dimensionality and sparsity issue. Relevant feature are employed for classification using ensemble learning algorithm [4].

In this paper, a novel ensemble text classification technique has been proposed as multistep learning model to categorize and predict the novel classes in the feature space extracted in aspect of feature evolution and concept evolution on subset. In this work, ensemble classifier framework including Principle Component Analysis [5], K Nearest Neighbour algorithm [6] and Expectation Maximization algorithm [7] has been employed in parallel to classify the evolving data streams. Best performing classifier is considered as training model for testing data. The performance of model is cross validated to compute the consistency of the classifier on class generation with curse of dimensionality and sparsity issues.

The rest of the paper is sectioned as follows: Section 2 discusses the related works employed to data classification for large scale data streams, Section 3 briefly defines the important key terms and discusses about the background of outlier detection. Section 4 defines proposed ensemble classification and Section 5 presents the experimental results on data sets along selected performance measures to cross validate. Section 6 discusses conclusions and future work.

II. RELATED WORK

In this section, state of art techniques employed to classify the large volume and velocity of data stream has been analysed against various strategies. Each of these techniques follows data classification is described as follows

A. Expected Maximization –Singular value Decomposition based learning for Outlier Detection with Imperfect Data Labels

In this method, data with outlier instances has been considered as Novel class on employing the expected maximization and SVD model. In this work, Apriori algorithm has been employed as supervised learning model to identify data features and classify the features space into discrete classes. Classes generated composed of normal instances and outlier instances.

B. Robust Adaptive Prototype Technique for Outlier Detection

In this method, Adaptive learning model using unsupervised learning is capable of predicting the Novel class on the data stream with drift. It effectively maps the data instance with imperfect labels using likelihood values and limited negative examples. It uses the two stages for classification, generation of feature space with latent instances and classification of the feature by likelihood values.

III. BACKGROUND

In this section, various background and key terms of existing learning model has been defined and analysed.

A. Feature Extraction and Reduction

Feature selection and reduction is defined as data pre-processing step in the classification and prediction learning model. It has been implemented to process the dataset to determine the relevant features and discarding the irrelevant ones as decomposition step. Further feature selection process compute the effective feature to determine the class labels on original category of feature extracted. Feature extraction model uses either of those methods such as filter, wrapper, embedded, and hybrid methods. Features reduction use specific measure named as Eigen matrix to obtain more instance which is relevant to specific class.

B. Class Determination

Gathering the similar patterns on the feature obtained is considered as class label. Collection of class label with members consist of features is considered as classifier. Classifier can be further sub sectioned into parametric and nonparametric classifiers. Nonparametric classifiers have received particular attention in this research as it mostly deals unsupervised model on data distributions.

C. Ensemble Data Classification

Ensemble classification is a model employed with more than two classifier for the data distribution. It uses the best data distribution classifier as training model to remaining classifier to indicate possible distribution employed to the streaming data with weights changing on specific intervals. Finally the model classifies the multiple data distribution according to their internal relevance on representative features.

D. Outlier Detection

Employing ensemble classifier to the Feature set determines the class boundaries to gather the feature points extracted to build a reprehensive class. In these constraints, the data points falling outside the decision boundary of the particular class are considered as outliers. The outliers of the data have been analyzed to verify the data point consisting of cohesion among themselves and separation from the existing class containing data instances. Finally it is mentioned as multiclass classifier

IV. PROPOSED MODEL

In this section, ensemble text classification framework has been described in detail on modelling objective function is as follows

A. Ensemble Text Classification Framework model

Ensemble text classification framework composed of classifier to classify the feature subset to generate the outlier classes. Framework is considered as multistep classification process. This model finds best classifier to train the further classifier during testing process. Proposed model is capable to determining the novel classes against the regularized classes. The Figure 1 describes the proposed architecture of framework towards data classification.

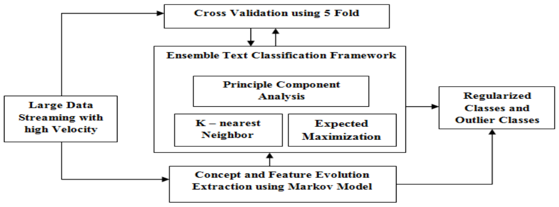
Cross Validation using 5 Fold
Ensemble Text Classification Framework
Principle Component Analysis
K – nearest Neighbor
Expected Maximization
Large Data Streaming with high Velocity
Regularized Classes and Outlier Classes
Concept and Feature Evolution Extraction using Markov Model

Fig 1: Architecture of proposed Ensemble Text classification Framework

B. Principle Component Analysis(DPCA)

PCA is employed to reduce the feature space containing the concept evolution and feature evolution as subset of features using orthogonal matrix. The matrix generates the Eigen value and Eigen vector on the feature to compute the correlate and covariance principal components. Data instance has been processed in the vectors to manage the multidimensional space.

C. K-Nearest Neighbour Algorithm

K-Nearest Neighbor is employed as classifier for the training data with subset of feature from PCA model. These feature set will be classified into regularized classes using k value. K clusters are built for the regularized class instances. Further instance computed against centroids, radius, and frequencies of data points of class.

D. Expected Maximization Algorithm(EM)

Expected maximization is employed to identify the novel classes from the outlier data using Iterative method for learning feature subset. Initially it uses probabilistic categorization using the unsupervised model [10]. Further categorized data points have been assumed with random assignment of categories using EM objective function. Those categorizes used to learn an outlier data against the feature evolved by estimating model parameters q on those unpartitioned data with expectation and maximization.

V. EXPERIMENTAL RESULTS

In section, experimental results of the proposed Ensemble text classification framework has been described against the existing approaches on forest cover, twitter and Reuter's dataset which collected from large scale data stream applications. The performance of the work has been computed using precision and recall measures as important factor accuracy.

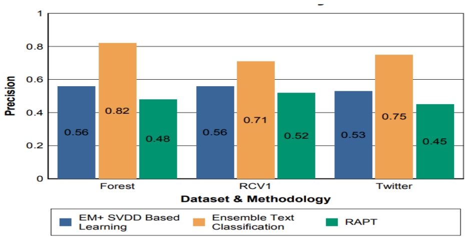

Fig 2: Performance Measures of the Proposed Model against Existing Model via Precision

It is observed that the proposed Ensemble text classification framework is always better on comparing with static data stream classification methods on precision and recall measures. Figure 2represents the Classification results on precision and figure 3 represents the classification result on recall values.

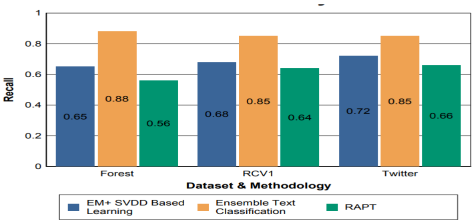

Fig 3: Performance Measures of the Proposed Model against Existing Model via Recall

However, after a certain point of the classification of the dataset, ensemble classifier will produces the accurate results on presence of curse of dimensionality and sparsity issues

CONCLUSION

Ensemble text classification framework has been designed and implemented to generate the regularized and novel classes on the outlier data from large dynamic data streams. Proposed model reduces the data sparsity and curse of dimensionality issues to much extent on comparing with state of art approaches. Further it has been revealed that an ensemble learning framework can be employed to classify huge corpus of dataset with high accuracy and scalability. The empirical evaluation shows that proposed learning framework outperforms the state of art methods especially on accuracy. Finally it has validated with like Accuracy in terms of precision and recall. In the future work, proposed framework can be enhanced to cope with various aspect drift on semantic data on differentiating two or more emerging new classes

REFERENCES

[1] R. Y. Lau, P. D. Bruza, and D. Song, “Towards a belief-revision based adaptive and context-sensitive information retrieval system,” ACM Transactions on Information Systems, vol. 26, no. 2, pp. 8.1-8.38, 2008.

[2] H.Liu & H.Motoda, “Feature selection for knowledge discovery and data mining”. Springer Science & Business Media, vol. 45, 2012

[3] K. Bunte, M. Biehl, & B. Hammer, “A General framework for dimensionality-reducing data visualization mapping,” Neural Computation, vol. 24, no. 3, pp. 771–804, 2012.

[4] W. Fan, “Systematic Data Selection to Mine Concept-Drifting Data Streams,” Proc. ACM SIGKDD 10th International Conference on Knowledge Discovery and Data Mining, pp. 128-137, 2004.

[5] Y. Li, A. Algarni, & N. Zhong, “Mining positive and negative patterns for relevance feature discovery,” in Proceedings of KDD’10, , pp. 753–762, 2010

[6] Hongxing Ma,Jianping Gou, Xili Wang, Jia Ke, Shaoning Zeng "Sparse Coefficient-Based k -Nearest Neighbor Classification “in IEEE Access , ,pp: 16618 – 16634, 2017

[7] Bhawna Nigam, Poorvi Ahirwal , Sonal Salve and Swati Vamney "Document Classification Using Expectation Maximization with Semi Supervised Learning "International Journal on Soft Computing ,Vol.2, No.4, November 2011

[8] Varun Mithal,Guruprasad Nayak, Ankush Khandelwal,Vipin Kumar, Nikunj C. Oza, Ramakrishna Nemani "RAPT: Rare Class Prediction in Absence of True Labels" IEEE Transactions on Knowledge and Data Engineering, Vol: 29, Issue: 11, 2017

[9] Z. Xu, I. King, M. R.-T. Lyu, and R. Jin, “Discriminative semisupervised feature selection via manifold regularization,” Neural Networks, IEEE Transactions on, vol. 21, no. 7, pp. 1033–1047, 2010.

[10] Bo Liu, Yanshan Xiao, Philip S. Yu, Zhifeng Hao and Longbing Cao "An Efficient Approach for Outlier Detection with Imperfect Data Labels" IEEE Transactions on Knowledge and Data Engineering in Vol: 26, Issue: 7, July 2014.