ARTIFICIAL INTELLIGENCE BASED EVOLVING ENSEMBLE LEARNING MODEL FOR EVOLVING DATA STREAM CLASSIFICATION

*S. Saravana Kumar**

Abstract

Data Stream classification is an emerging area in current data mining research. In recent years, Artificial intelligence technique has been employed for large data stream classification on exploration of feature extraction and knowledge representation process to categorize the data streams into one or more classes. Unsupervised data classification based on ensemble Artificial intelligence has been employed in this work to predict the classes for the outlier data in the data streams which is considered as imperfect labels from training samples for all features in the data stream on analysis.

In this paper, a novel artificial intelligence technique combination has been presented for data stream classification. It is named as Evolving Ensemble learning model as a multistep learning process utilizing the infrequent principle Component Analysis, K Nearest Neighbour and Expectation maximization Algorithm of the artificial intelligence technique to generate the new classes from its regularized classes with data outliers. In addition, Markov hidden model has been used in addition to detect the latent feature in the data stream to construct the feature set. It is capable of detecting the feature and concept of evolution on the feature space and label space of the classes. Further feature reduction technique has been employed to remove or eliminate the curse of dimensionality and sparsity issue in terms irrelevant, Noisy and redundant features and reduces the dimensionality of feature space. Experiment results explain the effectiveness of the proposed AI based ensemble model against the state of art approaches in large stream data classification. Proposed classification model outperforms the existing model on performance metrics like precision, recall, F measure and Accuracy and classification error.

Keywords: Artificial Intelligence, Data Stream Classification, Ensemble Technique, Feature extraction, Latent Features

I. INTRODUCTION

Due to the enormous growth of data streams from different online applications, data classification is one of the crucial technologies for handling the streaming data to achieve effective information management. Data classification utilizes the text mining technique, as it is becoming increasingly essential towards employing the supervised and unsupervised learning models [1]. Data classification plays a vital role in both managing and extracting the relevant information from exploring the use of feature selection and feature extraction techniques [2]. Moreover, data streams with the presence of unrelated, redundant and features contain noisy contents, the performance of the learning algorithm decreases in its classification accuracy [3]. Hence, it has become to devise a technique to reduce the dimensionality of the dataset to eliminate the curse of dimensionality issues and to improve both the efficiency of the data mining technique. Additionally, the data mining model has capability to determine the relation between normal and latent features for better visualization and generalization of data for improved understanding of the learning algorithms [4].

Department of Computer Science,
Karpagam Academy of Higher Education, Coimbatore, Tamil Nadu, India
*Corresponding Author

Feature selection technique has been employed to determine subset of most important features in the input set of features in the data streams. Data stream is a speed and continuous occurrence, it is assumed to have infinity on size of data and irrelevant features on a set of processes. Therefore, it was becoming impractical to manage and use every data of the dataset for the training process of the classifier. Best alternative for managing the large data streams is an incremental learning technique in the category of unsupervised classifiers [5]. However many data classification methods have failed to employ the incremental learning model towards extracted features to regularize the existing classes on composition capability to analyse new instances with similar characteristics.

In this paper, a novel artificial intelligence based data classification technique termed as Ensemble learning model (ELM). In this work, artificial intelligence based classifier includes infrequent principle Component Analysis, K Nearest Neighbour [6] and Expectation maximization algorithm [7] to classify the evolving concept in data streams as classes. Novel Class and Regularized class has been analysed on cross validation to estimate the consistency and cohesion among obtained class instances and its separation from the existing class features using similarity computations. Proposed model has capability to classify two or more unique classes on any kind of data.

The rest of the paper is sectioned as follows: Section 2 discusses the related works on data classification and presents its impacts against performing classification under feature evolution, Section 3 briefly defines the proposed model in terms of feature extraction technique and novel class prediction classifiers utilizing artificial intelligence techniques as ensemble model and Section 4 presents the experimental outcomes on a various sizes of data sets along performance measures. Section 5 discusses its conclusions and future work.

II. RELATED WORK

In this section, many techniques to classify the large stream data have been analysed against various constraints. Each of state of art techniques follows different kind of class categorization nearly related to the proposed defined framework, which is described as follows

II.1. Outlier Detection with Imperfect Data Labels

In this method, data with outlier instances has been considered as Novel class. Supervised learning model has been employed to identify data features and classify the features which are variant and inconsistent with the extracted set of features generated using feature extraction technique. However, in addition data streams containing normal instances, limited negative examples or outliers have been generated in many applications. In this novel class has been determined on the feature space containing the imperfect label. Artificial Intelligence base approach is capable of predicting the Novel class on the latent features in the data streams has been addressed in this work. It maps the data with imperfect labels using likelihood values and incorporates limited abnormal examples as a degree of membership into the learning process. The approach works in two steps. In the first step, a pseudo training dataset generates instances by likelihood values based on its available feature set.

III. PROPOSED MODEL

In this section, artificial intelligence based ensemble data classification model has been described using infrequent principle component analysis, K- Nearest Neighbour and Expected Maximization classification models along feature selection model as follows

III.1. Feature Selection using Information Gain

Feature Selection is the process employed to determine the instances on goodness criteria as threshold in the data streams. The criteria can be generated using the following process as follows

- **Information Gain**

It measures the bits of the data information obtained for several similar categories as prediction on determining the presence or absence of a word in a streaming of the data [8]. Information gain is calculated for each term and the best terms are extracted as features.

III.2. Ensemble Learning model

The proposed AI based technique applies the following classifier to generate the regularized class and novel class to the evolving data streams with evolving concepts and features in the feature space. The Figure 1 explains the proposed architecture encompassed of the multistep classifier as data stream classification the framework

- **K-Nearest Neighbour – Classifier 1**

K-Nearest Neighbor based classifier has been trained with a small partition of data streams initially. Processed small partitions of the data streams acts as raw training data for classification. Next K clusters are built for the remaining dataset trough usage of a semi-supervised K-means clustering model and those clusters obtained as information set of each cluster are utilized for classification. These pseudo-points compose the classification model [9]. The information of the clusters contains the information such as centroids, frequencies of data points and its radius of data streams belonging to each class.. The raw data points are eliminated after generating the cluster formation step for cluster and classification.

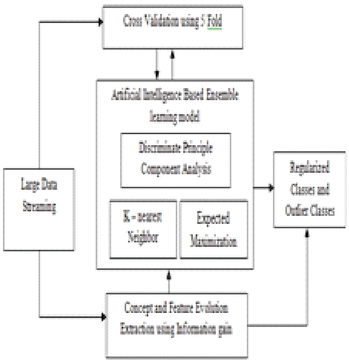
Cross Validation using 5 Fold
Artificial Intelligence Based Ensemble learning model
Discriminate Principle Component Analysis
K – nearest Neighbor
Expected Maximization
Large Data Streaming
Regularized Classes and Outlier Classes
Concept and Feature Evolution Extraction using Information gain

Figure 1: Architecture of proposed AI Model for Data Classification

- **Discriminate Principle Component Analysis(DPCA)**

DPCA employed the transformation of the feature space into an orthogonal matrix to compute the correlated and uncorrelated principal components using Eigenvalue and Eigenvector. Data points in the vectors represented as multidimensional space In addition, this discriminate model transfers a set of correlated variables of the matrix into a new set of uncorrelated variables for efficient processing and to use cross validation process. Data point considered as feature and concept. It is determined on the Eigenvalue and feature space is computed on the Eigenvector.

- **Expected Maximization Algorithm(EM)**

It is employed as an Iterative method for learning data streams as probabilistic categorization using the unsupervised model [10]. Initially data points have been assumed with random assignment of categories.

- **Expectation (E-step):** Compute P(ci | E) for each data point of the data streams by computed current model, and probabilistically re-label the examples based on these posterior probability estimates.

- **Maximization (M-step):** Re-estimate the model parameters,\, from the probabilistically re-labeled data of the data streams.

IV. EXPERIMENTAL RESULTS

In section, experimental results of the proposed AI based Ensemble learning model has been described against the existing approaches on forest cover, twitter and Reuter's dataset which collected from data stream applications. Although different varieties of dataset may have different issues on data points in novel class detection on employment of the data classification. The performance of the work has been computed using precision and recall measures to determine the accuracy of the model employed.

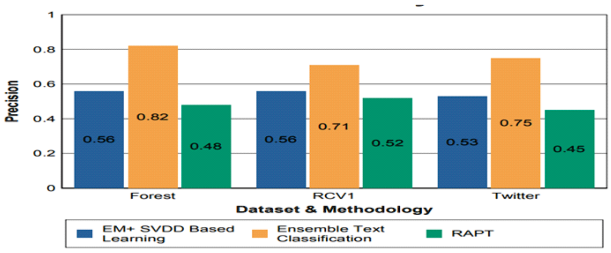

Figure 2: Performance Measures of the Proposed Model against Existing Model via Precision

It is measured that the proposed method always provides good accuracy when compared to feature selection methods and with ensemble classifiers on precision and recall measures, it has provided better or comparable results than existing state of art approaches. Figure 2 represents the outcome of precision and figure 3 represents the outcomes of the recall.

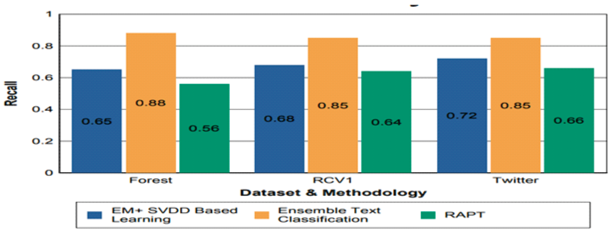

Figure 3: Performance Measures of the Proposed Model against Existing Model via Recall

However, after a certain point of the classification of the dataset, AI based technique will produce the accurate results on presence of curse of dimensionality and sparsity issues.

V. CONCLUSION

AI based document classification model has been designed and implemented to generate the regularized and novel classes on any best learning model on a set of ensemble classifiers. This work completely eliminates the data sparsity and curse of dimensionality issues after extracting the feature evolution in data streams. It has been revealed that an ensemble learning model can be applied to a huge corpus of dataset to classify effectively and efficiently. Major advantage of the framework is to detect new classes on a time varying dataset with high accuracy. The use of unsupervised learning models incorporated with the multiple classifiers is a new ability to differentiate between the normal and hidden features. The empirical evaluation shows that the proposed learning model outperforms the state of art methods. It has been cross validated with training instances for performance measures. In addition, the proposed framework works effectively under the limited memory requirement. In the future, proposed methods can be improved to deal with aspect drift on semantic data to differentiate 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 and H. Motoda, Feature selection for knowledge discovery and data mining. Springer Science & Business Media, vol. 45, 2012.

[3] K. Bunte, M. Biehl, and 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, and 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 semi supervised 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.