International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 22 (2017) pp. 12454-12459

Ensemble Text Classifier: A Document Classification Technique to Predict and Categorizes Regularised and Novel Classes Using Incremental Learning

[1]Mr.G. Silambarasan and [2]J.Anvar Shathik

[1]Research Scholar CMJ University, Meghalaya, India.
[2]Assistant Professor, KGISL Institute of Technology, Coimbatore, India.

Abstract

Document classification is employed through text classification techniques which have exploited many feature extraction and knowledge representation process in recent years in order to categorize the document into one or more classes. Unsupervised document classification is employed in this research in to predict the classes where there are no labels from the training samples but imperfect labels are available for all features in the dataset. In this paper, we present a novel technique termed as Ensemble Text Classifier. The Ensemble Text Classifier (ETC) is a multistep learning framework for classifying the novel classes from regularized classes in the document classification setting. The initial step of proposed framework is learning the document and extracting the features , feature evolutions , concept and concept evolution using feature extraction algorithm as it eliminates the irrelavent , Noisy and redundant features. As feature extraction algorithm reduces the dimensionality of the dataset, it becomes efficient to classify the features into novel classes and regularized classes from the available features. Proposed Framework uses the ensemble classifier to classify the feature to construct the feature set for class building through incremental learning model. Ensemble Classifier integrates the infrequent principle Component Analysis, K Nearest Neighbour and Expectation maximization Algorithm to detect the latent feature to construct the feature set. Experiment results shows that proposed System outperforms the state of art approaches in document classification against performance measures such as Accuracy , precision, recall, and F measure and classification error.

Keywords: Document Classification, Unsupervised Learning, Ensemble Classifier, Novel Class Prediction, Feature Extraction

INTRODUCTION

Due to the explosive growth of text documents, text classification is one of the crucial technologies for information management. Text mining technique is becoming increasingly important and attracting extensive attention in related research areas in recent years [1]. Text classification plays a key role in both organising and extracting the relevant information with use of feature selection and feature Extraction Algorithms from the huge text corpus and streaming of the data [2]. Moreover, with the presence of irrelevant, redundant and noisy features, the performance of the learning algorithm degrades [3]. Hence, it is crucial to reduce the dimensionality of the data to improve both the efficiency and effectiveness of most of the data mining algorithms. Also it is important for better visualization, data compression, noise removal, improved understanding ability, and generalization of the learning algorithms [4]. Feature selection aims at finding a subset of most useful features from the original set of features. Document is a fast and continuous phenomenon, it is assumed to have infinite length and irrelavent features. Therefore, it is impractical to store and use all the data of the dataset for training. The most obvious alternative is an incremental learning technique in terms of the unsupervised classifiers [5]. However many methods fails to employ the incremental learning model for the features extracted in order to regularize the existing classes with new instance with similar characteristics.

In this paper, we propose a novel framework termed as Ensemble Text Classifier (ETC) is a multistep learning to categorize and predict the regularised classes and novel classes utilizes the feature evolution extraction and concept evolution extraction methods. Ensemble Classifier integrates the infrequent principle Component Analysis, K Nearest Neighbour [6] and Expectation maximization algorithm [7] works parallel to classify the document in classes. Novel Class and Regularized class are analyzed to see if there is enough cohesion among them and separation from the existing class features. Proposed framework also allows for methods to distinguish among two or more novel classes.

With this approach, it is possible to distinguish different criteria for the classification of document into new class as document varies in terms of concept-drift, feature evolution. The model empirically shows the effectiveness of this approach. To the best of our knowledge, this is the first work that proposes these advanced techniques for novel class detection and classification to the evolving document. For Example, document in the cloud server will be updated by data owner for several reasons to attract the customers. Proposed technique is applied on a number of benchmark datasets.

The remainder of the paper is organized as follows: Section 2 discusses the related works in data classification and its impacts against the performing classification under feature

evolution, Section 3 briefly discusses the proposed technique in terms of feature extraction technique and novel class prediction ensemble classifiers and Section 4 presents the experimental results on a number of data sets. Section 5 discusses conclusions and future work.

RELATED WORKS

There exist many techniques to classify the documents are designed and implemented efficiently. Each of these techniques follows some sort of class categorization, among few performs nearly equivalent to the proposed framework, which is described as follows

RAPT Technique – Rare Class Prediction

RAPT Technique is a three step predictive modelling framework for classifying rare class [8]. The initial work of the technique learns a classifier that jointly optimizes precision and recall by only using imperfectly labelled training samples under certain assumptions on the imperfect labels, the quality of this classifier is almost as good as the one constructed using perfect labels. Finally it makes use of the fact that imperfect labels are available for all instances to further improve the precision and recall of the rare class. SMOTE is used for feature extraction. The New class is a very small fraction of the total number of samples (traditionally referred to as the rare class).

Outlier Detection with Imperfect Data Labels

In this outlier has been considered as Novel class which is to identify data features that are different from or inconsistent with the normal set of features. However, in addition to normal data, there also exist limited negative examples or outliers in many applications such that the novel class detection data is imperfectly labelled. These make novel class detection far more difficult than the traditional data classification technique under supervised model. Novel class detection approach presented in the work address data with imperfect labels and incorporates limited abnormal examples into learning.

To deal with data with imperfect labels, likelihood values for each input data which denote the degree of membership of an example toward the normal and abnormal classes respectively is been done using EM Algorithms. The approach works in two steps. In the first step, generates a pseudo training dataset by computing likelihood values of each instance based on its available feature set.

Along EM algorithm, kernel k -means clustering method and kernel LOF-based method is integrated to compute the likelihood values. In the second step, the generated likelihood values and limited abnormal examples into SVDD-based learning framework to build a more accurate classifier for global outlier detection is incorporated [9].

PROPOSED MODEL

In this section, we describe classifier which is an ensemble of infrequent principle component analysis, K- Nearest Neighbour and Expected Maximization classification models and feature selection model which is

FEATURE SELECTION

Feature Selection is a term goodness criteria which acts as threshold to extract the document corpus. The criteria can be derived using following measure as follows

- **Document Frequency**

Document frequency is the number of documents in which the term occurs. Document frequency for each unique term is the number of documents in which a term occurs.It is estimated as supervised technique[10]. The figure 3.1 describes the feature set formation

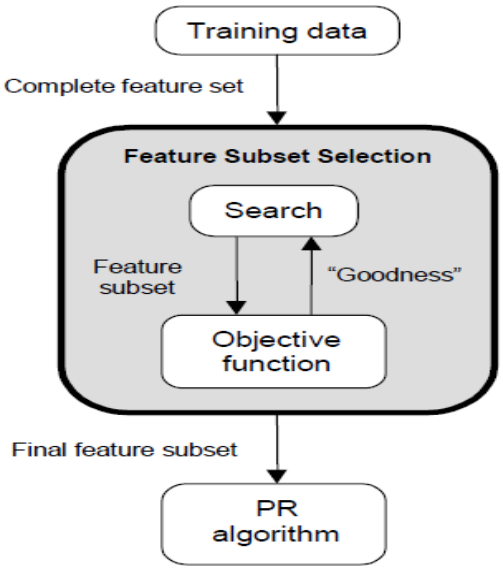

Figure 3.1: Feature extraction and Feature Subspace formation model

- **Information Gain**

It measures the (number of bits of) information obtained for category prediction by knowing the presence or absence of a term in a document[11]. information gain is calculated for each term and the best n terms are selected

$$G(t) = -\sum_{i=1}^{m} p(c_i)\log p(c_i) + p(t)\sum_{i=1}^{m} p(c_i \mid t)\log p(c_i \mid t) + p(\sim t)\sum_{i=1}^{m} p(c_i \mid \sim t)\log p(c_i \mid \sim t)$$

Where

T is the term

M is the number of categories

Term is selected into the feature space through probability estimation. Term is also represented as data point

T- {d1,d2,d3)

Where d1,d2,d3 is the datapoint of the term in the particular document

ENSEMBLE CLASSIFIER- CLASS REPRESENTATION

The proposed technique applies the following classifier to generate the regularized class and novel class to the evolving document with evolving concept and features. The Figure 2 describes the proposed architecture encompassed of the multistep classifier into the framework

- **K-NN**

A k-NN-based classifier is trained with the training data. Rather than storing the raw training data, K clusters are built using a semi-supervised K-means clustering, and the cluster summaries (mentioned as pseudopoints) of each cluster are saved. These pseudopoints constitute the classification model[12]. The summary contains the centroid, radius, and frequencies of data points belonging to each class. The radius of pseudopoints is equal to the distance between the centroid and the farthest data point in the cluster. The raw data points are discarded after creating the summary.

- **Infrequent Principle Component Analysis**

IPCA seeks a transformation to produce uncorrelated and orthogonal principal components. Also it Transfer a set of correlated variables into a new set of uncorrelated variables[13]. Data points are vectors in a multidimensional space. Data point which is considered as feature and concept is determined as Eigen value and feature space is termed ad Eigen vector

The class is constructed through eigen vector formation based eigen value derived through covariance and correlation matrix. The formulation to detect the novel class is given by

From k original variables: $x_1,x_2,...,x_k$:

Produce k new variables: $y_1,y_2,...,y_k$:

$y_1 = a_{11}x_1 + a_{12}x_2 + ... + a_{1k}x_k$

$y_2 = a_{21}x_1 + a_{22}x_2 + ... + a_{2k}x_k$

...

$y_k = a_{k1}x_1 + a_{k2}x_2 + ... + a_{kk}x_k$

Where

y_k's are uncorrelated (orthogonal)

y_1 explains as much as possible of original variance in data set

y_2 explains as much as possible of remaining variance

Each instance in the most recent unlabeled chunk is first examined by the ensemble of models to see if it is outside the decision boundary of the ensemble If it is inside the decision boundary, then it is classified normally (i.e., using majority voting) using the ensemble of models.

- **EM Algorithm**

It is Iterative method for learning probabilistic categorization model from unsupervised data [14]. Initially assume random assignment of examples to categories. Learn an initial probabilistic model by estimating model parameters θ from this randomly labeled data

■ Expectation (E-step): Compute P(c_i | E) for each example given the current model, and probabilistically re-label the examples based on these posterior probability estimates.

■ Maximization (M-step): Re-estimate the model parameters, θ, from the probabilistically re-labeled data.

Let's take a completely labeled corpus D, and randomly select a subset as D_K. Also use the set of unlabeled documents in the EM procedure. Correct classification of a document

Concealed class label = class with largest probability

Accuracy with unlabeled documents > accuracy without unlabeled documents

Criteria for initial Iteration is given by

$$\Pr(c_d \mid d) = 1 - \varepsilon \text{ and } \Pr(c' \mid d) = \varepsilon/(n-1) \text{ for all } c' \neq c_d$$

Let the class probabilities of the labeled documents is taken to re iteration based on the features extracted and features grouping for evolutional features. Keeping labeled set of same size Laplacian law for regularized class is given by

$$\theta_{c,t} = \frac{1 + \sum_{d \in D_c^K} n(d,t)}{|W| + \sum_{d \in D_c^K, \tau \in d} n(d,\tau)}$$

Re iterate Pr(c|d), for each feature and each document. It estimates class-conditional distribution which includes information from D

Once a new model is trained, it replaces one of the existing models in the ensemble. The candidate for replacement is chosen by evaluating each model on the latest training data, and selecting the model with the worst prediction error. This ensures that we have exactly L models in the ensemble at any given point of time. In this way, the infinite length problem is addressed because a constant amount of memory is required to store the ensemble. The concept-drift problem is addressed by keeping the ensemble up-to-date with the most recent concept.

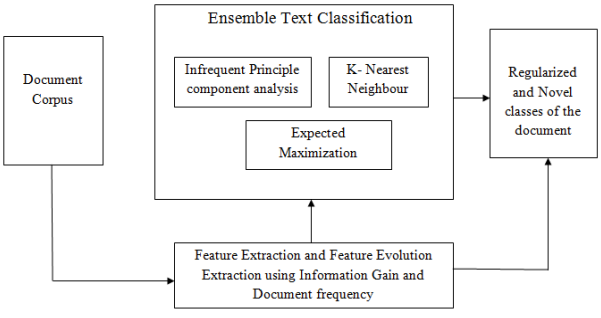

Figure 2: Architecture Diagram of the Ensemble Text Classification Framework

ALGORITHM: NOVEL CLASS PREDICTION

Input: Document Corpus

Process:

Apply Document Frequency and text Frequency for Feature creation

F1= Feature set

F1= {T1, T2, T3...}

Class = {f1, f2, f3}

Classes= {c1, c2, c3}

Where T1, T2 and T3 is conceptually relevant features

Update feature set f1

Ensemble (f1) = IPCA (f1) +Knn (f1) +Em (f1)

Class generation

If f4≠ c1|c2|c3

Generate

C4 = Novel class

Else

Group f4 into c1|c2|c3

End if

Each data point is transformed into pseudopoints data structure, which stores the centroid and weight (number of data points). The Algorithms is used predict any number of emerging new classes; they are grouped into one new class for the purpose of feature set update.

EXPERIMENTAL RESULTS

In section, we describe the experimental results of the proposed framework against the existing approaches

Dataset Description

We have done extensive experiments on 3 real datasets which is as follows

- **RCV1 (Reuters Corpus Volume I).**

This data set contains corpus of newswire describing the collection of the news.

- **Forest covers data set from UCI repository (Forest).**

The data set contains geospatial descriptions of different types of forests. We normalize the data set, and arrange the data so that new classes appear randomly.

- **Twitter**

This data set contains 340,000 Twitter messages (tweets) of different trends (classes) in different area and subjects.

EVALUATION

The proposed Framework is evaluated against the following measures against several preprocessing steps on those data sets [15]. The data set contains many multilabel documents

- **Precision**

Positive predictive value is the fraction of relevant instances among the retrieved instances. Precision is the number of correct feature divided by the number of all returned feature space.

$$\text{Precision} = \frac{\text{True positive}}{True\ positive + False\ Positive}$$

True positive is a number of real positive cases in the data and false negative is number of real negative cases in the data. The precision is evaluated against different dataset is depicted in the figure 3

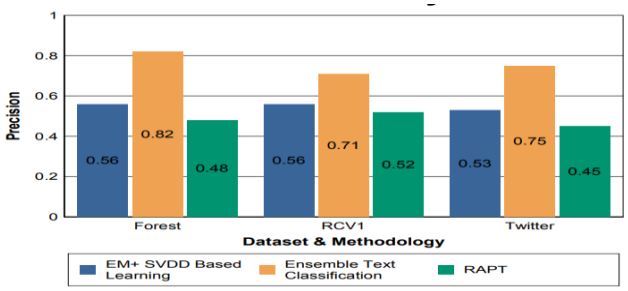

Figure 3: Performance Evaluation of the Precision towards technique against different datasets.

- **Recall**

It is the fraction of relevant instances that have been retrieved over the total amount of relevant instances. The recall is the part of the relevant documents that are successfully classified into the exact classes

$$\text{Recall} = \frac{\text{True positive}}{True\ positive + False\ negative}$$

True positive is a number of real positive cases in the data and false negative is number of real negative cases in the data. The recall is evaluated against different dataset is depicted in the figure 4

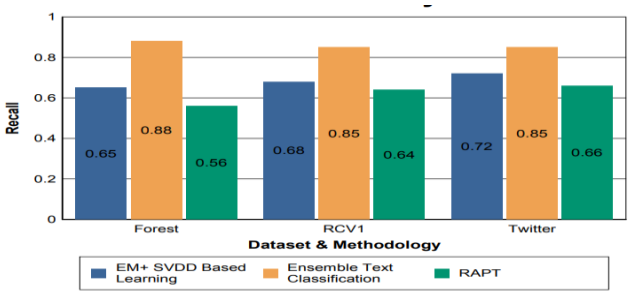

Figure 4: Performance Evaluation of the recall towards technique against different datasets.

A) True Positives: Observations where the actual and predicted Class were irrelevant

b) True Negatives: Observations where the actual and predicted class were irrelevant

c) False Positives: Observations where the actual class were irrelevant but predicted to be relevant

d) False Negatives: Observations where the actual class were irrelevant but weren't predicted to be relevant

- **Accuracy**

It is the number of correct class predictions to the single document to total number of predictions to whole document

Accuracy is given by

$$\frac{\text{True positive} + \text{True Negative}}{True\ positive + True\ Negative + false\ positive + False\ negative}$$

- **F Measure**

It is a measure of a test's accuracy and is defined as the weighted harmonic mean of the precision and recall of the test.

Although different document may have different impact on novel class detection, they are likely to have the same impact on classification. However, after a certain point, this improvement is diminished because of curse of dimensionality. On the other hand, for the Forest data set, the novel classes are less separable.

Table 1: Performance Evaluation of the Technique to Reuters Corpus Volume I dataset

Technique	Precision	Recall	F measure	Scalability	Accuracy
RAPT	0.45	0.66	0.54	125	0.64
EM+SVDD	0.53	0.72	0.61	165	0.73
ETC - Proposed	0.75	0.85	0.80	190	0.82

Table 2: Performance Evaluation of the Technique to twitter dataset

Technique	Precision	Recall	F measure	Scalability	Accuracy
RAPT	0.46	0.72	0.55	129	0.64
EM+SVDD	0.52	0.79	0.61	165	0.73
ETC - Proposed	0.79	0.86	0.80	180	0.82

Table 3: Performance Evaluation of the Technique to Reuters Forest dataset

Technique	Precision	Recall	F measure	Scalability	Accuracy
RAPT	0.46	0.66	0.64	135	0.64
EM+SVDD	0.54	0.72	0.63	145	0.73
ETC - Proposed	0.77	0.85	0.84	185	092

The evaluation of result is described in the table 1 for Reuters Corpus Volume I dataset, in table 2 for twitter dataset and table 3 for forest dataset.It is observed that the proposed method is always better when compared to feature selection methods and with ensemble classifier, it has provided better or comparable results.

CONCLUSION

We designed and implemented a novel document classification framework in terms of regularized and novel classes using unsupervised learning algorithms. This work has revealed that an ensemble classifier can be applied to huge corpus of dataset to classify effectively and efficiently. The Strength of the framework is its ability to detect new class with high accuracy. The use of an unsupervised learning models incorporated with the multiple classifiers is a new ability to differentiate between the features. The empirical evaluation shows that ETC outperforms the Existing methods despite the fact that it was not provides with training instance for performance measures like Accuracy , precision, recall, and F measure and classification error. In addition, proposed framework works effectively under the limited memory requirement. In the future, proposed method can be improved to deal with aspect drift differentiate two or more emerging new classes.

REFERENCES

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

[2] Y. Li, A. Algarni, and N. Zhong, "Mining positive and negative patterns for relevance feature discovery," in Proceedings of KDD'10, 2010, pp. 753–762

[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] H. Liu and H. Motoda, Feature selection for knowledge discovery and data mining. Springer Science & Business Media, 2012, vol. 454.

[5] W. Fan, "Systematic Data Selection to Mine Concept-Drifting Data Streams," Proc. ACM SIGKDD 10th Int'l Conf. Knowledge Discovery and Data Mining, pp. 128-137, 2004.

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

[7] Bhawna Nigam, Poorvi Ahirwal , Sonal Salve, Swati Vamney"Document Classification Using Expectation Maximization with Semi Supervised Learning "International Journal on Soft Computing (IJSC) 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 (Volume: 29, Issue: 11, Nov. 1 2017)

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

[10] 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.

[11] N. V. Chawla et al. Smoteboost: Improving prediction of the minority class in boosting. In Knowledge Discovery in Databases, pages 107–119. Springer, 2003.

[12] L. Zhang, Y. Li, Y. Xu, D. Tjondronegoro, and C. Sun, "Centroid training to achieve effective text classification," in 2014 International Conference on Data Science and Advanced Analytics, 2014, pp. 406–412.

[13] M. Masud, J. Gao, L. Khan, J. Han, and B. Thuraisingham. Classification and novel class detection in concept-drifting data streams under time constraints. IEEE Trans. Knowledge and Data Engineering, 23(6):859–874, 2011.

[14] E. J. Spinosa, A. C. P. de Leon Ferreira de Carvalho, and J. Gama. Novelty detection with application to data streams. Intelligent Data Analysis., 13(3):405–422, 2009.

[15] R. Jin and G. Agrawal. Efficient decision tree construction on streaming data. In Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 571–576, 2003.