TEST Engineering & Management

May – June 2020
ISSN: 0193-4120 Page No. 25246 – 25251

A Brief Survey of Machine Learning Algorithms for Text Document Classification on Incremental Database

Nihar M. Ranjan
Post Doc. Scholar
Lincoln University, Malaysia

Midhun Chakkaravarthy
Faculty of Engineering
Lincoln University, Malaysia

Article Info
Volume 83
Page Number: 25246 – 25251
Publication Issue:
May - June 2020

Article History
Article Received*: 11 May 2020*
Revised*: 19 May 2020*
Accepted*: 29 May 2020*
Publication*: 12 June 2020*

Abstract

Exponential growth of data in recent time is a very critical and challenging issue which requires significant attention. More than two third available information is stored in unstructured format largely in text formats. Knowledge can be extracted from many different available sources. Data which are mainly in unstructured format remain the largest readily available source of knowledge either online or offline so it must be attended very carefully. Text mining is believed to have a very high commercial potential value. Text classification is the process of classifying the text documents according to predefined categories. There are many databases which are dynamic and getting updated during course of time. To handle and classify these kind of datasets incremental learning is required which train the algorithm as per the arrival of new data. This paper covers different machine learning algorithms for text classification on the dynamic or incremental database also includes classifier architecture and Text Classification applications.

1. Introduction

Text Mining is extracting of valuable and yet hidden information from the text documents [2]. Text classification is one of the important research issues in the field of text mining also known as text categorization [6]. With the exponential increase in the amount of data contents available in digital forms from the different sources leads to a problem to manage this huge amount of online textual data [4]. So it has become necessary to analyze these voluminous data by applying classification/categorization algorithms and find some sense from these large texts (documents). Text Classification is a machine learning technique which assign a text document to one or more of a set of predefined classes [3]. At the outset, the text classification is carried in two ways i.e. manual and automatic. In the manual classification, the documents are categorized by concerning the knowledge and the interest of the person who predicts the facts to be considered as important. On the other hand, the automatic (using algorithms) text categorization schemes are capable of categorizing the document with prevalent importance considering time efficient and improved classification accuracy [7]. There are several methodologies available for the effective (automatic) text categorization. The purpose of this section is to identify, classify and discuss current research work. There are several

May – June 2020
ISSN: 0193-4120 Page No. 25246 – 25251

distinct methods available for the effective text classification. Text classification is the process of automatic sorting the set of available documents into some predefined categories [8]. It also includes automatic indexing, filling patterns, selection of dissemination, author attribution, survey of coding, grading of essay etc. The text classification is broadly divided into two categories, namely static data based learning and incremental data-based learning. Among the various text classification methodologies, the incremental (dynamic) learning has gained more interest because of its significance and diversified applicability [7].

2.0 Text Classification Based on Incremental Data-Based Learning

In this section we discussed different text classification algorithms based on incremental database learning. The machine learning algorithms such as Support Vector Machine, Neural Network, Random Forest, Fuzzy and Probabilistic are explored.

2.1 Random Forest

The existing text classification researchers utilizing the random forest algorithm are discussed in this subsection. S. Thamarai Selvi *et al.* [12] designed and developed a hybrid text classification algorithm by integrating the Rocchio algorithm and RF algorithm for implementing the multi-label text classification. The Rocchio algorithm is a supervised learning method which usage a set of vectors as an input dataset and classify it on the basis of cosine similarity. Stop word remover and word stemmer are used to overcome the limitations of this algorithm. The RF algorithm is an incremental database learning method based on decision tree. It randomly select datasets and features for the training purpose, and then assign the suitable category from the identified class. During text document classification, the input vector is passed through all the trees in the forest, the class with the maximum votes is termed as the output. This integrated method has overcome the limitations of both the algorithms present during individual implementation. This algorithm has more potential in comparison with other available methods like Naïve-Bayes, Multi-label KNN (ML-KNN) and fuzzy relevance clustering in terms of accuracy and effectiveness.

2.2 Neural Network

A neural network classifier is basically network of neurons. In the network input layer contains the feature terms, the output layer represents predefined target categories and the connections between the neuron units are represented by the assigned weights. The feature vector of the text document is assigned to the input layer neurons, activation function (sigmoid) is applied to these units and forwarded through the network. The value at the output layers determines the actual category of the input text documents. Feed forward and Back propagation are two typical methods of training the neural network[11].

ZhiHang Chen *et al.* [13] presented an Incremental Learning of Text Classification (ILTC). A framework is developed which learn the features of text class and then followed an incremental perceptron learning method. This algorithm has the capability of learning the new feature dimensions and new document classes incrementally. ILTC could learn from the new training data coming over period of time without referring or forgetting the previously learned training. This new learning method is economical and feasible for both temporal and spatial database as it does not require the need for pre-processing and storing the old instances. Author had discussed two learning phases of this method, first is the incremental learning of text classification features and second incremental neural learning of the features and newly introduced classes [5].

Patrick Marques Ciarelli *et al.* [21] proposed an incremental neural network. It is based on evolving Probabilistic Neural Network (EPNN), which take care the multi-label issue in text classification. EPNN is a neural network with compressed architecture supports an incremental learning algorithm. It has only one iteration for the training phase. EPNN had the capability of continuous learning with the reduced architecture. It always receive the training data even in non-availability of growth in the network architecture. The method required only a minimum number of weights and parameters and the architecture was also

May – June 2020
ISSN: 0193-4120 Page No. 25246 – 25251

comparatively stable as there was no proportional growth in a total number of weights to the total number of training instances. The computational cost is constant even with the increase in the number of training instances. EPNN performance is better than the other existing methods considered for evaluation with reference of the five parameters designed for multi-label issues in web pages. The complexity of the network architecture is low in comparison with other similar methods.

2.3 Using fuzzy system

In this subsection existing work on fuzzy system for text document classification is discussed. Nurfadhlina Mohd Sharefa and Trevor Martin [14] proposed an Evolving Fuzzy Grammar (EFG) method of text classification of crime related documents by using incremental learning. Incremental learning method was modelled on the basis of fuzzy grammar which is created by the transformation of a set of chosen text fragments. The identified grammars were integrated and used to find the matching with the learned fuzzy grammar and the testing dataset, further categorization was carried out on the basis of degree of parsing membership. As the derivation, parsing and grammar matching involved uncertainty; this fuzzy notion was used. The fuzzy union operator was used for integrating and transforming the individual text fragment grammars into more common representations of the already learned text fragments. The learned fuzzy grammar set was dependent on the existing pattern evolution. The results shown that the Evolving Fuzzy Grammar algorithm provided almost equivalent results and performance of the Machine learning algorithms. The EFG can be easily integrated into a more comprehensive grammar system; it was highly interpretable and has a low retraining adaptability time. The method was efficient and effective over the other related methodologies for text categorization, and it can considered as a potential technique for machine learning, text segment expression and representation.

2.4 Using probability theory

Bayesian classifier are also known as Naïve-Bayes classifier and it is based on Bayes probability theorey which has strong independent assumptions. Bayesian classifier estimates the joint probability of a given document belonging to a specific target class [9].

Renato M. Silva *et al.* [15] proposed a new algorithm of Minimum Description Length Text (MDLText) based on the principle of minimum description length for incremental text categorization. MDLText is a lightweight and fast multinomial algorithm which is highly scalable and efficient text classifier exhibiting fast incremental learning. A comprehensive grid search assuring a fair comparison was employed for the setting of best term weighting method and parameters for every dataset and method. MDLText method is found robust, overcome the issue of overfitting and have low computational cost. This method attained a significant balance between the computational efficiency of the algorithm and the predictive power. It has the better power of prediction and superior efficiency. All these features collectively made this method applicable in most of the online and real-world text classification on a large scale basis. This scheme was efficient in terms of time complexity, and the statistical evidence proved that it outperformed the other equivalent methods.

Farhad Pourpanah [16] proposed a multi-agent categorization system named Q-learning Multi-Agent Classifier System (QMACS) for minimizing the issues of data classification. The formulation of trust measurement is made by integrating the Bayesian formalism, belief functions and Q-learning. The big O-notation method is used for analyzing the time complexity of the algorithm. The bootstrap method with confidence levels of 95% is used for the statistical analysis of performance. For the negotiation between agents and agent teams of QMACS, a "sealed-bid first price auction" method is applied . The predictions from learning agents were combined for enhancing the QMACS's effectiveness and overall categorization performance. The results revealed the fact that the other schemes behave differently than the QMACS behaviour of integrating predictions for tackling distinct benchmark issues. This system had less time complexity in the training phase than the other

May – June 2020
ISSN: 0193-4120 Page No. 25246 – 25251

methods because of the involvement of many agents in architecture.

Renato M. Silva *et al.* [19] designed and developed a classifier Minimum Description Length Text (MDLText) based on the minimum description length for the filtration of disagreeable short text messages. This method was integrated with the incremental learning which made the predictive model scalable and continuously adaptive to the upcoming spamming methods. This method provides faster performance along with a linear increment in computational cost with the increase in a total number of features and samples. The outcome shows that the classifier with incremental learning enhanced the performance of the text classifier. The results statistics shows that the MDLText classifier provides have better performance than the Online Gradient Descent (OGD), Stochastic Gradient Descent (SGD), Approximate Large Margin Algorithm (ALMA), perceptron and Relaxed Online Maximum Margin Algorithm (ROMMA). This classifier was robust, overcome the overfitting issue during text categorization and had minimum computational complexity. The limitations of this method were high time complexity and dependence on the dictionary and TF-IDF method.

2.5 Using support vector machine

Support vector machine(SVM) is a linear classifier which is simple and effective as it work well with both positive and negative training data set to prepare the decision surface (hyper plane) that best separates the positive data from the negative one. This property is very much uncommon with the other classifier and it makes SVM a unique classifier. The data which is closest to the decision surface are called the support vec tor.

Wenbo Guo *et al*. [17] proposed an innovative and active learning algorithm integrated with incremental learning for solving the text categorization problems. SVM was used for estimating the information within the sample data, and it performed the learning of linear classifier with typical kernel-based feature space. The SVM has the accuracy with the optimal solution, and these things made it appropriate for active learning. Active learning was a machine learning algorithm, which was learned by the spontaneous selection of data, and utilized the distribution feature of datasets. The first step was the spectral clustering, which divided the dataset into two categories and the training of this classifier was done by using the labelled instances positioned at the category boundary. The incremental learning was integrated with the active learning for minimizing the computational cost and increasing the classification accuracy. This method was stable with minimal error rate and superior in terms of accuracy and effectiveness of the other state of the art active learning schemes.

2.6 Using hybrid model

Fabio Rangel *et al*. [18] developed a semi-supervised learning based on Wilkie, Stonham & Aleksander's Recognition Device (WiSARD) classifier (SSW) for text categorization. Because of the availability of variation in class distribution, the semi-supervised learning was satisfactory in the context of text categorization in social networks. The WiSARD classifier was a weightless neural network consists of individual classifiers and named discriminators assigned for learning the binary pattern of each category. This WiSARD acted as a one-shot classifier and permitted the incremental learning by adding training patterns to content. It performed text categorization in both unlabeled data and labelled data. The scheme was fifty times speedier than Expectation Maximization Naive Bayes (EM-NB) and Semi-supervised Support Vector Machine (S3VM) along with highly competitive accuracy. In all the tested datasets, the SSW portrayed a superior fitting time and better standard deviation.

Jianqiang Li *et al.* [20] presented a hybrid model, named Mixed Word Embedding (MWE), on the basis of word2vec toolbox for the text classification. This MWE integrated the two variants continuous skip gram model (SKIP-GRAM) and continuous bag-of-words model (CBOW) of word2vec. They both shared a common structure of encoding with the ability to capture the accurate syntax information of words. Additionally, MWE included the global text vector with the CBOW variant for capturing additional semantic information. The time

May – June 2020
ISSN: 0193-4120 Page No. 25246 – 25251

complexity of MWE was similar to the time complexity of SKIP-GRAM variant. The model was studied in the application and linguistic perspectives for the effective evaluation of MWE scheme. The empirical studies were conducted on the word similarities and word analogies for the linguistics. From the application point of view, the learned latent representations of sentiment analysis and document classification were regarded. MWE methodology was very competitive to most of the traditional classifiers, like a glove, SKIP-GRAM, and CBOW. The proximity and ambiguity among words were not considered, it was one of the limitations of this MWE scheme. In short, among the various text classification methodologies, the incremental (dynamic) learning has gained more interest because of its significance and diversified applicability [5].

3.0 Conclusions

Different existing works related to text document classification algorithms and the description of those works is briefly explained in this paper. The text document classification technique was evolved for organizing, structuring and classifying the large corpus of unstructured text. The text classifier assigns the test document with one or more predefined target categories. The curse of dimensionality and the broad semantic meaning along with the multiple context of the english words are some major challenges which needs to be addressed. Sparsity is one of the prevalent challenges of the text classification techniques. Moreover, the dimensionality remains high even after the removal of the stop words filtering and stemming. Because of the high dimensionality, the time complexity increases. These text classification methods enhance the classification accuracy and reduce the training and classification time by adopting different strategies and optimization algorithms.

4.0 References

[1] K. S. Deepashri and Ashwini Kamath, "Survey on Techniques of Data Mining and its Applications," International Journal of Emerging Research in Management & Technology, Vol.6, No.2, February 2017.

[2] Ramzan Talib, Hanify, Ayeshaz et al., "Text Mining: Techniques, Applications, and Issues," International Journal of Advanced Computer Science and Applications (IJACSA), Vol.7, No. 11, 2016.

[3] R. Sagayam, "A Survey of Text Mining: Retrieval, Extraction and Indexing Techniques," International Journal of Computational Engineering Research, Vol.2, No.5, 2012.

[4] W. Berry Michael, "Automatic Discovery of Similar Words: Survey of Text Mining: Clustering, Classification and Retrieval", Springer, PP. 24-43, 2004.

[5] ZhiHang Chen, Liping Huang et al., "Incremental Learning for Text Document Classification," In the Proceedings of International Joint Conference on Neural Networks, pp. 12-17, August 2007.

[6] Miji K Raju, Sneha T Subrahmanian et al. , "A Comparative Survey on Different Text Categorization Techniques," International Journal of Computer Science and Engineering Communications, Vol.5, No.3, pp. 1612-1618, 2017.

[7] Yung-Shen Lin, Jung-Yi Jiang et al., "A Similarity Measure for Text Classification and Clustering," IEEE Transactions on Knowledge and Data Engineering, Vol.26, No.7, July 2014.

[8] Said A. Salloum, Mostafa Al-Emran et al. "A Survey of Text Mining in Social Media: Facebook and Twitter Perspectives," Advances in Science, Technology and Engineering Systems Journal, Vol. 2, No. 1, pp. 127-133, 2017.

[9] B. Tang, H. He et al., "A Bayesian Classification Approach using Class-Specific Features for Text Categorization," in IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 6, pp. 1602-1606, June 1 2016.

[10] Chunting Zhou, Chonglin Sun et al., "A C-LSTM Neural Network for Text Classification," Computation and Language (cs.CL), November 2015.

[11] Alexis Conneau, Yann Le Cun et al., "Very Deep Convolutional Networks for Text Classification," In the Proceedings of 15th Conference of European Chapter of the Association for Computational Linguistics: Vol.1, pp. 1107–1116, April 3-7, 2017.

May – June 2020
ISSN: 0193-4120 Page No. 25246 – 25251

[12] S.Thamarai Selvi, P. Karthikeyan et al., "Text Categorization using Rocchio Algorithm and Random Forest Algorithm," In the Proceedings of IEEE Eighth International Conference on Advanced Computing (ICoAC), 2016.
[13] ZhiHang Chen, Liping Huang et al., "Incremental Learning for Text Document Classification," In the proceedings of International Joint Conference on Neural Networks, Orlando, Florida, USA, August 12-17, 2007.
[14] Nurfadhlina Mohd Sharef and Trevor Martin, "Evolving fuzzy grammar for crime text categorization," Journal of Applied Soft Computing, Vol. 28, pp. 175-187, March 2015.
[15] Renato M , Tiago A. et al., "MDLText: An Efficient and Lightweight Text Classifier," Knowledge-based Systems, Vol. 118, pp. 152-164, 15 February 2017.
[16] Farhad Pourpanah, Choo JunTan, Chee PengLim et al., "A Q-learning-based Multi-Agent System for Data Classification," Applied Soft Computing, Vol.52, pp. 519-531, March 2017.
[17] Wenbo Guo , Chun Zhong and Yupu Yang, "Spectral Clustering based Active Learning with Applications to Text Classification," In the Proceedings of 8th International Conference on Computer and Automation Engineering, Vol. 56, 2016.
[18] Fabio Rangel, Fabricio Firmino et al., "Semi-Supervised Classification of Social Textual Data Using WiSARD," In the proceedings of European Symposium on Artificial Neural Networks ESANN, Computational Intelligence and Machine Learning, 27-29 April 2016.
[19] Nihar M. Ranjan, Rajesh S. Prasad, " Automatic Text Classification using BP-Lion Neural Network and Semantic Word Processing" Imaging Science Journal, Taylor & Francis, ISSN: 1368-2199, Sept. 2017.
[20] Jianqiang Li, Jing Li et al., "Learning Distributed Word Representation with Multi-Contextual Mixed Embedding," Knowledge-Based Systems, Vol. 106, pp. 220-230, August 2016.
[21] Nihar M. Ranjan, Rajesh S. Prasad, " LFNN: Lion Fuzzy Neural Network based evolutionary model for text classification using context and sense based features", Applied Soft Computing Journal, Elsevier, PP 994-1008, ISSN: 1568-4946,July 2018.