Turkish Journal of Computer and Mathematics Education *Vol.12 No.01(2021), 830- 835*
Research Article

A REVIEW OF TEXT CATEGORIZATION ALGORITHMS USING THE MACHINE LEARNING PARADIGM

Narinder Gupta[1], Inderjeet Singh[2]
[1,2]Guru Kashi University, Talwandi Sabo

ABSTRACT

Automatic classification of text documents has recently been a hot topic in research. Information retrieval, machine learning, and natural language processing (NLP) techniques are required for proper classification of text materials. Our goal is to concentrate on three major approaches to automatic text classification using machine learning techniques: supervised, unsupervised, and semi-supervised. This paper reviews various text categorization algorithms using the machine learning paradigm in this study. We hope that our research will shed light on the relationships between various text classification techniques as well as the future research trend in this field.

Keywords: Automatic, text, classification, Information, Retrieval, NLP, Machine, Learning.

I. INTRODUCTION

We now have access to a vast amount of knowledge thanks to the widespread availability of internet text documents. For proper application, the available data must be structured in a methodical manner. The storage, searching, and retrieval of relevant text content for needy applications is made easier by systematic information organisation [9]. Text classification [3][4] is a useful technique for organising text documents into classes. Automatic Text classification appeals to businesses because it relieves them of the burden of manually arranging document repositories, which is not only costly, time-consuming, and error-prone. [1].

Automatic classification of text is a significant issue in numerous areas. It has numerous applications like robotized ordering of logical articles, spam sifting, ID of archive kind, creation attribution, computerized exposition evaluating, review coding, classification of news stories, and so forth This assignment falls at the intersection of information retrieval, NLP and machine learning [8].Information Retrieval is the finding of the archives which contain answers to the inquiries. Factual measures and strategies are utilized to accomplish the said goal[22] .Natural Language processing is utilized to improve comprehension of normal language by addressing reports semantically. This assists with improving classification results. Machine learning is worried about the plan and advancement of algorithms and procedures that permit PCs to "learn" in order to improve the normal future presentation. Algorithms used to prepare text classification frameworks in information retrieval are regularly specially appointed and inadequately comprehended. Specifically, almost no is thought about their speculation execution, that is, their conduct on archives outside the preparation information [14]. Machine learning procedures enjoy the benefit that they are better perceived from a hypothetical outlook, prompting execution assurances and direction in boundary settings.

II. SUPERVISED TEXT CLASSIFICATION ALGORITHMS

These algorithms use the training data, where each document is labeled by zero or more categories, to learn a classifier which classifies new texts. A document is considered as a positive example for all categories with which it is labeled, and as a negative example to all others. The task of a training algorithm for a text classifier is to find a weight vector which best classifies new text documents [21].

K-Nearest Neighbor classifier:

It's anything but a notable example acknowledgment calculation. Given a test report, the kNN calculation tracks down the k closest neighbors among the preparation records and uses the classifications of the k closest neighbors to weight the class competitors [2]. The comparability score of each neighbor archive to the test record is utilized as the heaviness of the classes of the neighbor report. This calculation depends with the understanding that the attributes of individuals from a similar class ought to be comparative. In this way perceptions found near one another in covariate space are individuals from a similar class.

It is appropriate for information streams. It doesn't assemble a classifier ahead of time.

Merits: This strategy is powerful, straightforward, non-parametric and simple to carry out.

Demerits: The significant downside of this technique is that it turns out to be moderate when size of preparing set develops The presence of unessential highlights seriously corrupts its precision.

Naïve Bayes Method:

The Bayesian method that makes independence assumption is termed as Naïve Bayes classifier [2]. It predicts by reading a set of examples in attribute value-representation and then by using the Bayes theorem to estimate the posterior probabilities of all qualifications. The independence assumptions of features make the features order irrelevant and presence of one feature does not affect other features in classification task.[22]

Merits: This method requires a small amount of training data to estimate the parameters necessary for classification. The classifiers based on this algorithm exhibited high accuracy and speed when applied to large databases.

Demerits: This method works well only if assumed features are independent; when dependency arises then it gives low performance.

Decision Trees:

The decision tree categorizes the training documents by constructing well-defined true/false queries in the form of tree structure. In this leaves represent the corresponding category of the text documents and branches represent conjunctions of features that lead to these categories [22].

Merits: This method works on data of any type. It is fastest even in the presence of large amounts of attributes.

Demerits: The major risk of implementation of decision tree is it over fits the training data with the occurrence of an alternative tree.

Decision Rules Classification:

This method uses the rule-based inference to classify documents to their annotated categories [22[These classifiers are useful for analyzing non-standard data. It constructs a rule set that describe the profile for each category. Rules are in the form of “If condition Then conclusion”, where condition portion is filled by features of the category, and conclusion portion is represented with the categories name or another rule to be tested.

Merits: This method is capable to perform semantic analysis.

Demerits: The major drawback of this method is the need of involvement of human experts to construct or update the rule set.

Support Vector Machines:

It is a statistical based learning algorithm [22]. This algorithm addresses the general problem of learning to discriminate between positive and negative members of a given class of n-dimensional vectors. It is based on the Structural Risk Minimization principle from computational learning theory. The SVM need both positive and negative training set which are uncommon for other classification methods. The performance of the SVM classification remains unchanged even if documents that do not belong to the support vectors are removed from the set of training data; this is one of its major advantages.

Merits: Amongst existing supervised learning algorithms for TC SVM has been recognized as one of the most effective text classification methods [7][21] as it is able to manage large spaces of features and high generalization ability.

Demerits: But this makes SVM algorithm relatively more complex which in turn demands high time and memory consumptions during training stage and classification stage.

III. UNSUPERVISED TEXT CLASSIFICATION ALGORITHMS / TEXT CLUSTERING

In unsupervised clustering, we have unlabelled collection of documents. The aim is to cluster the documents without additional knowledge or intervention such that documents within a cluster are more similar than documents between clusters. These are categorized into two major groups as partitioned and hierarchical.

Hierarchical algorithms produce nested partitions of data by splitting (divisive approach) or merging (agglomerative approach) clusters based on the similarity among them. Partitioned clustering algorithms group the data into un-nested non-overlapping partitions that usually locally optimize a clustering criterion.

Hierarchical Clustering Techniques:

Hierarchical clustering algorithms produce a cluster hierarchy in the form of tree structure named a dendrogram [11][21]. The root of the tree consists of single cluster containing all observations, and the leaves correspond to individual observations.

Merits: The major advantage of these techniques lies in its simplicity and ability to capture the information properly.

Demerits : It does not give discrete clusters, and we get different clustering for different experiment set.

This technique having following two subtypes:

- Divisive Hierarchical Clustering: Divisive algorithms start with one cluster at a time. During each iteration split the most appropriate cluster until a stopping criterion such as a requested number k of clusters is achieved. In this technique in each step the cluster with the largest diameter is split, i.e. the cluster containing the most distant pair of documents. As we use document similarity instead of distance as a proximity measure, the cluster to be split is the one containing the least similar pair of documents. Within this cluster the document with the least average similarity to the other documents

is removed to form a new singleton cluster. The algorithm proceeds by iteratively assigning the documents in the cluster being split to the new cluster if they have greater average similarity to the documents in the new cluster.

- Agglomerative Hierarchical Clustering: Agglomerative clustering algorithms start with each document in a separate cluster and at each iteration merge the most similar clusters until the stopping criterion is met. They are mainly categorized as single-link, complete-link and average-link depending on the method they define inter-cluster similarity [16][21]

Partitional Clustering Techniques:

In this clustering technique classes are mutually exclusive. Each object is the member of with which it is most similar. Each cluster may be represented by a centroid or a cluster representative; this is some sort of summary description of all the objects contained in a cluster. In this given the number of clusters k, an initial partition is constructed; next the clustering solution is refined iteratively by moving documents from one cluster to another [21].

Merits: It requires only one pass through dataset and therefore it is faster.

Demerits: In this resulting clusters are not independent of the order in which the documents are processed, with the first clusters formed usually being larger than those created later in the clustering run.

Kohonen's Self Organizing Network:

It uses a special type of neural network called Kohonen's self-organizing network. The novelty of the method is that it automatically detects the number of classes present in the given set of text documents and then it places each document in its appropriate class [20][21]. The method initially uses the Kohonen's self-organizing network to explore the location of possible groups in the feature space. Then it checks whether some of these groups can be merged on the basis of a suitable threshold resulting in expected clustering. These clusters represent the various groups or classes of texts present in the set of given text documents. Then these groups are labeled on the basis of frequency of the class titles found in the documents of each group. Merits: These algorithms can make it easy for humans to see relationships between vast amounts of data. These are commonly used in the applications for visualization aid.

Demerits: These are more complex and computationally extensive

VI. CONCLUSION

This review paper looked at the available research and looked at the primary classification strategies, as well as their benefits and drawbacks. According to the majority of the literature, the performance of a classification algorithm in text classification is heavily influenced by the quality of the data source and feature representation techniques, as irrelevant and redundant data features degrade the classifier's accuracy and performance. According to the results of the study, k-means and bisecting k-means perform the best among the unsupervised techniques in terms of time complexity and cluster quality. Support vector machines, on the other hand, perform the best among the supervised techniques, whereas naïve bayes performs the worst. Graph-based algorithms outperform co-training and Expectation-Maximization among semisupervised techniques, while a hybrid algorithm combining EM with nave bayes produces greater results. In many cases hybrid algorithms outperforms other and is gaining more research attention. However, more broad text categorization methods are required in order to effectively use large amounts of unlabeled

data. Such systems should additionally take into account the nature of the input text documents [e.g., single or multi-label, i.e., whether the text document belongs to one or more class labels]. The paper studied semi-supervised learning approaches and presents a new semi-supervised methodology for the Multi-Label Text classification issue, in which a new text document can be assigned to a more relevant category.

REFERENCES

[1]. A.Khan,B.Baharudin,Lan Hong Lee. A Review of Machine Learning Algorithms for Text-Documents Classification. Journal Of Advances in Information Technology, Vol. 1 , No. 1, Feb.2010.

[2]. Amarnag Subramanya, Jeff Bilmes , Soft-Supervised Learning for text classification. Proceedings of the Conference on Empirical Methods in Natural Language Processing. Pages: 1090-1099.2008.

[3]. Arturo Montejo-Raez . Automatic Text Categorization of documents in the High Energy Physics domain. Thesis submitted in 15 December, 2005.

[4]. Arzucan Ozgur. Supervised and unsupervised machine learning techniques for text document categorization.Thesis submitted in Department of Computer Science, Bogaziki University. 2004.

[5].Berkhin, P., "Survey of Clustering Data Mining Techniques", Research paper, Accrue Software, http://www.accrue.com/products/researchpapers.html, 2002.

[6]. Boyapati,V. Improving hierarchical text classification using unlabeled data. Proceedings of SIGIR, (2002).

[7]. Fabrizio Sebastiani , Machine Learning in Automated Text Categorization, ACM Computing Surveys, Vol. 34, No. 1, March 2002, pp. 1–47

[8]. Ganesh R.,Deepa P.,Byron Dom. A structure-sensitive framework for text categorization.CIKM International Conference on Information and Knowledge Management, Bremen, Germany, October 31 - November 5, 2005.

[9]. Goncalves, Paulo Quaresma. The impact of NLP techniques in the text multilabel classification problem.Hartigan, J., Clustering Algorithms, John Wiley & Sons, New York, NY, 1975.Dagan, Y. Karov, Dan Roth. Mistake- driven learning in text categorization

[10]. Kavi Narayana Murthy Advances in Automatic text categorization.DRTC Workshop on Semantic Web, Bangalore, India, 8-10 December, 2003.

[11]. L.Tang,S. Rajan,V.K. Narayanan. Large Scale Multi-Label Classification via MetaLabeler. In

[12]. Maribor, Slovenia. Text Categorization for Multi-label Documents and Many Categories. Twentieth IEEE International symposium on Computer-Based Medical Systems. June 20 2007.

[13]. N. Choudhary. An Unsupervised Text Classification Method using Kohonen's Self Organizing Network.N.M. Pise, Dr. Parag Kulkarni. A survey of Semi-Supervised Learning Methods". IEEE International conference on Computational Intelligence and Security. 2008.30-34.

[14]. Nigam, K., McCallum, A. K., Thrun, S., & Mitchell, T. M. (2000). Text classification from labeled and unlabeled documents using EM. Machine Learning,39, 103–134.

[15]. Nigam, K., McCallum, A., & Mitchell. T. (2006). Semi-supervised Text Classification Using EM. MIT Press.

[16]. Proceedings of the Data Mining and Learning 2009.

[17]. Tom M. Mitchell, The Discipline of Machine Learning , CMU-ML-06-108, July 2006.

[18]. Yu-Chuan Chang, S. Ming Chen. Multi-Label Text Classification based on a new linear classifier learning method and a category sensitive refinement method. ScienceDirect Expert Systems with Applications. Volume 34, Issue 3, April 2008. 1948-1953.

[19]. Z. Wang, X. Sun, D. Zhang. An optimal Text categorization algorithm based on SVM.

[20]. Zheng-Jun Zha, T. Mei.Graph-based Semi-Supervised Learning with Multi-label text classification. IEEE International conference on Multimedia.June 23 2008.1321-1324.