New Trends in Mathematical Sciences

http://dx.doi.org/10.20852/ntmsci.2019.334

Application of a collective semi-supervised method in text categorization: A case study of movie reviews

Nur Uylas Sati

Mugla Sitki Kocman University, Mugla, Turkey

Received: 7 August 2018, Accepted: 19 September 2018
Published online: 14 February 2019.

Abstract: Text categorization, namely text classification aims to predict the predefined labels (moods, subjects, etc.) of the given documents. It has been so difficult to categorize the text data in terms of their defined labels since the amount of online information increases day by day. Most of the real-world text datasets consist of too many unlabeled and a few labeled documents (instances). Therefore in this study we aim to use a collective semi-supervised approach that utilizes both unsupervised and supervised learning techniques for categorization of text. In numerical experiments we utilize The Movie Reviews text dataset and implement the suggested algorithm on it. MATLAB and WEKA software programs are used in the implementations. For performance evaluations we use accuracy results and running times. Obtained results are presented in tables.

Keywords: Text categorization, semi-supervised classification, *k*-means clustering, mathematical programming.

1 Introduction

Text categorization, namely text classification aims to predict the predefined labels (moods, subjects, authors, titles etc.) of the given documents. Due to the heavy increase of online information, it has been so difficult to categorize the text data as required. There are various researches presented for the aim of categorization of text data in the literature. The commonly used techniques are data mining techniques. Data mining, also called Knowledge Discovery in Databases (KDD), is the field of discovering novel and potentially useful information from large amounts of data [5]. Data classification defined in data mining, aims to classify the future data in terms of the predetermined labels by using the given data. Data classification techniques have three types according to the used datas' labels. If the current dataset consist of entirely labeled data, supervised classification techniques are used, if it has a few labeled and many unlabeled data, semi-supervised techniques are used and if it consisits of entirely unlabeled data, unsupervised techniques are used for data classification.

Since most of the real-world text datasets consist of too many unlabeled and a few labeled documents (instances), in this study we will study on semi-supervised classification for categorization of text. In the literature there can be found various semi-supervised classification techniques as self training [22], co training [3], transductive support vector machines [4], graph based methods [20]. A brief review of these techniques can be found in [15]. In addition to these approaches some researches collective text classifiers for more effective classification. The suggested collecitve approaches till 2016 are reviewed in [11].

In this study, we experiment the approach that combines an unsupervised and supervised technique. This approach was firstly defined in [16] but not experimented in text classfication. Here we aim to get good performance results by using this approach in text classification. In this approach labeled points are used for determining the center points of the classes by using an unsupervised technique, *k*-means clustering method. Then the unlabeled points are labeled according

* Corresponding author e-mail: nuruylas@gmail.com

to the obtained center points. After this process we have a dataset that we can make supervised learning on. In the following section the used clustering method and how the unlabeled points are labeled will be explained. Then the suggested algorithm will be presented step by step. In the third section, used dataset for the case study will be explained and the data preparation processes will be expressed. In the fourth section obtained results will be presented in tables and comments on the results are made. Finally in the last section the paper is concluded.

2 Text categorization

In this section, text categorization namely text classification will be explained. The main aim in text categorization is classifying the documents into a fixed number of predefined classes (labels). The effectiveness of the used algorithm in this process is so important but the proecesses before the implementation of the algorithm has also importance on getting satisfying good results. The steps of this process can be given as follows:

(1) Determining of text data collection: First thing to do in text categorization is determining the text data collection. These document collections consist of many words.

(2) Text preprocessing: The text documents are simplified. These words are cleaned out from stop words, conjunctions, meaningless expressions and then root of words are determined, e.g. classification to classify, application to apply. Commonly the steps taken in text preprocessing are Tokenization and Removing Stop Words like frequently occurring “the”, “a”, “an” etc. [2].

(3) Attribute selection: In attribute selection also known as feature selection, relevant words (every word can be thought as a feature in text categorization) in the preprocessed documents are detected and the rest of the words are left out from the process. An effective attribute selection increases the effectiveness of the algorithms and also decreases the running time.

(4) Text transformation: In text transformation, documents are defined with a goal- oriented suitable representation for learning algorithm. Namely unstructured data should be transformed into structured data. Here the aim is to reduce the complexity of the documents for an easy managing procedure by transforming the full text version of the document to a document vector. Vector space model (SMART) where documents are represented by vectors of words, is the commonly used document representation [2].

(5) Data mining: In data mining step, the most convenient method and algorithm is chosen and implemented to the transformed dataset. In the literature Naive Bayes [7], Rocchio’s method [10], *k*- nearest classifier [18], Support vector machine (SVM) [12], decision tree (DT) [9], polyhedral conic functions (PCFs) [17] are used for text classifaction. Besides, different classifiers are combined to classify the text data. The review of collective text classifiers were presented in [11].

(6) Evaluation: In performance evaluations, many measures have been used, such as F-measure, fallout, error, accuracy, cross validation etc. In this paper, accuracy values determined in results and finding section is used.

3 Methodology

In this section the used unsupervised and supervised techniques in data mining step of the suggested text categorization algorithm will be expressed. How the two techniques are combined will be explained and lastly the suggessted algorithm will be presented in step form. For labeling the unlabeled points we utilize from commonly used clustering method, *k*-means. *k*- means clustering method was firstly proposed by Mac Queen in 1967 to partition the unlabeled dataset into *k* parts in terms of the similarities. k-means algorithm is given as follows in [1].

Step 1. Choose a seed solution consisting of *k* centers (not necessarily belonging to A);
Step 2. Allocate data points to its closest center and obtain *k*-partition of A;
Step 3. Recompute centers for this new partition and go to Step 2 until no more data points change cluster.

In this study, this algorithm is used to obtain k center points of the classes by using the labeled data. We implement this algorithm for each of the class and we obtain k center points for each of one all. Then we label each unlabeled data by the closest center's label. We use euclidian distance in the calculations as follows.

If $a = (a_1, a_2,..., a_n)$ and $b = (b_1, b_2,..., b_n)$ are n-dimensional two data, then the distance (d) from a to b is given by the formula:

$$d = \sum_{i=1}^{n} \sqrt{(a_i - b_i)^2}.$$

After labeling the unlabeled points we obtain a dataset that we can implement all supervised classification algorithms on. In this study we experiment J48, Logistics, ClassificationViaRegression and NaiveBayes algorithms since they are easily accessible in Waikato Environment Knowledge Analysis (WEKA). These methods are expressed briefly below and also a detailed review of these methods can be found in [19].

J48: (J48) is an algorithm used to generate a decision tree developed by Ross Quinlan mentioned earlier [14].

Logistics: Logistic regression is an alternative method to the Linear Discriminant Analysis that generates classifier functions to separate two or more groups by minimizing the misclassification cost [13].

ClassificationViaRegression: It uses regression methods for classification. Class is binarized and one regression model is built for each class value [8].

NaiveBayes: Bayesian classifiers assign the most likely class to a given point described by its feature vector.

The step form of the suggested algorithm can be given as follows [16].

Step 0. (Initilization): Determine the number of clusters that will be used in the clustering algorithm and seperate the given dataset in terms of their labels (datasets for each of the labels and one for the unlabeled).
Step 1. Find center points of each of the labeled datasets via k-means clustering algorithm.
Step 2. Label each unlabeled point as the closest center points' label by using euclidian distance.
Step 3. Redetermine the given dataset in accordance with the new labeled data.
Step 4. Implement a supervised data classification algorithm on the redetermined dataset.
Step 5. Define the obtained function or model that separates the classes and STOP.

4 Dataset preparations

For the text categorization experiment on the suggested algorithm, we use a predefined "Movie Reviews" text dataset that is constituted by Kaya A. and Amasyalı M. F. in 2010 and accessible in (www.kemik.yildiz.edu.tr). It has three different classes as "negative, positive and neutral" For each of the class it has 35 movie review documents. Totally it has 105 instances. It has 6440 attributes that is found by text2arff software program, accessible in (www.kemik.yildiz.edu.tr). The attributes of the instances (review documents) are defined by the number of every word stem (w_i) exists in the document. These numerical details are also shown in Table 1.

Table 1: Details of movie reviews dataset.

Number of attributes	Number of review documents	Number of Negative Reviews	Number of Positive Reviews	Number of Neutral Reviews
6440	105	35	35	35

We decrease the dimension of the problem by using a feature selection algorithm, InfoGainAttributeEval in WEKA machine learning software program. This algorithm evaluates the worth of an attribute by measuring the information gain with respect to the class. We choose the attributes whose rank values are bigger than 0 (>0). Thus in the experiments, we use just 130 attributes. The other details are left alone. Since we study on a numerical dataset, we assign the labels positive, negative and neutral respectively as 0, 1 and 2. In the implementations %5, 10 %, 15% and % 20 of the whole data are used as labeled and the remain as unlabeled data since we are making semi-supervised learning.

5 Results and findings

In the numerical experiments we use accuarcy and running times as performance evaluations. We use the whole (10 %0) dataset for testing. Accuracy value is calculated as follows:

$$accuracy\text{ value} = \frac{number\ of\ correctly\ classified\ documents \times 100}{number\ of\ all\ documents}.$$

WEKA (Waikato Environment Knowledge Analysis) and MATLAB (Matrix Laboratory) software programs are used for implementations. For step 1 and 2 the implementations are done on MATLAB by writing codes and for step 4 and 5 we use WEKA software program since it has currently ready to use codes of the supervised classification algorithms. We implement J48 decision tree, Logistics, ClassificationViaRegression and NaiveBayes algorithms that are accessible in WEKA classifiers package for the supervised classification processes. For observing the effects of the unlabeled points, we also implement the supervised classification algorithms just on the each 5%, 10 %, 15% and 20% of the whole data as labeled data. Obtained results are given in Table 2 and 3. Besides these implementations, we applied different collective semi-supervised methods in literature on WEKA for more confidential results on the case study. The used collective semi-supervised methods are expressed briefly as follows and the results are presented in Table 4.

YATSI: "Yet another two stage idea" was defined in 2006 by Driessens and his friends [6]. It is a collective classifier that uses the given classifier to train on the training set and labeling the unlabeled data. The chose of classifier and nearest neighbour search algorithm is made by the experimenter.

LLGC: LLGC (Learning with local and global consistency), was presented in 2003 by Zhou and his friends [21]. It is a collective classifier that generates a smooth classifier function for labeled and unlabelled data.

Weighting: It is a collective classifier that uses one classifier for labeling the test data after training on the train set. In the applications J48 classifier is used for labeling the test data.

Table 2: Results of suggested semi-supervised algorithms.

A*	5%		10%		15%		20%	
	Accuracy (%)	Time (sec)	Accuracy (%)	Time (sec)	Accuracy (%)	Time (sec)	Accuracy (%)	Time (sec)
J48	82.85	0.03	82.85	0.02	39.04	0.01	82.85	0.05
Logistics	**97.14**	0.09	**97.14**	0.11	42.85	0.09	**97.14**	0.2
ClassificationviaRegr.	74.28	0.25	74.28	0.07	37.14	0.04	74.28	0.14
Naive Bayes	85.71	0.01	85.71	0.01	**52.38**	0.01	85.71	0.01

A*=k-means on labeled points and labeling unlabeled ones + Supervised classification technique on the new dataset.

Table 3: Results of supervised algorithms.

B*	5%		10%		15%		20%	
	Accuracy (%)	Time (sec)	Accuracy (%)	Time (sec)	Accuracy (%)	Time (sec)	Accuracy (%)	Time (sec)
J48	36.19	0.02	47.61	0.01	54.28	0.01	48.57	0.02
Logistics	46.66	0.01	55.23	0.01	**65.71**	0.01	63.80	0.09
ClassificationviaRegr.	34.28	0.02	49.52	0.01	57.14	0.02	54.28	0.15
Naive Bayes	**49.52**	0.02	**57.14**	0.01	64.76	0.01	**68.57**	0.01

B*=Supervised classification technique on labeled points.

Table 4: Results of WEKA collective classification algorithms.

C*	5%		10%		15%		20%	
	Accuracy (%)	Time (sec)	Accuracy (%)	Time (sec)	Accuracy (%)	Time (sec)	Accuracy (%)	Time (sec)
YATSI	**32.00**	0.01	**38.29**	0.01	**37.07**	0.01	40.47	0.01
LLGC	**32.00**	0.01	30.85	0.01	30.33	0.01	29.76	0.01
Weighting	**32.00**	0.01	35.10	0.01	30.33	0.01	**44.04**	0.01

C*=Collective Classification Algorithms.

The comments on the results are given as follows.

(1) On the whole implementations (except 15%), when we compare the Table 2 and Table 3 results, we get better accuracy results by using unlabeled documents in the classification process, in other words by using suggested semi-supervised learning algorithm for text categorization.

(2) In the implementation of 15% labeled data, the *k*- means clustering algorithm can not reflect the real structure of the dataset so we cannot get effective results. It should not be forgotten that the chose of the labeled points are made randomly for each of the implementations.

(3) The running times are so close to each other so we cannot make any strict comments on the compilation time of the algorithms.

(4) When we compare the accuracy results of WEKA collective classification algorithms in Table 4 with the suggested method's results in Table 2, it is seen that higher accuracy results are obtained by the suggested semi-supervised algorithm. When we compare the running times, it is seen that WEKA collective classification algorithms are more effective since WEKA is a collection of ready to use Java codes of machine learning algorithms.

6 Conclusion

In this study a semi-supervised algorithm is experimented for text categorization. Two data mining methods are combined. Firstly for labeling the unlabeled data, k-means method is used then a chosen supervised classification algorithm is implemented on the rearranged dataset. In the numerical experiments, we utilized a text dataset, Movie Reviews after making feature selection and data preparations. Since it is a labeled dataset, in the implementations 5%, 10%, 15% and 20% of the points are used as labeled and the remain as unlabeled points. For performance evaluations, accuracy and running times are presented in tables. To see the effect of the unlabeled points on the suggested approach, supervised classification algorithms are also implemented on given labeled points separately and the original dataset is used in testing. According to the obtained results, suggested collective semi-supervised classification approach get convincing good results in terms of the accuracy values in text classification.

Competing interests

The authors declare that they have no competing interests.

Authors’ contributions

All authors have contributed to all parts of the article. All authors read and approved the final manuscript.

References

[1] Bagirov, A. M.; Mardaneh, K.: Modified global k-means Algorithm for Clustering in Gene Expression Data Sets. Intelligent systems for Bioinformatics 2006, Hobart, Australia, Australian Computer Society (ACS), (2006).

[2] Bhumika, Sehra, S,; Nayyar, A: A Review Paper On Algorithms Used For Text Classification, International Journal of Application or Innovation in Engineering & Management (IJAIEM), Volume 2, Issue 3, March, (2013).

[3] Blum, A. and Mitchell, T.: Combining labeled and unlabeled data with co-training, Proceeding COLT’ 98 Proceedings of the eleventh annual conference on Computational learning theory, Pages 92-100, (1998)

[4] Bruzzone, L.; Chi, M.; Marconcini, M.: A Novel Transductive SVM for Semisupervised Classification of Remote-Sensing Images, IEEE Transactıons on Geoscience and Remote Sensing, vol. 44, no. 11, november, (2006).

[5] Chen, M.S. , Han J. , and Yu P. S.: Data mining: An overview from a database perspective. IEEE Trans. Knowledge and Data Engineering , 8:866–883, (1996).

[6] Driessens, K., Reuteman, P., Pfahringer, B. and Leschi, C. Using Weighted Nearest Neighbor to Benefit from Unlabeled Data. In: Proc. 10th Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 60–69. (2006).

[7] Eui-Hong SH,; George K,; Vipin K: Text Categorization Using Weighted Adjusted k-Nearest Neighbor Classification, Department of Computer Science and Engineering. Army HPC Research Centre, University of Minnesota, Minneapolis, USA. (1999).

[8] Frank E., Wang Y, Inglis S, Holmes G, Witten I H. “Using model trees for classification”, Machine Learning. 32(1):63-76, (1998).

[9] Hmeidi I,; Hawashin B,; El-Quawasmeh E.: Performance of KNN and SVM classifiers on full word Arabic articles, Advanced Engineering Informatics, 22, 106-111, (2008).

[10] Ittner, D,; Lewis, D,; Ahn, D.: Text Categorization of Low Quality Images, In: Symposium on Document Analysis and Information Retrieval, Las Vegas, NV .pp. 301-315, (1995).

[11] Jain, A,; Mandowara, J: Text Classification by Combining Text Classifiers to Improve the Efficiency of Classification, International Journal of Computer Application (2250-1797), Volume 6– No.2, March- April (2016).

[12] Joachims, T.: Text categorization with support vector machines: learning with many relevant features, Universit£t Dortmund lnformatik LS8, Baroper Str. 301, (1999).

[13] Press S J, Wilson S. “Choosing between Logistic Regression and Discriminant Analysis”, Journal of the American Statistical Association, Volume 73, - Issue 364, (1978).

[14] Quinlan R. “C4.5: Programs for Machine Learning”. Morgan Kaufmann Publishers, San Mateo, CA. (1993).

[15] Sadarangani, A. and Jivanni, A.: A Survey of Semi-Supervised Learning, International Journal of Engineering Sciences & Research Technology, October, 5(10), (2016).

[16] Uylaş Satı, N.: A collective learning approach for semi-supervised data classification, Pamukkale University Journal of Engineering Sciences, doi: 10.5505/pajes.2017.44341, (2018).

[17] Uylaş Satı N.; Ordin B.: “Application of the Polyhedral Conic Functions Method in the Text Classification and Comparative Analysis,” Scientific Programming, vol. 2018, Article ID 5349284, 11 pages, https://doi.org/10.1155/2018/5349284, (2018).

[18] Tam, V,; Santoso, A,; Setiono, R.: A comparative study of centroid-based, neighborhood-based and statistical approaches for effective document categorization, Proceedings of the 16th International Conference on Pattern Recognition, pp.235–238, (2002).

[19] Witten. I, H, Frank. E, Trigg. L, Hall. M, Holmes. G, Cunningham. S, J.: Weka: Practical machine learning tools and techniques with Java implementations. (Working paper 99/11) (1999).

[20] Zha, Z.; Mei, T.; Wang, J.; Wang, Z.; Hua, X.: Graph-based semi-supervised learning with multiple labels, Journal of Visual Communication and Image Representation, Vol. 20 Issue 2, February, Pages 97-103, (2009).

[21] Zhou, D., Bousquet, O., Lal, N. T., Westor, J. and Schölkopf, B. Learning with local and global consistency, Max Planck Institute for Biological Cybernetics, technical Report No:TR-112, June. (2003).

[22] Zhu, X.: Semi- Supervised Learning Literature Survey. Computer Sciences TR 1530 University of Wisconsin – Madison, Last modified on December 14, (2007).