Journal of Engineering Sciences and Information Technology
Volume (4), Issue (2) : 30 June 2020
P: 110 - 120

AJSRP
ISSN: 2522-3321

مجلة العلوم الهندسية وتكنولوجيا المعلومات
المجلد (4)، العدد (2) : 30 يونيو 2020م
ص: 110 -120

English Text Classification Using Improved Recursive Feature Elimination (IRFE) Algorithm

Esraa H. Abd Al-Ameer

Ahmed H. Aliwy

Faculty Education for Girls || University of Kufa || Iraq

Abstract: Documents classification is from most important fields for Natural language processing and text mining. There are many algorithms can be used for this task. In this paper, focuses on improving Text Classification by feature selection. This means determine some of the original features without affecting the accuracy of the work, where our work is a new feature selection method was suggested which can be a general formulation and mathematical model of Recursive Feature Elimination (RFE). The used method was compared with other two well-known feature selection methods: Chi-square and threshold. The results proved that the new method is comparable with the other methods, The best results were 83% when 60% of features used, 82% when 40% of features used, and 82% when 20% of features used. The tests were done with the Naïve Bayes (NB) and decision tree (DT) classification algorithms , where the used dataset is a well-known English data set "20 newsgroups text" consists of approximately 18846 files. The results showed that our suggested feature selection method is comparable with standard Like Chi-square.

Keywords: Decision Tree; Naïve Bayes, Text Classification, features selection.

تصنيف النص الإنجليزي باستخدام الخوارزمية العودية المحسنة لإزالة الخواص (IRFE)

إسراء حسين عبد الأمير

أحمد حسين عليوي

كلية التربية للبنات || جامعة الكوفة || العراق

الملخص: تصنيف الوثائق هو أحد أهم المجالات في معالجة اللغات الطبيعية وتنقيب النصوص. هناك العديد من الخوارزميات التي يمكن استخدامها لهذه المهمة. في هذه الورقة، نركز على تحسين TC عن طريق اختيار الخواص. وهذا يعني تحديد مجموعة فرعية من الخواص الأصلية دون التأثير على دقة العمل، حيث اقترحنا طريقة جديدة لاختيار الخواص والتي يمكن أن تكون صيغة عامة ونموذج رياضي الى (RFE). قارنا الطريقة المستخدمة مع طريقتين أخرى معروفة لاختيار الخواص هي : Chi-square, and Threshold. أثبتت النتائج ان الطريقة الجديدة منافسة للطرق الأخرى المعروفة، وكانت أفضل النتائج 83% عند استخدام 60% من الخواص، و 82% عند استخدام 40% من الخواص، و 82% عند استخدام 20% من الخواص. تم إجراء الاختبارات باستخدام خوارزميتي التصنيف (DT) Decision Tree و(NB) Naïve Bayes . حيث مجموعة البيانات المستخدمة هي مجموعة بيانات معروفة(20 newsgroups text) تتكون من (18846) مستند. وأظهرت النتائج بأن طريقة اختيار الخواص المقترحة قابلة للمقارنة مع الطريقة القياسية لاختيار الخواص مثل (Chi-square).

الكلمات المفتاحية: تصنيف النص، شجرة القرار، نايف بايز، اختيار الخواص.

DOI: https://doi.org/10.26389/AJSRP.R080420

1. Introduction

The text classification is very important because of massive of numbers of documents.. Each document contains irrelevant information that reduces accuracy of text classification. feature selection task in text classification focuses on identifying a relevant information without affected on accuracy of the work. The aim for feature selection is find useful styles in text documents. feature selection task will reduce original features to new features by applying some functions where the new feature set has features or dimensions lower than the original set [1].

Text Classification is achieved by classifying documents based on their content (or/and its topic) into predefined categories [2]. TC is too important; therefore, many methods and algorithms, different in their efficiency and computation accuracy, were used to solve it. TC could be utilized for document indexing, web browsing, and e-mail filtering. Moreover, it is important and active area for machine learning and information retrieval (IR) intersect [3].

Here, literature review for some existent methods for improving TC by feature selection: In Uysal, A. K. (2016) [4] used the proposed improved global feature selection scheme (IGFSS) for text classification and compared it with classical feature selection techniques are Information gain (IG), Gini index (GI),Distinguishing feature selector (DFS),Odds ratio (OR).used three datasets , the first dataset consists of the top-10 classes of the celebrated Reuters-21578 ModApte split , the second dataset is another popular benchmark collection namely WebKB which has four classes, and The third dataset is Classic3 whose class distribution is nearly homogenous among three classes , where used 70% of documents for training and 30% for testing. using Support vector machine (SVM) and Naïve bayes (NB) as classifiers for texts. in Pandya & Pandya (2015) [5] used C5.0, ID3, and C4.5 algorithms and compared between them , reduced error pruning technique was used with decision tree, where the accuracy is from1 to 3%. The algorithm was implemented using WEKA packages. in Singhal & Sharma (2014) [6] used correlation based feature selection (CFS) algorithm using the Naive Bayes classifier. They improving the selected dataset from the Tuned IT repository of machine learning databases. In Dey Sarkar, S., Goswami, S., Agarwal, A., & Aktar, J. (2014) [7] suggested the two step feature selection technique upon firstly a univariate feature selection and then feature clustering, to classify over 13 datasets by using naive Bayes algorithm. in Galathiya, Ganatra & Bhensdadia (2012) [8] Compared among ID3, C4.5 and C5.0 , where used cross validation, reduced error pruning, feature selection, and model complexity along for classification. The utilized dataset are Zoo dataset, Ionosphere, Contact-lenses, Au1_1000, Breast Cancer, iris, Annealing and Weather nominal dataset. in Galathiya, Ganatra & Bhensdadia (2012) [9] used C5.0 to implement used cross validation, reduced error pruning, feature selection, and model complexity of the original C5.0 in order to reduce the error ratio. The reduced error pruning technique was used in the decision tree to solve over fitting problem. The used dataset were RGUI with WEKA packages. in Yuan (2010) [10] implemented conditional probabilities to finding dependency between features and apply it to Naïve Bayes classifier.

the used data is 10 data sets obtained from UCI machine learning repository and LIBSVM. in Harrag, El-Qawasmeh & Pichappan (2009) [11] improving Naïve Bayes text classification by “calculating posterior probability and reducing dimension of feature words of text". The used data set was "the Starter Edition text classification data made by Sogou laboratory which has17910 documents of 9 categories". In Wang, Y., & Wang, X. J. (2005) [12] used variance-mean based feature filtering technique and compared it with DF, CHI and IG classical feature selection techniques , applied to Chinese language text by Ronglu Li published on the website http://www.nlp.org.cn, where the corpus divide into two sets a training set having 10 categories with 100 texts in each and a test set having the same 10 categories with another 100 texts in each also. used "the open source Chinese lexical processing software ICTCLAS made by ICT, CAS (Institute of Computing Technology, Chinese Academy of Sciences)" , Using SVM as the classifier for text classification. In Liao, S., & Jiang, M. (2005) [13] use term frequency and cross entropy (TF-CE) as feature selection. applied by using HowNet which refered to Chinese words dictionary to extract concept attributes from the words in the text by using C-Tree algorithm. In Rogati, M., & Yang, Y. (2002) [14] suggested using filter methods which include the $\chi 2$ statistic, combining them with DF or IG, and eliminating the rare words. the datasets used are two benchmark collections Reuters-21578 and part of RCV1,where the experiments were by using for all four classifiers are Naive Bayes, Rocchio, K-Nearest Neighbor and Support Vector Machines. In Lewis & Ringuette (1994) [15] used decision tree learning and Bayesian classifier algorithm on two text categorization data sets. The first of them was a set of 21,450 Reuter's newswire stories. The second data set included of 1,500 documents of the U.S. Foreign Broadcast Information Service (FBIS) that used in the MUC-3 evaluation of natural language processing systems. Documents used a set of 8,876 binary features corresponding to English words occurring in 2 or more training documents. The features ranking for each category by using the information gain measure. This Techniques achieve optimizing in text classification. In This paper introduces an investigation on the performance of two widely used feature selection namely Chi-square, Threshold and suggested way Improved recursive feature elimination (IRFE). The experiments are conducted using decision tree (DT), and Naïve Bayes (NB) classifiers to classify a published English corpus which It involves about 18846 newsgroups positions upon 20 subjects divided into two groups: one for training and others for testing. The training files consist of 11314 document and the testing files consist of 7532 where the numbers of all features are 101322 without reduction, and the two algorithms were implemented using Python version 3.4. It is a high-level programming language.

2. Methodology

2.1. Naive Bayes Classifier

Naïve Bayes classifier is easy probabilistic classifiers upon on a common assumption, where all the features are independence of each other according to the class variable [16]. the binary independence classifier is from of the best known approaches to Naive Bayes classifier which used binary- valued representations of documents [17]. it is fast and easy for implemented [18]. it is effective sufficient to text classification in many fields, and although it is less accurate than other discriminative methods as, a Support Vector Machines (SVM) [19]. NB classifier mostly utilize Bayes' rule [20]:

$$P\ (c_i \mid d) = (P\ (c_i)\ p\ (d \mid c_i))\ /\ p\ (d)$$

Where:

$p\ (c_i \mid d)$ = probability of class, i given a document d,

$p\ (c_i)$ = probability of class, i which calculating by:

$$P\ (c_i) = N_i\ /N$$

Where:

Ni = is the number of documents in class i and N is the number of the all classes,

p (d | ci) = is the probability of a document d given a class i, p (d) is the probability of document d.

2.2. Decision tree (DT) Classifier

When decision tree consist of tree, where internal node is label by term, branches represent weight, and leaf represent the class, it implemented the classification through the query structure from root until it reaches to the goal for the classification of the document (a certain leaf), [16]. Most of training data was not fit in memory, decision tree construction it is ineffectual because of swapping of training tuples [21]. it is the widely used inductive learning methods, and learned from labeled training documents. ID3 is from of the most used decision tree learning algorithms, and it has stretching as C4.5 and C5. DT which is a flowchart such as tree structures, where the internal node indicated test in document, the branch represents outcome to the test, and the leaf node holds a category label , from decision tree Advantages is (i)capable to learn disjunctive terms and (ii) their robustness to the noisy data is convenient for document classification, while it disadvantages is (i) learning of decision tree algorithms cannot warranty to return the optimal decision tree [22]. Data comes in records as follow [23].

$$(X, Y) =(x_1, x_2, x_3\dots x_k, Y)$$

Where:

vector Y= is the target variable.

vector x= is composed of the input variables, x1, x2, x3,...etc.

decision tree linearizes into decision rules. the result from this is the contents to the leaf node, and conditions along the path are formed in relation to the requirement condition [24]. In general, the rules is as following "If condition1 and condition2 and condition3 then outcome". Decision rules creates by building association rules with the aim variable in the right, also It indicate temporal or causal relationships [25].

2.3. Feature Selection techniques

2.3.1. Chi Square technique

Chi Square (χ2) [26] defines as a fashion of selecting folk features which evaluating the features individually via calculating (χ2 square) statistics with respect with the layers.in other words, chi-square try to analyzed the dependent between the expression and the category. If they are independent, the score is 0 otherwise 1. a term with a higher chi-squared score is more informational. The formulation for CHI is [27]:

$$X^2(c,t) = \frac{N*(AD-BC)}{(A+C)(B+C)(A+B)(C+D)}$$

Where:

A. frequency of t and c occurrences,

B. frequency of t occurrences without c,

C. frequency of c without t,

D. frequency of non-occurrence of both c , t and N is the quantity of document.

2.3.2. The Suggested Improved Recursive Feature Elimination (IRFE) technique

The improved recursive feature elimination (IRFE) is suggested to solve the problem of slow RFE instead of eliminating one feature with many trainings, group of features will be eliminated for little trainings. This is done by partitioning the features into p equal parts. Then p training will be done with eliminating features equal to size of one part in each training and testing.

This method in this way has a problem with feature elimination because 100 features will take same error value (the same ranking). One group, may be, has the most important feature and the less important feature therefore the overlap between two groups can be taken as feature elimination not takes the part only. Each one part will be examined more than once but the test number remains the same. In this case number of parts of feature elimination (e) will be introduced.

In the explained method, neglecting the parts will affect the performance of the system this affect will be the error for these parts. The worst part will give very small error. Each part will take its errors from summation of e's neglecting tests. Then each part will have a rank according to these errors.

Another problem will be raised where all features in one group will have same error rate.

So, this problem can be solved by multiplying the value of each feature in term frequency (tf) which breaks the equality for the same group. Different values of the features will be obtained. The features which have lowest ranks will be eliminated. This will be repeated till the specified number of features will be remained. The neglected features should be less than a half of the value of the segment and should be as little as possible.

Then the rank of feature i is:

$$\breve{R}(f_i) = (\sum_{t_j \in T_i} t_j^2)/e$$

The final rank of feature will be:

$$R(f_i) = tf_i(\sum_{t_j \in T_i} t_j^2)/e$$

Where $R(fi)$ represented rank of each feature in each group , tfi represented value of term frequency for each feature in groups, tj represented value of error for each test , Ti represented list of errors and e represented eliminated parts. The skeleton of The IRFE algorithm and IFE is given in Figure 3.2& Figure 3.3 respectively.

Algorithm (1) Improved Recursive Feature Elimination

```
Input: F: set of features where F= {f1, f2 … f n}, P: the number of parts (groups)
 where: 1<p≤ n, e: the number of eliminating,
 where: 1<e≤ p/2. D: Dataset,
 n: number of the required features.

Output: FN: the best features which are subset of F.

Step 1: FN=F .

Step 2: FN=Improved Feature Elimination (FN, P, e, D)://Return output R.

Step 3: if n<=|FN|-|FN|/P then

 FN=FN - worst features (|FN|/P): //eliminate |FN|/P worst features

 Else: FN =FN - worst features (|FN|-n): //eliminate |FN|- n worst features .

Step 4: repeat step2 to step3 until |FN| =n .

End
```

3. The used dataset

A well-known English data set, 20 newsgroups text data set, was used for testing the proposed approaches. It involves about 18846 newsgroups positions upon 20 subjects divided into two groups: one for training and others for testing. The training files consist of 11314 document and the testing files consist of 7532 where the numbers of all features are 101322 without reduction.

In the implementation, the whole data set (DataSet1) was taken in all tests and then parts (DataSet2) of the categories were taken for the same tests. This is done for study the classification for high and low numbers of classes. The parts of data set testing consist of four classes where the training files consist of 2034 document and the testing files consist of 1353 and the numbers of all features are 26576 without reduction.

1. Performance Measures

[28] F-measure (combines precision and recall) used to evaluate the useful of feature selection techniques.

In classification problem, the system can give four states for a class c:

- Return some of documents from the class c (true positive).
- Return some of documents not from the class c (false positive).
- Not return other documents from the class c (false negative).
- Not return other documents from other classes (true negative).

Precision defined as the ratio of correct categorization of documents into the total number of tried classifications.

$$\text{Precision} = \frac{\text{true positive}}{\text{true positive} + \text{false positive}}$$

Recall defined as the ratio of correct categorization of documents into the total number of labeled data in the testing set :

$$\text{Recall} = \frac{true\ positive}{true\ positive + false\ negative}$$

F1-measure defined as the harmonic mean of precision and recall. the good classifier has a high F1-measure, which means the classifier do well with respect to precision and recall :

$$\text{F1-measure} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

4. Results

To compare the performance of the previously mentioned feature selection techniques decision Tree (DT) and Naïve Bayes (NB) classifiers are used, we focuses on, the results obtaining from (i) the implemented of the two algorithms without applying the feature selection techniques, (ii) the results obtaining after applying each of the three techniques to feature selection (Chi-square, threshold, and improved recursive feature elimination (RFE)) on the used algorithms. The whole experiments were done using the same corpus namely 20 newsgroups text dataset and the two algorithms were implemented using Python version 3.4. It is a high-level programming language. For comparison of the results from the two classifiers with feature selection, the average of each classifier output is shown in Table1.1 for the tests. the results of classification implemented on DataSet2 and DataSet1 for two classification algorithms with three feature selection techniques Chi square, threshold, and IRFE.

Table (1) the average of each classifier output for the tests

Implementation cases	NB	DT
without FS + 20 Categories	0.73	0.48
without FS + 4 Categories	0.78	0.64
Chi2 (20% FS + 20 categories)	0.69	0.45
Chi2 (20% FS + 4 categories)	0.78	0.61
Chi2 (40% FS + 4 categories)	0.78	0.59
Chi2 (60% FS+ 4 categories)	0.78	0.58
Threshold (20% FS + 20 categories)	0.64	0.44
Threshold (20% FS + 4 categories)	0.77	0.59
Threshold (40% FS + 4 categories)	0.78	0.58
Threshold (60% FS+ 4 categories)	0.78	0.59
Suggested (20% FS + parts=26 left_ parts=2 + 20 categories)	0.60	0.40
Suggested (20% FS + parts=26 left_ parts=2 + 4 categories)	0.77	0.59
Suggested (20% FS + parts=26 left_ parts=3 + 4 categories)	0.77	0.59
Suggested (20% FS + parts=16 left_ parts=2 + 4 categories)	0.77	0.58
Suggested (20% FS + parts=16 left_ parts=3 + 4 categories)	0.77	0.59
Suggested (20% FS + parts=32 left_ parts=2 + 4 categories)	0.77	0.58
Suggested (20% FS + parts=32 left_ parts=3 + 4 categories)	0.77	0.60
Suggested (20% FS + parts=64 left_ parts=2 + 4 categories)	0.77	0.58
Suggested (20% FS + parts=64 left_ parts=3 + 4 categories)	0.77	0.59
Suggested (40% FS + parts=26 left_ parts=2 + 4 categories)	0.78	0.60
Suggested (60% FS + parts=26 left_ parts=2 + 4 categories)	0.78	0.59

5. Conclusions and Discussion

Our work is improvement of feature selection and then testing this improvement with two well-known classifiers, Where implemented a new approach to feature selection via the IRFE feature selection technique to reduce the features number observably, which can be a general formulation and mathematical model of Recursive Feature Elimination (RFE). This method is tested in many variant tests with the used two classifiers. The suggested approach not only reduces the number of dimensions, as well as the results prove that a new method is comparable with the other known techniques. The results showed that ours feature selection approach best compared with other traditional of feature selection techniques, where the suggested feature selection technique is comparable with standard Like Chi-square. The best results were 83% when 60% of features used, 82% when 40% of features used, and 82% when 20% of features used.

6. Suggestions for Future Works

Drawling upon the results of this investigation of the following could be explored farther.

1. To beat issues in the existing algorithms, effective algorithm needs to be developed.
2. Improving text classification by using new feature selection methods.
3. Utilizing suggested method (IRFE) with other algorithms of text mining and with other feature selection methods to text classification.
4. Future works can be devotes to many other problems of text classification such as, detecting spam, defining the topic of a news article, or chosing the correct mining of a multi-valued word.

References

[1]. kumbhar, P., & Mali, M. (2016). A survey on feature selection techniques and classification algorithms for efficient text classification. International Journal of Science and Research, 5(5), 9.

[2]. Purohit, A., Atre, D., Jaswani, P., & Asawara, P. (2015). Text Classification in Data Mining. International Journal of Scientific and Research Publications, 5(6), 1-7.

[3]. Scott, S., & Matwin, S. (1999, June). Feature engineering for text classification. In ICML (Vol. 99, pp. 379-388).

[4]. Uysal, A. K. (2016). An improved global feature selection scheme for text classification. Expert systems with Applications, 43, 82-92.

[5]. Pandya, R., & Pandya, J. (2015). C5. 0 algorithms to improved decision tree with feature selection and reduced error pruning. International Journal of Computer Applications, 117(16).

[6]. Maneesh Singhal#1, Ramashankar Sharma#2(2014). Optimization of Naïve Bayes Data Mining Classification Algorithm. International Journal for research in applied Science and Engineering Technology (I JRAS ET). Vol. 2 Issues VIII, August 2014.

[7]. Dey Sarkar, S., Goswami, S., Agarwal, A., & Aktar, J. (2014). A novel feature selection technique for text classification using Naive Bayes. International scholarly research notices, 2014.

[8]. Galathiya, A. S., Ganatra, A. P., & Bhensdadia, C. K. (2012). Classification with an improved Decision Tree Algorithm. International Journal of Computer Applications, 46(23), 1-6.

[9]. Galathiya, A. S., Ganatra, A. P., & Bhensdadia, C. K. (2012). Improved Decision Tree Induction Algorithm with Feature Selection , Cross Validation, Model Complexity and Reduced Error Pruning. International Journal of Computer Science and Information Technologies, 3(2), 3427-3431.

[10]. Yuan, L. (2010). An improved Naive Bayes text classification algorithm in Chinese information processing. Science, 267-269.

[11]. Harrag, F., El-Qawasmeh, E., & Pichappan, P. (2009, July). Improving Arabic text categorization using decision trees. In Networked Digital Technologies , 2009. NDT'09. First International Conference on (pp. 110-115). IEEE.

[12].Wang, Y., & Wang, X. J. (2005, August). A new approach to feature selection in text classification. In 2005 International conference on machine learning and cybernetics (Vol. 6, pp. 3814-3819). IEEE.

[13]. Liao, S., & Jiang, M. (2005, August). An improved method of feature selection based on concept attributes in text classification. In International Conference on Natural Computation (pp. 1140-1149). Springer, Berlin, Heidelberg.

[14]. Rogati, M., & Yang, Y. (2002, November). High-performing feature selection for text classification. In Proceedings of the eleventh international conference on Information and knowledge management (pp. 659-661).

[15]. Lewis, D. D., & Ringuette, M. (1994, April). A comparison of two learning algorithms for text categorization. In Third annual symposium on document analysis and information retrieval (Vol. 33, pp. 81-93).

[16]. Aliwy, A. H., & Ameer, E. H. A. (2017). Comparative study of five text classification algorithms with their improvements. International Journal of Applied Engineering Research, 12(14), 4309-4319.

[17]. Cachopo, A. M. D. J. C. (2007). Improving methods for single-label text categorization (Doctoral dissertation, Universidade Técnicade Lisboa).

[18]. Rennie, J. D., Shih, L., Teevan, J., & Karger, D. R. (2003, August). Tackling the poor assumptions of naive bayes text classifiers. In ICML (Vol. 3, pp. 616-623).

[19]. Ting, S. L., Ip, W. H., & Tsang, A. H. (2011).Is Naive Bayes a good classifier for document classification. International Journal of Software Engineering and Its Applications, 5(3), 37-46.

[20]. Adel, A., Omar, N., & Al-Shabi, A. (2014). A comparative study of combined feature selection methods for Arabic text classification. Journal of Computer Science, 10(11), 2232-2239.

[21]. Korde, V., & Mahender, C. N. (2012). Text classification and classifiers: A survey. International Journal of Artificial Intelligence & Applications, 3(2), 85.

[22]. Nalini, K., & Sheela, L. J. (2014). Survey on Text Classification. International Journal of Innovative Research in Advanced Engineering, 1(6), 412-417.

[23]. Decision tree learning. (2017, August 4). In Wikipedia , the Free Encyclopedia. Retrieved

[24]. Quinlan, J. R. (1987). Simplifying decision trees. International journal of man-machine studies, 27(3), 221-234.

[25]. Karimi, K., & Hamilton, H. J. (2010). Generation and Interpretation of Temporal Decision Rules. ArXiv preprint arXiv: 1004.3334.

[26]. Ali, S. I., & Shahzad, W. (2012, October). A feature subset selection method based on symmetric uncertainty and ant colony optimization. In Emerging Technologies (ICET) , 2012 International Conference on (pp. 1-6). IEEE.

[27]. Adel, A., Omar, N., & Al-Shabi, A. (2014). A comparative study of combined feature selection methods for Arabic text classification. Journal of Computer Science, 10(11), 2232-2239.

[28]. Shabir, M., & Naz, M. (2011). On soft topological spaces. Computers & Mathematics with Applications, 61(7), 1786-1799.