Comparison of the accuracy of SVM kernel functions in text classification

1st Neli Kalcheva
Department of Software and Internet Technologies
Technical University of Varna
Varna, Bulgaria
n_kalcheva@abv.bg

2nd Milena Karova
Department of Computer Science and Engineering
Technical University of Varna
Varna, Bulgaria
mkarova@tu-varna.bg

3rd Ivaylo Penev
Department of Computer Science and Engineering
Technical University of Varna
Varna, Bulgaria
ivailo.penev@tu-varna.bg

***Abstract*— The objective of this paper is to compare the accuracy of different kernel functions of the SVM method for text classification. As a basis for the research film reviews are used. The authors try to detect the kernel functions and their parameters to achieve high accuracy in movie reviews classification. The studied kernel functions are: polynomial kernel of degree 2, a linear kernel and a radial base kernel. The achieved accuracy is higher than 83%. The experiments show that the sigmoid radial kernel is an inappropriate choice in text classification.**

Keywords*— Support Vector Machines, SVM, Support Vector Classification, SVC, text classification, machine learning, kernel functions, kernels poly, rbf, sigmoid, linear

I. Introduction

Text classification is one of the important and typical tasks in supervised machine learning. It is applied to different kinds of documents: web pages, library books, media articles, social posts etc. The number of text classification applications include: spam filtering, email routing, sentiment analysis, text searching, text sorting.

Text classification is a difficult task due to the availability of high-dimensional feature vectors comprising noisy and irrelevant features. Various feature reduction methods have been proposed for eliminating irrelevant features as well as for reducing the dimension of the feature vectors.

In machine learning, classification is accepted as a task for determining the class of an unknown object on the basis of empirical data. The reference object - class is unknown. The analysis of the nature and characteristics of the objects are directly related to the synthesis of the classification model. No universal and unambiguous method for classification of objects has been found.

The classification of texts into predefined categories is an inductive process based only on endogenous data. Nowadays, the classification is directly related to the selective or to the adaptive document's organization and its popularity increases both in research and in business areas.

II. Presentation

A. Literature review

Support vector machine (SVMs) algorithms have strong theoretical foundations and excellent empirical successes. They are seen to be more interpretable than deep neural networks, and it had been mostly used in many classification problems such as handwritten digit recognition, object recognition, and text classification. The support vector machines are widely used due to their relatively powerful performance over many different areas.

The authors in [6] explore two main problems: text analysis in natural language processing and finding the best algorithm in machine learning. Several of the most well-known classification methods are considered: the Bayesian method, the Support Vector Machines method (SVM), and the K nearest neighbor method. Experiments on the classification of a collection of texts in English, Chinese and Russian prove that SVM has an advantage over other methods in accuracy and completeness.

Analysis of texts in Russian and English on the basis of machine learning methods is presented in [7]. The aim of the study is to test and compare different methods of machine learning, of which: Bayesian method, K nearest neighbor method, Rocchio method, SVM, key classification method words and its combination with SVM. In all approaches, the weights of the words in the text are determined by the TF.IDF scheme and English and Russian "stop words" are excluded. The metrics used for comparisons are accuracy, precision, completeness and F-measure. Studies show that SVM and the combined method give the best results.

The authors in [8] present a study of the most common methods for constructing classifiers in the period from 2011 to 2016. The methods considered are: Naive Bayes (NB) method, K Nearest Neighbors (KNN) method, Decision Trees (DT) method, SVM, logistic regression, a method based on artificial neural networks. Accuracy, completeness and F-measure metrics are used to compare the properties of the algorithms. The training and text samples are selected in a 70/30 ratio. Based on the research, the author came to the conclusion that the method of the support vectors and the neural network has the best characteristics of the selected metrics.

An empirical study of three text classification algorithms is presented by the authors in [9], using two data sets. The Naive Bayes Classifier, the Support Vector Machine Method, and the Solution Tree (C4.5) are studied by training instances of the Weka tool datasets. The results are compared based on the values of precision and sensitivity for each of the algorithms. Studies show that SVM is the most effective among the three classifiers.

According to the authors in [10], the best classification algorithms are: Naive Bayes, Random Forest, and Support Vector Machine, which is why many researchers combine

978-1-7281-7073-2/20/$31.00 ©2020 IEEE

them to improve accuracy. The researchers tested these algorithms with Rapid Miner software with 500 sample data. The results they receive show that the Support Vector Machine has the highest accuracy.

Pang, Lee and Vaithyanathan compare the accuracy of three classifiers: the naive Bayesian classifier, the maximum entropy, and the reference vector method [11]. Randomly selected 700 documents with positive and 700 documents with negative movie reviews are used for the study. The data are divided into three equal-sized groups, maintaining a balanced distribution of classes in each group. The results obtained are the average three times the accuracy values from the cross-validation of these data. In this study, the researchers focused on characteristics based on unigrams (negatively marked) and bigrams. The results for unigrams are in favor of the multinomial Naive Bays classifier (78.7%). In cases where bigrams are reported, the results of the three algorithms are similar with differences below 1 percent. But when they combined unigrams with bigrams, the reference vectors method (82.7%) outperformed the other two algorithms by about 2 percent.

SVM classification algorithm is proposed by Vapnik [3] to solve two-class problems. This classical machine learning algorithm is the method of reference vectors, which presents the training examples as points in n-dimensional space. The examples are designed in space in such a way that they are linearly separable. Working with two classes, a drawing line method is found that divides the data into the two classes, shown in Figure 1. The separated data line is called the hyper-separation plane. This hyperplane must be chosen in such a way as to be as far as possible from the both classes examples. The SVM is trained using preclassified documents.

The linear classification function f (x) is presented as (1):

$$f(x) = w^T x + b \quad (1)$$

where $\mathbf{w^T}$ is a weight vector and **b** is a deviation.

The goal is to find the values of $\mathbf{w^T}$ and **b** in order to determine the classifier. It is necessary to find the points with the smallest deviation that must be maximized (Fig. 1).

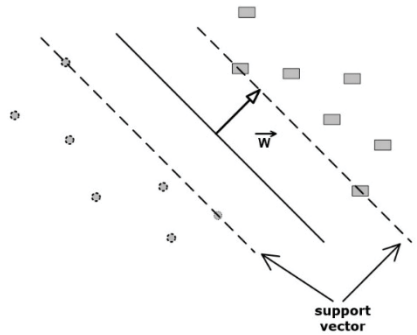

Fig. 1. Example of SVM hyperplane pattern

In nonlinearly separable data, the basic idea is to achieve linear separation by moving the data to another higher measurable functional space by transforming with a function of the input nonlinear data. This is done by the so-called kernel function K, which is defined as follows:

$$K(x_i,x_j) = f(x_i).f(x_j) \quad (2)$$

Some of the most used kernel functions are presented in [4, 5, 6].

- Polynomial kernel function

The polynomial kernel function uses one of the most popular methods of nonlinear modeling, [2a]

$$K(x, y) = \langle x, y\rangle^d \quad (3)$$

$$K(x, y) = (\langle x, y\rangle + 1)^d \quad (4)$$

The second core avoid problems at values equal to zero.

- Radial basis function

$$K(x,y) = exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right) \quad (5)$$

- Sigmoid function [5]

$$K(x,x') = \tanh(k1(x,x') + k2 \quad (6)$$

where

$$\tanh x = \frac{e^x - e^{-x}}{e^x - e^{-x}}$$

The SVM algorithm is trained using preclassified documents. It has good performance on large data sets. In all shown experiments (see Section 4), the classification methods are implemented using cinema reviews data. The output file of the SVMs contains decision function output value for each classified document: positive or negative reviews.

B. Evaluation of the Accuracy classification

The text classification process uses training and test data. The training data "trains" the test data. The cross-validation method is used to evaluate the classifier. This method guarantees an equal number of participations of each object in the training example and exactly one participation in the test example. The nature of the method consists in the following: all sets are divided into **k** parts, each of which appears as a test. In this case, it is important to choose **k**. According to researchers in the field of text classification**, k = 5** or **k = 10** is preferred.

The classification model divides the respective object into one of the possible classes: True (Posive Reviws) or False (Negative Reviews). Four classification options are possible with this model:

- True Positive (TP) - the Positive class is valid and correctly classified as the Positive class

$$TP\ Rate = \frac{TP}{TP+FN} \quad (7)$$

- True Negative (TN) – the Negative class is valid and correctly classified as the Negative class

$$TN\ Rate = \frac{TN}{TN+FP} \quad (8)$$

- False Positive (FP) – the Negative class is valid and the Positive class is incorrectly classified

$$FP\ Rate = \frac{FP}{FP+TN} \quad (9)$$

- False Negative (FN) – the class Positive is valid and incorrectly classified as Negative

$$FN\ Rate = \frac{FN}{FN+TP} \quad (10)$$

As solving the classification problem, the classifier work is evaluated through the accuracy of prediction.

The Accuracy measure is calculated as the ratio of correctly classified examples to the total number of objects in the test set. It is usually presented in percentages.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (11)$$

III. Experiments

A. Test data set selection

A data set consisting of 1000 documents representing moview reviews written in English are used. 500 reviews are positive containing 291,125 words and 500 reviews are negative containing 249,025 words. The data are taken from the standard open source library of Python - NLTK module - Natural Language Toolkit and provided in 2004 by Bo Pang and Lillian Lee (Department of Computer Science. Cornell University. Ithaca, NY).

B. Implementation tools

The Python Scikit-learn module is used to implement SVM. In SVM, the most important parameter is the kernel function (the kernel) and different kernels can lead to different values of the classifier accuracy. For instance, sklearn's SVM implementation svm.SVC has a kernel parameter which can take on linear, poly, rbf [4].

The following abbreviations in the experimental results are used:

- Support Vector Classification with polynomial kernel - SVC_poly
- Support Vector Classification with radial basic kernel - SVC_rbf
- Support Vector Classification with sigmoid radial kernel - SVC_sigmoid
- Support Vector Classification with linear kernel - SVC_linear

A 5-fold cross-validation is used to evaluate the classification, in which the total data set is randomly divided into 5 parts and the algorithm is executed 5 times. One-fifth of the data is used for the test set and four-fifths for the training set. The result of the classification is the total average accuracy for all pairs of examples after the classifier training and testing.

C. Results and analyses

The objective of the classification is to compare the accuracy in determining the positive and negative film reviews from the data set with different kernels of the SVM method. Initially, in the study, the algorithms of the SVM method are tested with default parameters.

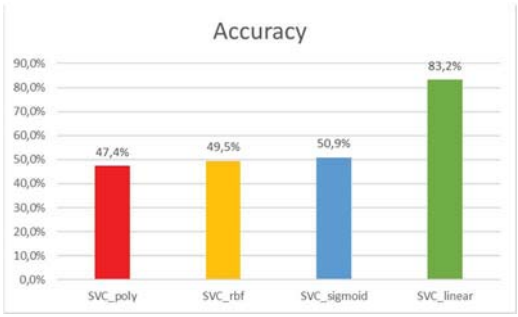

Fig. 2. The Accuracy of classifiers with different SVC cores with default parameters

The results show that the SVM with linear kernel is classified with the highest accuracy (Fig. 2). The classifier with a polynomial kernel is about 2% more accurate than the classifier with a radial basis kernel function and more than 3% more accurate than the sigmoidal kernel classifier. These three classifiers are more inaccurate than the linear kernel classifier by more than 32-35%.

Fig. 3. Classifiers time in ms with different SVM cores with default parameter values

On Fig. 3 the results show that the linear kernel algorithm is the fastest one. It classifies the algorithm with a polynomial kernel with the weakest time. The difference compared to the fastest is about 7 ms.

Carrying out the experimental tests, the classifiers are implemented with the following parameters:

- Support Vector Classification, kernel = 'poly', C = 10, gamma = auto
- Support Vector Classification, kernel = 'rbf', C = 10, gamma = auto
- Support Vector Classification, kernel = ‘sigmoid, C = 10, gamma = auto
- Support Vector Classification**, kernel = 'linear', C = 10**

where in [4] the used parameters are described:

C - Regularization parameter; **gamma** - kernel coefficient for ‘rbf’, ‘poly’ and ‘sigmoid’; **degree** – Degree of the polynomial kernel function (‘poly’). Ignored by all other kernels; **coef0** – Independent term in kernel function. It is only significant in ‘poly’ and ‘sigmoid’; **kernel** - the kernel type to be used in the algorithm. It must be one of ‘linear’, ‘poly’, ‘rbf’, ‘sigmoid’.

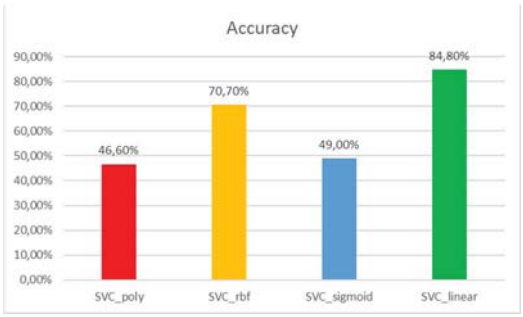

Fig. 4. The Accuracy of classifiers with different SVM cores at C = 10

The results show that the linear kernel has the highest accuracy (Fig. 4). The polynomial and the sigmoid radial kernel have lower accuracy compared to the previous study. The classifier with the radial basis function kernel is increased the accuracy by about 20%.

The time is studied at C = 10.

Fig. 5. The time of classifiers in ms with different SVM kernels at C = 10

The results show that the linear kernel is the fastest one (Fig. 5). It classifies the algorithm with a polynomial kernel with the weakest time, and the difference from the fastest is about 5 ms.

The best results are obtained after an experimental selection and research. The classifiers are implemented with the following parameters:

- Support Vector Classification, kernel='poly', C=10, coef0=10, gamma= 'auto'
- Support Vector Classification, kernel='rbf', C=100, gamma= 'auto'
- Support Vector Classification, kernel= 'sigmoid, C=100, gamma= 'auto'
- Support Vector Classification, kernel='linear', C=10

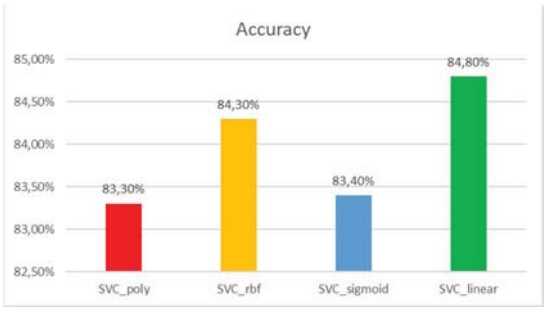

Fig. 6. The Accuracy of classifiers with different SVM cores with parameters above

The results on Figure 6, show the classified film reviews in English after changing the parameters with the highest accuracy classifies the linear kernel. The algorithms with polynomial, sigmoid radial and radial basis kernel functions are classified with an accuracy approximately equal to each other, with a difference of about 1-2%.

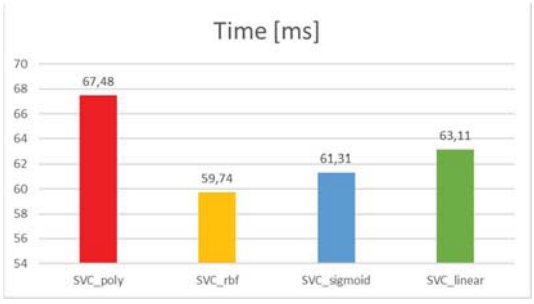

Fig. 7. The time of classifiers in ms with different SVM kernels with parameters above

The results show that the fastest is the algorithm with radial basis kernel function, followed by the classifier with a linear kernel with a difference of about 1.5 ms (Fig. 7). It classifies the algorithm with a polynomial kernel with the weakest time, with a difference compared to the fastest by about 8 ms.

IV. Conclusion and Future Work

This paper studies the effect of using the SVM method with different kernels for text classification. Four SVM classifiers are presented:

- Support Vector Classification with polynomial kernel - SVC_poly
- Support Vector Classification with radial basic function - SVC_rbf
- Support Vector Classification with sigmoid radial core - SVC_sigmoid
- Support Vector Classification with linear core - SVC_linear. Compared

The correct choice of the kernel parameters is crucial to achieve good results. In practice, this means that an in-depth search of the kernel parameter space, which often complicates the task, must be performed on methods for effective automatic parameter selection.

Classifying texts in English, representing film reviews, it uses the SVM method with a polynomial kernel of degree 2, with a linear kernel and a radial base kernel. They are classified with an accuracy higher than 83%. The experiments show that the use of the sigmoid radial kernel of SVC is an inappropriate choice in text classification.

While the results are encouraging, there are still much improvement to be made. More context features can be experimented to examine their usefulness in text classification. In particular, it is possible to include in experiments texts in different languages and to analyse a comparison between the algorithms depending on languages.

Acknowledgments

This paper is supported by the National Scientific Program "Information and Communication Technologies for a Single Digital Market in Science, Education and Security (ICTinSES)" (grant agreement DO1-205/23.11.18), financed by the Ministry of Education and Science.

References

[1] Aggarwal, C., Z. ChengXiang. Mining Text Data, Springer, 2012.

[2] A. Basu, C. Watters, and M. Shepherd, Support Vector Machines for Text Categorization, Proceedings of the 36th Annual Hawaii International Conference on System Science, 2003.

[3] Vapnik, The Nature of Statistical Learning Theory. Springer, Berlin, 1995.

[4] "Scikit-learn: sklearn.svm.SVC Documentation", https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html.

[5] Золотых, Н. Ю. Машинное обучение и анализ даных 2013, http://www.uic.unn.ru/~zny/ml.

[6] Зайцев, В., Л. Линь. Способы повышения эффективности классификации документов для конечного множества языков. Вісник НТУУ «КПІ» Інформатика, управління та обчислювальна техніка №50, 2009, с. 100-104.

[7] Feng, M., G. Wu. A distributed chinese Naive Bayes classifier based on word embedding. International Conference on Machinery Materials and Computing Technology (ICMMCT 2016), 2016, pp. 1121-1127.

[8] Батура, Т. Методы автоматической классификации текстов. Программные продукты и системы 30, no. 1, 2017.

[9] Trivedi, M., S. Sharma, N. Soni, S. Nair. Comparison of text classification algorithms. International Journal of Engineering Research & Technology (IJERT) 4, no. 02, 2015.

[10] Sheshasaayee, A., G. Thailambal. Comparison of classification algorithms in text mining. International journal of pure and applied mathematics 116, no. 22, 2017, pp. 425-433.

[11] Pang, B., L. Lee, S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing, vol. 10, Association for Computational Linguistics, 2002, pp. 79-86.