УДК 004.912: 004.632

V.V. LYTVYNOV[*], O.P. MOYSEENKO[*]

THE MECHANISMS OF TEACHING AND EVALUATION OF THE QUALITY OF PERFORMANCE OF THE TEXT DOCUMENTS CLASSIFIER

[*]Chernihiv National University of Technology, Chernihiv, Ukraine

Анотація. *Описані механізми навчання та оцінки якості роботи класифікатора в розроблюваній системі автоматизованої обробки великих об'ємів текстової інформації. Класифікатор базується на вільній бібліотеці LibSVM та методі опорних векторів. Система виконує функції пошуку, класифікації, рубрикації та кластеризації текстових документів за запитами користувача.*
Ключові слова: *класифікація, рубрикація, кластеризація, обробка текстових документів.*

Аннотация. *Описываются механизмы обучения и оценки качества работы классификатора в разрабатываемой системе автоматизированной обработки больших объемов текстовой информации. Классификатор базируется на свободной программной библиотеке LibSVM и методе опорных векторов. Система выполняет функции поиска, классификации, рубрикации и кластеризации текстовых документов по запросам пользователя.*
Ключевые слова: *классификация, рубрикация, кластеризация, обработка текстовых документов.*

Abstract. *The mechanisms of teaching and evaluation of the performance of the classifier in the developing system of the automated processing of large volumes of textual information are described. The classifier is based on the free software library LibSVM and support vector machines. The system performs the functions of search, classification, categorization and clustering of text documents at the request of the user.*
Keywords: *classification, categorization, clusterization, processing of text documents.*

1. Introduction

The aim of classification (thematic categorization) of electronic natural language documents, i.e. classification of the texts content to one or several thematic sections, is currently important due to the continuous growth of stored or transmitted text data.

In theory, the solution of the documents classification task involves the presence of a certain plurality of electronic documents D={di}, that has to be separated into several nonintersecting, thematically homogeneous subset (classes, C) and defining to which class each document from the total mass of documents to be processed should be classified [1].

$$C = \{C_i\} \bigcup_{d \in C_i}^{\forall i} d = D \times C_i \cap C_j = 0 (i \neq j). \quad (1)$$

2. Problem statement

The objects of the research are:

• a relatively large text collection of several hundred documents, previously separated by content into thematic groups (classes/sub-collections);

• the mechanisms of text data analysis in the system of natural language documents processing.

The tools are:

• the developed system of processing of multilingual, dynamic flows of text data on the base of support vector machine algorithm (SVM) [2], implemented in the free library LibSVM;

• implementation of SVM in the module Machine Learning (Support Vectors), the product of the company StatSoft, STATISTIKA 8.0.

ISSN 1028-9763. Математичні машини і системи, 2014, № 4

As a result of theoretical and practical experiments, it will be possible to investigate more thoroughly the processes of study and testing of the classifier in the system of "Processing of high-speed information flows of text data".

3. Problem solution

By the example of the method of support vector machines (fig. 1), the model of the text documents classifier can be presented as:

$$R =< D, C, F, Rc, f >, \tag{2}$$

where D – plurality of documents that need to be classified;

C – plurality of thematic rubrics (classes) $C = \{ci\}$, $i = 1..Nc$, N – number of possible rubrics;

F – plurality of rubrics descriptions. Each class Ci has its distinctive description Fi;

Rc – ratio $C \times F$, to check the single description of each rubric. $\forall c_i \in C \exists! F_i \in F : (c_i, F_i) \in R_c$;

f – function $\exists d \in D : f(\mathrm{d}) = \mathrm{C}_d \subset C \cap |C_t| > 1$, i.e. the process of classification of objects d Є D in the result of which the correspondence of a specific document d to one of the descriptions Fi and its assignment to the rubric Ci are defined. According to this function, elements of the plurality of documents can be assigned to several thematic rubrics at the same time. To minimize the number of such cases the classifier has to be properly taught before usage.

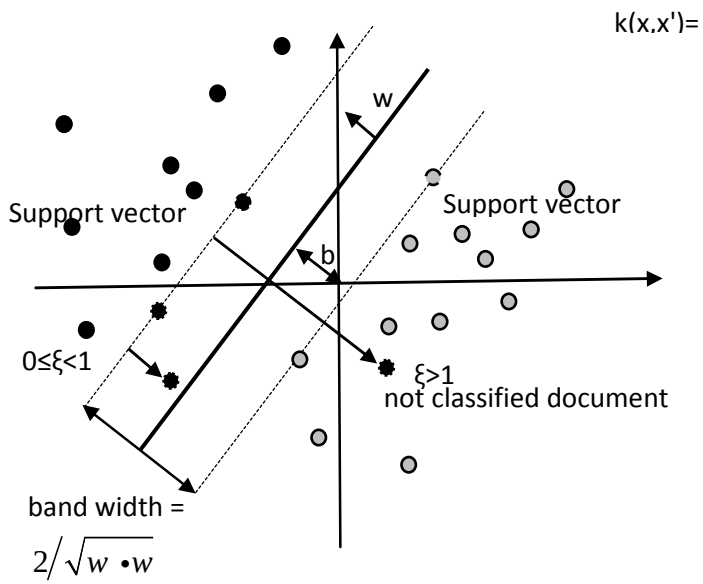

Fig. 1. The general scheme of work of SVM classifier where, dots are the vector representation of two thematically different subsets, pluralities, classified NL documents; k – some function of the nucleus, that allows to separate thematic classes so that a separating plane could be drawn; w – support vectors on the base of bordering documents; ξ – the introduced variable error to assess the classifier; b – the distance between the separating pluralities plane and the beginning of coordinates; w – support vector

The popular in text data classification tasks collection of English short financial and stock documents Reuters-21578 [3] of the eponymous information agency has been used in the research. As it is seen from the name, this collection consists of 21 578 documents. Some of the documents are marked as not properly categorized, that is why only 12902 documents are used in practice. The corpus of texts is presented in the form of both txt and xml files. The collection is a part of the first volume of categorized documents of the information agency Reuters that is abbreviated as RCV1 (Reuters Corpus Volume 1) [4]. In its original form the set of text documents of Reuters-21578 includes 135 thematic rubrics, 56 names of organizations, 267 different personalities and 175 geographical names. The documents are collected in 21 xml-files and are presented in the following way:

```
<REUTERS TOPICS="NO" LEWISSPLIT="TRAIN" CGISPLIT="TRAINING-
SET" OLDID="5545" NEWID="2">
```

<DATE> date of publication </DATE>
<TOPICS/>
<PLACES>
<D> location </D>
</PLACES>
<PEOPLE/>
<ORGS/>
<EXCHANGES/>
<COMPANIES/>
<UNKNOWN>more information </UNKNOWN>
<TEXT>
<TITLE> topic </TITLE>
<DATELINE> origin </DATELINE>
<BODY>text</BODY>
</TEXT>
</REUTERS>;

For teaching and evaluation of the quality of performance of SVM classifier the method ModApte split has been used, that involves the separation of documents plurality of Reuters-21578 collection into the subset for teaching – 9603 documents (74% of the total amount) and the subset for testing of the chosen method of machine teaching with 3299 documents (26% off the total amount). ModApte separation is recommended to use to compare results of work of several classifiers.

4. Experimental part

The developing system is based on one of the variety of existing implementations of the support vector machines method, namely the free library with nonlinear nucleuses – LibSVM [8, 9]. Preference to this library to the library of the same developers LibLinear, that is implementing a quick linear classifier SVM, was given due to the work with small text corpuses and the possibility of occurrence of the situation of linear inseparability after a change of documents collections by including documents in other NL. The mechanisms of SVM algorithm implementation in the program product Statistika are not known. In the available software version there is one module that implements this method for the tasks of classification and categorization for any text corpuses.

The quality of the classifier work depends on the correct presentation of processed documents in the form of a vector model [10, 11]. Each document from the collection of such model is presented as a plurality of terms (words, word combinations, numbers and other elements of which a document consists). According to the mentioned laws of Zipf, a certain weight can be specified to the terms from the collection, i.e. how important this term is for the document characteristic. For the presentation of a document in the vector space, the weights of all terms of the collection in regard to this document are denoted. The dimension of the document vector will be equal to the total amount of all terms outlined from the collection.

$$d_j = (w_{1j}, w_{2j}, ..., w_{nj}), \tag{3}$$

where d_j – vector presentation of j document, w_{ij} – weight of i term in j document, n – total amount of terms.

Thanks to such presentation of documents they can be compared by finding the distance between vectors of the space (Euclidean distance or Mahalanobis distance). The smaller the distance is, the greater probability of thematic similarity between the documents.

In the system of automatic processing of text data flows on the base of LibSVM library the following functions of nucleus are possible that implement the linear separation of classified subjects:

<Label> – nucleus identifier. Examples of functions:

0 – linear $k(x,w) = sign(< x \cdot w >)$.

1 – polynomial $k(x,x') = (x \cdot x')^d$.

2 – radial basis function, $k(x,x') = \exp(-\gamma \| x - x' \|^2)$, for $\gamma > 0$.

3 – sigmoid $k(x,x') = th(kx \cdot x' + c)$ for $k > 0, c < 0$ б

where K – nucleus function, x·x' – scalar product of vectors, y – mapping of a vector from the space of features R^n into another space, d – degree, к and c – parameters, w – weights of features.

<Index1>:<Value1>

<Index2>:<Value>

…

Index – number of the vector coordinates, Value – value of the vector.

There are several standard ways of weight determination of a term in a document:

a) Boolean weight – 1, if the term is in the document, and 0 if it doesn't occur;

b) Term Frequency (TF) – the frequency of the term occurrence in the document;

c) Term Frequency – Inverse Document Frequency (TF-IDF) – the frequency of the term occurrence in the document at the amount that is inverse to the number of documents in which this term occurs;

d) Pointwise Mutual Information (PMI) – all negative weights are replaced by zero.

For cleanliness of the experiment in the developed system the tf-idf method of determining terms weights is used as it is used in the software Statistika [8]:

$$w_{ki} = \frac{(1 + \log(N_{ik}) \cdot \log(\frac{|D|}{N_k}))}{\sqrt{\sum_{s \neq k} (\log(N_{is}) + 1)^2}}, \quad (4)$$

where N_{ki} – number of occurrence of k term in the i document, N_k – number of occurrence of k term in all documents, $|D|$ – number of documents in the collection.

Taking into account the possibility of cases of linear inseparability of classified objects into the equation describing the hyperplane that separates classes of documents in the space D, the variable error is introduced $\xi_i \geq 0$.

$$y_i(\omega \bullet d_i - b) > 1 - \xi_i, \quad (5)$$

where y_i – number equal to 1 in the case the vector di refers to the rubric we are interested in and -1 if it doesn't;

w – support vector;

b – boundary value of the distance between the separating hyperplane and the beginning of the coordinates;

w·di > b ⇒ yi = 1;

w·di < b ⇒ di = −1.

It is supposed that if $\xi_i = 0$, there is no error in the document di. If $\xi_i > 1$, there is an error in the document. If $0 < \xi_i < 1$, the object di falls within the band of the separating plane.

The task of the classifier teaching is to solve the issue of optimisation of the function separating plane using the method of Lagrange [13]:

if at the point x relative minimum of the original objective function is achieved, then under condition there is the equality 0 derivatives with respect to x of the new objective function,

there exists a set λ_i, that at the same point x the minimum of the new objective function is attained, but globally for all x. At that for each λ_i the following is true:

either λ_i is equal to 0 and the corresponding constraint is not active, or λ_i is not equal to 0 and the corresponding constraint is satisfied, but then this is already the equation.

Formulating this task in terms of Lagrange method, it turns out that it is necessary to find the minimum of w, b, ξ and the maximum of λ_i of the function:

$$\frac{1}{2}\omega\bullet\omega + C\sum_i \xi_i - \sum_i \lambda_i(\xi_i + y_i(\omega\bullet d_i - b) - 1) \text{ при } \xi_i \geq 0, \lambda_i \geq 0. \quad (6)$$

If $\lambda_i > 0$, then the document of the teaching collection d_i is called the support vector.

After these manipulations the optimized separating hyperplane equation looks as follows:

$$\sum_i \lambda_i y_i d_i \bullet d - b = 0, \quad (7)$$

where d_i – document to be categorized.

As a numerical evaluation of the classification by both systems, the traditional set of metrics for a given issue was used: Accuracy (A), Precision (P) and Recall (R).

The first metric shows the general picture of the classifier performance, calculating the ratio of documents properly distributed by the classifier to total.

$$A = \frac{M}{N} \cdot 100\%, \quad (8)$$

where M – the amount of correctly classified documents, N – the total amount of documents.

The metric of precision indicates the relation of correctly classified documents to a particular class and of all documents referred to this class.

$$P = \frac{TP}{TP + FP} \cdot 100\%. \quad (9)$$

The metric of recall is the relation of correctly classified documents to a particular class and all documents belonging to this class in the test sample.

$$R = \frac{TP}{TP + FN} \cdot 100\%. \quad (10)$$

The formulas of recall and precision metrics are constructed on the basis of contingency tables compiled for each of the possible classes.

Table 1. Variant of the classifier evaluation

Evaluation of the results by the classifier	Evaluation of the classification results by an expert		
		True	False
	True	TP (true-positive)	FP (false-positive)
	False	FN (false-negative)	TN (true-negative)

The calculation of recall and precision is conducted separately, not joining them in the popular metric of F-measure (11), which shows generalized assessment of the classifier performance.

$$F = 2 \cdot \frac{P \cdot R}{P + R} \cdot 100\% . \quad (11)$$

5. Results

After teaching and test categorization the classifiers of the tested systems showed the following results.

For the texts corpus of 3299 documents from Reuters-21578 collection the developed system based on the free library LibSVM and program product Statistica has given the evaluation.

Table 2. Results of the evaluation

System	Accuracy, %	Precision, %	Recall, %
developed	93	80	94
Statistika	89	75	75

In the table there are average values of the metrics for the developed system with step-by-step application of nucleus functions mentioned previously and realized in the library LibSVM. The classifier based on the support vector machines algorithm, implemented in the product of StatSoft company allows automatically determine the most suitable nucleus function for classification of concrete objects, thus the figures obtained are considered as average and optimal for this classifier.

6. Conclusions

The classifier of the developed system of processing text data flows on the base of free library LibSVM has shown better results in comparison to the module Machine Learning (Support Vectors) of the system Statistika. This may be caused by both: difference of approaches to texts processing (markup, normalization) and choice of the nucleus function. It is planned to improve the classifier performance evaluation on mixed collections.

REFERENCES

1. Joachims T. Text Categorization with Support Vector Machines: Learning with Many Relevant Features / T. Joachims // Proc. of ECML-98, 10th European Conference on Machine Learning. – Dortmund, 1998. – P. 137 – 142.
2. Вапник В.Н. Восстановление зависимостей по эмпирическим данным / Вапник В.Н. – М.: Наука, 1979. – 448 с.
3. Коллекция документов Рейтерс [Электронный ресурс]. – Режим доступа: http://ronaldo.cs.tcd.ie/esslli07/data/reuters21578-xml.
4. Коллекция документов Рейтерс [Электронный ресурс]. – Режим доступа: http://www.ai.mit.edu/projects/jmlr/papers/volume5/lewis04a/lyrl2004_rcv1v2_README.htm.
5. Новостная коллекция РОМИП [Электронный ресурс]. – Режим доступа: http://romip.ru/ru/collections/news-collection.html.
6. Емпирические законы Зипфа [Электронный ресурс]. – Режим доступа: http://artprom.net/article/read/zakon_Zipf.html.
7. Куняев Н.Н. Конфиденциальное делопроизводство и защищенный электронный документооборот / Куняев Н.Н., Демушкин А.С., Фабричнов А.Г. – М.: Логос, 2011. – 452 с.
8. Библиотека LibSVM [Электронный ресурс]. – Режим доступа: http://www.csie.ntu.edu.tw/~cjlin/libsvm.
9. Литвинов В.В. SVM при классификации мультиязычных текстов / В.В. Литвинов, О.П. Мойсеенко // Весник ЧНТУ. – 2013. – № 4. – С 59 – 64.
10. Векторная модель коллекции документов [Электронный ресурс]. – Режим доступа: http://www.machinelearning.ru/wiki/index.php?title=Векторная_модель.

11. Нейлор К. Как построить свою экспертную систему / Нейлор К. – М.: Энергоатомиздат, 1991. – 286 с.
12. Боровиков В.П. Программа STATISTICA для студентов и инженеров. – [2-е изд.]. – М.: КомпьютерПресс, 2001. – 301 с.
13. Лифшиц Ю. Метод опорных векторов (Слайды) — лекция № 7 из курса «Алгоритмы для Интернета» [Электронный ресурс]. – Режим доступа: yury.name/internet/07iah.pdf.
14. Крулькевич М.И. Информационная деятельность в организациях / М.И. Крулькевич, Е.М. Сынкова. – Донецк: ДонНУ Украины, 2001. – 176 с.

Стаття надійшла до редакції 20.08.2014