УДК 004.8

Крак Ю.В.[1], д.ф.-м.н., проф.,
Корлюк О.С.[1], м.н.с.

Доповнення навчальної вибірки для класифікації засобами кластеризації

В даній статті розглядається підхід розширення навчальної вибірки засобами кластеризації. Даний підхід дозволяє підвищити якість класифікації на обмежених наборах даних навчальної вибірки.

Ключові слова: кластеризація, класифікація, тестова вибірка,навчання.

[1] Інститут кібернетики імені В.М.Глушкова Національної академії наук України, 03680, МСП, м. Київ, пр-т. Глушкова 40,
e-mail: arheys@gmail.com

Iurii Krak, Ph. D., Prof.,
Olexandr Korlyuk[1], junior researcher

Addition of the training set for the classification of clustering methods

This article describes an approach of extending the training sample by clustering methods. This approach improves the quality of classification on a limited set of data the test sample.

Keywords: clustering, classification, training sample, training.

[1]Glushkov Institute of Cybernetic of NAS of Ukraine, 40 Glushkova ave., Kyiv, Ukraine, 03680,
e-mail: arheys@gmail.com

This paper was presented by Fedir Garashcenko, Ph . D., Prof

Introduction

The number of documents in electronic form is steadily increasing, especially for web pages and e-mail. Systematization is necessary to facilitate the search in large volumes of data, for example, classification is identification as some comprehensible content.

The classification of full-text print documents is the process of determining the document belongs to one or more categories which are the most appropriate to its content.

The development of new methods of classification is an actual problem, due to the emergence of the new types of documents such as social networks websites and short messages (Twitter). The text classification needs to determine the list of named categories for classifying new documents as one of them. The quality of classification methods depends on the quality and the size of document sets which are pre-prepared for algorithm classification training. In the case of documents in Ukrainian, the verification of classification methods is complicated by the absence of significant document arrays as a training sample.

In case of a small number of documents for study, the accuracy of the classifier usually is not high. In practice, it’s possible to prepare the minimum necessary training sample by hand, but considering the fact that the larger training sample the higher classification accuracy, it is necessary to develop automated methods of updating the training sample for classification problem. Some approaches of semiautomatic classification are posted in the papers [1- 4].

This paper presents a technique which allows to extend primary sample using the clustering of unnamed documents, which belong to the pre-formed clusters consisting of classified documents set.

Clustering-based classification

Let we have a set of named categories

$$D = 1...N \qquad (1)$$

and certain document sets that are included to each of these categories by expert assessment

$$\begin{aligned} &D_1 = \{x : x(i_l), l = \overline{1, n_1}\}, \\ &D_2 = \{x : x(j_s), s = \overline{1, n_2}\}, \ldots, \qquad (2) \\ &D_N = \{x : x(k_r), r = \overline{1, n_m}\}. \end{aligned}$$

Using the clustering method [5] with the use of special distance measure, let’s form the clusters of extended training sample for the classification. For the set of clusters (2) determine the appropriate matrix $\tilde{X}(1), \tilde{X}(2) \ldots \tilde{X}(N)$ which are accordingly defined as

$$\begin{aligned}
&\tilde{X}(1)=(\tilde{x}(i_1)\vdots\ldots\vdots\tilde{x}(i_{n_1})),\\
&\tilde{X}(2)=(\tilde{x}(j_1)\vdots\ldots\vdots\tilde{x}(j_{n_2})),\ldots,\\
&\tilde{X}(N)=(\tilde{x}(k_1)\vdots\ldots\vdots\tilde{x}(k_{n_m})).
\end{aligned}\tag{3}$$

where

$$\begin{aligned}
&\tilde{x}(i_l)=x(i_l)-\hat{x}(1),\\
&\tilde{x}(j_s)=x(j_s)-\hat{x}(2),\ldots,\\
&\tilde{x}(k_r)=x(i_r)-\hat{x}(n),\\
&l=\overline{1,n_1},\ s=\overline{1,n_2},r=\overline{1,n_m}.
\end{aligned}$$

$$\begin{aligned}
&\hat{x}(1)=\tfrac{1}{n_1}\sum_{l=1}^{n_1}x(i_l),\\
&\hat{x}(2)=\tfrac{1}{n_2}\sum_{s=1}^{n_2}x(j_s),\ldots,\\
&\hat{x}(n)=\tfrac{1}{n_m}\sum_{r=1}^{n_m}x(k_r).
\end{aligned}\tag{4}$$

The distance from the new document to each cluster $\Omega_1,\Omega_2,\ldots,\Omega_n$ will be calculated using the following formulas:

$$\begin{aligned}
&\rho(x,\Omega_1)=\tfrac{1}{d_1}((x-\hat{x}(1))^T R(\tilde{X}^T(1))(x-\hat{x}(1))^{\frac{1}{2}},\\
&\rho(x,\Omega_2)=\tfrac{1}{d_2}((x-\hat{x}(2))^T R(\tilde{X}^T(2))(x-\hat{x}(2))^{\frac{1}{2}},\\
&\rho(x,\Omega_n)=\tfrac{1}{d_n}((x-\hat{x}(N))^T R(\tilde{X}^T(N))(x-\hat{x}(N))^{\frac{1}{2}}.
\end{aligned}\tag{5}$$

where

$$\begin{aligned}
&d_1=\max_{l=1,n_1}(x(i_l)-\hat{x}(1))^T R(\tilde{X}^T(1))(x(i_l)-\hat{x}(1)),\\
&d_2=\max_{s=1,n_2}(x(j_s)-\hat{x}(2))^T R(\tilde{X}^T(2))(x(j_s)-\hat{x}(2)),...,\\
&d_n=\max_{r=1,n_m}(x(k_r)-\hat{x}(N))^T R(\tilde{X}^T(N))(x(k_r)-\hat{x}(N)).
\end{aligned}\tag{6}$$

and accordingly

$$\begin{aligned}
&R(\tilde{X}^T(1))=\tilde{X}^{+^T}(1)\tilde{X}^{+}(1),\\
&R(\tilde{X}^T(2))=\tilde{X}^{+^T}(2)\tilde{X}^{+}(2),\ldots,\\
&R(\tilde{X}^T(N))=\tilde{X}^{+^T}(N)\tilde{X}^{+}(N).
\end{aligned}$$

This approach allows to reorganize points of the set of unnamed documents and high level of confidence to extend the document training sample for each category for the following classification.

So, we have extended documents training sample for applying of any classification algorithm. A new set of points represents, geometrically, ellipsoidal containers, therefore apply the verification of disjointness created containers to reduce the influence of distortion of formed training sample.

Considering (2) for the 1st, 2nd and N-grade appropriately, ellipsoidal containers can be represented as follows

$$\begin{aligned}
&V_1(x)=d_1^{-2}(x-\hat{\hat{x}}(1))^T R(\tilde{X}^T(1))(x-\hat{x}(1))\le 1,\\
&V_2(x)=d_2^{-2}(x-\hat{\hat{x}}(2))^T R(\tilde{X}^T(2))(x-\hat{x}(2))\le 1,\\
&V_N(x)=d_n^{-2}(x-\hat{\hat{x}}(N))^T R(\tilde{X}^T(N))(x-\hat{x}(N))\le 1.
\end{aligned}\tag{7}$$

disjointness of these containers could be made with checking the following conditions

$$\begin{aligned}
&x(i_l)\overline{\in}\{x:V_2(x),\ldots,V_N(x)\le 1\},l=\overline{1,n_1},\\
&x(j_s)\overline{\in}\{x:V_1(x),V_3(x),\ldots,V_N(x)\le 1\},s=\overline{1,n_2},\\
&x(k_r)\overline{\in}\{x:V_1(x),V_2(x),\ldots,V_{N-1}(x)\le 1\},r=\overline{1,n_m}.
\end{aligned}\tag{8}$$

Correlation (8) is only the requirement of disjointness ellipsoidal containers (7), but if

$$\begin{aligned}
&V_1(x(j_s),x(k_r))>\Delta,s=\overline{1,n_2},r=\overline{1,n_m},\\
&V_2(x(i_k),x(k_r))>\Delta,k=\overline{1,n_1},r=\overline{1,n_m},\\
&V_N(x(i_k),x(j_s))>\Delta,k=\overline{1,n_1},s=\overline{1,n_2}.
\end{aligned}\tag{9}$$

where Δ is much greater than 1, then these conditions almost provide the disjointness of containers. At this stage, expected that data in the training sample is complete, that is achieved including outlined above way,

$$\det R(\tilde{X}^T(D))>0,D=1,2\ldots N.$$

Otherwise, should be applied in determining the functions $V_1(x),V_2(x),\ldots V_n(x)$ instead of matrix $R(\tilde{X}^T(D))$ regularized matrix. If the ellipsoidal containers cross, we can use the way of sequential filtration of the most similar objects using recurrence relations which are proposed in [6].

Taking into account foregoing, there is the following combined algorithm of the classification that is based on clustering.

Step 1.Clustering data sets, including both named and unnamed data. As a result, we have an extended set of named data (training sample).

Step 2. Classification using the extended training set.

This approach engaged to use positive aspects of clustering and classification, it can reduce the impact of small training sample on classification results.

A set of documents for testing

We have a set of documents that consists of collection of 8300 essays, each document is classified to one of 65 categories.

For classifier testing purpose, it can be used the part of the collection and compare the results of the classifier with the results that were obtained through essays.

To assess the classifier, we use several metrics:

P = kr / n – (precision) – the ratio of the number of documents that have been assigned to the correct category of the total number of documents which belongs to this category.

R = kr / r – (recall) – the ratio of the number of documents that have been assigned to the correct category of the total number of documents which belongs to a particular category by results of the classification.

In Figure 1 we can see this dependence of evaluation parameters of the classifier on the completeness of the training sample.

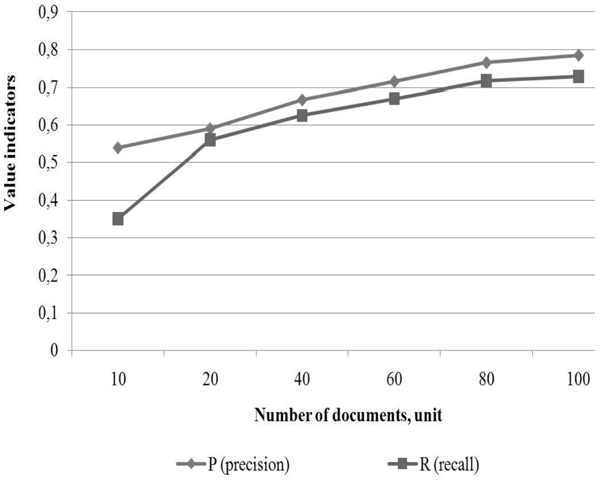

Fig.1.Test results of the classifier.

Conclusions

The test results show that the documents classification strongly depends on the quality and the size of the training sample. The proposed method combines the positive qualities of the ways of clustering and classification, which, at testing, have showed very good results of the classification of essays collection in ukrainian language. Considering almost the absence of training and test samples for printed documents in Ukrainian language the outlined approach can significantly improve the development and testing of new algorithms of classification. Further it's possible to apply this approach to the web pages, for this should be improved the automatic collection and pre-processing incoming web-documents.

References

1. *Nigam K.* Text Classification from Labeled and Unlabeled Documents using EM / Nigam K., Mccallum A., Thrun S., Mitchell T. // Machine Learning. – 1999. – P.103 –134.

2. *Wang J.* On transductive support vector machines / Wang J. Xiaotong S., Wei P. // Prediction and Discovery American Mathematical Society – 2007.– vol. 444 – P. 7 – 19.

3. *Zeng H.* CBC: Clustering Based Text Classification Requiring Minimal Labeled Data / Zeng H., Wang X., Chen Z., Ma W. // ICDM. Third IEEE International conference : Beijing. – 2003. – P. 443 – 450.

4. *Korlyuk A.* Hyperplane classifier of full-text documents / A. Korlyuk // Bulletin of Taras Shevchenko National University of Kyiv. Series: Physics&Mathematics- 2012. - №4. - C. 136 – 138.

5. *Kirichenko M. F.* Method of K-hyperplane clustering of text information / M. F. Kirichenko, O. S. Korlyuk // USiM. – 2008. – № 5. – P. 79 –75 (in russian).

6. *Kirichenko M. F.* Container Funds clastering and classification of signals / M. F. Kirichenko, O.S. Korlyuk // Cybernetics and systems analysis. – 2009. – № 5. – P. 111–118 (in russian).

Arriving to an editorial board 03.02.13.