ALGORITHM FOR CLASSIFYING DOCUMENTS OF A SCIENTIFIC AND EDUCATIONAL ORGANIZATION USING MACHINE LEARNING METHODS

Zebiniso Abdulxamidovna Abduvalieva
Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, graduate student

Ergashev Sirojiddin Baxtiyorovich
Jizzakh Branch of the National University of Uzbekistan, assistant

Annotation: Intelligent analysis is used in almost all areas of technology. Machine learning does not stand still and is constantly evolving. Given the transition in modern society to electronic document management, the main assumption in them is that the training and test data must be in the same feature space and follow the same distribution. In real applications, this is not always the case. In this case, the role of transfer learning can be distinguished since transfer learning does not make the same distributional assumptions as traditional machine learning and reduces dependencies on the target task and training data, and has a wider knowledge migration. The article proposes a transfer learning algorithm for document categorization based on clustering. An experiment is also used to test the algorithm. The experiment shows that the algorithm proposed in this article has its advantages.

Keywords: *transfer learning, machine learning; classification of documents; data mining, based spatial clustering*

I. Introduction

Most machine learning and data mining algorithms usually assume that the training and test data have the same feature space and data distribution, but in a real application, these two factors often change, so the trained model becomes outdated very easily.When the existing training data is outdated and there is very little new data, or labeling the new data is expensive, you might consider using the existing training data but a different distribution with test data to help the new data learn what transfer learning is.

Cluster analysis comes from many areas of research, including data mining, statistics, biology, and machine learning. The main clustering methods include partitioning methods, hierarchical methods, density-based methods, grid-based methods, and model-based methods.

In the algorithm considered and used below, D is a set (collection) of text documents of the department, W is a set (dictionary) of all words used in them, C is a set of document categories fixed in advance. Each document d ∈ D is a sequence of words слов (w1, . . .

, wnd)$w_{1,\dots,}w_d$ from dictionary W, where d is the length of the document in words. The same word can be repeated many times in a document. The categorization problem is the problem of assigning a boolean value to each pair $\{d, c\} \in D * C$.. Boolean value 1 means document d belongs to category c, while a value of 0 means the opposite. More formally, the categorization problem is the problem of recovering an unknown objective function $\Phi: D * C \to \{1,0\}$.

In many document categories, the training data set from the source field is always out of date, but if some existing stale data is similar to the test data in the target field, then we might consider using clustering technology to find them to help train targets and goals.

In this article, the density-based spatial clustering (DSP) algorithm is applied to classify documents in higher education institutions. In this case, the documentation of the Department of "System and Software" at the Institute of Tashkent University of Information Technologies. The essence of classification is to reduce and filter characters and repeat words. Information in documents is treated as character labels and has no additional meaning.

One of the difficulties of categorizing documents is the high dimension of the feature element space. Characteristic elements in the categorization of documents mainly refer to words obtained as a result of text processing, and the dimension of a functional element is equal to the number of different words. In this article, density-based spatial clustering The problem of non-hierarchical categorization can be considered as a multi-class classification problem, for which the set of classes is the set of categories C, the set of objects is the set of documents D, and the set of precedents is a previously known set of pairs {d,c}, where $d \in D$, $c \in C$.

II. Material and methods

There are three approaches to solving the text classification problem: supervised learning, unsupervised learning, and confirmation learning. One of the popular approaches of most interest is classification based on machine learning. With this approach, the training of the classifier (a system of naming objects, each of which corresponds to a unique identifier) is carried out on a set of initial training data in the form of documents with categories assigned

$$\begin{array}{c} \\ w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_n \end{array} \begin{array}{c} \begin{array}{ccccc} w_0 & w_1 & w_2 & \dots\dots & w_n \end{array} \\ \begin{bmatrix} a_{00} & a_{01} & a_{02} & \dots & a_{0n} \\ & a_{11} & a_{12} & \dots & a_{1n} \\ & & a_{22} & \dots & a_{2n} \\ & & & \ddots & \vdots \\ & & & & a_{nn} \end{bmatrix} \end{array} \qquad (1)$$

In formula (1), the vertical and horizontal coordinates are the keyword list index. $\mathbf{a_{ij}}$ $(\mathbf{i \neq j})$ indicates the number of words i_w and j_w occurring together in the same document. $\mathbf{a_{ij}}$ $(\mathbf{i - 1, 2, \dots n})$ expresses the frequency of words. The distance between words can be determined using the word matching matrix. The higher the frequency of matching two words, the smaller the distance between them. The transformation formula used here is:

$$d(w_i, w_j) = \frac{1}{(1+co\_words(w_i,w_j))} \quad (2)$$

$co\_words(w_i, w_j)$ in formula (2) shows the number of words w_i, w_j in the same document, denominator plus 1 to eliminate the possibility of infinite distance while protecting the standardized requirements. Using formula (1), formula $2\ a_{ij}$ can be converted into a distance to obtain a word matching matrix of documents.

To initialize the Eps radius, a minimum number of points is considered using a "density-based spatial clustering" algorithm to achieve word clustering. The cluster should be output after clustering and cluster processing. An isolated point is removed. The more closely the words are connected in a cluster, the farther it is.

Definitions in the "Density Based Spatial Clustering" algorithm:

dense area: for each point in the cluster, the circle with radius contains at least the minimum number of points (Min points).

The Epsilon neighborhood of a point P in the database is determined by the following formula:

$$N(p) = \{q \epsilon D | dist(p, q) \leq \epsilon\} \quad (3)$$

With feature clustering, the new model of the vector space of documents can be expressed as the i-th cluster.

If we consider N as the number of documents in collection D.

The weight formula will look like this[4]:

$$w_i(d) = \frac{TF * \log_2(N/N_t + \beta)}{\sqrt{\sum_{i=1}^{n} TF * \left[\log_2(N/N_t + \beta)\right]}} \quad (5)$$

Formula (5) is directly related to the vector space model, where TF is the word frequency of the characteristic word in the tth feature cluster, and $N_t$refers to the number of documents in the collection. The characteristic word appears in the tth characteristic cluster. The feature cluster weight can also be obtained by accumulating each feature word in a member of the cluster. The similarity between two documents can be obtained by calculating the cosine of the angle between the two vectors, assuming that the two documents

$$d_1 = (t_1, w_1, t_2, w_2, \dots, t_n, w_n)$$

$$\text{и} \quad d_1 = (t_1, x_1, t_2, x_2, \dots, t_n, x_n)$$

similarities between formulas:

$$\text{sim}(d_1, d_2) = \cos \propto = \frac{\sum_{i=1}^{n} \varpi_i * x_i}{(\sum_{i=1}^{n} \varpi^2 * \sum_{i=1}^{n} x_i^2)^{\frac{1}{2}}} \qquad (6)$$

The categorization task means assigning a boolean value to each pair of $f_d$fd

$$\{d. c\} \in D * C$$

A boolean value of 1 means that document d belongs to the given category, while a value of 0 means the opposite. The categorization problem is the problem of recovering an unknown objective function : $\Phi: D * C \rightarrow \{1,0\}$

All categories can be considered symbolic labels, and their meaning does not have any additional meaning.

When categorizing documents, first of all, the secondary training data with the target training data are combined for clustering. The secondary training data that is not collected together with the training target data in the same cluster is filtered out. The rest is higher than the target data, and it will be trained along with the training target data. This will greatly improve the classification performance. Some definitions will be given for the main characters used in the article.

If you set the target pattern space to F' and F for the secondary pattern space, $Y = \{0, 1\}$ for the class space. test data set: $S^c = \{(x_1^c, x_2^c, \dots, x_p^c)\}$, $x_k^c \epsilon F'$, $k = 1.2.3, \dots p.$

The training dataset consists of two parts : the target training dataset: $D' = \{(x_1^t, y_1^t), (x_2^t, y_2^t), \dots, (x_n^t, y_n^t)\}$ $x_i^t \epsilon F^t$, $y_i^t \epsilon Y (i = 1,2, \dots n)$ and a set of secondary training data : $D^s = \{(x_1^s, y_1^s), (x_2^s, y_2^s), \dots, (x_n^s, y_n^s)\}$ $x_j^s \epsilon F^s$, $y_j^s \epsilon Y (i = 1,2, \dots m)$ weight of samples in D' is: $w_1^t, w_2^t, \dots, w_n^t \{\}$; prediction objective function: $h(x_i): x_i \rightarrow y_i$, $\widetilde{P}(x_i^1)$~ is the prior probability of x_i^1 в D'.

Algorithm steps:

The training data sets $D^S$and the target data set D^t, the test data set S are fed into the input. Output: classified $h'(x')$

a) into the input D^S and D^t set $\tilde{p}(x_j) = w_i / \sum_{i=1}^{n_i} w_i$;

b) according to the class standard, the training data can be divided into N classes: D_i $(i = 1,2,3, \dots N)$, Di $(i = 1, 2, \grave{} \grave{} \grave{} N)$, где D_i where D_i means the set of instance classes labeled i;

c) for $i = 1$ to N;

d) The k means clustering algorithm for D_i clustering is invoked and returns the clustering results;

e) scan D_i deleting a cluster of instances in secondary data that were not collected along with the target data;

f) end of this part

g) derived a classification model from the filtered training data and test data S by calling the KNN algorithm $h(x^i)(x^i \in F^t \cup F^S)$;

h) calculate the error rate $h(x_i^t)$on D^t

$$\delta^t = \sum_{j=1}^{n_i} \tilde{p}\,|(x_i^t - y_i^t)|$$

set $\beta = {}^{1}/_{2\ln(1-\delta^t)/\delta^t}$;

i) update the weight vector of the target training data, the weight of the first $K+1$ iterations is

$$w_i^t(k+1) = \begin{cases} w_i^t(k)e^{-\beta}, h(x_i^t) = y_i^t \\ w_i^t(k)e^{-\beta}, h(x_i^t) \neq y_i^t \end{cases} (6)$$

j) at the entrance:

$$h_i(x_t) \begin{cases} 1, \prod_{t=[N/2]}^{N} \beta_t^{-h_i^t}(x) \geq \prod_{t=[N/2]}^{N} \beta_t^{-1/2} \\ 0, \text{otherwise} \end{cases} \quad (7)$$

III. Application implementation

In the experiments, I used the documents of the Department of Systematic Practical Programming of the Tashkent University of Information Technologies. The data set contains categories of documents of the department, including information about the department, about students, articles, etc. Each large class also contains several sub-categories below and includes a total of 500 documents. In the experiment, the main categories are selected: department documents; documents related to teachers; and documents related to students, including the annual report, plans, and rating. In each selected category, there are subcategories. We chose the main categories because the objectives of the class and secondary documents are outdated. For example, the annual report of the department means that we select the report for 2021 and 2022 as target categories, for example, financial (wealth, consumption). The specific data distribution is shown in Table 1.

Table 1

Department data distribution

Target dataset Initial	Initial training data	Auxiliary data
Students' documents	Themes of diploma work and master's theses Written work by part-time students and reviews Documents on student practice	Documents on student practice

Plan	document control/ documents to test students' knowledge (bases of written and oral control questions, written and electronic test questions, options for written work, block modules, and so on).	report on spiritual and educational work.
Scientific works	Reviews of scientific works	Diplomas, dissertations
Teachers documents	Documents of cooperation with professional colleges and enterprises of the department	articles of teachers of the department / Individual work plans of teachers of the department
Department documents	Orders of the dean of the faculty and information on their implementation / Decisions of the University Council, Methodological Council, Faculty Academic Council (copy)	Certificates and instructions of the rector and vice-rectors of the university on the activities of the department and their implementation

The study compared the transfer learning algorithm, which is based on the application of knowledge gained from other studies, with the density-based spatial clustering algorithm.

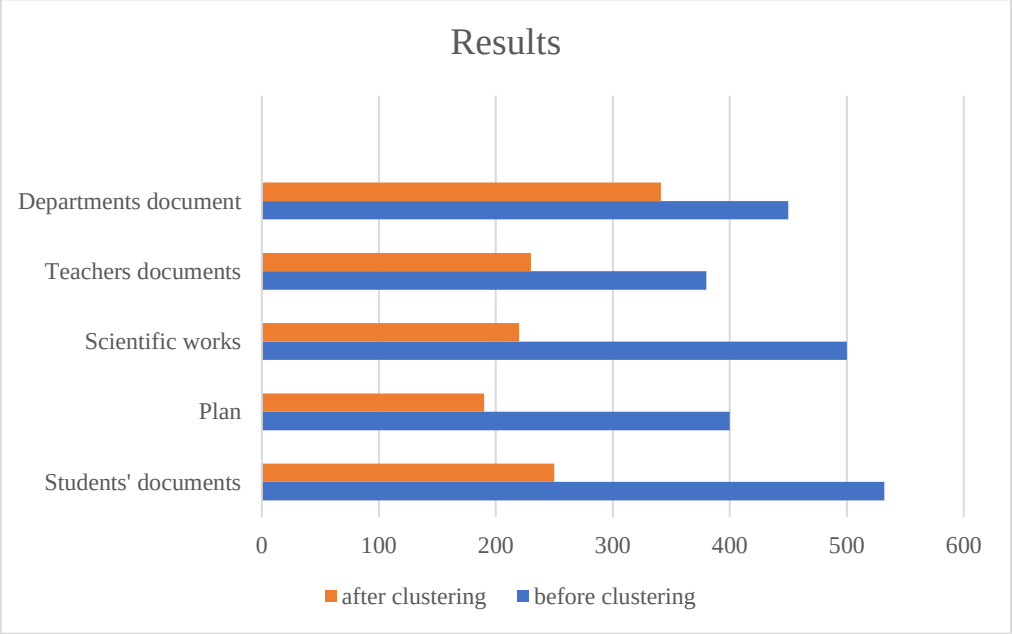
The article proposes a transfer learning algorithm based on clustering. The feature cluster is first achieved by the algorithm, then the algorithm is used for the dataset cluster, and finally, after the weight adjustment strategy, the K-means algorithm is used to classify the documents into different classes.

IV. Results

Table 2

The result after CLUSTERING

Data set	Before clustering	After clustering
Students' documents	532	250
Plan	400	190
Scientific works	500	220
Teachers documents	380	230
Departments document	450	341



Acknowledgment

The work is carried out with the Department of System and Practical Programming of TUIT. based on the documentation of the department.

VI. Conclusion

This article considers the task of classifying documents in the electronic document management system of a scientific and educational institution. A comparative analysis of existing approaches to machine learning was carried out, on the basis of which it was concluded that the use of transfer learning gives a more effective result in the classification of documents. educational institution to improve the quality of classification and. For solving the classification problem it is also necessary to select certain to which the initial set of documents will be distributed, for which the algorithm presented in the article is proposed. Thus, the algorithmic support presented in the article can be used as a theoretical basis for the integration of machine learning methods in the analysis and classification of documents of a scientific and educational institution.

References

1. Marszalek M., Schmid S., Harzalla H., Van de Weyer J. Learning object representations for class recognition of visual objects. In: Seminar on ICCV Visual Recognition Problems. 2007. [1]

2. Mikhalkova L., Mooney R.J. Transfer learning by matching against minimum target data. In: Proc. assoc. to advance the workshop on artificial intelligence transfer learning for complex tasks. 2008 [2]

3. Moreno O., Shapira B., Rokach L., Shani G. (2012) TALMUD—transfer learning for several areas. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management. 2012. [3]

4. Nam J., Kim S. (2015) Prediction of heterogeneous defects. In: Proceedings of the 10th 2015 Joint Meeting on Fundamentals of Software Engineering. 2015. [4]

5. Ng MK, Wu Q, Ye Y. Collaborative learning with a method based on a joint transition probability graph. In: Proceedings of the 1st International Workshop on Revealing Cross-Domain Knowledge in Web and Social Networks. 2012. [5]

6. W. Dai, Q. Yang, G. Xue and Y.Yu, “Boosting for Transfer learning”, Proc.24 International Conference. Machine Learning, pp.193-200, June 2007. [6]

7. Xinno Jialing Pan, Yang K. Survey on transfer learning. IEEE TKDE, 2009. [7]

8. Dai V, Yang Q, Xue GR, Yu Yu. Boosting for transfer learning. Proceedings of the Twenty-fourth International Conference on Machine Learning, 2007: 193 × 200. [8]

9. Dai Wee, Chen WK, Xue GR, Yang Q, Yu Yu. Transfer learning: Transferring learning to different functional spaces. Advances in Neural Information Processing Systems 21, 2009. [9]