Algorithm of Co-Authors Selection for Preparing Scientific Works based on Gaussian Models and Data from the E-library Web Resource

Nikita Andriyanov
Data Analysis and Machine Learning Department
Financial University under the Government of the Russian Federation
Moscow, Russia
nikita-and-nov@mail.ru,
ORCID: 0000-0003-0735-7697

Alexandr Tashlinskii
Radioengineering Department
Ulyanovsk State Technical University
Ulyanovsk, Russia
ORCID: 0000-0003-4732-0418

Vitalii Dementyev
Telecomunication Department
Ulyanovsk State Technical University
Ulyanovsk, Russia
ORCID: 0000-0002-4880-0433

Abstract—**The article deals with the urgent task of selecting co-authors of scientific works using cluster analysis methods. In particular, on the basis of the resources of the scientific electronic library in Russia, E-library, test data on scientists were selected. These data were used to unite scientists into clusters according to interests and topics of their publication. To solve this problem, a clustering method based on Gaussian mixture models (GMM) was used. The result of the research groups selection showed that the algorithm is able to qualitatively select scientists with common interests. To assess the effectiveness of the algorithm, the clustering results were checked, where the groups of scientists who already had common publications were chosen as the base. The obtained clustering accuracy was 100\% according expert assessment and exceeded the indicators obtained using the K-means algorithm.**

Keywords— Cluster Analysis, Web Engineering, Gaussian Model of Mixtures, K-means, Research Teams, E-library, Education.

I. Introduction

Today, for effective work in higher educational institutions, teachers need to conduct active publishing activities [1]. In the context of the rapid development of information technology, teachers of disciplines in the field of computer science are also forced to actively work on improving the courses they teach. On the other hand, in the humanities, the workload of teachers is also increasing. This, in turn, leads to a deterioration in publication activity. In addition, scientific foundations often hold competitions, for the submission of applications for which a large number of competencies are required, which are not always available within one university department. The most important condition for maintaining publication activity with high-quality material that allows the publication of high-level scientific works is the breakdown of the work on the article between co-authors. The selection of co-authors today often takes place within one structural unit, sometimes with the involvement of graduate students and students. However, analysis shows that the most successful articles are prepared by distributed teams [2]. However, finding coauthors "from outside" is time consuming. Therefore, it is desirable to reduce the initial reduction in the circle of potential scientific partners using data mining algorithms.

Moreover, today there are many international competitions for scientific projects [3]. Usually well-coordinated teams from different countries, between which there is already a connection, take part in such competitions. These connections most often arise in the framework of international conferences. However, during the period of restrictions introduced due to the spread of coronavirus infection, the likelihood of such contact is significantly reduced. Thus, an urgent task is to develop an algorithm that will analyze the publication activity of authors and, on the basis of such analysis, propose research teams.

This article discusses an algorithm for clustering multivariate data based on Gaussian mixture models (GMM), which will be described in detail in the next section. As data for the analysis, characteristics from the Russian scientific electronic library E-library \cite{4} were selected. E-library allows authorized users to view information about authors, articles and citations.

II. Gaussian Mixture Models

One of the popular clustering algorithms is the GMM [5-7]. The popularity of this model is due to the use of a normal distribution, with which most of the real data can be described. Another advantage of the model is its applicability to multidimensional data. The GMM, like the K-means algorithm, requires a preliminary determination of the number of clusters, as well as several more parameters. On the other hand, there are information criteria, on the basis of which the optimal GMM can be chosen for the given parameter options. These criteria include the Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC) [8]. The selected parameters of the model include the following:

1) Number of clusters K.

2) Characteristics of the covariance matrix.

A covariance matrix is classified by the relationship between parameters within one class into a matrix with full and diagonal structure. According to the relationship between different clusters, covariance matrices are divided into shared and unshared. This classification takes into account the orientation and size of the clusters. With the standard approach to the clustering problem, ellipsoids describing clusters can be oriented in multidimensional space at any angle, which is provided by the full structure of the covariance matrix. For a diagonal structure, the orientation must be strictly perpendicular and parallel to the axes of the main parameters. Since the data can be heterogeneous, and the clusters contain a completely different number of objects, they use covariance matrices with a unshared structure. The shared structure implies that all ellipsoids will have the same dimensions along each axis and the same orientation in space.

3) Regularization parameter R.

This parameter usually takes values of tenths, hundredths or thousandths. Regularization allows the Gaussian model not to fall apart when obtaining an unsuccessful covariance matrix, since this parameter provides a positive value for the determinant of the covariance matrix.

Thus, if the distribution of parameters describing the object is approximated by a Gaussian distribution, the probability of high accuracy of clustering will be quite large.

In the general case, the system can have K clusters, to which, based on the analysis of N parameters, M objects should be assigned. The simplest case describes the situation when there are two classes into which objects with only one property X should be distributed. Thus, the solution to the clustering problem is represented in the form of a classical Bayesian detector.

There is a distribution of the parameter X under two hypotheses: H_1 is about belonging to class №1, H_2 is about belonging to class №2.

Based on the current value X_i, describing the property X for the $i-$th object, it is required to determine the closest distribution. It is clear that this can be done by estimating the probabilities that X_i is an object of classes No. 1 and No. 2. For this, it is possible to construct the probability distribution density function (PDF) of the parameter X for both classes. The Gaussian distribution is described using mean and variance. The one-dimensional normal PDF is known to have the form (1).

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left(-(x-m_x)^2 \Big/ 2\sigma_x^2 \right), \qquad (1)$$

where m_x is average value of the parameter X, σ_x^2 is variance of the parameter X.

The difference between distributions (1) for hypotheses H_1 and H_2 consists in different values of the distribution parameters.

However, if the total distribution is constructed based on distributions of the form (1) for the global case of the presence of K clusters and, accordingly, K hypotheses, then it is possible to write an expression of the form (2), which describes the model of Gaussian mixtures.

$$f_{GMM}(x) = \frac{1}{K}\sum_{i=1}^{K} f_i(x), \qquad (2)$$

where K is the total number of clusters in the task.

For example, let's plot such a mixture for two clusters. In this case, the first cluster is described by a normal distribution with parameters $m_{x1} = 0$, $\sigma_{x1}^2 = 1$ and the second cluster has $m_{x1} = 7$, $\sigma_{x1}^2 = 2$.

Fig. 1 shows a plot of the PDF of the mixture.

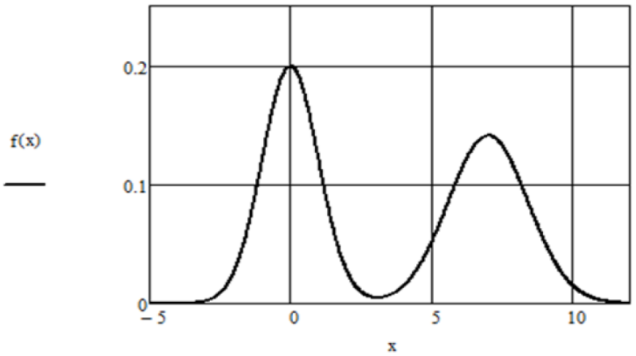

Fig. 1. Mixture of two Gaussian distributions

Analysis of expression (2) shows that the area under each figure will be equal to 0.5. On the other hand, the variance for the second distribution is greater; therefore, intuitively, the probability of an object falling into the second cluster should also be greater. However, for Fig. 1, the probabilities of belonging to each cluster coincide. To simulate a situation in which the areas under the curve formed by the PDF will be different, it is necessary to produce a weighted mixture of distributions. In this case, the weights can be the probabilities of belonging to each cluster $p_1, p_2, ..., p_K$. It should be remembered that the sum of these probabilities should be equal to 1.

Let's rewrite the mixture model in the form (3).

$$f_{GMM}(x) = \sum_{i=1}^{N} p_i f_i(x) \qquad (3)$$

To illustrate, let us mix two distributions in such a way that the probability of hitting the first cluster is 75%, and the probability of hitting the second cluster is 25%. The distribution parameters correspond to the distribution parameters from Fig. 1. So Fig. 2 shows a mixture of normal distributions with different proportions.

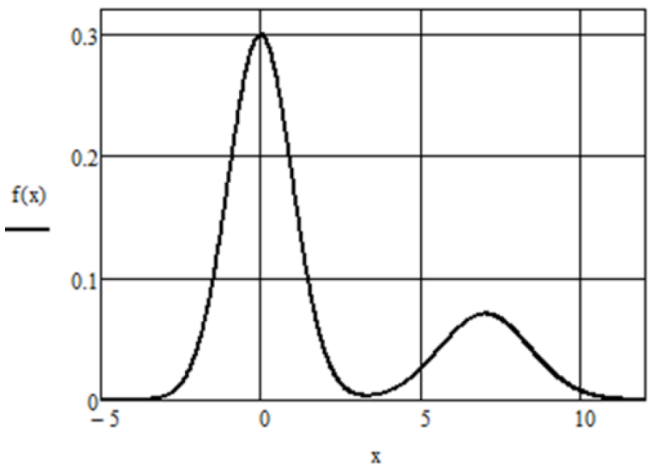

Fig. 2. Proportional mixture of Gaussian models

Analysis of Fig. 2 shows that the general distribution of the variable X is such that the PDF of the mixture has 2 maxima. Depending on the distance to the nearest maximum, the classification of any object described by the parameter X can be easily performed. And the average values for the distributions of the parameter X in clusters 1 and 2 will be the centers of the clusters. And since variances determine the spread of values, they allow us to choose the correct cluster sizes.

Obviously, the further in space the maxima themselves diverge, the easier and more accurate the clustering will be. In addition to the X parameter, a whole vector of parameters $X = (X_1 \quad X_2 \quad \ldots \quad X_N)$ can be used to describe each object. Then it is necessary to make a transition to a multidimensional GMM. The convenience of such a model lies in the fact that it is quite simple to generalize to the multidimensional case. If for the considered example, clusters are considered in the form of segments on the axis of the parameter X, then in the two-dimensional case the sections of Gaussian distributions are ellipses. Such ellipses are used to describe clusters. In the three-dimensional case, GMM provides ellipsoids, etc. It is also important that the model, for which the specified type of covariance matrix, regularization and the number of clusters are specified, is able to independently select normal distributions and build data clusters from the available data. Thus, when applying the Gaussian model, neither data markup nor training is required.

It should be noted that an increase in the number of parameters does not always lead to an increase in the efficiency of clustering. Sometimes some properties can introduce additional complications, so the choice of the main characteristic properties of objects is important. Despite this, the model based on Gaussian mixtures was chosen to develop the algorithm for clustering scientists. The choice of algorithm is easy to understand since the choice of the scientists parameters was made in manual mode, and they were chosen in such a way that the multidimensionality should not cause a decrease in the efficiency of the algorithm.

III. Source Data Sampling and Clustering Results

For the research 10 scientists were selected using the E-library system. This choice was made based on four articles with 4, 3, 2 and 1 authors. The first 2 keywords were selected from each article. Further, for each of the selected scientists, the share of the use of each keyword in his works was calculated. It is important to understand, since several keywords can be used in one work, the restrictions on the total value of the share will not be one, but the number of keywords being checked.

Table 1 shows the results of the preparation of the test sample. In order to protect information, the personal data of scientists were anonymized, and the keywords were indicated as they really are. It should be noted that, according to the expert opinion, only the group of coauthors from the third article (2 coauthors) has a low probability of overlapping the area of interest with other coauthors.

After data collection a covariance matrix was calculated to analyze the relationship between topics for the selected scientist. Table 2 shows the results obtained from covariance matrix. However the covariance was normalized so Table 2 provides correlation coefficients between topics.

TABLE I. Distribution of Topics by Scientists

Topic/Scientist	CV	ML	IP	AR	LC	RP	CNN	AUG
Scientist1 (Paper1)	0.226	0.34	0.787	0.52	0	0	0.333	0.189
Scientist2 (Paper1)	0.189	0.2	0.654	0.16	0.05	0.05	0.255	0
Scientist3 (Paper1)	0.614	0.135	0.59	0.12	0	0	0.614	0.018
Scientist4 (Paper1)	0.578	0.642	0.435	0.372	0.037	0.037	0.656	0.382
Scientist5 (Paper2)	0.218	0	0.805	0.92	0.12	0.12	0.182	0.012
Scientist6 (Paper2)	0.189	0.236	0.732	0.514	0.08	0	0.165	0.12
Scientist7 (Paper2)	0.756	0.522	0.792	0.2	0	0	0.718	0.365
Scientist8 (Paper3)	0.108	0.151	0.332	0	0.5	0.3	0.102	0.067
Scientist9 (Paper3)	0	0	0.5	0	0.75	0.75	0	0
Scientist10 (Paper4)	0.614	0.756	0.614	0.21	0	0	0.614	0.432

TABLE II. Correlation Matrix of Research Topics

Topic	CV	ML	IP	AR	LC	RP	CNN	AUG
CV	1	0.720	0.159	-0.058	-0.672	-0.619	0.981	0.720
ML	0.720	1	-0.009	-0.076	-0.541	-0.525	0.767	0.966
IP	0.159	-0.009	1	0.643	-0.567	-0.471	0.136	0.041
AR	-0.058	-0.076	0.643	1	-0.437	-0.403	-0.061	0.002
LC	-0.672	-0.541	-0.567	-0.437	1	0.968	-0.713	-0.459
RP	-0.619	-0.525	-0.471	-0.403	0.968	1	-0.651	-0.440
CNN	0.981	0.767	0.136	-0.061	-0.713	-0.651	1	0.742
AUG	0.720	0.966	0.041	0.002	-0.459	-0.440	0.742	1

The following abbreviations are used in Tables 1 and 2: CV - Computer Vision; ML - Machine Learning; IP - Image Processing; AR - Autoregression; LC - Laser Coagulation; RP, Retinopathy; CNN - Convolutional Neural Network; AUG - Augmenation.

Analysis of the results obtained shows that Table 1 contains 10 objects described using 8 parameters. Accordingly, these objects can be clustered using GMM. Table 2 shows which topics are most related to each other.

However, this analysis is recommended not for a small sample of 10 scientists, but for the entire system. Moreover, Table 2 shows that scientists working with laser coagulation can only cooperate with scientists studying retinopathy. Indeed, retinopathy is a disease that can be treated using laser coagulation.

In general, depending on the level of correlation between topics, it is possible to search for co-authors by specifying the topic and choosing other topics with a level above a certain threshold, for example, with a correlation greater than 0.5. Then, going to Table 1, it is possible to select a scientist whose share of work in the field of related topics also exceeds a certain threshold, for example, with a share of keywords greater than 0.5. The higher the selected thresholds, the narrower the circle of specialists will be selected by such a filter.

Finally, let’s consider the clustering of scientists into 2 classes using a GMM, K-means clustering and implying reference assignment of authors №8 and №9 to a separate cluster.

Table 3 shows the clustering results.

TABLE III. CLUSTERING OF SCIENTISTS

Topic/Scientist	GMM	K-means	Expert
Scientist1 (Paper1)	Cluster1	Cluster1	Cluster1
Scientist2 (Paper1)	Cluster1	Cluster1	Cluster1
Scientist3 (Paper1)	Cluster1	Cluster1	Cluster1
Scientist4 (Paper1)	Cluster1	Cluster1	Cluster1
Scientist5 (Paper2)	Cluster1	Cluster2	Cluster1
Scientist6 (Paper2)	Cluster1	Cluster1	Cluster1
Scientist7 (Paper2)	Cluster1	Cluster1	Cluster1
Scientist8 (Paper3)	Cluster2	Cluster2	Cluster2
Scientist9 (Paper3)	Cluster2	Cluster2	Cluster2
Scientist10 (Paper4)	Cluster1	Cluster1	Cluster1

Thus, the analysis shows that scientists №8 and №9 can be combined into one scientific group. In addition, there is the potential for a joint paper to be written for the remaining scientists. This is confirmed by expert analysis. However, the K-Means algorithm also assigned scientist № 5 to the second cluster. This is probably due to the third level of his share of publications in the field of RP and LC.

IV. CONCLUSIONS

The article presents an approach to the preparation of a database and its processing in order to identify potential colleagues in scientific work using open web resources. An algorithm based on Gaussian mixture models was used to unite scientists into groups. This approach made it possible on a test sample of 10 scientists to obtain a classification that fully corresponds to the classification proposed by the expert. At the same time, the K-means algorithm resulted in a discrepancy with expert assessment for the same data. In the future, it is planned to use correlations of research topics to improve the quality of the algorithm.

ACKNOWLEDGMENT

The research was funded by the RFBR and the Government of the Ulyanovsk region, Project №19-47-730011, and also partially funded by the RFBR, Project №19-29-09048.

REFERENCES

[1] I.N. Kim, “On the measures contributing to the successful formation of the publication career of a university teacher,” Engineering Education, vol. 18, pp. 64-78, 2015.

[2] L. Puentes, A. Raina, J. Cagan, C. McComb, “Modeling a strategic human engineering design process: human-inspired heuristic guidance through learned visual design agents,” Proceedings of the Design Society, Cambridge: Cambridge University Press, 2020, p. 355-364, doi: 10.1017/dsd.2020.42.

[3] Funds progamms, url: https://narfu.ru/international/projects/funds programms/, last accessed 2021/02/27.

[4] Electronic Library, url: https://www.elibrary.ru/, last accessed 2021/02/2021.

[5] T. Athey, B. Pedigo, T. Liu, J. Vogelstein, “AutoGMM: Automatic and Hierarchical Gaussian Mixture Modeling in Python,” url: https://arxiv.org/pdf/1909.02688.pdf, last accessed 2021/02/26.

[6] N. Andriyanov, A. Tashlinsky, V. Dementiev, “Detailed Clustering Based on Gaussian Mixture Models,” Advances in Intelligent Systems and Computing (AISC), vol. 1251, pp. 437-448, 2021, doi: 10.1007/978-3-030-55187-2 34.

[7] N. Andriyanov, “Comparative analysis of football statistics data clustering algorithms based on deep learning and Gaussian mixture model,” CEUR Workshop Proceedings, vol. 2667, pp. 71-74, 2020.

[8] H. Akaike, “A new look at the statistical model identification,” IEEE Transactions on Automatic Control, vol. 19, pp. 716-723, 1974.