Comparative analysis of football statistics data clustering algorithms based on deep learning and Gaussian mixture model

Nikita Andriyanov
JSC "RPC "Istok" named after Shokin
Fryazino, Moscow Region, Russia
Telecommunication department
Ulyanovsk State Technical University
Ulyanovsk, Russia
nikita-and-nov@mail.ru

***Abstract*—The paper considers the Gaussian mixtures model and the possibilities of its application for solving clustering tasks. First, the case is considered when the Gaussian mixtures model is formed in such a way that all the parameters of the model are known. Next, the case is considered when the approximation of normally distributed data occurs using the Gaussian mixtures model. Finally, the article presents a study of the accuracy of clustering two-dimensional data of football statistics of medal-position teams, middle-table teams and worst teams of the top 5 European football championships such as English Premier League, Spanish La Liga, German Bundesliga, Italian Serie A and French League One. The results of the algorithm based on the Gaussian mixtures models are compared with the results of clustering performed using neural networks.**

Keywords—Gaussian mixture models; machine learning; data clustering; data analysis; football statitstics

I. Introduction

Today data mining as intelligent analysis allows specialists in various fields to greatly simplify their work. For example, on the basis of such an analysis, deliberately non-solvent customers who apply for a loan to the bank can be eliminated, and data on the number of taxi service orders can be predicted [1,2]. Indeed, digitalization of various areas of the economy and areas of state activity on an ongoing basis provides significant amounts of information. In this regard the range of tasks solved using data mining is so wide.

One of the most interesting tasks in this area is the problem of data clustering [3,4], which should be associated with the recognition, classification or segmentation tasks [5-9]. However, in these tasks it is usually possible to distinguish several groups of objects. The simplest example is the choice of male students and female students in a group. Every person here can be described by their height and weight. Each object in such sample can be displayed at a specific point on the plane. In this case this plane is two-dimensional. It is possible to expand dimensions of the plane if the new parameter, for example, a hair length will be introduced. Then the solution of the clustering task will be simplified. Each group of objects can be represented by some ellipsoid at the plane. Then the clustering decision for a particular new object will depend on which ellipsoid is closest to the point characterizing this object.

So the further research considers a clustering algorithm based on Gaussian mixtures models (GMM) [10, 11], because quite often real data can be well approximated by Gaussian distributions. And the comparison algorithm is trained neural network clustering. It should be noted that for the first time a comparison of the GMM and trained neural networks will be performed as part of the task of analyzing football statistics. In addition, a combination of the proposed clustering methods can lead to a new type of clustering bases simultaneously on supervised and unsupervised learning.

II. Brief Classification of Clustering Algorithms

Known clustering algorithms [3] can be divided according to 2 basic principles. Let consider main features for them.

First, clustering can be crisp or fuzzy. In the first case each object as a result of clustering is assigned exactly one group. With fuzzy clustering a set of values is usually determined that characterize the belonging probability of each object to each group, i.e. such clustering gives some probability distribution.

Secondly, cluster analysis can be flat single-level or hierarchical multi-level. In the first case the initial selection of objects according to some criterion is divided into several classes in the form of a single partition. For example, clustering the university students again only by gender. If the further clustering considers that male students and female students will be separated, keeping the first level, then a deeper clustering will be obtained, in particular, the original object in the sample can be characterized not just as a male student or female student, but as an excellent (“A”) male student, excellent (“A”) female student, bad (“F”) male student or bad (“F”) female student. This separation provides hierarchical clustering. It should be noted that the deep Gaussian mixtures model (DGMM) considered in [11] copes well with the goals of hierarchical clustering. Moreover, the assignment of an object to a particular group is carried out according to the principle of crisp clustering.

Finally, neural networks are gaining more and more popularity in clustering problems [12]. Depending on the training parameters and type of networks, various models for clustering can be obtained. And now a deep learning is a very perspective tool for mentioned tasks.

Thus, before choosing a clustering algorithm, it is necessary to first formulate the clustering problem itself, and then perform the data splitting.

III. Gaussian Mixture Model

The application of flat, crisp clustering is considered on the example of analysis of football statistics from the Top 5 European Championships (England, Spain, Germany, Italy, France). Since the problem of multilevel clustering is not posed, it is possible to use GMM [10]. This is such a model, the probability density function (PDF) of which is described by the sum of the PDFs of Gaussian distributions. The number of terms in the sum is the number of clusters. Thus, the total distribution has several peaks, and for each of the objects during clustering the proximity to each peak is considered and the peak with the smallest distance is selected. Moreover, each object can be characterized not by one but by several parameters, for which multidimensional PDFs are found. Fig. 1 presents an example of the PDF of the GMM of three distributions with two parameters.

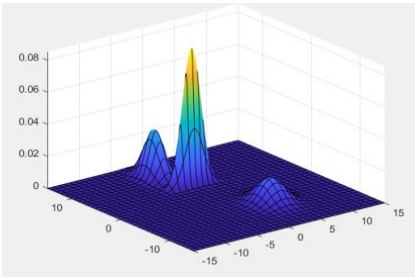


Fig. 1. PDF of 3 distribution GMM.

An analysis of Fig. 1 allows to conclude that there are two groups of objects that are characterized by a large variance along one of the axes (ordinates or abscissas), and one group with approximately the same variance along both axes. In addition, three characteristic peaks or mathematical expectations can be seen in Fig. 1.

The advantage of using the GMM is that for a given number of objects, the model itself performs estimates of the component distributions. This allows the approximation of real data using such a model. However, even if the number of clusters is not known in advance, it is possible to build several models of mixtures and choose the optimal one according to some criterion. Most often, the Akaikeian information criterion (AIC) [13] and the Bayesian information criterion (BIC) [14] are used. Application of these criteria allows to cope with the problem of a priori uncertainty regarding the number of classes.

IV. Clustering with a Gaussian Mixture Model

Consider an example of the GMM application in the clustering of teams playing in the European football championships in England, Spain, Germany, Italy and France. Only 2 parameters will be included in the initial sample. It is goals scored and points. However, in order to make it more convenient to check the accuracy of clustering, it is good idea to exclude some teams from the selection. Thus, the thinning done will include 3 teams in the upper part of the tournament table (1 - 3 places), 3 teams in the middle of the table (9/8 - 11 /10 places) and 3 teams in the lower part of the table (18/16 - 20/18 places). Such thinning is done for each championship. In addition, a statistics on such teams not only for the last season, but also for the previous 2 seasons is taken. This, on the one hand, will increase the information content of the sample, and on the other hand, it can also lead to an increase in anomalous points (“too successful”, “too unsuccessful” or “strange” season for one team or in general). Fig. 2 shows the collected statistics. The points are plotted on the abscissa axes, and goals scored are plotted on the ordinate axes.

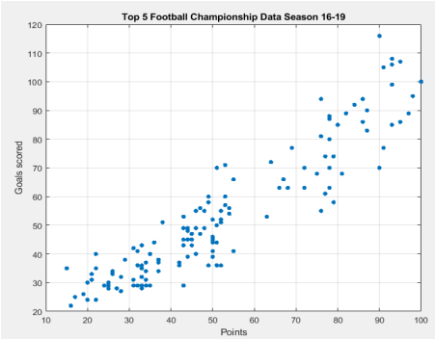


Fig. 2. Statistics of the top 5 football championships for the seasons 2016/2017, 2017/ 2018 and 2018/ 2019.

From Fig. 2 it can be seen that the selected parameters have an almost linear relationship and visually the most preferable division seems to be simply dividing by lines along the abscissa (points). In this case, the numbers 40 (points) and 60 (points) can be chosen as the visual threshold. In fact, such a division will provide only one erroneously clustered point. Fig. 3 shows 3 clusters according to real championship tables.

An analysis of Fig. 3 shows that there is a point in the 3rd cluster which is closer to the center and other points of the 1st cluster than to the cluster to which it really belongs.

Next it is necessary to approximate the statistics of Fig. 2 by GMMs with various parameters. Let use the following parameters:

1) The number of clusters k=1…5.

2) Covariance matrix (CM) which can be described by the following statements: diagonal/full and shared/unshared. The diagonal or full structure of CM characterizes the relationships between the parameters of one cluster, and the shared or unshared structure of CM characterizes the relationships between different classes. For the diagonal structure of the CM, the axes of the ellipse are parallel or perpendicular to the axes of abscissas and ordinates, and for the shared structure, the dimensions and orientation of all ellipses are the same.

3) The regularization parameter R = 0.01 or R = 0.1 is introduced to provide a positive determinant of the CM.

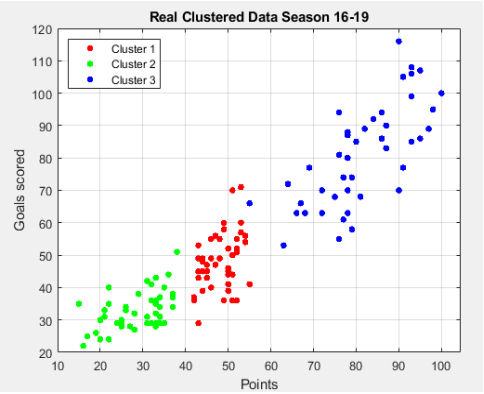
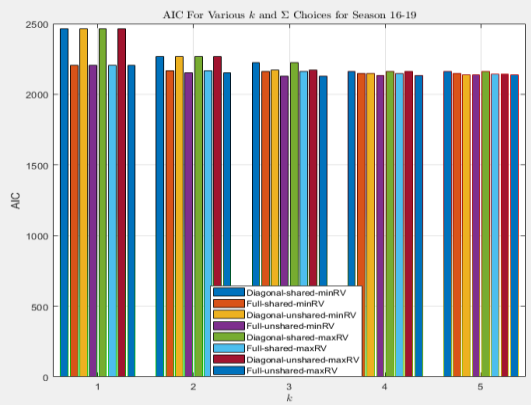
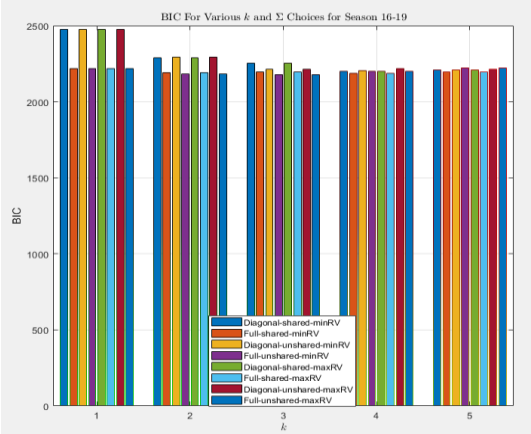


Fig. 3. Clustering of teams into classes according to the championship’s tables.



a)



b)

Fig. 4. AIC and BIC for various models.

By changing the above parameters, one can obtain several distributions of Gaussian mixtures, for which then it is possible to calculate the AIC and BIC coefficients presented. Fig. 4a and Fig. 4b shows AIC and BIC coefficients respectively for investigated football statistics with different parameters.

According to Fig. 4, the minimum values of AIC and BIC are provided by the model for k = 3 clusters, which has a full and unshared CM structure with a regularization parameter
R = 0.01. Fig. 5 shows the PDF of this model, and Fig. 6 shows the result of clustering using this model.

Comparison with the clustering presented in Fig. 3 shows that the clustering error was 1.48% or 2 incorrect assignment of teams to the group. Thus, high accuracy was obtained during clustering using the GMM.

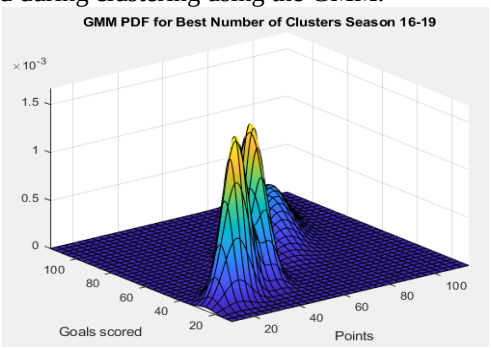


Fig. 5. PDF of the best approximation GMM.

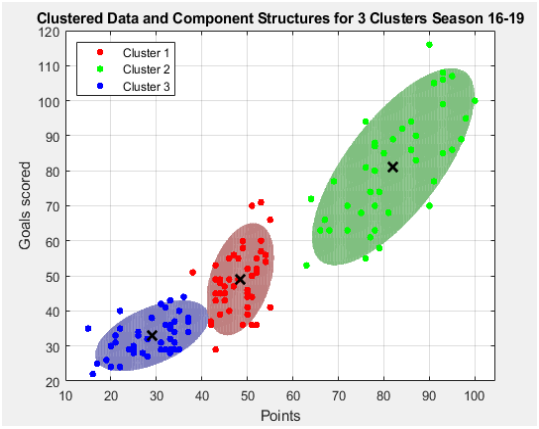


Fig. 6. Data clustering using GMM.

V. Clustering Using Neural Networks

In this section clustering based on neural networks is performed. Since the sample size is small, a feed forward network with the back propagation of error, consisting of 1 layer of 15 neurons, is used. For such a network, training based on data for the seasons 2016/2017 (train dataset) and 2017/2018 (validation dataset) is carried out. For test dataset statistics of season 2015/2016 is used. A pair of parameters, such as goals scored and points, is fed to the input of such a network, and the cluster number is obtained at the output. Fig. 7 shows the structure of the neural network, and Fig. 8 shows the learning process.

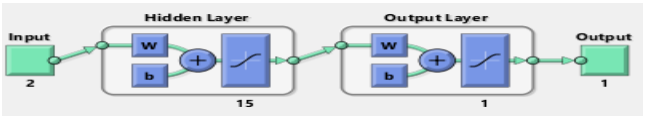


Fig. 7. Neural network structure.

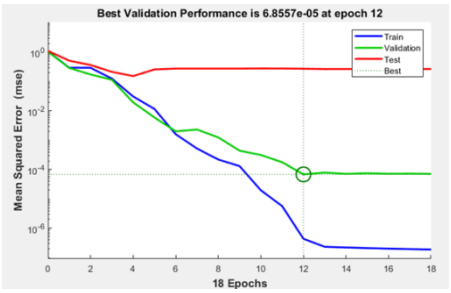
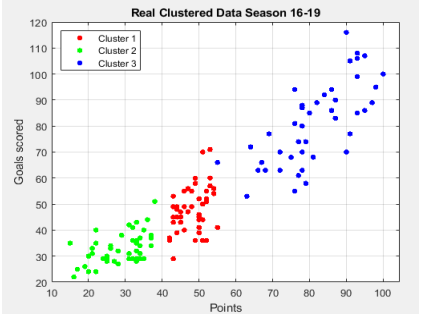


Fig. 8. Neural network training.

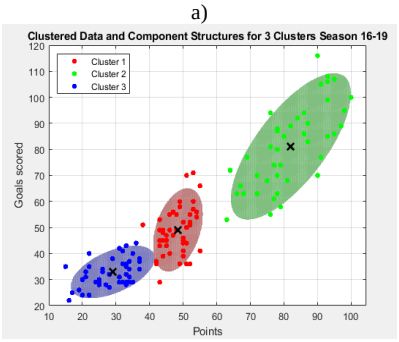
The analysis of Fig. 8 shows that the network converges quite quickly by the 12th epoch, achieving minimal error on the validation data. Fig. 9 shows the correct clustering (a), clustering using GMM (b) and clustering by the neural network (c).

So Fig. 9 shows that the neural network also provides satisfactory clustering, for which the error percentage is 1.48% or 2 objects (teams). Moreover, if the Gaussian mixture model mistakenly assigned one team from the group of outsiders (worst teams) to the middle-table teams and one team from the group of leaders (medal-position teams) to the middle-table teams, then the neural network incorrectly assigned two teams from the middle of the table (middle-table teams) to the teams of the upper part (medal position teams). It should also be noted that the use of deep learning

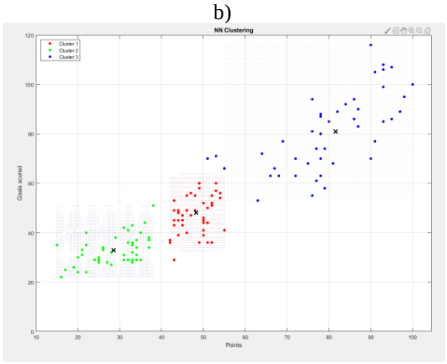
(increasing the number of layers to 5, and the number of neurons to 128) does not lead to improved results.



a)



b)



c)

Fig. 9. Comparison of clustering results.

VI. Conclusion

The paper studies data clustering algorithms using the example of clustering football statistics. The clustering algorithms based on the GMM and the neural network algorithm are considered. A comparative analysis of the accuracy of clustering showed that for the presented example, both algorithms provide the same result. Moreover, the clustering error is only 1.48%. However, the model of Gaussian mixtures looks preferable for several reasons. Firstly, it can determine the number of clusters by some information criterion. Secondly, when training the neural network, the data included in the data for which clustering was performed was used. Thirdly, in the neural network algorithm there were insignificant computational costs for training. The results obtained indicate that with the use of intelligent clustering algorithms it is possible to build a more adequate team rating, since, for example, the FIFA rating existing today does not reflect the actual strength of teams. Thus, the use of GMM for data mining is currently advisable. Moreover, in the future it is also planned to investigate the operation of the DGMM.

Acknowledgment

This work was supported by the RFBR and the Government of the Ulyanovsk Region Grant, Project No. 19-47-730011 and partly RFBR Grant, Project No. 19-29-09048.

References

[1] A.N. Danilov, N.A. Andriyanov and P.T. Azanov, “Ensuring the effectiveness of the taxi order service by mathematical modeling and machine learning,” Journal of Physics: Conference Series, vol. 1096, pp. 1-8, 2018. DOI:10.1088/1742-6596/1096/1/012188.

[2] N.A. Andriyanov and V.A. Sonin, “Using mathematical modeling of time series for forecasting taxi service orders amount,” CEUR Workshop Proceedings, vol. 2258, pp. 462-472, 2018.

[3] K.V. Vorontsov, “Clustering and multidimensional scaling algorithms,” Lecture course. Moscow State University, 2007. [Online]. URL: http://www.ccas.ru/voron/download/Clustering.pdf.

[4] I.A. Rytsarev, D.V. Kirsh and A.V. Kupriyanov, “Clustering media content from social networks using BigData technology,” Computer Optics, vol. 42, no. 5, pp. 921-927, 2018. DOI: 10.18287/2412-6179-2018-42-5-921-927.

[5] V.B. Nemirovsky and A.K. Stoyanov, “Clustering face images,” Computer Optics, vol. 41, no. 1, pp. 59-66, 2017. DOI: 10.18287/2412-6179-2017-41-1-59-66.

[6] Y. Tarabalka, J.A. Benediktsson and J. Chanussot, “Spectral–spatial classification of hyperspectral imagery based on partitional clustering techniques,” IEEE Transactions on Geoscience and Remote Sensing, vol. 47, no. 8, pp. 2973-2987, 2009.

[7] N.A. Andriyanov and V.E. Dementiev, “Developing and studying the algorithm for segmentation of simple images using detectors based on doubly stochastic random fields,” Pattern Recognition and Image Analysis, vol. 29, no. 1, pp. 1-9, 2019. DOI: 10.1134/S105466181901005X

[8] N.A. Andriyanov and V.E. Dement'ev, “Application of mixed models of random fields for the segmentation of satellite images,” CEUR Workshop Proceedings, vol. 2210, pp. 219-226, 2018.

[9] K.K. Vasiliev, V.E. Dementyiev and N.A. Andriyanov, “Using probabilistic statistics to determine the parameters of doubly stochastic models based on autoregression with multiple roots,” Journal of Physics: Conference Series, vol. 1368, pp. 1-7, 2019. DOI: 10.1088/1742-6596/1368/3/032019.

[10] Y.A. Philin and A.A. Lependin, “Application of the Gaussian mixture model for speaker verification by arbitrary speech and counteracting spoofing attacks,” Multicore processors, parallel programming, FPGAs, signal processing systems, vol. 1, no. 6, pp. 64-66, 2016.

[11] C. Viroli and G.J. McLachlan, “Deep Gaussian mixture models,” Stat Comput, vol. 29, pp. 43-51, 2019. DOI:10.1007/s11222-017-9793-z.

[12] J. Guérin and B. Boots, “Improving Image Clustering With Multiple Pretrained CNN Feature Extractors,” ArXiv Preprint: 1807.07760.

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

[14] H.S. Bhat and N. Kumar, “On the derivation of the Bayesian Information Criterion” [Online]. URL: https://faculty.ucmerced.edu/hbhat/BICderivation.pdf.