Сравнительный анализ алгоритмов кластеризации данных футбольной статистики на основе глубокого обучения и модели гауссовых смесей

Н.А. Андриянов[1,2]

[1]АО «НПП «Исток» им. Шокина», Вокзальная 2а, к.1, Фрязино, Московская область, Россия, 141190
[2]Ульяновский государственный технический университет, Северный Венец 32, Ульяновск, Россия, 432027

Аннотация. В работе рассмотрена модель гауссовых смесей и возможности ее применения для решения задач кластеризации. Сначала рассмотрен случай, когда модель гауссовых смесей формируется таким образом, что все параметры модели известны. Далее рассмотрен случай, когда происходит аппроксимация нормально распределенных данных с помощью модели гауссовых смесей. Наконец, в статье приведено исследование точности кластеризации двумерных данных футбольной статистики команд-призеров, команд-середняков и команд-аутсайдеров топ-5 европейских чемпионатов. Результаты работы алгоритма на базе моделей гауссовых смесей сравниваются с результатами кластеризации, выполненной с помощью нейронных сетей.

1. Введение

Интеллектуальный анализ данных позволяет сегодня специалистам в различных областях значительно упростить свою работу. Например, на базе такого анализа могут быть отсеяны заведомо не платежеспособные клиенты, подающие заявку на кредит в банк, а также могут быть спрогнозированы данные числа заказов службы такси [1,2]. Действительно, цифровизация различных областей хозяйства и сфер государственной деятельности на постоянной основе обеспечивает значительные объемы информации. В связи с этим спектр задач, решаемых с помощью интеллектуального анализа данных, так широк.

Одной из наиболее интересных задач в данной области является задача кластеризации данных [3,4], которая остается актуальной и при обработке изображений и может быть тесно связана с задачей сегментации [5-9]. Однако в рамках задач именно анализа данных обычно можно выделить несколько групп объектов, описываемых несколькими параметрами. Простейший пример – это выборка студентов и студенток в группе, которые могут быть описаны с помощью их роста и веса. Каждый объект в выборке может быть отображен отдельной точкой на плоскости. В данном случае – на двумерной. Если, например, добавить третий параметр – длина волос, то решение задачи кластеризации упростится. При этом каждая группа объектов может быть представлена на плоскости некоторым эллипсоидом. Тогда решение кластеризации для конкретного нового объекта будет зависеть от того, к какому эллипсоиду ближе всего окажется точка, характеризующая этот объект. В данной работе будет

рассмотрен алгоритм кластеризации на основе моделей гауссовых смесей [10,11], потому что достаточно часто реальные данные удается хорошо аппроксимировать гауссовыми распределениями.

2. Краткая классификация алгоритмов кластеризации

Известные алгоритмы кластеризации [3] можно разделить по 2 принципам. Рассмотрим их основные особенности.

Во-первых, разбиение на кластеры может быть четким или нечетким. В случае четкого разделения каждому объекту в результате кластеризации ставится в соответствие строго одна группа. При нечеткой кластеризации обычно определяется набор значений, которые характеризуют принадлежность каждого объекта к каждой группе, т.е. такая кластеризация дает некоторое вероятностное распределение.

Во-вторых, кластерный анализ может быть плоским одноуровневым или иерархическим многоуровневым. В первом случае исходная выборка объектов по какому-то критерию разделяется на несколько классов в виде одного разбиения. Например, кластеризация тех же учащихся вузов только по полу. Если же далее уже разделять студентов и студенток по оценкам, сохраняя первый уровень, то получится более глубокая кластеризация, в частности исходный объект в выборке может быть охарактеризован не просто, как студент или студентка, а как студентка-отличница или студент-двоечник. Такое разделение обеспечивает иерархическая кластеризация. В частности, глубокая модель гауссовых смесей, рассмотренная в работе [11], хорошо справляется с целями иерархической кластеризации. При этом отнесение объекта к той или иной группе выполняется по принципу четкой кластеризации.

Наконец, все большую популярность в задачах кластеризации набирают нейронные сети [12]. В зависимости от параметров обучения и типа сетей можно получить различные модели для кластеризации. А сейчас перспективным направлением является глубокое обучение.

Таким образом, перед выбором алгоритма кластеризации необходимо предварительно сформулировать саму задачу кластеризации, а затем выполнить разбиение данных.

3. Модель гауссовых смесей

Рассмотрим применение плоской четкой кластеризации на примере анализа данных футбольной статистики Топ-5 европейских чемпионатов (Испания, Англия, Италия, Германия, Франция). Поскольку не ставится задача многоуровневой кластеризации, то будем использовать модель гауссовых смесей [10]. Это такая модель, плотность распределения вероятностей (ПРВ) которой описывается суммой ПРВ гауссовых распределений. Количество слагаемых в сумме – и есть число кластеров. Таким образом, суммарное распределение имеет несколько пиков, а для каждого из объектов считается близость к каждому пику и выбирается пик с наименьшим удалением. При этом каждый объект может характеризоваться не одним, а несколькими параметрами, для чего строятся многомерные ПРВ. Пример функции ПРВ модели гауссовой смеси трех распределений с двумя параметрами представлен на рисунке 1.

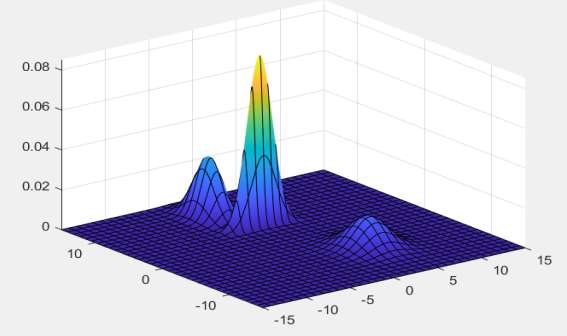


Рисунок 1. ПРВ гауссовой смеси.

Анализ рисунка 1 позволяет сделать вывод о том, что есть две группы объектов, для которых характерна большая дисперсия по одной из осей (ординат или абсцисс), и одна группа

с примерно одинаковой дисперсией по обеим осям. Кроме того, можно определить три характерных математических ожидания, которые соответствуют каждому из трех пиков.

Преимуществом применения модели гауссовой смеси является то, что при заданном числе объектов модель сама выполняет оценки составляющих распределений. Это позволяет выполнять аппроксимацию реальных данных с помощью такой модели. Однако даже в случае неизвестности числа кластеров заранее, можно построить несколько моделей смесей и выбрать оптимальную по некоторому критерию. Чаще всего используются критерий Акаике [13] и информационный критерий Байеса [14]. Применение данных критериев позволяет справиться с проблемой априорной неопределенности относительно числа классов.

4. Кластеризация с помощью модели гауссовых смесей

Рассмотрим пример применения модели гауссовых смесей при кластеризации команд, выступающих в европейских футбольных чемпионатах Англии, Испании, Германии, Италии и Франции. В исходную выборку будем включать 2 параметра: забитые голы и набранные очки. Однако для того, чтобы было удобнее проверять точность кластеризации, исключим из выборки некоторые команды. Таким образом, сделанное прореживание будет включать в себя 3 команды в верхней части турнирной таблицы (1 – 3 места), 3 в середине таблицы (9(8) – 11(10) места) и 3 команды в нижней части таблицы (18(16) – 20(18) места). Такое прореживание делается для каждого чемпионата. Кроме того, возьмем статистику по таким командам не только за последний сезон, но и за предыдущие 2 сезона. Это, с одной стороны, позволит увеличить информативность выборки, а с другой стороны, может привести также к увеличению аномальных точек («слишком удачный», «слишком неудачный» или «странный» сезон). На рисунке 2 представлена собранная статистика. По оси абсцисс отложены набранные очки, по оси ординат – забитые голы.

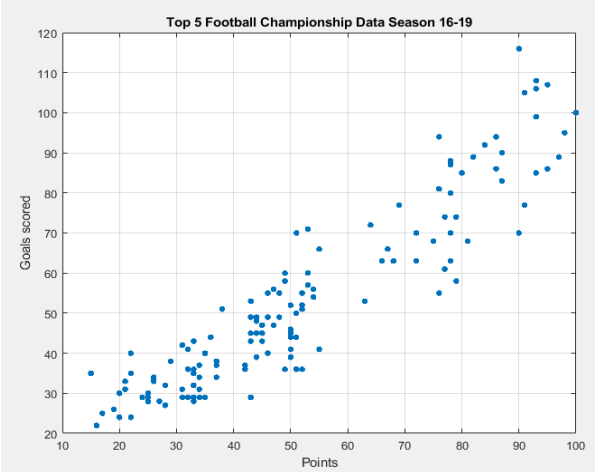


Рисунок 2. Статистика топ-5 футбольных чемпионатов за сезоны 2016-2017, 2017-2018 и 2018-2019.

Из рисунка 2 видно, что выбранные параметры имеют практически линейную зависимость и визуально наиболее предпочтительным разделением кажется просто деление прямыми по оси абсцисс (очков). При этом порогом могут быть цифры 40 и 60. На самом деле такое деление обеспечит лишь одну ошибочно кластеризованную точку. На рисунке 3 показаны 3 кластера в соответствии с реальными таблицами чемпионатов.

Анализ рисунка 3 показывает, что точка 3 кластера, находится ближе к центру и другим точкам 1 кластера, чем к своему настоящему. Аппроксимируем статистику рисунка 2 моделями гауссовых смесей с различными параметрами. К таким параметрам отнесем следующие:

1) Число кластеров k=1…5.

Ковариационная матрица: диагональная или полная и общая или необщая.

2) Диагональная или полная структура характеризует связи между параметрами одного кластера, а общая или необщая – между разными классами. Для диагональной структуры

ковариационной матрицы оси эллипса параллельны или перпендикулярны осям абсцисс и ординат, а для общей структуры размеры и ориентация всех эллипсов одинаковы.

3) Параметр регуляризации R=0.01 или R=0.1. Вводится для обеспечения положительного определителя ковариационной матрицы.

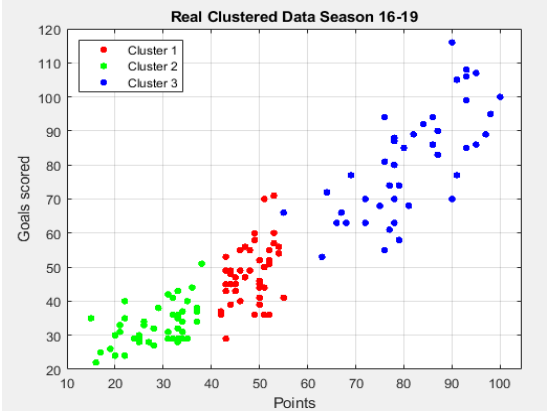
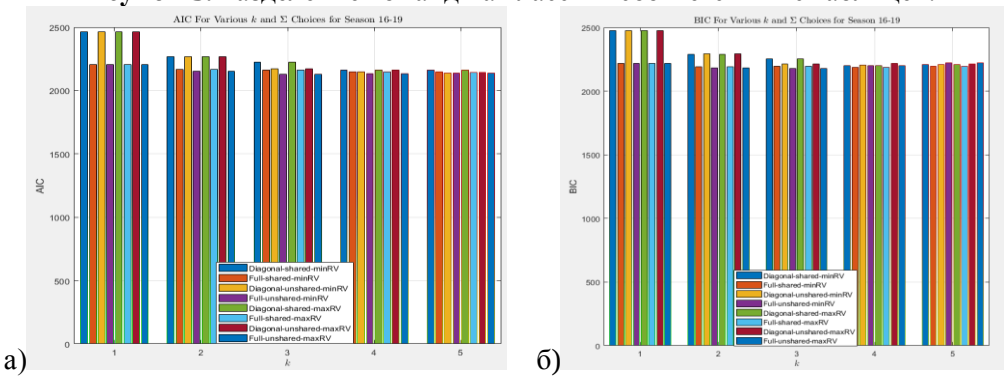


Рисунок 3. Разделение команд на классы в соответствии с таблицей.



а) б)

Рисунок 4. AIC и BIC для различных моделей

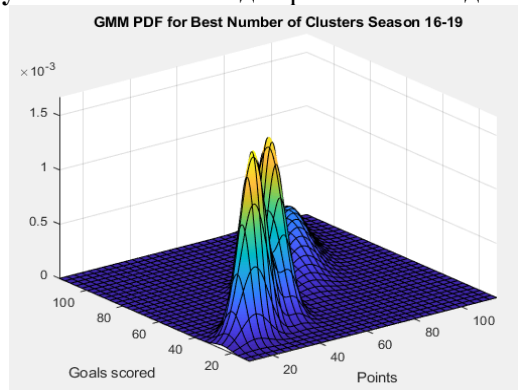


Рисунок 5. ПРВ наилучшей модели гауссовой смеси.

Меняя перечисленные параметры, получим несколько распределений гауссовых смесей, для которых затем посчитаем AIC и BIC коэффициенты, представленные на рисунке 4а и рисунке 4б соответственно.

Согласно рисунку 4, минимальные значения AIC и BIC обеспечивает модель для k=3 кластеров, имеющая полную и необщую ковариационную структуру с параметром регуляризации R=0.01. На рисунке 5 показана ПРВ данной модели, а на рисунке 6 – результат кластеризации с помощью данной модели.

Сравнение с кластеризацией, представленной на рисунке 3, показывает, что ошибка кластеризации составила 1.48% или 2 неправильных отнесения команд к группе. Таким

образом, получена высокая точность при кластеризации с использованием модели гауссовых смесей.

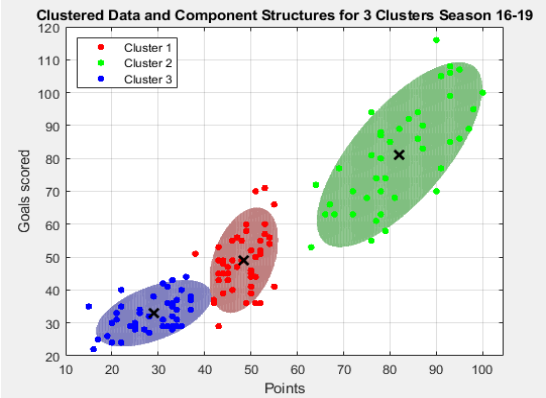


Рисунок 6. Кластеризация данных с помощью модели гауссовых смесей.

5. Кластеризация с помощью нейронных сетей

Выполним также кластеризацию на основе нейронных сетей. Поскольку объём выборки незначительный, будем использовать прямую сеть с обратным распространением ошибки, состоящую из 1 слоя в 15 нейронов. Для такой сети выполним обучение на основе данных за сезоны 2016 – 2017 и 2017 – 2018. На вход такой сети подается пара значений голы – очки, а на выходе получается номер кластера. На рисунке 7 приведена структура нейронной сети, а на рисунке 8 – процесс обучения.

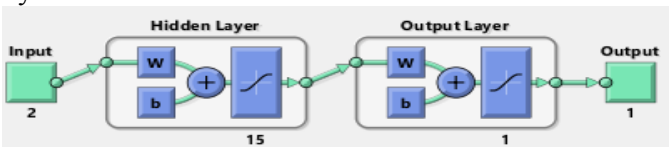


Рисунок 7. Структура нейронной сети.

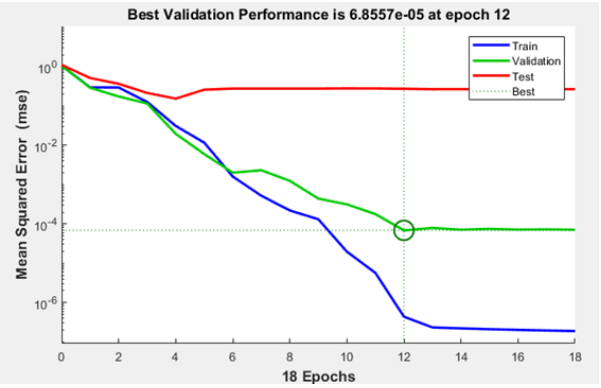
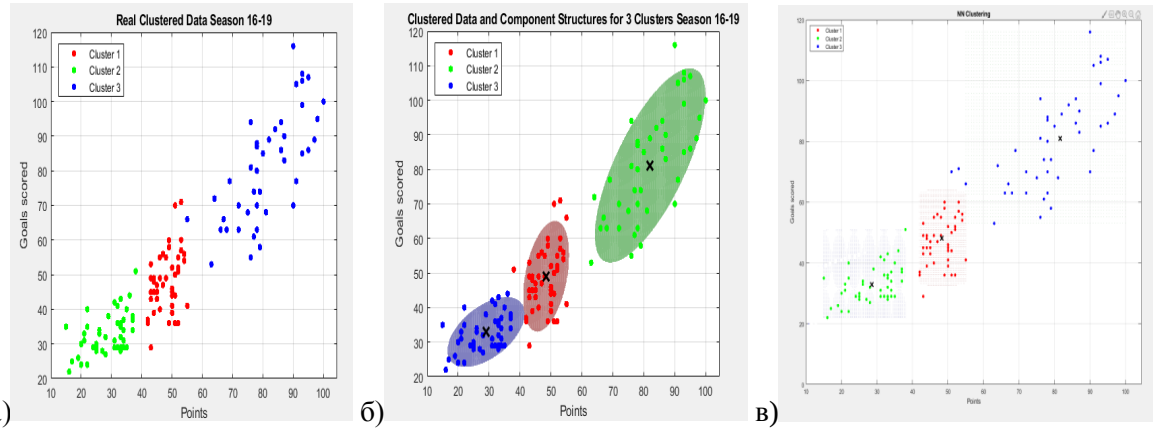


Рисунок 8. Обучение нейронной сети.

Из рисунка 8 видно, что сеть достаточно быстро сходится к 12 эпохе, достигая минимальной ошибки на проверочных данных. Рисунок 9 показывает правильную кластеризацию (а) и кластеризацию с помощью нейронной сети (б) и модели гауссовых смесей (в).

Из рисунка 9 видно, что нейронная сеть также обеспечивает удовлетворительную кластеризацию, для которой процент ошибки равен 1.48% или 2 объекта. При этом если модель гауссовой смеси ошибочно отнесла одну команду из группы аутсайдеров к середнякам и одну команду из группы лидеров к середнякам, то нейронная сеть неправильно отнесла две команды из средней части таблицы к командам верхней части. Следует также отметить, что применение глубокого обучения (увеличение числа слоев до 5, а числа нейронов до 128) не приводит к улучшению результатов.



а) б) в)

Рисунок 9. Сравнение результатов кластеризации.

6. Заключение

В статье проведено исследование алгоритмов кластеризации данных на примере кластеризации данных футбольной статистики. Рассмотрены алгоритмы кластеризации на основе модели гауссовой смеси и нейросетевой алгоритм. Сравнительный анализ точности кластеризации показал, что для представленного примера оба алгоритма обеспечивают одинаковый результат. При этом ошибка кластеризации составляет всего 1.48%. Однако модель гауссовых смесей смотрится предпочтительнее в силу ряда причин. Во-первых, она позволила сама определить число кластеров. Во-вторых, при обучении нейронной сети использовались данные, входящие в состав тех данных, для которых выполнялась кластеризация. В-третьих, в нейросетевом алгоритме были незначительные вычислительные затраты на проведение обучения. Таким образом, применение моделей гауссовых смесей для интеллектуального анализа данных в настоящее время целесообразно. Более того, в будущем планируется также исследовать работу глубокой модели гауссовых смесей.

7. Благодарности

Работа выполнена при поддержке Гранта РФФИ и Правительства Ульяновской области, Проект № 19-47-730011.

8. Литература

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

[2] Andriyanov, N.A. Using mathematical modeling of time series for forecasting taxi service orders amount / N.A. Andriyanov, V.A. Sonin // CEUR Workshop Proceedings. – 2018. – Vol. 2258. – P. 462-472.

[3] Воронцов, К.В. Алгоритмы кластеризации и многомерного шкалирования. Курс лекций – МГУ, 2007 [Электронный ресурс]. – Режим доступа: http://www.ccas.ru/voron/download/Clustering.pdf (21.12.2019).

[4] Рыцарев, И.А. Кластеризация медиа-контента из социальных сетей с использованием технологии BigData / И.А. Рыцарев, Д.В. Кирш, А.В. Куприянов // Компьютерная оптика. – 2018. – Т. 42, №5. – с. 921–927. DOI: 10.18287/2412-6179-2018-42-5- 921-927.

[5] Немировский, В.Б. Кластеризация изображений лиц / В.Б. Немировский, А.К. Стоянов // Компьютерная оптика. – 2017. – Т. 41, №1. – С. 59-66. DOI: 10.18287/2412-6179-2017-41-1-59-66.

[6] Tarabalka, Y. Spectral–spatial classification of hyperspectral imagery based on partitional clustering techniques / Y. Tarabalka, J.A. Benediktsson, J. Chanussot // IEEE Transactions on Geoscience and Remote Sensing. – 2009. – Vol. 47(8). – P. 2973-2987.

[7] Andriyanov, N.A. Developing and studying the algorithm for segmentation of simple images using detectors based on doubly stochastic random fields / N.A. Andriyanov, V.E. Dementiev // Pattern Recognition and Image Analysis. – 2019. – Vol. 29(1). – P. 1-9. DOI: 10.1134/S105466181901005X.
[8] Andriyanov, N.A. Application of mixed models of random fields for the segmentation of satellite images / N.A. Andriyanov, V.E. Dement'ev // CEUR Workshop Proceedings. – 2018. – Vol. 2210. – P. 219-226.
[9] Андриянов, Н.А. Сегментация изображений на основе оценивания параметров дважды стохастической модели // Современные проблемы проектирования, производства и эксплуатации радиотехнических систем. – 2017. – № 1-2(10). – С. 83-87.
[10] Филин, Я.А. Применение модели гауссовых смесей для верификации диктора по произвольной речи и противодействия спуфинг-атакам / Я.А. Филин, А.А. Лепендин // Многоядерные процессоры, параллельное программирование, ПЛИС, системы обработки сигналов. – 2016. – Т. 1, № 6. – С. 64-66.
[11] Viroli, C. Deep Gaussian mixture models / C. Viroli, G.J. McLachlan // Stat Comput. – 2019. – Vol. 29. – P. 43-51. DOI: 10.1007/s11222-017-9793-z.
[12] Guérin, J. Improving Image Clustering With Multiple Pretrained CNN Feature Extractors / J. Guérin, B. Boots [Электронный ресурс]. – Режим доступа: https://arxiv.org/abs/1807.07760 (21.12.2019).
[13] Akaike, H. A new look at the statistical model identification // IEEE Transactions on Automatic Control. – 1974. – Vol. 19. – P. 716-723
[14] Bhat, H.S. On the derivation of the Bayesian Information Criterion / H.S. Bhat, N. Kumar [Электронный ресурс]. – Режим доступа: https://faculty.ucmerced.edu/hbhat /BICderivation.pdf (21.12.2019).

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

N.A. Andriyanov[1,2]

[1]JSC "RPC "Istok" named after Shokin", Vokzalnaya street 2a, b.1, Fryazino, Moscow Region, Russia, 141190
[2]Ulyanovsk State Technical University, Severny Venets street 32, Ulyanovsk, Russia, 432027

Abstract. The paper considers the Gaussian mixture models and the possibilities of its application for solving clustering problems. First, we considered the case when the Gaussian mixture models are formed in such a way that all the parameters of the model are known. Next, we consider the case when the approximation of normally distributed data occurs using the Gaussian mixture model. Finally, the article presents a study of the accuracy of clustering two-dimensional data of football statistics of prize-winning teams, middle teams and outsider teams of the top 5 European championships. The results of the algorithm based on the Gaussian mixture models are compared with the results of clustering performed using neural networks.