УДК 519.237.8

МОДЕЛЬ НЕЧЕТКОГО БАЙЕСОВСКОГО КЛАССИФИКАТОРА ДЛЯ ОБРАБОТКИ ИНФОРМАЦИИ

Певнева А.Г., Обухов А.В., Зимовец А.И.

ФГБВОУ ВО «Военно-космическая Краснознамённая академия имени А.Ф. Можайского» Министерства обороны Российской Федерации, Санкт-Петербург, e-mail: vka@mil.ru

Представлена вероятностная модель процесса классификации на основе байесовского подхода. Совместное распределение данных по классам и прототипов классов является вероятностной моделью байесовского нечеткого классификатора. Неизвестные значения принадлежностей множества и прототипы множеств как центры классов трактуются как случайные величины. Далее выбирается некоторое эмпирическое распределение для каждой из них и устанавливается наиболее вероятное значение для сделанной выборки. Вероятностная модель в рамках байесовского подхода задается тремя распределениями: совместным распределением многомерных случайных величин, определяющих принадлежности каждого элемента данных какому-либо классу, распределением прототипов классов и априорным распределением прототипов по классам. Проводятся параллели данного подхода с процессом нечеткого вывода на основе продукций в рамках вероятностной логики. Данная модель может иметь более широкое применение по сравнению с традиционным нечетким подходом нечетких с-средних, из-за отсутствия ограничений на параметры кластеризации. Проведено сравнение модели с моделью разделения гауссовой смеси, так как методологические подходы к конструкции алгоритма нечеткой кластеризации и метода разделения гауссовой смеси кажутся внешне схожими. Установлены базовые различия между вероятностными моделями этих задач. Приведены некоторые численные результаты, подтверждающие работоспособность модели для стандартного и расширенного диапазона значений параметров.

Ключевые слова: байесовский подход, вероятностная модель, методы кластеризации, классификация, нечеткий вывод, метод с-средних

MODEL OF A FUZZY BAYESIAN CLASSIFIER FOR INFORMATION PROCESSING

Pevneva A.G., Obukhov A.V., Zimovets A.I.

Military Space Academy named after A.F. Mozhaiskiy of the Ministry of Defense of the Russian Federation, Saint Petersburg, e-mail: vka@mil.ru

A probabilistic model of the classification process based on the Bayesian approach is presented. The joint distribution of data by classes and class prototypes is a probabilistic model of a Bayesian fuzzy classifier. Unknown values of set accessories and prototypes of sets as class centers are treated as random variables. Next, some empirical distribution is selected for each of them and the most probable value for the sample is set. The probabilistic model within the Bayesian approach is defined by three distributions: the joint distribution of multidimensional random variables that determine whether each data element belongs to a class, the distribution of class prototypes and the a priori distribution of prototypes by classes. Parallels are drawn between this approach and the process of fuzzy inference based on products within the framework of probabilistic logic. his model may have a wider application compared to the traditional fuzzy approach of fuzzy c-means, due to the lack of restrictions on clustering parameters. The model is compared with the Gaussian mixture separation model, since the methodological approaches to the construction of the fuzzy clustering algorithm and the Gaussian mixture separation method seem superficially similar. The basic differences between the probabilistic models of these problems are established. Some numerical results confirming the operability of the model for the standard and extended range of parameter values are presented.

Keywords: bayesian approach, probabilistic model, clustering methods, classification, fuzzy inference, c-means method

Прикладной характер исследований в области классификации и кластеризации данных, широчайшая область задач, решение которых сводится к эвристическому выбору метода и подбору параметров, оставляет в тени теоретические обоснования этих методов. Между тем сильные расхождения в эффективности одного и того же алгоритма на разных наборах данных можно интерпретировать только на основании математически формализованной модели. Этим обусловлена актуальность работы по построению и исследованию теоретической модели процесса нечеткой классификации в рамках Байесовского подхода (НКБ).

Методологические подходы к конструкции алгоритма нечеткой кластеризации и вероятностного метода разделения Гауссовой смеси (англ. GMM) кажутся внешне схожими. В некоторых публикациях демонстрируется сходство получаемых результатов [1–3]. Однако в данных работах различие иллюстрируется на уровне методик применения, а не на уровне математически строгого определения вероятности.

Общность понятий «принадлежности» в методологии нечеткого логического вы-

вода и «вероятности» в задачах разделения распределений ощущается исследователями на интуитивном уровне. В [4] анализ сходства этих понятий подвергается развернутому анализу и даже используется неформальное определение: «нечеткость – это замаскированная вероятность», однако сходство это только внешнее, так как «принадлежность» связана с реализацией случайной величины посредством понятия лингвистической переменной, выбор которой субъективен.

В работе [5] объясняется возможность встраивания байесовского статистического подхода в систему нечеткого вывода в рамках вероятностной логики, построив тем самым логико-вероятностную модель классификатора. Работа [6] в рамках этого направления посвящена результирующему этапу процесса нечеткого вывода – фаззификации. В этих работах множество лингвистических переменных ассоциировано с множеством байесовских гипотез. Каждая гипотеза соответствует утверждению об определенном значении выходной лингвистической переменной из своего терм-множества. В [5] приводится также сравнение результатов дефаззификации на основе апостериорной байесовской вероятности с алгоритмом нечеткого вывода Мамдани.

Целью исследования, таким образом, является построение вероятностной модели нечеткой классификации, чтобы синтезировать алгоритм, объединяющий нечеткий и вероятностный подход в рамках теории вероятности. В данной работе описывается математическая вероятностная модель, которая служит основой расчетного алгоритма, приводятся результаты предварительного вычислительного эксперимента на модельных данных. Также с помощью вероятностной формализации модели нечеткого вывода устанавливаются глубокие различия между задачей нечеткой классификации и задачей разделения смеси распределений, так как им соответствуют разные вероятностные модели [6].

Материалы и методы исследования

Задача четкой классификации является задачей построения разбиения множества данных X. Нечеткую классификацию можно представить как задачу построения открытого покрытия $\bigcup_j U_j$ множества данных, так как каждый элемент множества U может содержаться сразу в нескольких множествах покрытия U_j с различными значениями функции принадлежности. Отличие от вероятностного подхода в задаче разделения распределений, состоит в том, что ни одно из этих множеств не является выборкой из генеральной совокупности, имеющей теоретическое распределение.

Каждому элементу покрытия U_j сопоставим так называемый центр y_j. Каждому элементу x_n исходного множества данных сопоставимо значение принадлежности $pr_{i,j}$ какому-либо элементу покрытия. В терминах нечеткой логики это значение функции принадлежности. Под принадлежностью понимается мера близости этого элемента к центру класса y_j. Каждый элемент покрытия соответствует множеству термов выходной логической переменной. Искомое открытое покрытие должно обладать свойством оптимальности, в том смысле, что общее взвешенное расстояние принадлежности каждой точки данных x_n к прототипу каждого множества y_i должно быть минимальным.

Таким образом, принадлежности трактуются как аргументы вероятностных логических функций в системе вероятностной логики. Нечеткие продукции в системе логического вывода преобразуются в вероятностные логические функции, а их значения понимаются как условные вероятности в определении апостериорного распределения. Понятие условной вероятности близко определению продукции, которая является импликативной логической функцией: при выполнении посылок продуктивного правила заключения отражают степень уверенности в том, что выходная лингвистическая переменная принимает то или иное значение из множества термов.

Формальное преобразование логической функции, представленной в нормальной форме, в функцию вероятностной логики следующее: логическим переменным ставятся в соответствие p_1, p_2, …, p_m элементарных событий, соответствующим атомарным формулам. При этом инверсиям переменных соответствуют значения $1 - p_i$. Конъюнкции и дизъюнкции соответствуют арифметическим операциям умножения и сложения.

На этапе дефаззификации [6] для каждой выходной лингвистической переменной на множестве гипотез о принятии значения из своего терм-множества по формуле Байеса вычисляется распределение апостериорных вероятностей.

Вероятностные модели байесовского нечеткого классификатора (НКБ) и гауссовой смеси

Неизвестные значения принадлежностей множества pr_{nj} и прототипы множеств y_j (центры классов) трактуются как случайные величины $X|U$ и Y соответ-

ственно. Далее выбирается некоторое эмпирическое распределение для каждой из них и устанавливается наиболее вероятное значение для сделанной выборки.

Пусть N – количество точек данных, D – размерность пространства данных, J – количество классов. Обозначим pr_{nj} принадлежность точки x_n классу j, m – параметр фаззификации [6–8], y_j прототип (центр) класса с номером j. Ясно, что значения наблюдаемых данных представляются матрицей размера $D \times N$, значения принадлежностей U представляются матрицей размера $J \times N$, а значения прототипов представляются матрицей размера $D \times J$.

Совместное распределение данных по классам и центров классов по множеству данных является вероятностной моделью БНК. Оно задается тремя распределениями: совместным распределением многомерных случайных величин $X|U$ и Y, априорным распределением прототипов и классов (вероятность того, что определенный прототип задает класс) и распределением прототипов по данным, т.е. вероятность того, что точка из множества данных является прототипом класса.

$$p(X,U,Y) = p(X|U,Y)\,\tilde{p}(U|Y)\,p(Y). \qquad (1)$$

Определим каждый из этих сомножителей.

$$P(X|U,Y) = \prod_{n=1}^{N} p_F\left(x_n|u_n,Y\right) = \qquad (2)$$

$$= \prod_{n=1}^{N} \frac{1}{Z\left(x_n,m,Y\right)} \prod_{j=1}^{J} F_G(x_n|\mu = y_c, \Lambda = pr_{n,j}^m).$$

Это фактически означает, что каждое наблюдаемое значение попадает в какую-либо нормально распределенную выборку с некоторой «уверенностью» $\Lambda = pr_{n,j}^m$ различной для каждого значения x_n. Количество таких выборок равно числу классов J. $Z(u_n, m, Y)$ – это нормирующий множитель, зависящий от параметров выборочного нормального распределения и параметра фаззификации.

Прототип полагается нормально распределенной величиной

$$P(Y) = \prod_{j=1}^{J} F_G(y_j \,|\, \mu_y, \Sigma_y) \qquad (3)$$

с параметрами

$$\mu_y = \frac{1}{N}\sum_{n=1}^{N} x_n,$$

$$\Sigma_y = \frac{1}{N}\sum_{n=1}^{N} (x_n - \mu_y)(x_n - \mu_y)^T.$$

Далее, априорная вероятность того, что определенный класс задается определенным прототипом, имеет распределение:

$$\tilde{P}(U|Y) = \prod_{n=1}^{N} \tilde{p}\left(u_n|Y\right) = \qquad (4)$$

$$= \prod_{n=1}^{N} Z\left(u_n,m,|Y\right)\left(\prod_{j=1}^{J} u_{n,j}^{-\frac{mD}{2}}\right) Dirichlet\left(u_n|\alpha\right).$$

Здесь первый множитель – произведение констант нормализации, при этом использование евклидовой нормы не является обязательным [6, 7].

Относительно второго произведения необходимо отметить, что оно относится к известному классу «неправильных приоров», так как распределение (4) не может быть нормализовано на интервале [0, 1], по крайней мере, для многих соответствующих значений m и D) [7].

Третий сомножитель в (4) – априорная вероятность прототипов классов, для ее представления используется распределение Дирихле, параметризованное вектором α.

Модель гауссовой смеси имеет принципиально иную вероятностную интерпретацию, чем предложенная модель НКБ.

Распределение гауссовой смеси представлено выражением

$$p(X|Z,Y) = \prod_{n=1}^{N}\prod_{j=1}^{J} F_G(x_n \,|\, \mu_j, \Sigma_j)^{z_{n,j}}, \qquad (5)$$

где $z_{n,j} \in \{0,1\}$, $\Sigma\, z_{n,\mathrm{j}} = 1$.

Сравнение вероятностных моделей

Переменные z_{nj}, для каждой точки x_n представляют вероятность того, что данная точка лежит в классе с прототипом в точке μ_i. При этом остальные значения этих переменных равны нулю, так как в основе разделения смесей лежит предположение, что каждая точка является элементом выборки только одной генеральной совокупности с некоторым распределением. Сравнивая вероятность в (5) с выражением (1), очевидно, что нечеткий классификатор Байеса использует произведение всех вероятностей компонентов, что означает учет возможности попадания значения x_{n} в различные классы. В этом состоит основное отличие рассматриваемых задач.

Так как обратная ковариация использует принадлежность как значение весовой переменной для каждого значения x_n, то значения, полученные по модели НКБ, независимы, но их распределение неравномерно на множестве данных. В отличие от этого, модель разделения смеси предполагает независимость, но значения, получаемые по (5) в результате итерационного алгоритма, называемого EM [8, 9], оказываются равномерно распределенными на исходном множестве.

Результаты вычислительного эксперимента на тестовом множестве точек данных

Целью проведенных расчетов является проверка работоспособности расчетной модели на различных значениях параметров алгоритма. На первом этапе эксперимент проводился на модельном наборе данных. В [7] подробно разобран алгоритм, который не использует итерационный пересчет значений на множестве данных, следовательно, расширяются возможности использования значения параметров метода, неразрешенных в методе с-средних. Реализация алгоритма представлена в [10]. Там же представлена реализация традиционного алгоритма с-средних.

Модельный набор данных состоял из двух двумерных нормальных выборок, содержащих 250 точек с различными параметрами. Параметры распределений: $\mu1 = (1, 2{,}5)$, и $\mu2 = (3, 1)$, ковариационные матрицы $\Sigma1$, $\Sigma2$ единичные. Параметр фазификации был взят равным $m = 1$, а параметр распределения Дирихле $\alpha = 1$.

При $m = 3$, $m = 5$ для тех же параметров нормальных выборок значения принадлежностей незначительно уменьшаются. Установка значения этого параметра должна связываться с параметрами нормального распределения. Для распределений с единичной ковариационной матрицей влияние изменения этого параметра не выявляется. Расчеты производились для единичного значения. Число пересчетов принадлежностей также не терпит критических изменений: использовалось 124 и 131 итерация для $m = 3$, $m = 5$ соответственно.

Было установлено, что реализация стандартного метода с-средних без реализации параллельных вычислений и реализация предложенной модели НКБ имеют схожие эмпирические показатели производительности, оба использовали чуть более 100 итераций.

Вид функции принадлежности при расчетах по методу НКБ для изменённых значений $m = -5$ и $m = -10$, показанный на рис. 2, заставляет сделать вывод о том, что отрицательные значения параметра, определяющего, собственно, количественную оценку нечеткости классификации позволяют в некотором смысле «усреднить» значения принадлежностей для каждой точки модельного набора. С увеличением по модулю этого параметра растет число требуемых итераций и классы становятся более «размытыми» даже на модельном наборе. Можно предположить, что варьирование этого параметра должно проводиться совместно с изменением параметра α априорного распределения Дирихле. Взаимовлияние параметров распределений составляет самостоятельное направление исследований в этом направлении.

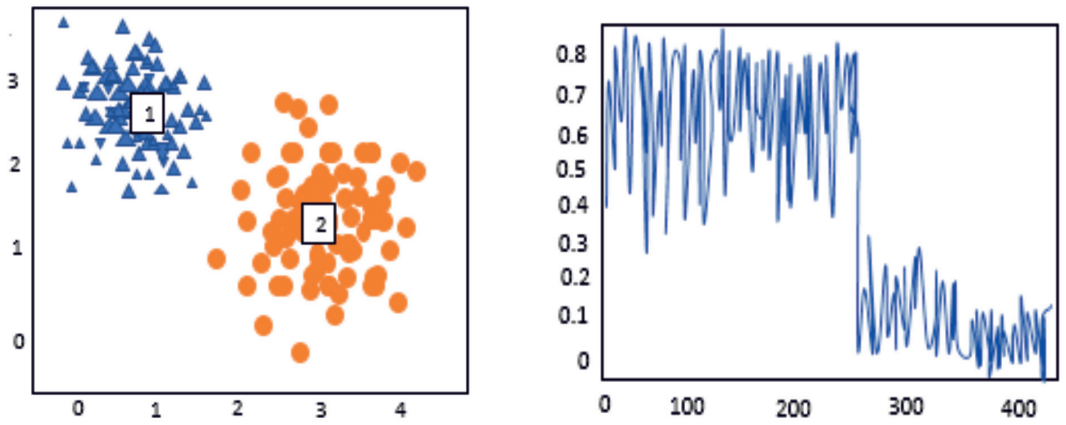

Рис. 1. Результаты кластеризации модельных данных

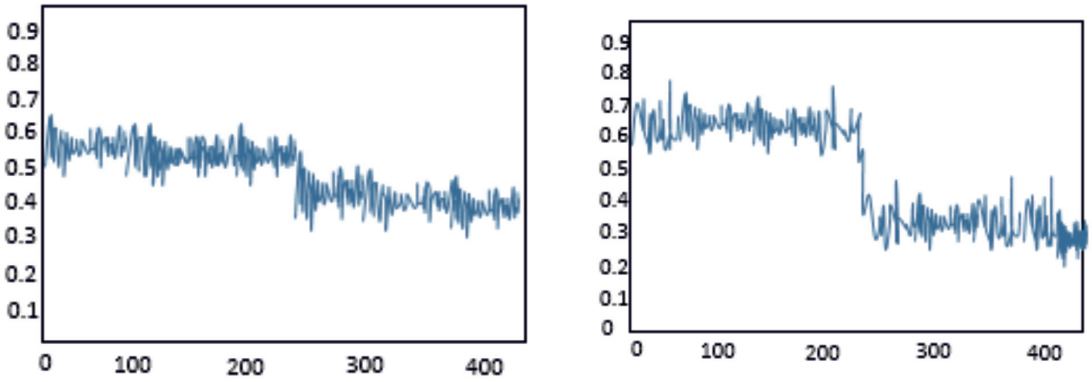

Рис. 2. Функции принадлежности для m = -10 (слева) , m = -5 (справа)

Заключение

Синтез байесовского подхода и нечеткого вывода является современным направлением в исследованиях решения задач классификации и кластеризации, этому посвящено немало публикаций в мировых научных изданиях.

Построенная вероятностная модель байесовской классификации имеет работоспособность не хуже традиционного метода с-средних на тестовых данных, кроме того, она имеет расширенные возможности в отношении значений параметров классификации.

Перспективные направления представленного исследования состоят прежде всего в применении построенной модели распределения в качестве инструмента дефаззификации в процессе нечеткого вывода. Представляет интерес теоретическая формализация связи построенной модели с алгоритмом нечеткого вывода, в особенности строгое представление связи между продукционными вероятностными функциями и параметрами распределения Дирихле при определении априорных вероятностей. Также возможен переход от классификации к более широкой задаче нечеткой кластеризации, так как прототипы классов в модели являются случайными величинами.

Другим направлением развития является детальное исследование влияния параметров на результаты алгоритма, в частности влияние параметров распределения Дирихле, используемого в качестве априорного распределения. Вычислительный эксперимент с изменением значения параметра *m*, определяющего «нечеткость» процесса, показал, что отрицательные значения *m* могут ухудшить работу алгоритма в целом, однако возможность гибкого изменения этого параметра полезна в усовершенствованном алгоритме для решения задачи кластеризации, когда число кластеров неизвестно заранее.

Проведенное сравнение модели задачи нечеткой байесовской классификации с задачей разделения смеси выявляет различие на уровне математической формализации. Однако построенная модель не лишена основного недостатка в сравнении с алгоритмом GMM: алгоритм на основе данной модели применяется лишь для «симметричных по координатам» классов данных. В дальнейших исследованиях целесообразно провести сравнение эффективности расчетных алгоритмов для обеих моделей с целью выявления влияния их различий на результаты классификации объектов с признаками различных типов: символьных, текстовых или числовых.

Список литературы

1. Кубарев А.И., Поддубный В.В. Байесовская классификация обучением на основе использования копула-функций // Информационные технологии и математическое моделирование. (ИТММ-2013): материалы XII Всероссийской научно-практической конференции с международным участием им. А.Ф. Терпугова (29–30 ноября 2013 г.). Томск: Изд-во Том. ун-та, 2013. Ч. 2. С. 126–130.

2. Ключихин Г.А. Вероятностный метод обработки информации в поле когнитивной науки. Модель наивного байесовского классификатора // Мягкие вычисления и измерения: материалы XXI международной конференции. СПб.: Изд-во Санкт-Петербургского электротехнического университета «ЛЭТИ», 2017. Т. 3. С. 22–26.

3. Кубарев А.И., Поддубный В.В. Адаптивная байесовская классификация объектов в метрическом пространстве // Новые информационные технологии в исследовании сложных структур: материалы Десятой российской конференции с международным участием. Томск: Издательство Томского государственного университета, 2014. С. 115–116.

4. Ланге Ф. Нечеткая логика. СПб.: Страта, 2018. 116 с.

5. Кожомбердиева Г.И. Байесовская логико-вероятностная модель нечеткого вывода // Международная конференция по мягким вычислениям и измерениям (Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина), 2015. С. 35–38.

6. Бураков Д.П. Этап дефаззификации нечеткого вывода: традиционный и байесовский логико-вероятностный подходы // Международная конференция по мягким вычислениям и измерениям. 2019. Т. 1. С. 39–42.

7. Taylor C. Glenn, Alina Zare, Paul D. Gader Bayesian Fuzzy Clustering. Transactions of fuzzy systems. 2017. V. 23. No. 5. P. 1222–1246.

8. Модель Гауссовой смеси [Электронный ресурс]. URL: http://espressocode.top/gaussian-mixture-model/ (дата обращения 15.11.2021).

9. Петренко А. Разбираем EM-algorithm на маленькие кирпичики. 2020 [Электронный ресурс]. URL://httpshabr.com/ru/post/501850/ (дата обращения 13.11.2021).

10. Реализация на Python алгоритмов GMM и EM. 2019 [Электронный ресурс]. URL: https://russianblogs.com/article/41131145085/ (дата обращения 13.11.2021).