Информационные процессы, Том 21, № 3, 2021, стр. 149–161

ТЕОРИЯ И МЕТОДЫ ОБРАБОТКИ ИНФОРМАЦИИ

О влиянии энтропии внутри класса на точность и полноту классификации

П.Р. Акимова, В.В. Мацулевич, Н.М. Замурий, М.О. Украинцев, С.А. Баринова, А.Б. Дубов, В.В. Аниканова, Ф.Р. Занченко, Ф.И. Иванов

Национальный исследовательский университет «Высшая школа экономики»

Поступила в редколлегию 06.07.2021

Аннотация—В данной работе исследуется зависимость между нормализованной энтропией (на один объект) внутри класса и качеством классификации. Исследуется влияние энтропии на такие основные метрики, как точность и полнота классификации для заданного класса. Для получения основных результатов используются различные методы классификации а также три основных подхода к восстановлению плотности распределения: непараметрический, параметрический и на основе восстановления смеси гауссовских распределений.

КЛЮЧЕВЫЕ СЛОВА: Энтропия, статистическая классификация, восстановление плотности распределения.

DOI: 10.53921/18195822_2021_21_3_149

1. ВВЕДЕНИЕ

В современном мире методы машинного обучения играют все большую роль в экономической и повседневной жизни. Данные алгоритмы выполняют множество задач: от построения моделей предсказания погоды [1], до распознавания речи [2] и обнаружения некоторых объектов (например, гос. номеров автомобилей) в видеопотоках [3]. Причем следует отметить, что зачастую алгоритмы машинного обучения способны решать такого рода задачи лучше человека. Машинное обучение используется в голосовых помощниках [4], системах навигации [5], медицине [6], в сфере безопасности [7] и многих других. Преимущества его использования настолько велики и очевидны, что человечеству крайне сложно представить жизнь без подобных алгоритмов. Данная область знаний, несомненно, продолжит свое интенсивное развитие и в будущем.

Следует отметить, что задачи в области знаний, связанной с машинным обучением и искусственным интеллектом, не ограничиваются построением эффективных алгоритмов классификации/кластеризации или восстановления регрессии. Ключевую роль играет также предобработка (препроцессинг) тех данных, которые впоследствие поступают на вход алгоритмам машинного обучения [8].

В частности, актуальной является задача проверки входных данных на способность алгоритмов на них обучиться. Для решения такой проблемы существует несколько способов, применимых для конкретных задач: метод главных компонент в задаче регрессии [9], метод отступа в задаче классификации [10] и многие другие.

В данной статье разработан метод оценки пригодности выборки для обучения на ней статистических алгоритмов классификации. В основе предложенного метода лежит наблюдение о связи нормализованной энтропии [11] (энтропии на один объект) внутри класса с показателями точности и полноты обучения для данного класса. Для вычисления нормализованной

энтропии решается задача восстановления плотности распределения признаковых описаний объектов класса. В работе рассматриваются различные алгоритмы восстановления плотности: параметрические [13], непараметрические [12], а также метод на основе разделения смеси гауссовых распределений [14]. Основной результат статьи состоит в том, что обнаружена взаимосвязь таких метрик, как точность и полнота классификации заданнного класса, с величиной нормализованной энтропии на классе. Причем данная взаимосвязь обнаруживается для широкого класса статистических алгоритмов классификации.

2. МЕТОДЫ ВОССТАНОВЛЕНИЯ ПЛОТНОСТИ

В данном разделе дано краткое описание известных методов восстановления плотности распределения $p(x)$ для некоторой наблюдаемой выборки $X = \{x_1, \ldots, x_n\}$, $x_i \in \mathbb{R}^m$.

2.1. Непараметрический метод

В одномерном непрерывном случае, если $P[a, b]$ – вероятностная мера отрезка $[a, b]$, то плотность $p(x)$ распределения случайной величины X, заданной на $[a, b]$, рассчитывается как:

$$p(x) = \lim_{h \to 0} \frac{1}{2h} P[x - h, x + h].$$

Эмпирическая оценка плотности по окну ширины h рассчитывается как:

$$\widehat{p}_h(x) = \frac{1}{2mh} \sum_{i=1}^{m} \frac{1}{2} \left[\frac{|x - x_i|}{h} < 1 \right].$$

Сформулируем теперь обобщенную оценку Парзена-Розенблатта [15]:

$$\widehat{p}_h(x) = \frac{1}{2mh} \sum_{i=1}^{m} K\left(\frac{x - x_i}{h} \right), \tag{1}$$

где $K(r)$ – ядро, (четная нормированная функция).

На одномерном случае можно наглядно продемонстрировать, что на восстановление плотности (вид функции $\widehat{p}_h(x)$) существенно влияет ширина окна h. Для демонстрации данной зависимости создадим выборку из 2000 элементов, распределенных по нормальному закону, и восстановим плотность с помощью формулы (1) для различных значений ширины окна:

Из Рис. 1 видно, что с увеличением ширины окна график плотности приближается к своему истинному виду, то есть к плотности нормального распределения. Обобщим теперь формулу (1) на многомерный случай. Если на числовом множестве X задана фукция расстояния $\rho(x, x')$, $x, x' \in X$, то парзеновская оценка плотности по каждому классу принимает вид:

$$\widehat{p}_{y,h}(x) = \frac{1}{l_y V(h)} \sum_{i : y_i = y} K\left(\frac{\rho(x, x_i)}{h} \right), \tag{2}$$

где

$$V(h) = \int_X K\left(\frac{\rho(x, x_i)}{h} \right) dx$$

– нормирующий множитель.

В формуле (2) в качестве функции расстояния в статье используется Евклидова функция $\rho(x, x') = ||x - x'||_{L_2}$. Для подбора ширины окна входная выборка случайным образом делится

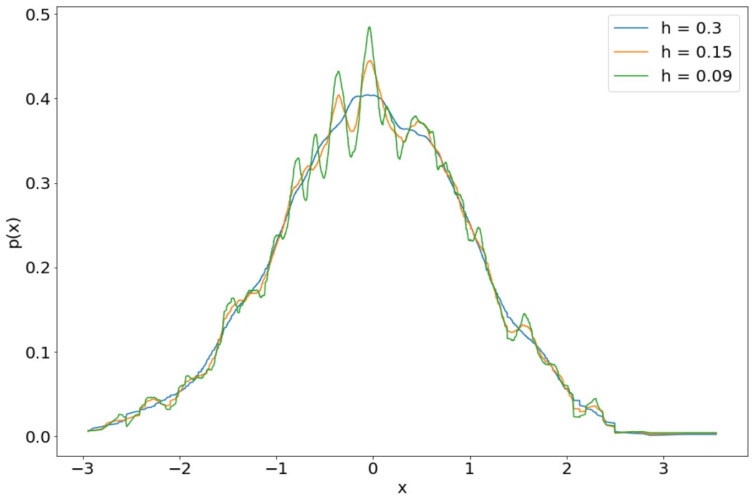

Рис. 1. Влияние ширины окна на восстановление плотности распределения.

на несколько частей, на каждой из которых восстанавливается плотность с различным значением ширины окна h. Выбирается ширина окна, соответствующая лучшему из полученных результатов.

Ниже приведены функции ядра, наиболее часто использующиеся на практике:

$$E(r) = \frac{3}{4}\left(1 - r^2\right)[|r| \le 1]$$

– ядро Епанечникова,

$$Q(r) = \frac{15}{16}\left(1 - r^2\right)^2[|r| \le 1]$$

– квартическое,

$$T(r) = (1 - |r|)\,[|r| \le 1]$$

– треугольное,

$$P(r) = \frac{1}{2}\,[|r| \le 1]$$

– прямоугольное,

$$G(r) = 2\pi^{-\frac{1}{2}} \exp\left(-\frac{1}{2}r^2\right)$$

– гауссовское.

Зависимость значения плотности от выбора ядра показана на рисунке ниже:

На качество восстановления плотности выбор ядра влияет незначительно. В дальнейшем будет использоваться ядро Епанечникова.

2.2. Параметрическое восстановление плотности

Плотность распределения на классе y при параметрическом восстановлении плотности при предположении о нормальности распределения объектоов класса6 определяется формулой:

$$p_y(x) = \frac{e^{-1/2(x-\mu_y)^T \Sigma_y^{-1}(x-\mu_y)}}{\sqrt{(2\pi)^n \det \Sigma_y}}, \tag{3}$$

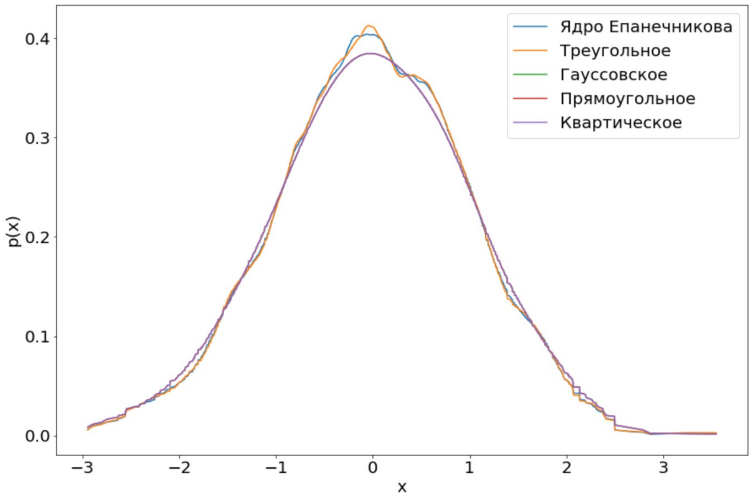

Рис. 2. Функции ядра.

где μ_y — центр, а Σ_y — матрица ковариаций, выборочные оценки которых определяются как:

$$\widehat{\mu}_y = \frac{1}{G_y}\sum g_i x_i,$$

$$\widehat{\Sigma}_y = \frac{1}{G_y}\sum g_i (x_i - \widehat{\mu}_y)(x_i - \widehat{\mu}_y)^T,$$

где

$$G_y = \sum g_i.$$

2.3. Восстановление смеси распределений с помощью EM алгоритма

В некоторых случаях следует моделировать плотность вероятности $p(x)$ в виде смеси гауссовых распределений:

$$p(x) = \sum_{j=1}^{k} w_j p_j(x), \quad \sum_{j=1}^{k} w_j = 1, \quad w_j \geq 0,$$

где $p_j(x) = \phi(x;\theta_j)$ – гауссова функция плотности j-й компоненты смеси (3); w_j – её априорная вероятность; k – число компонент смеси.

Задача восстановления плотности в данном случае заключается в оценке параметра k и вектора параметров $\Theta = (w_1, \ldots, w_k, \theta_1, \ldots, \theta_k)$, где θ_j – параметры j-й компоненты смеси – вектора средних значений $\mu_j = (\mu_{j1}, \ldots \mu_{jn})^T$ и ковариационной матрицы компоненты Σ_j размера $n \times n$: $\theta_j = (\mu_j, \Sigma_j)$.

Данную задачу можно решать различными методами, в частности методом максимального правдоподобия. Чтобы максимизировать логарифмическую функция правдоподобия

$$L(\theta) = \sum_{i=1}^{m} \log \sum_{j=1}^{k} w_j p_j(x_i),$$

требуется решить многоэкстремальную задачу с $3k - 1$ переменных. Известен итерационный алгоритм Expectation–Maximization (EM) [25] с последовательным добавлением компонент, который существенно упрощает задачу оптимизации $L(\theta)$ за счет введения матрицы G скрытых переменных.

Определим алгоритм EM более детально.

Пусть $q_{ij} = P(\theta_j|x_i)$ – вероятность того, что объект x_i получен из j-й компоненты смеси. Т. к. каждый объект принадлежит одной и только одной компоненте смеси, то $\sum_{j=1}^{k} q_{ij} = 1$ для любого $i = 1, .., m$. Воспользуемся формулой Байеса и получим, что:

$$g_{ij} = \frac{w_j p_j(x_i)}{\sum_{s=1}^{k} w_s p_s(x_i)}. \tag{4}$$

Считая известной матрицу скрытых переменных G, максимизируем логарифмическую функцию правдоподобия

$$L(\theta) = \sum_{i=1}^{m} \log \sum_{j=1}^{k} w_j p_j(x_i) \longrightarrow \max_{\theta}.$$

При этом мы получаем k независимых задач максимизаций правдоподобия для каждой компоненты смеси:

$$\sum_{i=1}^{m} g_{ij} \log(\phi(x_i; \theta)) \longrightarrow \max_{\theta_j},$$

где m – число элементов в выборке. Тогда

$$w_j = \frac{1}{m} \sum_{i=1}^{m} q_{ij}, j = 1, \ldots, k.$$

Функцию правдоподобия максимизируется путем чередования обновления скрытых переменных (М-шаг) и обновления параметров модели (Е-шаг) до сходимости. Изначально матрица G задается либо произвольно, либо в соответствие с некоторыми эвристиками [16]

Для реализации классического EM-алгоритма требуется задать количество компонент смеси k. Поэтому в данной работе используется обобщение этого алгоритма - EM-алгоритм с последовательным добавлением компонент [17]. Начиная с $k = 1$, выполняются вышеописанные действия последовательно для каждого k, после чего находится количество элементов, значение функции плотности которых меньше пороговой величины и, если это количество меньше некоторого заданного критического значения m_0, то алгоритм останавливаемся на данном k. Пороговая плотность определяется как ϕ_{max}/R, где R - гиперпараметр. Таким образом, EM алгоритм с последовательным добавлением компонент требует задания двух параметров — R и m_0. Их целесообразно подбирать для каждой выборки. При этом обычно $R \in [2; 10], m_0 \geq 10$. Покажем, как меняются оценки распределения при разлинчных R:

Можно заметить, что чем меньше параметр R, тем хуже обобщающая способность алгоритма. Также результат чувствителен к начальному приближению матрицы G.

В случае $x \in \mathbb{R}^2$, то для наглядности значение плотности вероятности каждой точки пространства можно выделить цветом (чем интенсивнее красный – тем больше значение плотности) и проиллюстрировать модель плотности:

3. ЭНТРОПИЯ НА МНОЖЕСТВЕ ОБЪЕКТОВ КЛАССА

3.1. Определение энтропии

Восстановление плотности распределения требуется для того, чтобы рассчитать энтропию внутри класса. Сформулируем понятие энтропии.

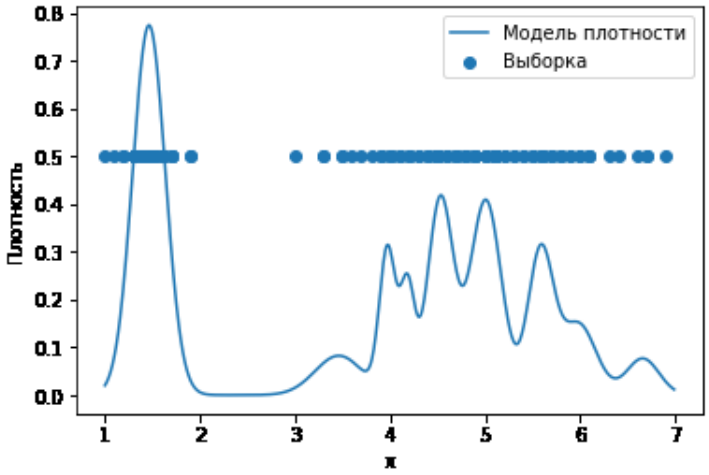

Рис. 3. Восстановление плотности распределения при $R = 1.5$

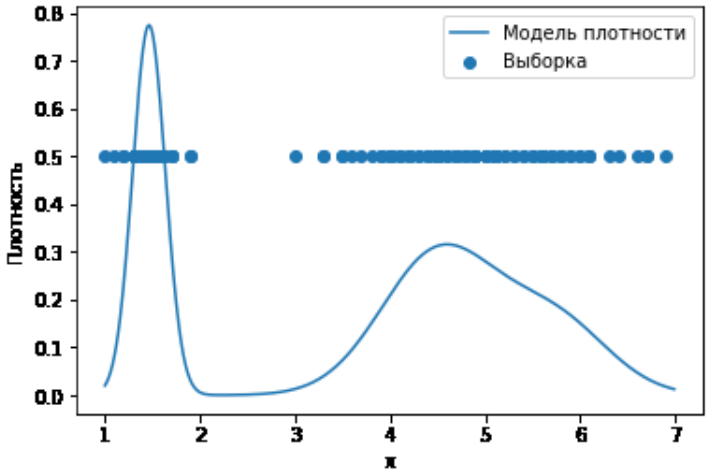

Рис. 4. Восстановление плотности распределения при $R = 4$

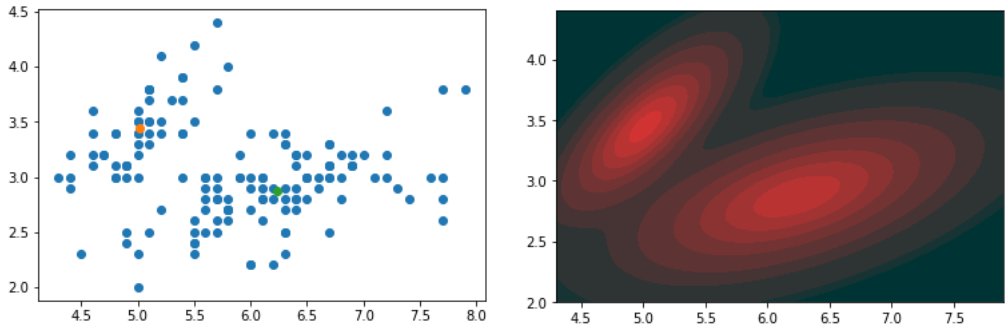

Рис. 5. Восстановление плотности распределения двумерной выборки в виде смеси из двух компонент

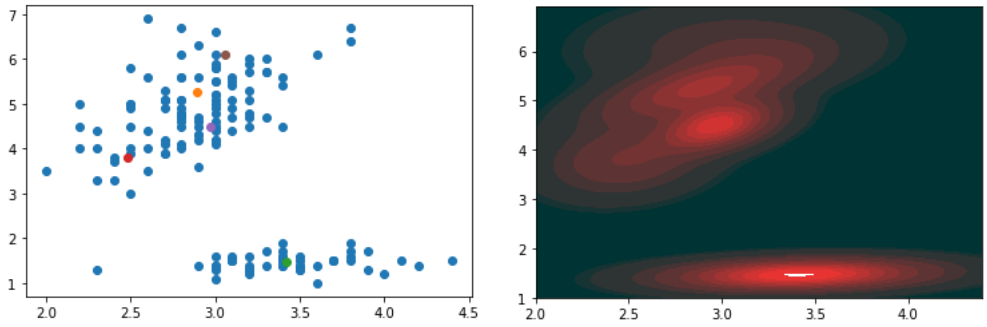

Рис. 6. Восстановление плотности распределения двумерной выборки в виде смеси из пяти компонент

Определение 1. Дифференциальная энтропия – функционал, заданный на множестве абсолютно непрерывных распределений вероятностей. Для непрерывной случайной величины ξ, распределенной на $X \subseteq R^n (n < \infty)$, дифференциальная энтропия вычисляется по формуле:

$$H(\xi) = -\int_X f(x) \log f(x) dx,$$

где $f(x)$ – плотность распределения случайной величины.

Определение 2. Для дискретной случайной величины ξ, распределенной по закону $p_i = p(x_i), (i = 1, \ldots, n)$, энтропия вычисляется по формуле:

$$H(x) = -\sum_{i=1}^{n} p(x_i) \log p(x_i),$$

где $p(x_i)$ – плотность распределения случайной величины.

Введем также понятие нормализованной энтропии (энтропия на объект множества $\mathcal{X}$) для дискретной случайной величины X, заданной на $\mathcal{X}$:

$$H(x) = -\frac{1}{|\mathcal{X}|} \sum_{i:y_i=y} p(x_i) \log p(x_i), \tag{5}$$

где $p(x_i)$ – плотность распределения случайной величины X.

Для исследования влияния энтропии на качество классификации, необходимо рассчитать величину энтропии для каждого класса в обучающей выборке.

В качестве примера далее рассматривается датасет качества вина (Wine Quality Data Set) [18]. В нем приведена информация о более чем 1500 измерений содержания в вине различных веществ и показателей кислотности. Целевая переменная – показатель качества вина (от 3 до 8). Данная выборка позволяет проводить исследования алгоритмов мультиклассовой классификации. Стоит упомянуть, что объектов классов 3, 4 и 8 довольно мало по сравнению с классами 5 или 6 (не более 10%). Классы несбалансированны, однородность данных многочисленных классов довольно высокая.

Для восстановления плотности распределения, необходимой для применения (5), мы применяем описанные выше методы восстановления на основе параметрического, непараметрического подходов, а также на основе смеси гауссовских распределений.

3.2. Непараметрическое восстановление плотности

Для непараметрического метода восстановления плотности использовалась формула (1) для восстановления плотности и (5) для оценки энтропии. Полученные значения нормализованной энтропии (в долях от максимального значения) по каждому классу представлены на рисунке ниже:

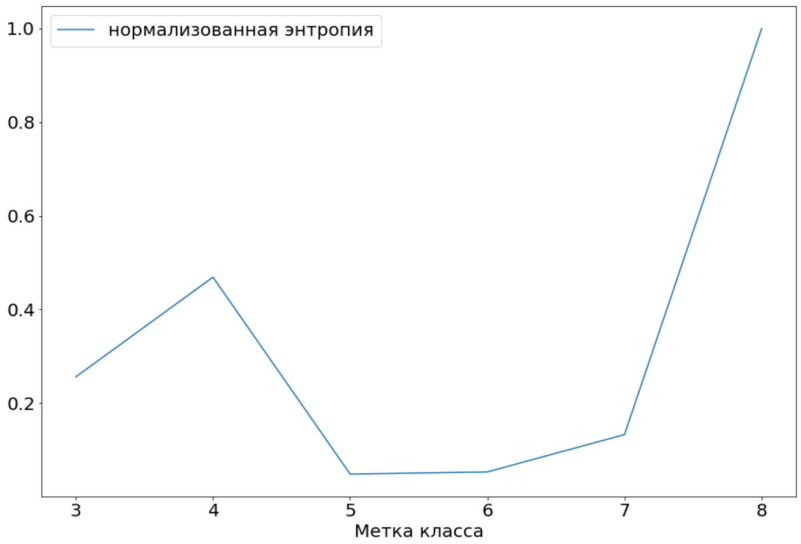

Рис. 7. Оценка нормализованной энтропии при оценке плотности методом Парзена-Розенблатта.

3.3. Параметрическое восстановление плотности

Известно, что энтропия случайной величины, заданной многомерным нормальным распределением, определяется как:

$$H(Y) = \frac{1}{2}\ln[(2\pi e)^n|\widehat{\Sigma}_y|],$$

где $\widehat{\Sigma}_y$ – матрица ковариаций, n – размер выборки.

Как видно из Рис. 8, характер зависимости нормализованной энтропии от метки класса коррелирует с результатами, полученными для непараметрического восстановления плотности методом окна Парзена (оценка Парзена–Розенблатта), хотя полученные оценки отличаются в абсолютных значениях.

3.4. Оценка энтропии смеси гауссовых распределений

Для смеси гауссовых распределений аналитической формулы для вычисления энтропии не существует [26]. Поэтому, были использованы аналитические формулы для нижней и верхней оценок энтропии $H_l \leq H(X) \leq H_u$ [27]:

$$\begin{aligned} H_l &= -\sum_{i=1}^{k} w_i \log \sum_{j=1}^{k} w_j \phi(\mu_i; \mu_j, \Sigma_i + \Sigma_j), \\ H_u &= \sum_{i=1}^{k} w_i \left(-\log w_i + \frac{1}{2}\log((2\pi e)^n \det \Sigma_i) \right). \end{aligned} \tag{6}$$

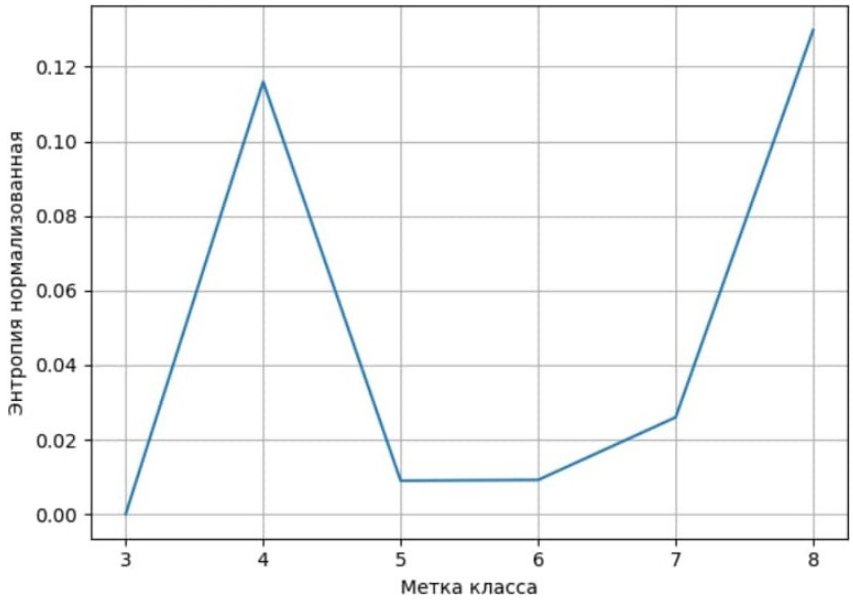

Рис. 8. Оценка нормализованной энтропии при восстановлении плотности параметрическим методом при предположении о нормальности распределения внутри класса.

Результаты экспериментов показали, что по крайней мере для датасета [18] значения нижней и верхней оценок практически не отличались, поэтому мы будем использовать нижнюю оценку из (6) для приближенного подсчёта энтропии смеси гауссовых распределений.

4. КЛАССИФИКАЦИЯ

При классификации данных из [18] использовались следующие алгоритмы машинного обучения: логистическая регрессия, метод опорных векторов, модель на основе гауссовой смеси, классификатор на основе стохастического градиентного спуска.

- Логистическая регрессия [19]. Данный алгоритм стремится предсказать вероятность принадлежности объекта некоторому классу, минимизируя верхнюю оценку, в качестве которой могут выступать логистическая функция потерь или сигмоида.
- Метод опорных векторов (SVC) [20]. Алгоритм стремится разделить все пространство объектов на классы с помощью гиперплоскости, а затем максимизировать отступ объектов от нее.
- Модель смеси гауссовских распределений (GMM) [21] — это вероятностная модель, которая предполагает, что все точки данных генерируются из смеси конечного числа нормальных распределений с неизвестными параметрами. Для максимизации вероятности и схождению к локальному минимуму используется EM алгоритм.
- Классификатор на основе стохастического градиентного спуска (SGD) [22] — это простой, но очень эффективный подход к оптимизации линейных классификаторов под выпуклые функции потерь.
- Классификатор на основе дерева решений [29] — непараметрический классификатор цель которого состоит в том, чтобы создать модель, которая предсказывает значение целевой переменной, изучая простые правила принятия решений, выведенные из признаковых описаний данных. Дерево можно рассматривать как кусочно-постоянное приближение.
- Классификатор по ближайшим соседям [23] — метрический алгоритм для автоматической классификации, при котором объект присваивается тому классу, который является наиболее распространённым среди k соседей данного элемента, классы которых уже известны.

Для оценки качества классификации использовались метрики точности (Precision) и полноты (Recall).

5. ВЗАИМОСВЯЗЬ ТОЧНОСТИ, ПОЛНОТЫ И НОРМАЛИЗОВАННОЙ ЭНТРОПИИ

5.1. Непараметрическое восстановление плотности

Покажем, что по значению нормализованной энтропии можно предсказать точность и полноту классификации по каждому из классов датасета. На Рис. 9 приведены графики зависимости нормализованной энтропии и метрик точности и полноты на каждом классе:

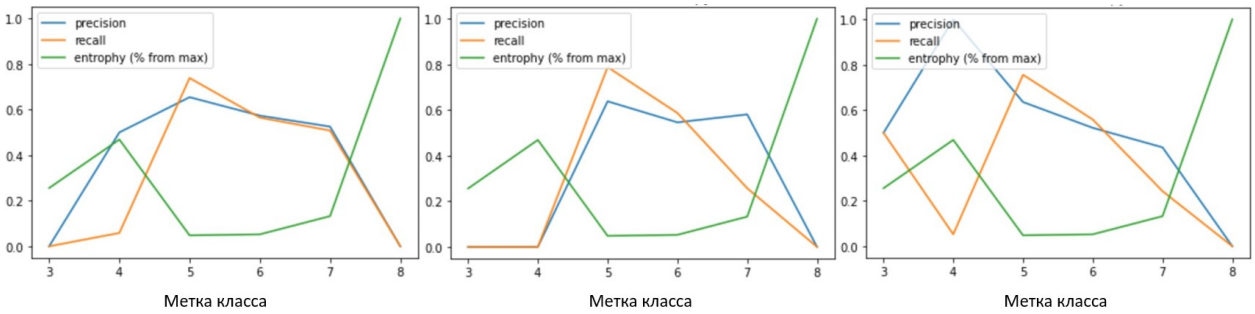

Рис. 9. Взаимосвязь энтропии, точности и полноты для парзеновского метода, SVC и логистической регрессии при непараметрическом восстановлении плотности.

Ввиду того, что нормализованная энтропия указывает на величину неопределенности внутри класса, то классификаторы лучше справляются с задачей в случае, если эта неопределенность мала. Именно такие результаты отражены на Рис. 9. Алгоритмы смогли достаточно обучиться, чтобы разделить классы 5, 6 и 7 с достаточно высокими показателями точности и полноты. При этом доля от максимальной нормализованной энтропии на этих классах принимает значения не более 0.15. Результаты показывают, что нормализованная энтропия, основанная на непараметрическом восстановлении плотности, может быть полезной предварительной оценкой для классификаторов.

5.2. Параметрическое восстановление плотности

Покажем, что по значению нормализованной энтропии, рассчитанной на основе параметрических методов восстановления плотности при предположении о нормальности распределения внутри классов, можно предсказать точность и полноту классификации по каждому из классов датасета. На Рис. 10 приведены графики зависимости нормализованной энтропии и метрик точности и полноты на каждом классе.

В данном случае в логистической регрессии оптимизировалась другая функция потерь, поэтому результат классификации немного отличается от предыдущих графиков (не распознаются классы 3 и 4). Прослеживается та же закономерность, что и при непараметрическом восстановлении плотности: энтропия на классах 5, 6, 7 небольшая, и моделям удается неплохо их отделять от других классов и друг от друга, с классами 4 и 8 ситуация обратная — большая энтропия на данных классах приводит к тому, что алгоритмы не способны их отделить. Однако все же заметны отличия от непараметрических методов: у объектов класса 3 нормализованная энтропия равна нулю, что выбивается из общей тенденции. Величина энтропии очень близка к максимальному значению на классе 4, хотя некоторые классификаторы все же способны его предсказывать.

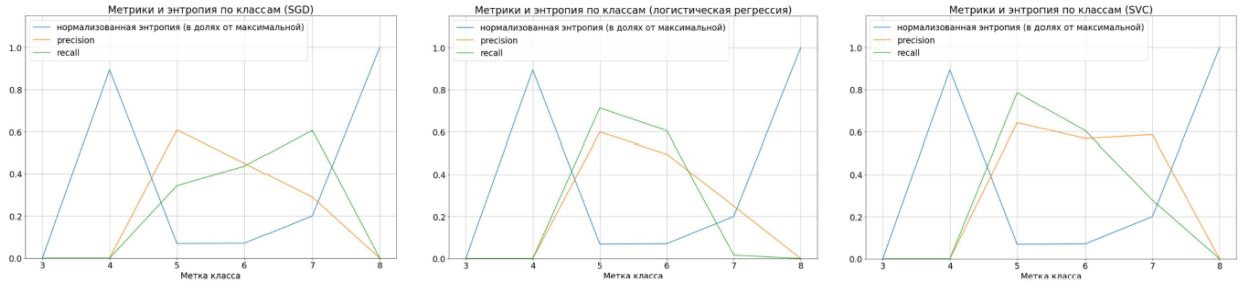

Рис. 10. Взаимосвязь энтропии, точности и полноты для SGD, логистической регрессии и SVC при параметрическом восстановлении плотности.

5.3. Восстановление смеси распределений с помощью EM алгоритма

Результаты классификации, полученные при восстановлении плотности из предположения о модели смеси гауссовских распределений, представлены на Рис. 11:

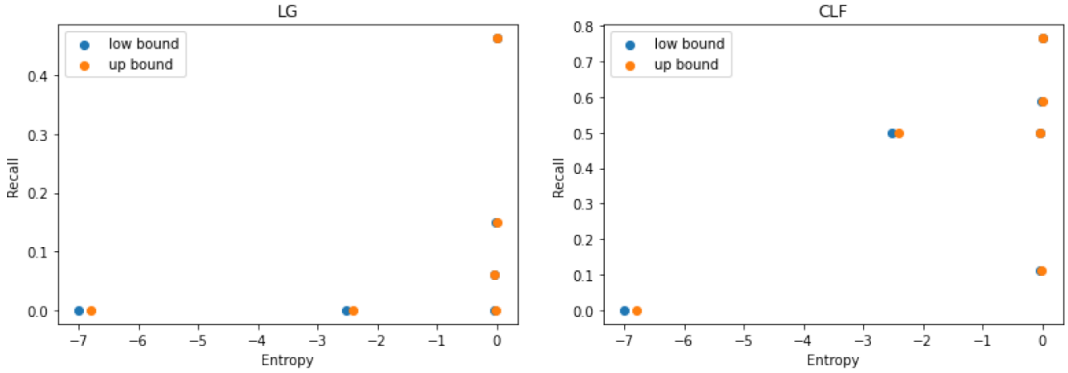

Рис. 11. Зависимость полноты от нормализованной внутриклассовой энтропии для логистической регрессии и решающего дерева

Точность и полнота каждого из классификаторов убывает с увеличением нормализованной энтропии класса. Чтобы проверить зависимость, требуется проанализировать изменение взаимосвязи при изменении содержания классов, а также провести тесты на другой выборке.

Возьмем класс, на котором presicion и recall были наибольшими для всех классификаторов, и будем удалять из него объеты, плотность вероятности для которых меньше некоторого значения p_{thr}. График изменения при этом нормализованной внутриклассовой энтропии представлен на Рис. 12:

Из Рис. 12 видно, что с удалением объектов с наименьшей вероятностью точность и полнота возрастают до некоторого значения, а далее удаление объектов только уменьшает качество классификации. Аналогичные результаты были получены и для остальных классов. Эту процедуру можно интерпретировать как удаление шумовых объектов. Отсюда следует, что нормализованная внутриклассовая энтропия не должна быть слишком большой, чтобы не брать во внимание шумовые объекты и не должна быть слишком маленькой, чтобы сохранить разброс выборки.

6. ЗАКЛЮЧЕНИЕ

В данной работе исследована зависимость между нормализованной энтропией (на один объект) внутри класса и точностью/полнотой классификации. Показано, что для широкого класса

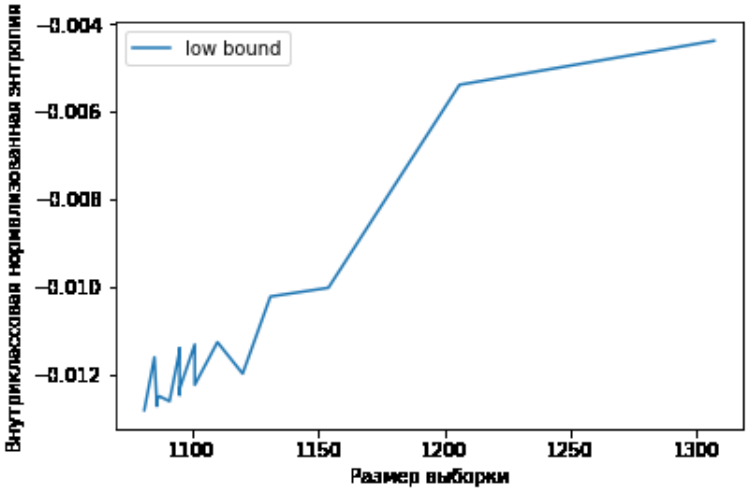

Рис. 12. Зависимость нормализованной внутриклассовой энтропии от размера выборки при удалении объектов из класса

статистических алгоритмов классификации наблюдается сильная корреляция между величиной нормализованной энтропии и точностью/полнотой классификации. На основе полученных результатов построен алгоритм отброса объектов обучающей выборки, повышающий качество классификации.

СПИСОК ЛИТЕРАТУРЫ

1. Holmstrom M., Liu D., Vo C. Machine learning applied to weather forecasting //Meteorol. Appl. – 2016. – pp. 1–5.
2. Deng L., Li X. Machine learning paradigms for speech recognition: An overview //IEEE Transactions on Audio, Speech, and Language Processing. – 2013. – vol. 21. – no. 5. – pp. 1060–1089.
3. Srivastava N., Mansimov E., Salakhudinov R. Unsupervised learning of video representations using lstms //International conference on machine learning. – PMLR, 2015. – pp. 843–852.
4. Zhang X. L., Wu J. Deep belief networks based voice activity detection //IEEE Transactions on Audio, Speech, and Language Processing. – 2012. – vol. 21. – no. 4. – pp. 697–710.
5. Bagnell J. A. et al. Learning for autonomous navigation //IEEE Robotics and Automation Magazine. – 2010. – vol. 17. – no. 2. – pp. 74–84.
6. Rajkomar A., Dean J., Kohane I. Machine learning in medicine //New England Journal of Medicine. – 2019. – vol. 380. – no. 14. – pp. 1347–1358.
7. Stamp M. A survey of machine learning algorithms and their application in information security //Guide to Vulnerability Analysis for Computer Networks and Systems. – Springer, Cham, 2018. – pp. 33–55.
8. Famili A. et al. Data preprocessing and intelligent data analysis //Intelligent data analysis. – 1997. – vol. 1. – no. 1. – pp. 3–23.
9. Pearson K. LIII. On lines and planes of closest fit to systems of points in space //The London, Edinburgh, and Dublin philosophical magazine and journal of science. – 1901. – vol. 2. – no. 11. – pp. 559–572.
10. Загоруйко Н. Г. Прикладные методы анализа данных и знаний. – 1999.
11. Kumar U., Kumar V., Kapur J. N. Normalized measures of entropy //International Journal Of General System. – 1986. – vol. 12. – no. 1. – pp. 55–69.
12. Kwak N., Choi C. H. Input feature selection by mutual information based on Parzen window //IEEE transactions on pattern analysis and machine intelligence. – 2002. – vol. 24. – no. 12. – pp. 1667–1671.

13. Silverman B. W. Density estimation for statistics and data analysis. – Routledge, 2018.

14. McLachlan G. J., Krishnan T. The EM algorithm and extensions. – John Wiley & Sons, 2007. – vol. 382.

15. Bickel P. J., Rosenblatt M. On some global measures of the deviations of density function estimates //The Annals of Statistics. – 1973. – pp. 1071–1095.

16. Melnykov V., Melnykov I. Initializing the EM algorithm in Gaussian mixture models with an unknown number of components //Computational Statistics & Data Analysis. – 2012. – vol. 56. – no. 6. – pp. 1381–1395.

17. Richardson S., Green P. J. On Bayesian analysis of mixtures with an unknown number of components (with discussion) //Journal of the Royal Statistical Society: series B (statistical methodology). – 1997. – vol. 59. – no. 4. – pp. 731–792.

18. P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547–553, 2009.

19. Wright R. E. Logistic regression. – 1995.

20. Drucker H. et al. Support vector regression machines //Advances in neural information processing systems. – 1997. – vol. 9. – pp. 155–161.

21. Reynolds D. A. Gaussian mixture models //Encyclopedia of biometrics. – 2009. – vol. 741. – pp. 659–663.

22. Tsuruoka Y., Tsujii J., Ananiadou S. Stochastic gradient descent training for l1-regularized log-linear models with cumulative penalty //Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. – 2009. – pp. 477–485.

23. Jiang L. et al. Survey of improving k-nearest-neighbor for classification //Fourth international conference on fuzzy systems and knowledge discovery (FSKD 2007). – IEEE, 2007. – vol. 1. – pp. 679–683.

24. David J.C. MacKay, "Information Theory, Inference, and Learning Algorithms". — Cambridge University Press, 2003. — С. 30-—35.

25. Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome. 8.5 The EM algorithm // The Elements of Statistical Learning (неопр.). — New York: Springer, 2001. — С. 236-—243.

26. P. Viola and W. M. Wells III, “Alignment by Maximization of Mutual Information”, International Journal of Computer Vision, vol. 24, no. 2, pp. 137-–154, 1997

27. M. F. Huber, T. Bailey, H. Durrant-Whyte, and U. D. Hanebeck, “On entropy approximation for Gaussian mixture random vectors”, in Proc. IEEE Int’l Conf. Multisensor Fusion and Integration for Intelligent Systems (MFI), Aug. 2008

28. SM Kim, TT Do, TJ Oechtering, G Peters, “On the Entropy Computation of Large Complex Gaussian Mixture Distributions” in Proc. IEEE Transactions on Signal Processing, Feb. 2015

29. Safavian S. R., Landgrebe D. A survey of decision tree classifier methodology //IEEE transactions on systems, man, and cybernetics. – 1991. – vol. 21. – no. 3. – pp. 660–674.

On the Influence of Entropy within a Class on the Precision and Recall of Classification

P.R. Akimova, V.V. Matsulevich, N.M. Zamuriy, M.O. Ukraintsev, S.A. Barinova, A.B. Dubov, V.V. Anikanova, F.R. Zanchenko, F.I. Ivanov

This paper deals with the relationship between the normalized entropy (per object) within a class and the quality of the classification. The influence of entropy on such basic metrics as precision and recall of classification for a given class is investigated. To obtain the main results, various classification methods are used, as well as three main approaches to reconstructing the distribution probability density function: nonparametric, parametric, and based on reconstructing a mixture of Gaussian distributions.

KEYWORDS: Entropy, statistical classification, distribution density reconstruction