УДК 004.93'11

А.В. АКИМОВ, А.О. ДОНСКИХ, А.А. СИРОТА

A.V. AKIMOV, A.O. DONSKIKH, A.A. SIROTA

**АНАЛИЗ СВОЙСТВ СМЕШАННЫХ ЯДЕРНЫХ ОЦЕНОК ПЛОТНОСТИ РАСПРЕДЕЛЕНИЯ ВЕРОЯТНОСТЕЙ В ИНТЕРЕСАХ РЕШЕНИЯ ЗАДАЧ МАШИННОГО ОБУЧЕНИЯ**

**AN ANALYSIS OF MIXED KERNEL DENSITY ESTIMATION METHOD WITH APPLICATION TO MACHINE LEARNING**

***Аннотация.*** *Исследуются свойства смешанных оценок плотности распределения, получаемых по независимым ядерным оценкам неизвестной плотности распределения для вектора информативных признаков и известной плотности распределения вектора помеховых искажений. Приводятся результаты сравнения смешанных оценок с традиционной ядерной оценкой Парзена, применяемой непосредственно к обучающим данным, на примере задачи восстановления плотностей одномерных и двумерных случайных величин с равномерными и гауссовскими распределениями. Предлагается алгоритм размножения обучающих данных на основе смешанных оценок и исследуется его эффективность на примере задачи классификации многомерных гауссовских величин с помощью искусственных нейронных сетей.*

***Ключевые слова:*** *ядерная оценка плотности распределения, смешанные оценки, машинное обучение, искусственное размножение данных.*

***Abstract.*** *A mixed density estimation method combining density estimators calculated independently for informative features and non-informative noise is analyzed. The method is compared with traditional Parzen window method, applied directly to the training data. One-dimensional and two-dimensional Gaussian and uniform random variables are used as datasets for comparison. A new data augmentation algorithm based on mixed density estimators is proposed.* *The method is tested using artificial neural networks and multivariate Gaussian mixtures as testing data sets. The results of the testing are provided.*

***Keywords:*** *kernel density estimation, mixed density estimator, machine learning, data augmentation.*

**Введение.** Одним из способов решения проблемы наличия искажений в задачах машинной обработки изображений и сигналов (например, в задачах классификации элементов движущегося потока биологических объектов в поточных спектральных анализаторах [1-2]) является применение алгоритмов аугментации – искусственного размножения данных при формировании обучающей выборки на основе использования имеющейся априорной информации о характере помеховых искажений. В частности, в [3] был предложен метод размножения данных в обучающей выборке на основе восстановления многомерной плотности распределения вероятностей вектора признаков, описывающего объекты в задаче распознавания, с применением ядерных оценок (оценок Парзена) по отношению к обучающим данным и генерации новых векторов с использованием метода исключений.

Целью данной работы является исследование свойств предложенного ранее [4] метода восстановления многомерной плотности распределения вероятностей для данных, подверженных различного рода искажениям, основанного на использовании независимых оценок плотности распределения для вектора информативных признаков и известной плотности распределения вектора помеховых искажений, на основании которых затем вычисляется общая (смешанная) оценка как свертка этих плотностей. Наиболее подробно будет рассмотрен вопрос применения смешанных оценок в составе алгоритмов искусственного размножения данных и сравнение полученных алгоритмов с ранее описанным в [3].

**Материалы и методы.** Рассмотрим следующую постановку применительно к использованию ядерных оценок в задачах классификации образов или аугментации данных для произвольных алгоритмов машинного обучения. Пусть регистрируемый вектор наблюдаемых признаков формируется на основе смеси из двух статистически независимых составляющих **,** где  **–** случайный вектор, являющийся информативным для решаемой задачи и описываемый характерным для каждого класса образов  распределением ; **–** случайный вектор помеховой составляющей, не несущий полезную информацию и описываемый общим для всех классов образов распределением ; , – значения случайных векторов и**.**

Ограничимся рассмотрением аддитивной модели смеси. В этом случае выполняется

**,** , (1)

т.е. результирующая плотность распределения формируется как свертка плотностей распределения составляющих смеси. Будем также считать, что при построении распределения смеси плотность  неизвестна, а плотность известна (например, получена путем многократных измерений признаков для одного и того же образца класса). При этом для построения плотности  задана обучающая выборка в виде совокупности реализаций . Тогда получение искомой оценки функции плотности предлагается проводить в два этапа. На первом этапе на основе обучающей выборки может быть получена стандартная оценка Парзена в виде

. (2)

На втором этапе осуществляется свертка полученной оценки распределения (2)  с учетом и распределения аддитивного шума в соответствии с соотношением:

 (3)

,

**,**

т.е. определяется конечной суммой, элементы которой являются свертками ядерных функций и плотности **,** центрированных относительно точек совокупности .

Рассмотрим пример, когда при построении оценки (2) используются ядра в виде гауссианы



и присутствует аддитивный гауссовский шум, описываемый распределением . Тогда при использовании предлагаемого подхода получается оценка в виде

. (4)

Полученные таким образом оценки (3), (4) будем назвать смешанными, акцентируя внимание на том, что одна из составляющих смеси получена на основе непараметрической оценки, а другая определена изначально в аналитическом виде.

Какизвестно [5-6], непосредственные оценки Парзена на основе соотношения (2) при определенных условиях являются несмещенными и состоятельными. В [4] было доказано, что если оценка  является несмещенной состоятельной при , то смешанные оценки вида (3) также обладают этими свойствами. Кроме того, в [4] было доказано следующее утверждение:

**Утверждение 1.** Пусть , где каждый является реализацией . Тогда при известной плотности распределения  использование смешанной оценки эквивалентно реализации процедуры искусственного размножения обучающих данных, в ходе которой для каждого вектора  проводится генерация подвыборки, состоящей из  образов на основе распределения , где , причем  – случайная величина с распределением, плотность которого определяемым одним из возможных ядер, используемых при получении смешанной оценки, . Здесь  значение, выбираемое исходя из заданного . Отметим, что плотность распределения  для , если известна определяется выражением .

Следствием данного утверждения является то, что при достаточно больших искусственное размножение данных может осуществляться как , так как  при этом стремится к нулю, а стремится к дельта-функции.

Известная методика [7] определения оптимального значения параметра  для обычной (не смешанной) оценки основана на нахождении максимума функции логарифма отношения правдоподобия при кросс-валидации. Показано [7], что при использовании гауссовского ядра с диагональной матрицей  область расположения максимума функции правдоподобиядля поиска оптимального значения путем переборанаходится в границах

 **,** (5)

где  расстояния между элементами обучающей выборки  при их попарном сравнении.

При использовании ядра с диагональной матрицей  в ядерной оценке для любой пары  оценки (5) для границ диапазона значений  будут верны с тем только отличием, что вместо в (5) следует использовать

.

**Результаты и их обсуждение.** В качестве числового критерия для сравнения плотностей и оценки качества восстановления была выбрана широко используемая в задачах восстановления плотности интегральная квадратическая ошибка (ISE) [8]:



Для экспериментальной проверки сходимости смешанной оценки и изучения ее эффективности было проведено несколько экспериментов с использованием искусственно сгенерированных одномерных и двумерных данных, имеющих различные распределения.

Для одномерных данных было рассмотрено два случая:

* распределение – смесь из двух гауссовских распределений с математическими ожиданиями  и дисперсиями ,  - гауссовское распределение с нулевым математическим ожиданием  и дисперсией **** (рисунок 1, а-в);
* распределение – равномерное в диапазоне  с математическим ожиданием и дисперсией ****,  - равномерное распределение с нулевым математическим ожиданием и дисперсией;

Для двумерных данных рассматривался один случай: – смесь из двух гауссовских распределений с математическими ожиданиями  и матрицами ковариации ,  - гауссовское распределение с нулевым математическим ожиданием и матрицей ковариации , где  имеют вид

.

В ходе исследования проверялось, что при  смешанная оценка , вычисляемая по восстановленной плотности и известной плотности , сходится к и проводилось сравнение смешанной оценки  с обычной оценкой . В начале каждого эксперимента генерировалось  значений случайной величины , по которым методом Парзена восстанавливалась плотность распределения  и вычислялась смешанная оценка . Также с использованием имеющихся значений величины  генерировалось  значений случайной величины , по которым вычислялась обычная оценка Парзена . Значение параметра  в обоих случаях подбиралось методом кросс-валидации путем поиска максимума функции правдоподобия. Исследование проводилось для значений  от 50 до 5000, при этом для каждого значения выполнялось по 100 экспериментов (для двумерного случая – по 10 экспериментов), в которых вычислялись значения  для оценок  и , после чего полученные значения усреднялись.

Полученные зависимости интегральной квадратичной ошибки оценок плотностей  и  от объема исходных данных приведены на рисунке 1.

а) б) в)

Рисунок 1. Зависимости интегральной квадратичной ошибки оценок плотностей  и  от объема исходных данных для случая одномерного мультигауссовского распределения (а) и для случая одномерного равномерного распределения (б); для случая двумерного мультигауссовского распределения (в).

Видно, что для всех трех рассмотренных случаев при увеличении объема данных  среднеквадратичная ошибка монотонно убывает как для , так и для . При этом величина ошибки для смешанной оценки  оказывается примерно вдвое меньше при любых значениях . Также стоит отметить, что величина ошибки для обеих оценок в случае равномерного распределения значительно превышает величину ошибки в случае гауссовских распределений, что, вероятно, обусловлено тем, что при восстановлении плотности использовались гауссовские ядра, сильно отличающиеся по форме от равномерного распределения.

Для исследования возможности использования смешанных ядерных оценок как альтернативы методам искусственного размножения данных, было проведен эксперимент, направленный на сравнение эффективности применения смешанных оценок в задачах классификации образов по сравнению с методом размножения данных, описанным в [3]. В ходе эксперимента сравнивалась точность классификации случайных величин нескольких классов, описываемы случайными векторами вида , с помощью следующих классификаторов:

* нейронная сеть, обученная по малой обучающей выборке;
* нейронная сеть, обученная по большой обучающей выборке;
* искусственная нейронная сеть, обученная по малой обучающей выборке, искусственно размноженной до объема большой с помощью метода, описанного в [3];
* искусственная нейронная сеть, обученная по малой обучающей выборке, искусственно размноженной до объема большой путем добавления случайно генерируемых величин с плотностью  к имеющимся значениям величины  каждого класса;
* байесовский классификатор, синтезированный для известной плотности распределения , используемый для оценки нижней границы вероятности ошибок.

В качестве  использовались распределения, имеющие вид смесей гауссовских случайных величин (ГСВ). В ходе эксперимента генерировались случайные векторы, соответствующие трем классов образов, представленных смесями с несколькими компонентами смеси внутри каждого класса (по две компоненты для первого и второго классов и три компоненты для третьего) со следующими параметрами:



, , , ,



где  – одномерная равномерно распределенная случайная величина в диапазоне ;  - параметр, определяющий степень рассредоточенности (пересечения) компонентов смесей каждого класса;  – коэффициент корреляции, используемый при задании матрицы ковариаций, выбираемый случайным образом в диапазоне . В качестве  для всех классов использовалось гауссовское распределением с нулевым математическим ожиданием и единичной матрицей ковариации .

В начале каждого эксперимента генерировалось ,  и  значений случайной величины  для каждого класса, составляющие, соответственно, малую обучающую выборку, большую обучающую выборку и тестовую выборку. Каждый образ описывается  признаками, количество которых сокращалось при обучении нейронных сетей до 5 с помощью метода главных компонент. Для каждого значения выполнялось по 100 экспериментов, после чего полученные значения усреднялись. В качестве обучаемого классификатора использовалась нейронная сеть класса MLP (многослойный персептрон). Сеть содержит один скрытый слой с сигмоидальной функцией активации и один выходной слой. Количество входных контактов сети соответствует количеству используемых признаков распознавания, т.е. 5, а количество нейронов в выходном слое  равно числу классов (в данном случае 3), при этом значение «1» на выходе нейрона означает, что рассматриваемый образец принадлежит к соответствующему классу, а «0» – не принадлежит. Количество нейронов в скрытом слое  выбиралось из диапазона значений . Сеть создавалась и тестировалась в среде MATLAB, для обучения использовался алгоритм Левенберга-Марквардта.

В таблице 1 приведены результаты, полученные для смесей ГСВ со слабо коррелированными признаками ().

Таблица 1 – Вероятности ошибочного распознавания для смеси ГСВ при  и различных объемах искусственно размноженной обучающей выборки

|  |  |  |
| --- | --- | --- |
|  | Классификатор | Вероятность ошибочного распознавания |
|  | 300 | 400 | 500 | 600 | 700 | 800 | 900 | 1000 |
| 0 | НС (малая выборка) | 0.1716 | 0.1806 | 0.1791 | 0.1786 | 0.1854 | 0.1764 | 0.1837 | 0.1744 | 0.1731 |
| НС (большая выборка) | 0.1613 | 0.1563 | 0.1572 | 0.1562 | 0.1576 | 0.1525 | 0.1518 | 0.1504 | 0.1525 |
| НС (размн. по плотн.) | 0.1674 | 0.1707 | 0.1646 | 0.1593 | 0.1585 | 0.1593 | 0.1596 | 0.1582 | 0.1586 |
| НС (размн. добавл. V) | 0.1616 | 0.1644 | 0.1611 | 0.1600 | 0.1579 | 0.1607 | 0.1591 | 0.1594 | 0.1573 |
| Байесовский классиф. | 0.1466 | 0.1480 | 0.1474 | 0.1474 | 0.1470 | 0.1482 | 0.1477 | 0.1469 | 0.1474 |
| 4 | НС (малая выборка) | 0.0954 | 0.0918 | 0.0883 | 0.0920 | 0.0889 | 0.0931 | 0.0930 | 0.0895 | 0.0928 |
| НС (большая выборка) | 0.0707 | 0.0624 | 0.0613 | 0.0642 | 0.0574 | 0.0543 | 0.0591 | 0.0539 | 0.0536 |
| НС (размн. по плотн.) | 0.0806 | 0.0742 | 0.0692 | 0.0734 | 0.0684 | 0.0664 | 0.0673 | 0.0653 | 0.0685 |
| НС (размн. добавл. V) | 0.0757 | 0.0750 | 0.0678 | 0.0648 | 0.0637 | 0.0621 | 0.0666 | 0.0625 | 0.0646 |
| Байесовский классиф. | 0.0100 | 0.0096 | 0.0098 | 0.0097 | 0.0097 | 0.0095 | 0.0100 | 0.0095 | 0.0098 |

Видно, что метод искусственного размножения данных путем добавления шума  показывает результаты, близкие к методу размножения данных по восстановленной плотности, и несколько превосходит его в случае сильно рассредоточенных данных . При этом оба метода позволяют сократить вероятность ошибочного распознавания.

В таблице 2 приведены результаты, полученные для смесей ГСВ с сильно коррелированными признаками (). Оба метода искусственного размножения данных так же позволяют сократить вероятность ошибочного распознавания, при этом в случае сильно рассредоточенных данных  метод искусственного размножения данных путем добавления шума  снова оказывается эффективнее.

Таблица 2 – Вероятности ошибочного распознавания для смеси ГСВ при  и различных объемах искусственно размноженной обучающей выборки

|  |  |  |
| --- | --- | --- |
|  | Классификатор | Вероятность ошибочного распознавания |
|  | 300 | 400 | 500 | 600 | 700 | 800 | 900 | 1000 |
| 0 | НС (малая выборка) | 0.3541 | 0.3449 | 0.3533 | 0.3505 | 0.3494 | 0.3463 | 0.3504 | 0.3533 | 0.3519 |
| НС (большая выборка) | 0.3300 | 0.3261 | 0.3208 | 0.3155 | 0.3151 | 0.3144 | 0.3120 | 0.3125 | 0.3130 |
| НС (размн. по плотн.) | 0.3411 | 0.3368 | 0.3286 | 0.3267 | 0.3263 | 0.3240 | 0.3227 | 0.3272 | 0.3252 |
| НС (размн. добавл. V) | 0.3377 | 0.3332 | 0.3287 | 0.3238 | 0.3263 | 0.3241 | 0.3214 | 0.3248 | 0.3238 |
| Байесовский классиф. | 0.3065 | 0.3069 | 0.3059 | 0.3057 | 0.3057 | 0.3042 | 0.3037 | 0.3058 | 0.3052 |
| 4 | НС (малая выборка) | 0.0699 | 0.0623 | 0.0639 | 0.0715 | 0.0663 | 0.0692 | 0.0687 | 0.0667 | 0.0648 |
| НС (большая выборка) | 0.0517 | 0.0477 | 0.0443 | 0.0412 | 0.0403 | 0.0399 | 0.0394 | 0.0389 | 0.0377 |
| НС (размн. по плотн.) | 0.0668 | 0.0565 | 0.0560 | 0.0570 | 0.0542 | 0.0545 | 0.0538 | 0.0514 | 0.0514 |
| НС (размн. добавл. V) | 0.0551 | 0.0533 | 0.0508 | 0.0513 | 0.0494 | 0.0467 | 0.0478 | 0.0461 | 0.0471 |
| Байесовский классиф. | 0.0014 | 0.0014 | 0.0014 | 0.0014 | 0.0013 | 0.0013 | 0.0014 | 0.0013 | 0.0014 |

**Заключение.** В рамках данной работы описан метод восстановления функции плотности распределения данных в условиях помеховых искажений с использованием смешанных непараметрических ядерных оценок функций правдоподобия классов. Проведено сравнение предложенного метода с традиционными ядерными оценками, применяемыми напрямую к исходной обучающей выборке, экспериментально показана сходимость смешанных оценок при увеличении объема обучающей выборки. Также предложен метод искусственного размножения данных, основанный на идее смешанных оценок, и проведено исследование его эффективности. Анализ полученных результатов позволяет сделать вывод о том, что предлагаемый метод позволяет снизить вероятность ошибок классификации в задачах машинного обучения.

Результаты работы получены в рамках выполнения государственного задания Минобрнауки России по проекту № 8.3844.2017/4.6 «Разработка средств экспресс-анализа и классификации элементов неоднородного потока зерновых смесей с патологиями на основе интеграции методов спектрального анализа и машинного обучения».

**СПИСОК ЛИТЕРАТУРЫ**

1. Donskikh, A.O. Methods of analysis and classification of the components of grain mixtures based on measuring the reflection and transmission spectra / A.O. Donskikh, D.A. Minakov, A.A. Sirota, V.A. Shulgin // Scientific Study and Research: Chemistry and Chemical Engineering, Biotechnology, Food Industry – 2017. – vol. 18(3) – pp. 291-302.

2. Donskikh, A.O. Optical methods of identifying the varieties of the components of grain mixtures based on using artifcial neural networks for data analysis / A.O. Donskikh, D.A. Minakov, A.A. Sirota // Journal of Theoretical and Applied Information Technology – 2018. – vol. 96 (2) – pp. 534-542.

3. Донских, А.О. Метод искусственного размножения данных в задачах машинного обучения с использованием непараметрических ядерных оценок плотности распределения вероятностей / А.О. Донских, А.А. Сирота // Вестник Воронежского государственного университета Серия: Системный анализ и информационные технологии. – 2017. – №3. – С. 142-155.

4. Акимов, А.В., Донских, А.О., Минаков, Д.А., Сирота, А.А. Смешанные ядерные оценки многомерных распределений и их применение в задачах машинного обучения // Математическое моделирование и информационные технологии в инженерных и бизнес-приложениях: сборник материалов междунар. науч. конф. – Воронеж: 2018. – С. 133-149.

5. Крянев, А.В. Математические методы обработки неопределенных данных / А.В. Крянев, Г.В. Лукин – М.: Физматлит, 2003. – 216 с.

6. Фукунага, К. Введение в статистическую теорию распознавания образов / К. Фукунага – М.: Наука, 1979. – 368 с.

7. Кривенко, М.П. Непараметрическое оценивание элементов байесовского классификатора / М.П. Кривенко // Информатика и ее применения. – 2010. – Т. 4. Вып. 2 – С. 13-24.

8. Gramacki, A. Nonparametric Kernel Density Estimation and Its Computational Aspects / A. Gramacki, Springer International Publishing. – 2018 – pp. 42-49.

**Акимов Алексей Викторович**

Воронежский государственный университет, г. Воронеж

Исследователь, преподаватель-исследователь, младший научный сотрудник кафедры технологий обработки и защиты информации

Тел.: +7(960)125-71-30

E-mail: akimov@vsu.ru

**Донских Артём Олегович**

Воронежский государственный университет, г. Воронеж

Аспирант кафедры технологий обработки и защиты информации, младший научный сотрудник физической лаборатории факультета компьютерных наук

Тел.: +7(960)125-81-11

E-mail: a.donskikh@outlook.com

**Сирота Александр Анатольевич**

Воронежский государственный университет, г. Воронеж

Д. т. н., профессор, заведующий кафедрой технологий обработки и защиты информации.

Тел.: +7(903)030-69-43

E-mail: sir@cs.vsu.ru