УДК 004.891

ПАРАМЕТРИЗАЦИЯ ГИПЕРПАРАМЕТРОВ В ПРИКЛАДНЫХ ЗАДАЧАХ МАШИННОГО ОБУЧЕНИЯ НА ОСНОВЕ ЯДЕРНЫХ ФУНКЦИЙ

Майтак Р. В., бакалавр
Пылов П. А., магистр
Научный руководитель: Дягилева А. В., к.т.н., доцент
Кузбасский государственный технический университет
имени Т.Ф. Горбачева

Метод парзеновского окна – это усовершенствованный алгоритм метода ближайших соседей. Он подразумевает, что вес признака не должен зависеть от порядкового номера объекта в ранжированном списке, но должен зависеть от расстояния.

В такой вариации вес признака оценивается следующим образом **(1)**:

$$w(i,x) = K\left(\frac{p(x,\ x^{(i)})}{h}\right) (1)$$

где h – это ширина окна, а K(r) – это ядро.

Функция K(r) в формуле (1) представляется убывающей функцией от расстояния между x и его $x^{(i)}$ соседом. Функция является финитной, то есть не возрастает и положительна на [0; 1]. За пределами отрезка она равна нулю.

Обозначим формализованный вид обобщенного классификатора **(2)** для решения задач регрессии и классификации (на котором основан метод окна Парзена).

$$x^{(i)} - \text{это } i - \text{ый сосед объекта } x \text{ среди } x_1, \dots, x_l;$$
$$y^{(i)} - \text{это ответ на } i - \text{ом соседе объекта } x.$$

Пользуясь данными обозначениями, можно записать общую конструкцию классификатора **(2)**, который оценивает, насколько объект x близок к объектам класса y.

$$a(x; X^l) = arg\max_{y \in Y} \sum_{i=1}^{l} [y^{(i)} = y] w(i,x) \quad (2)$$

где: $w(i,x) - \text{это вес (степень важности) } i -$ го соседа объекта x; $\sum_{i=1}^{l} [y^{(i)} = y] w(i,x) -$ это оценка близости объекта x к классу y.

Если подставить весовую функцию **(1)** в общую конструкцию классификатора **(2)**, то получим следующую конструкцию нового частного случая классификатора **(3)**.

$$a(x; X^l, h, K) = arg \max_{y \in Y} \sum_{i=1}^{l} [y_i = y] K\left(\frac{p(x,\ x^{(i)})}{h}\right) \quad (3)$$

Фактически, вместо параметра k из алгоритма k-ближайших соседей появляется другой параметр – ширина окна. В зависимости от того, сколько объектов попадают в окно, по этим объектам принимается решение о том, к какому классу отнести *x*.

Опасность данного подхода состоит в том, что объект *x* может оказаться в значительном удалении от объектов обучающей выборки, на расстоянии, большем, чем *h*.

Первым решением данной проблемы может служить уход от использования финитных ядер и использование только ядра, которые имеют значения строго большее нуля в диапазоне значений [0; +∞].

Вторым подходом решения проблемы является совершенствование метода ближайших соседей **(4)**.

$$a(x; X^l, k, K) = arg \max_{y \in Y} \sum_{i=1}^{l} [y_i = y] K\left(\frac{p(x,\ x^{(i)})}{p(x,\ x^{(k+1)})}\right) \quad (4)$$

В формуле **(4)** используется окно Парзена переменной ширины. Тогда окно Парзена будет определяться так, чтобы в окно всегда попадало *k* соседей. В формуле **(4)** реализован этот механизм через взятие расстояние для *k+1* соседа в качестве *h*.

Рассмотрим функционал алгоритма парзеновского окна на примере двумерной выборки (рисунок 1).

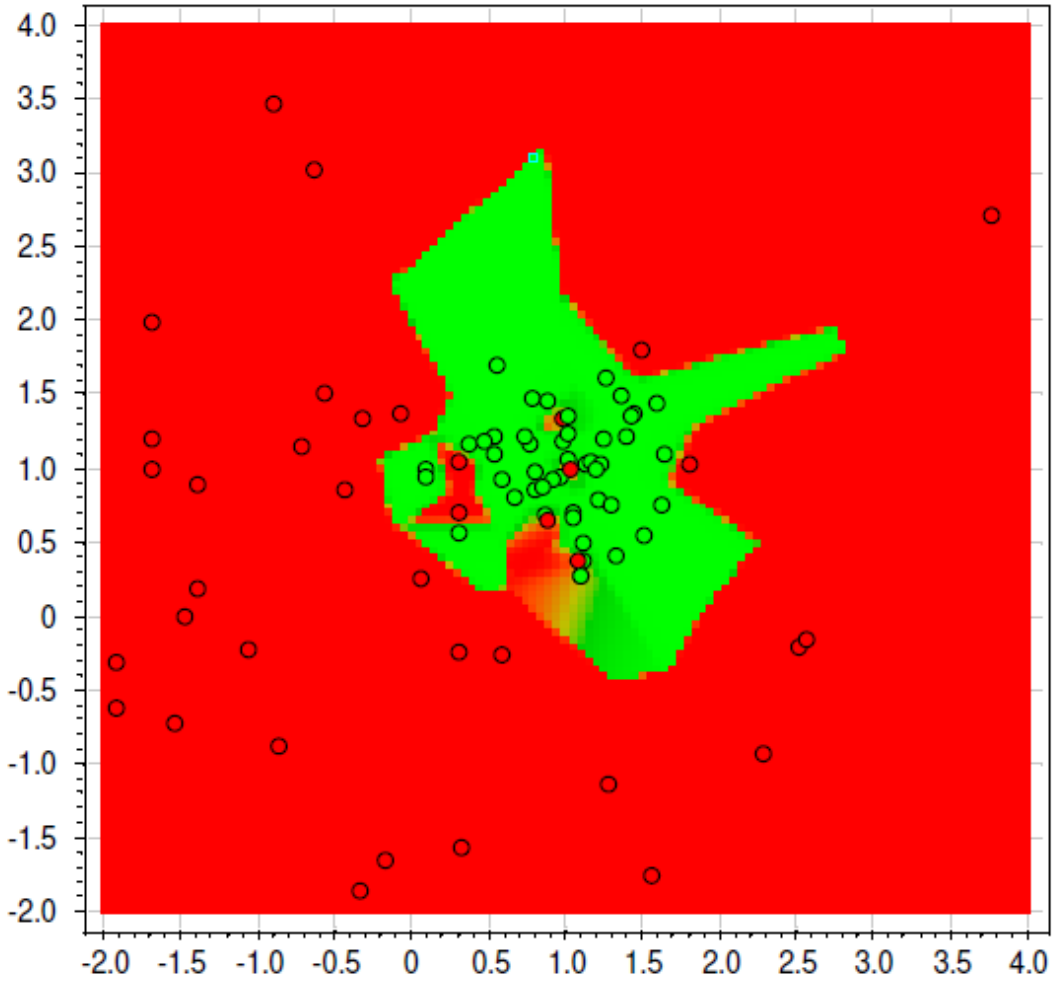

Рисунок 1 – Минимальный размер ширины окна *h*

На рисунке 1 разные классы представляются разными цветами (зеленым и красным цветом).

Ширина окна *h* была принята равной 0,05. Это очень маленькое значение, то есть для классификации практически всегда используется один ближайший сосед. Из рисунка 1 очевидно, что разделяющая поверхность очень похожа на кусочно-линейную поверхность, причем каждый отрезок линии связан с тем, что разделяющая плоскость проходит перпендикуляром посередине между двумя объектами разных классов. Иными словами, каждая такая линия определена какой-то парой точек двух разных классов.

При увеличении окна *h* (рисунок 2) до 0,2 (при том же самом масштабе) увеличивается окрестность и уже больше соседей захватывается ею. Разделяющая поверхность перестает быть похожей на кусочно-линейную и сильнее приближается к гладкой функции.

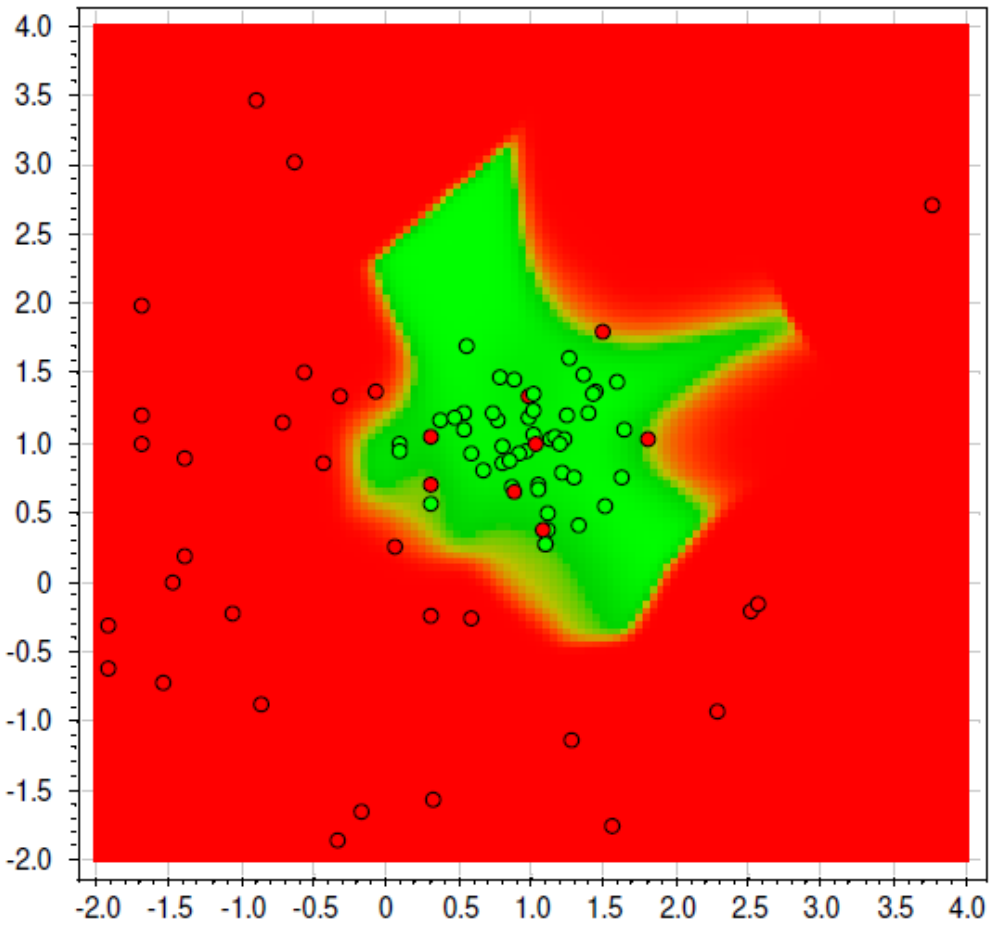

Рисунок 2 – Увеличение размеров окна *h* при сохранении масштаба

При переходе окна к размеру 0,5 (рисунок 3) выборка одного класса (зеленых элементов) уже явственно выделяется на фоне выборки второго класса (красных элементов).

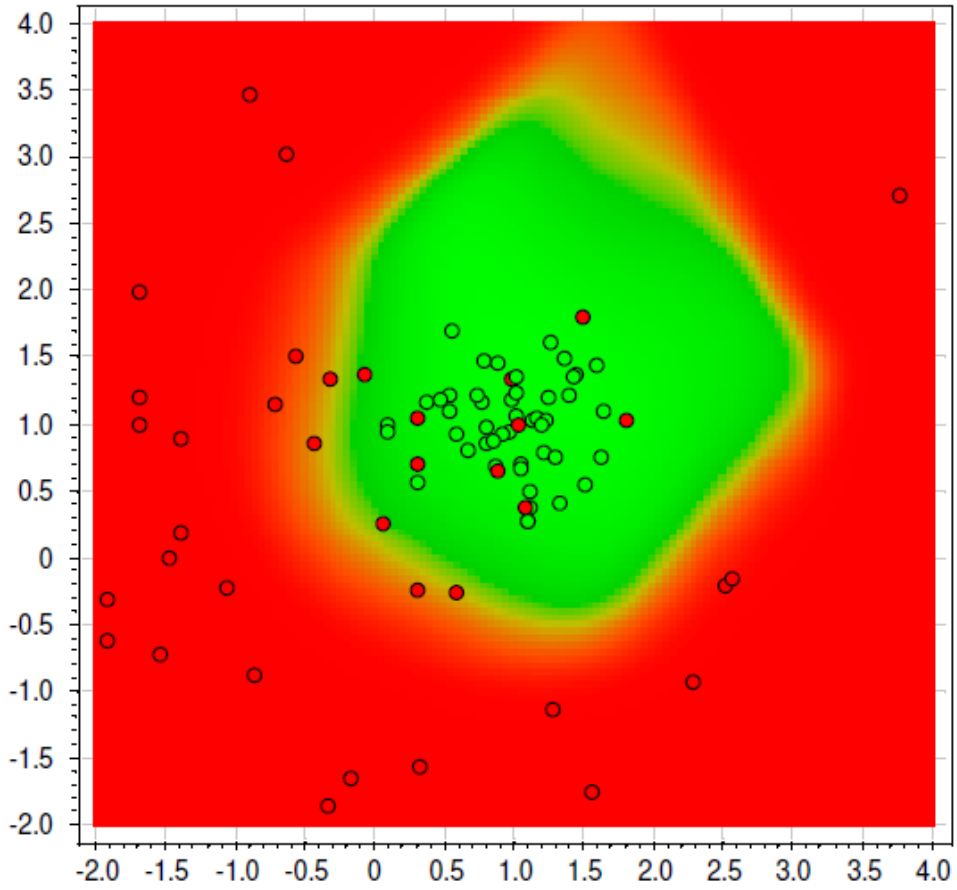

Рисунок 3 – Размер ширины окна h = 0,5

При выборе большего значения ширины окна разделяющая поверхность все сильнее упрощает свою форму.

В пределе (при стремлении ширины окна к $+\infty$) метод парзеновского окна начинает стремиться к линейному классификатору.

Наиболее сложная, изрезанная разделяющая поверхность получается в том случае, когда принимается окрестность очень маленького радиуса (малое значение h).

Метод окна Парзена отлично подходит для прикладных задач, в том числе для тех, в которых признаки несбалансированны [1]. Такой эффект достигается благодаря адаптивному подбору параметры переменной ширины окна, изменяющейся в рамках финитного ядра.

Однако, стоит отметить, что метод парзеновского окна с должной степенью справляется с задачей классификации, но он в полной мере не подходит для задачи регрессии, так как отнесение объекта к заданному классу совершается самим алгоритмом изнутри, не выводя при этом для исследователя данных о вероятности отнесения объекта к заданному классу.

Список используемой литературы:

1. Томас Кормен, Чарльз Лейзерсон. Алгоритмы: построение и анализ, 3-е издание – М.: ООО И.Д. Вильямс. 2013. – 1328 с.