УДК 004.855.5

ПОВЫШЕНИЕ ТОЧНОСТИ КЛАССИФИКАЦИИ ДАННЫХ С ИСПОЛЬЗОВАНИЕМ АЛГОРИТМА *K*-БЛИЖАЙШИХ СОСЕДЕЙ НА ОСНОВЕ ПРЕКЛАСТЕРИЗАЦИИ ОБУЧАЮЩИХ ДАННЫХ

В. И. Орешков, к.т.н., доцент кафедры САПР ВС РГРТУ, Рязань, Россия;
orcid.org/0000-0003-0316-4927, e-mail: vyacheslav.oreshkov@yandex.ru

Рассматривается задача классификации данных с использованием метрического метода машинного обучения k-ближайших соседей. ***Целью работы*** *является разработка методики адаптивной подстройки параметра соседства алгоритма k-ближайших соседей с целью повышения точности алгоритма за счёт учёта неоднородности распределения обучающих примеров в пространстве признаков. В тех областях пространства признаков, где плотность обучающих примеров высока, алгоритм хорошо работает даже при больших значениях параметра соседства, поскольку все определяемые им примеры расположены компактно и, как правило, относятся к одному классу. В разреженных областях пространства признаков большие значения параметра соседства приводят к вовлечению в процесс классификации примеров, расположенных на большом расстоянии, которые могут относиться к различным классам, что ухудшает точность классификации. Для решения данной проблемы в статье предлагается использовать значение параметра соседства, которое не задаётся одинаковым для всех примеров, а может изменяться в зависимости от их плотности в области пространства признаков. Для этого производится предварительная кластеризация обучающего набора и определяется плотность каждого кластера как среднеквадратичное расстояние от примеров кластера до его центроида. На основе полученного значения плотности вычисляется параметр соседства для кластера, который используется для классификации любого объекта в пределах кластера. Эксперимент показал более точную работу модифицированного алгоритма в процессе перекрёстной проверки*

Ключевые слова: *интеллектуальный анализ данных, машинное обучение, обучение с учителем, обучающий пример, классификация, кластеризация, класс, кластер, центроид, перекрёстная проверка.*

DOI: 10.21667/1995-4565-2021-76-65-73

Введение

Одной из важнейших задач интеллектуального анализа данных является классификация – отнесение объектов предметной области к заранее определённым группам, называемым классами [3]. При этом каждому классу должны принадлежать объекты, близкие по своим свойствам. Обобщая свойства известных объектов класса на новые, отнесённые к нему объекты, можно получать знания о них.

Задача классификации решается с помощью аналитических моделей, называемых классификаторами. Классифицировать объект означает предъявить набор его признаков (обычно представленных в виде вектора) на вход модели-классификатора, которая должна присвоить ему метку или номер класса. В настоящее время разработано большое количество различных видов классификаторов, для построения которых используются как статистические методы (логистическая регрессия, дискриминантный анализ), так и методы машинного обучения (нейронные сети, деревья решений и др.). Необходимость использования в анализе данных большого числа разнообразных методов классификации обусловлена тем, что решаемые с её помощью задачи могут иметь свои особенности, связанные, например, с представлением исходных данных, их количеством и качеством, что требует выбора адекватного классификатора. Поэтому выбор классификатора, соответствующего особенностям решаемой задачи анализа, является важным фактором получения правильного решения.

Отнесение классификации к интеллектуальным технологиям анализа данных обусловлено тем, что в повседневной жизни сознание человека, в поле которого постоянно попадают новые объекты окружающего мира, сопоставляет их с уже известными объектами и оценивает степень их сходства. Затем на основе этой оценки объект ассоциируется с определённой группой (классом). Таким образом, классификация является наиболее «естественным» для человеческого интеллекта способом получения знаний о процессах и явлениях, происходящих в окружающем мире [1].

Все методы классификации в том или ином виде будут использовать формализованное понятие «сходства», мера которого будет оцениваться с помощью некоторой функции. В статистических методах анализа мерой сходства является вероятность принадлежности объекта классу, которая оценивается для каждого класса, после чего выбирается тот из них, для которого эта вероятность наибольшая.

В метрических методах мерой сходства является расстояние (например, евклидово) в векторном пространстве, где каждый объект представлен своим вектором признаков. Логика здесь проста: новый объект скорее всего принадлежит к тому же классу, что и большинство соседних с ним объектов. Метрические методы, как правило, используются для построения классификаторов на основе машинного обучения.

Преимуществом статистических методов является их хорошая математическая обоснованность, недостатком – низкая объясняющая способность. Использование вероятностных оценок позволяет с высокой точностью предсказать, к какому классу относится объект, но не позволяет сказать почему. Поэтому результаты статистических методов классификации могут оказаться сложными для понимания и интерпретации результатов.

Недостатком метрических методов является их эвристический характер – они могут дать неточное и неоднозначное решение, считающееся приемлемым в большинстве практически значимых случаев. Но при этом они имеют высокую объясняющую способность и поэтому их результаты проще интерпретировать. В простейшем случае можно использовать правило: объект относится к тому же классу, что и большинство его ближайших соседей.

Типичным представителем методов классификации, использующих эту логику, является метод *k*-ближайших соседей (*k*-nearest neighbors – *k*-NN). Метод относится к классу непараметрических, т.е. не требует предположений о том, из какого статистического распределения была сформирована обучающая выборка. Следовательно, классификационные модели, построенные с помощью метода *k*-NN также будут непараметрическими. Это означает, что структура модели не задаётся жёстко изначально, а определяется данными.

Теоретическая часть

Пусть имеется обучающий набор данных $\{a_1, a_2, \ldots, a_n\}$, где каждый объект a_i $(i = 1..n)$ представляет собой обучающий пример (X_i, c_j), который содержит вектор признаков $X_i = (x_1, x_2, \ldots, x_m)$ и метку класса c_j $(i = 1..k)$. Алгоритм *k*-NN содержит две фазы: обучение и классификацию. Процесс обучения, в отличие от большинства других методов машинного обучения, не представляет собой итеративную подстройку параметров модели, а заключается в простом запоминании обучающих примеров.

На фазе классификации алгоритму предъявляется классифицируемый объект (a, X), для него вычисляются расстояния (обычно, евклидовы), до каждого обучающего примера (X_i, c_j). Затем примеры сортируются

$$(X_1, c_j) \le (X_1, c_j) \le \ldots \le (X_n, c_j),$$

и объекту (a, X) присваивается класс, к которому принадлежит большинство из k – ближайших к нему примеров, т.е.

$$(a, X) = argmax \sum_{i=1}^{k} [c_j = c].$$

Здесь число соседей k, по которым определяется класс объекта, является гиперпараметром алгоритма (параметром соседства).

Постановка задачи

Основной проблемой при решении задачи классификации с помощью алгоритма *k*-NN является выбор параметра соседства *k*. Действительно, неудачный выбор параметра может привести к снижению точности и обобщающей способности алгоритма. Если значение параметра соседства будет слишком малым, то классификация будет определяться локальными неоднородностями пространственного распределения точек данных и классификатор будет иметь низкую обобщающую способность. В этом случае даже похожие новые наблюдения могут быть отнесены к разным классам.

Напротив, если значение *k* окажется слишком большим, то в «соседи» классифицируемого наблюдения попадут объекты другого класса, «голосование» которых также может привести к ошибочной классификации.

Следствием указанных проблем является то, что при низких значениях параметра *k* повышается выраженность границ классов, но при этом растёт влияние шумов (локальных особенностей) в данных. При увеличении параметра влияние шумов уменьшается, но происходит «размытие» границ классов.

Проиллюстрируем сказанное графически. На рисунке 1 представлено распределение объектов двух классов (круги и квадраты) в пространстве признаков, которое разделено на два класса (для простоты разделимость классов условно считаем линейной). Треугольником представлен классифицируемый объект.

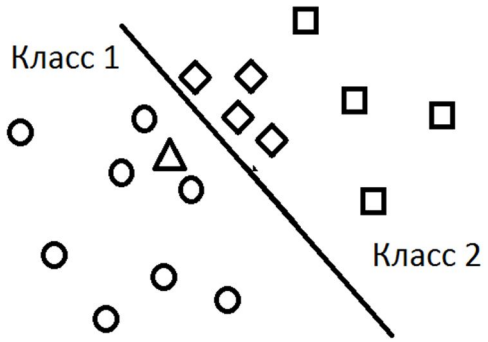

Рисунок 1 – Изменение класса объекта при изменении значения параметра соседства
Figure 1 – Changing the object class when changing the value of neighborhood parameter

При $k = 3$ классифицируемый объект будет отнесён к классу «кругов», а при $k = 7$ к классу «квадратов». Т.е. возникает неопределённость классификации объектов, расположенных вблизи границ классов, с увеличением *k*.

В этом смысле можно рассматривать выбор параметра как масштабирование. При большом масштабе теряются детали, а при малом – детали мешают рассмотреть общую картину. Иными словами, чем больше *k*, тем проще разделяющая поверхность.

Хотя в настоящее время разработано значительное количество рекомендаций по выбору параметров соседства и модификаций алгоритма *k*-NN, проблема выбора параметра *k* попрежнему остаётся актуальной.

Например, в статье [4] выбор параметра *k* предлагается производить на основе соотношений количества объектов различных классов в исходном наборе данных в задаче анализа текстов. В работе [5] предложена концепция адаптивного параметра соседства, который определяется отдельно для каждого объекта на основе свойств локальных структур данных, но задача рассмотрена только для случая бинарной классификации. В работе [6] предлагается для наборов данных, содержащих более 100 элементов выбирать $k = \sqrt{n}$, где n – число элементов. Но при этом выбор параметра на выборках меньшего размера не рассматривается.

Таким образом, обзор работ, в которых предлагаются решения данной задачи, показывает, что большинство из них сосредоточены на некоторых частных случаях использования алгоритма *k*-NN и зачастую требуют от аналитика достаточно высокой математической подготовки. В то же время, если алгоритм предполагается применять в составе аналитических платформ, он должен легко встраиваться пользователем в сценарии аналитической обработки данных, а настраиваемые параметры должны оптимизироваться автоматически, результаты должны быть понятны и интерпретируемы.

Поэтому авторы в данной работе ставят целью предложить модификацию алгоритма *k*-NN, в которой параметр соседства является адаптивным, т.е. определяется для каждого классифицируемого объекта. Кроме этого, подход должен быть универсальным в плане применимости к любым данным и приложениям.

Ещё одним фактором, который влияет на точность классификации с использованием алгоритма *k*-NN, является «плотность данных» в области векторного пространства, где расположен классифицируемый объект. Действительно, если параметр соседства будет одинаковым для двух классифицируемых объектов, один из которых расположен в «разреженной» области векторного пространства, а другой в области, где точки объектов расположены плотно, то алгоритм будет работать по-разному.

В более «плотной» области пространства *k* объектов соседей будут расположены ближе к классифицируемому объекту, чем в «разреженной» и, следовательно, точность классификации в ней будет выше. Таким образом, чтобы избежать снижения точности классификации в разреженных областях, для параметра *k* следует ввести корректирующий коэффициент, значение которого будет меняться в зависимости от плотности данных.

Проиллюстрируем сказанное с помощью рисунка на котором примеры представлены в виде точек, а классифицируемый объект – кружком. Плотность кластера 1 низкая и при $k = 5$ в классификации принимают участие удалённые объекты, которые, возможно, принадлежат разным классам. Во втором кластере, где плотность примеров высокая, при том же значении параметра соседства решение принимается на основе классов близких объектов, которые расположены компактно вблизи классифицируемого объекта, что делает вероятность их принадлежности к разным классам ниже, чем для первого кластера. Это позволяет увеличить значение параметра соседства без ущерба качеству классификации.

В качестве оценки плотности данных в области пространства классифицируемого объекта может быть выбрана плотность кластера, покрывающего данную область пространства.

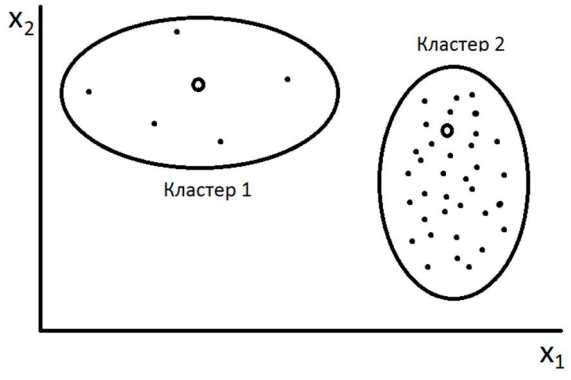

Рисунок 2 – Кластеры с различной плотностью данных
Figure 2 – Clusters with different data densities

Под плотностью кластера будем понимать среднее расстояние между объектами кластера и его центроидом. Центроид (центр масс, центр тяжести) представляет собой вектор, число элементов которого равно размерности пространства признаков, а значение каждого элемента – среднее значение признака по всем объектам, попавшим в кластер, т.е.

$$\mathbf{C}_j = \{e_{1j}, e_{2j}, \ldots, e_{pj}\},$$

$$e_{tj} = \frac{1}{m_j}\sum_{i=1}^{m_j} x_{tj},$$

где $\mathbf{C}_j$ – центроид j-го кластера, e_{tj} – элемент центроида j-го кластера, x_{tj} – соответствующий t-му измерению данных. Тогда плотность кластера будет определена следующим образом:

$$D_j = \frac{1}{m_j}\sum_{i=1}^{m_j}(a_{ij} - \mathbf{C}_j)^2,$$

где a_{ij} – i-й объект j-го кластера.

Поскольку на практике расстояние между объектами в пространстве признаков может значительно изменяться в зависимости от природы данных и особенностей задачи анализа, необходимо выполнить нормализацию плотности кластеров. Для этого будем использовать минимаксный метод:

$$D_{j\text{н}} = \frac{D_j - D_{jmin}}{D_{jmax} - D_{jmin}}.$$

В результате плотность кластеров будет лежать в диапазоне от 0 до 1.

Таким образом, внутри каждого кластера параметр соседства k будет различным и определяться в зависимости от плотности кластера. При этом если плотность кластера высокая, то увеличение значения параметра соседства приводит к увеличению качества классификации за счёт вовлечения в «голосование» большего числа примеров, расположенных близко к классифицируемому объекту. Напротив, увеличение параметра k в случае разреженного кластера приводит к участию в классификации примеры, удалённые от классифицируемого объекта, что негативно сказывается на точности классификации. Таким образом, зависимость между параметром k и плотностью кластера должна быть обратной.

При определении значения параметра k для кластера необходимо учесть следующие соображения. Минимальное значение, принимаемое параметром, $k_{min} = 1$. В этом случае имеет место так называемый наивный k-NN классификатор, который присваивает любому объекту класс, которому принадлежит его ближайший сосед. Максимальное значение $k_{max} = m_j$, т.е. числу примеров, распределённых в кластер. Тогда учёт плотности при вычислении параметра может быть произведён следующим образом:

$$k_j = m_j^{D_j}.$$

Несложно увидеть, что если в кластер попал единственный объект, который, очевидно, является и его центроидом, то плотность будет равна 0, а $k_j = 1$. В этом случае имеем наивный k-NN классификатор, который присваивает любому объекту из области пространства признаков, покрываемой кластером, класс единственного объекта.

Если плотность кластера растёт и $D_{j\text{н}} \to 1$, то $k_j \to m_j$, т.е. число примеров, с использованием которых производится классификация, стремится к числу объектов в кластере. Иными словами, чем выше плотность кластера, тем больше примеров мы можем использовать для классификации без снижения её качества из-за участия в «голосовании» примеров, удалённых в пространстве признаков от классифицируемого.

Экспериментальные исследования

Экспериментальные исследования точности предлагаемого подхода производились на примере практической задачи классификации полей по уровню урожайности ячменя на основе данных агрохимического исследования почв [3]. Входными признаками классификационной модели на основе k-NN-алгоритма являлись кислотность почвы (pH), а также содержание макроэлементов: азота (мг/100 г), фосфора (мг/100 г) и калия (мг/100 г). Исходный

набор данных содержит 56 полей, для которых указанные характеристики известны, а также известна урожайность, изменяющаяся от 1 до 12 ц/га. Для задачи классификации урожайность была категорирована на классы: от 1 до 4 ц/га – низкая, от 5до 8 ц/га – средняя, 9 – 12 ц/га – высокая. Фрагмент соответствующего набора данных представлен на рисунке 3.

№ поля	Площадь, га	Кислотность, pH	Азот, мг/100г	Калий, мг/100г	Фосфор, мг/100г	Урожайность (класс)
1	183,01	4,8	3	19,3	18,52	Низкая
2	185,99	5,5	5	10,83	19,15	Средняя
3	47,52	5,6	6	12,76	16	Средняя
4	63,04	5,6	6	12,76	16	Средняя
5	141,08	5,2	4	24,42	19,55	Низкая
6	100,06	5,2	4	14,89	10,8	Низкая
7	18,15	5	3	7,4	6	Низкая
8	99,64	5,4	5	15,97	25,2	Средняя
9	91,36	5,1	4	9,31	6,94	Низкая
10	118,06	5,6	6	11,17	13,73	Средняя
11	73,24	4,9	3	13,15	12,32	Низкая
12	128,73	6,1	9	6,91	16,11	Средняя
13	40,87	6,1	9	6,43	12,93	Средняя

Рисунок 3 – Обучающий набор данных для предсказания класса урожайности
Figure 3 – Yield class prediction training dataset

Целью анализа является выявление зависимости урожайности от агрохимических параметров для разработки комплекса агрохимических мероприятий, направленных на её повышение. Кроме этого, для новых полей, для которых измерены приведённые характеристики, может быть предсказан класс урожайности и по результатам принято решение о включении поля в севооборот.

Для визуальной оценки пространственного распределения объектов построим диаграмму размещения (рисунок 4). Для лучшего визуального восприятия приведены только три измерения данных. Тем не менее даже простой визуальный анализ позволяет увидеть неоднородность набора данных по пространственной плотности: в области 1 данные расположены более разреженно, чем в области 2.

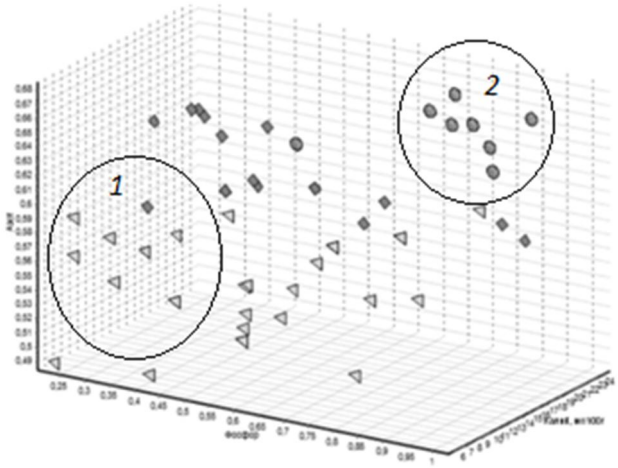

Рисунок 4 – Диаграмма размещения точек данных в пространстве признаков
Figure 4 – Diagram of data points placement in the feature space

Для оценки точности классического алгоритма *k*-NN сначала была проведена перекрестная проверка по методу «one leave – one out». На каждой итерации один объект исходного набора данных предъявлялся модели для классификации, а остальные примеры использовались в качестве обучающих. Перекрёстная проверка повторялась для значений параметров соседства, изменявшихся от 2 до 10. На втором этапе эксперимента перекрёстная проверка проводилась для модифицированного алгоритма, после чего результаты сравнивались. Кро-

ме этого, проводилось исследование зависимости точности модифицированного алгоритма от числа кластеров.

Прекластеризация исходных данных производилась с помощью алгоритма *k*-средних для числа кластеров, равного трём (начальное приближение числа кластеров взято равным числу классов). На рис. 5 представлена сформированная кластерная структура, визуализированная с помощью самооранизующейся карты признаков (карты Кохонена).

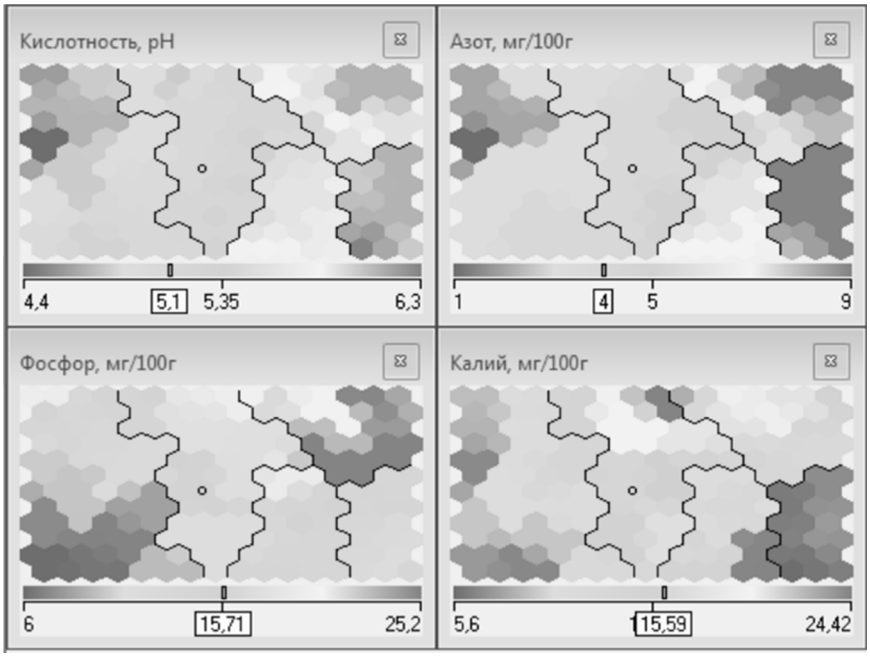

Рисунок 5 – Кластерная структура, используемая в алгоритме *k*-NN, визуализированная с помощью карты Кохонена

Figure 5 – Cluster structure used in *k*-NN algorithm visualized with Kohonen map

Для оценки качества классификации введём в рассмотрение функционал ошибки вида:

$$Q(a, X) = \frac{1}{N}\sum_{i=1}^{N}\left[a(x_i) \neq y_i\right].$$

Фактически, в соответствии с нотацией Айверсона, он возвращает 1 если пример не распознан, и 0 – в противном случае и усредняет результаты по всем классифицированным наблюдениям. Сравнения результатов перекрестной проверки для исходного алгоритма (пунктирная линия) и алгоритма с прекластеризацей (сплошная линия)

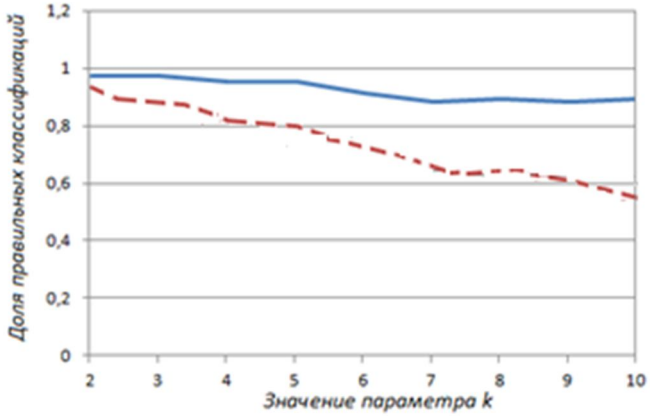

Рисунок 6 – Сравнение точности классификации для простого алгоритма *k*-NN и с использованием прекластеризаци

Figure 6 – Comparison of classification accuracy for a simple *k*-NN algorithm and using pre-clustering

Из графика видно, что алгоритм *k*-NN с прекластеризацией обеспечивает большую долю правильно классифицированных примеров по результатам перекрёстной проверки, чем классический алгоритм. При этом, когда значение параметра соседства становится больше 5, доля правильно распознанных примеров на обычном алгоритме падает, в то время как на модифицированном алгоритме поддерживается неизменной.

Таким образом, алгоритм с прекластеризацией позволяет использовать в работе *k*-NN классификаторов большие значения параметра соседства, что повышает обобщающую способность модели при работе с новыми данными. Это связано с тем, что принятие решения о принадлежности к классу производится с помощью большего числа примеров, которые отражают значимые структуры и зависимости от данных.

В качестве недостатка предлагаемого подхода следует отметить, что увеличиваются вычислительные затраты, поскольку кроме, собственно, классификации требуется выполнить кластеризацию данных. Тем не менее данный недостаток нельзя назвать критическим, поскольку алгоритм *k*-NN изначально ориентирован на решение задач, в которых быстродействие не является главным фактором, а на первое место выходит простота реализации и интерпретируемость результатов. Кроме этого, алгоритм позволяет легко добавлять новые обучающие данные без перестроения всей модели, может работать с любым числом классов и при этом сам использует всего два параметра - параметр соседства и расстояния между объектами. Ещё одним немаловажным положительным аспектом предложенной модификации алгоритма является то, что пользователь избавляется от процедуры самостоятельного определения параметра соседства. Это позволяет легко встраивать алгоритм в сценарии анализа данных, реализуемые в программных аналитических платформах.

Заключение

Таким образом, в статье предложен алгоритм адаптивного расчёта параметра соседства для алгоритма классификации *k*-ближайших соседей в зависимости от плотности обучающих примеров в области пространства признаков, куда попадает классифицируемый объект. Показаны причины снижения точности классификации для разреженных областей пространства признаков, а также потери обобщающей способности классификатора при использовании малых значений параметра соседства.

Расчёт параметра производится на основе определения плотности точек данных, которая оценивается как их среднеквадратичное расстояние до центроида кластера, покрывающего данную область пространства.

Предложенная методика позволяет использовать большие значения параметра соседства без снижения точности классификации. Это, в свою очередь, повышает обобщающую способность модели за счёт привлечения к процессу классификации большего числа обучающих примеров.

Работоспособность модифицированного алгоритма экспериментально проверена на практическом наборе данных для классификации полей по уровню урожайности на основе данных агрохимического обследования почв. При этом был показан выигрыш в точности модифицированного алгоритма относительно классического на 3 – 10 % для малых значений параметра соседства и на 20 – 30 % для больших. В качестве недостатка алгоритма отмечена необходимость проведения прекластеризации данных, что увеличивает вычислительные затраты. Преимуществами являются повышение точности и обобщающей способности классификатора и автоматический выбор значения параметра соседства.

Библиографический список

1. **Бакулева М. А., Корячко В. П., Орешков В. И.** Интеллектуальные системы и нечеткая логика. М.:КУРС, НИЦ ИНФРА-М, 2017. 352 с.

2. **Васильев Е. П., Орешков В. И.** Моделирование урожайности на основе данных агрохимического обследования почв с помощью метода ассоциативного анализа. Вестник РГАТУ. 2012. № 4 (16). С. 8-13.

3. **Паклин Н. Б., Орешков В. И**. Бизнес-аналитика: от данных к знаниям. 2-е изд., перераб. и доп. СПб.: Питер, 2010. 704 с.: ил.

4. **Baoli Li, Shiwen Yu, Qin Lu** An Improved k-Nearest Neighbor Algorithm for Text Categorization Expert Systems with Applications 39(1):1503-1509.

5. **Luo X, Yu J.X., Li Z. (Eds.):** kNN Algorithm with Data-Driven *k* Value. International Conference on Advanced Data Mining and Applications – ADMA 2014. Springer, Cham. pp 499-512. ISBN 978-3-319-14716-1.

6. **Lall U., Sharma A.:** A nearest neighbor bootstrap for resampling hydrologic time series Water Resources Research, vol. 32(Issue 3), pp. 679-693 (1996). ISSN:1944-7973.

UDC 004.855.5

INCREASING DATA CLASSIFICATION ACCURACY USING k NEAREST NEIGHBORHOOD ALGORITHM BASED ON TRAINING DATA PRE-CLUSTERIZATION

V. I. Oreshkov, Ph.D. (in technical sciences.), associate professor, CAD department, RSREU, Ryazan, Russia; orcid.org/0000-0003-0316-4927, e-mail: vyacheslav.oreshkov@yandex.ru

The problem of data classification using the metric method of machine learning of k nearest neighbors is considered. ***The aim of the work*** *is to develop a methodology for adaptive adjustment of the algorithm's k nearest neighbors neighborhood parameter in order to improve the accuracy of the algorithm by taking into account the heterogeneity of the distribution of training examples in feature space. In those areas of the feature space where the density of training examples is high, the algorithm works well even for any values of neighborhood parameter, since all the examples it defines are compactly located and, as a rule, belong to the same class. In sparse areas of the feature space, large values of neighborhood parameter lead to the examples located at a large distance, which may belong to different classes, being involved in the classification process which degrades the classification accuracy. To solve this problem, the article offers to use the value of neighborhood parameter, which is not set similar for all examples, but can change depending on their density in the area of the feature space. For this, preliminary clustering of the training set is performed and the density of each cluster is determined as a root-mean-square distance from examples of a cluster to its centroid. Based on the obtained density value, a neighborhood parameter for the cluster being used to classify any object within the cluster is calculates. The experiment showed more accurate operation of the modified algorithm in cross-validation process.*

Key words: *data mining, machine learning, supervised learning, training example, classification, clustering, class, cluster, centroid, cross validation.*

DOI: 10.21667/1995-4565-2021-76-65-73

References

1. **Bakuleva M. A., Korjachko V. P., Oreshkov V. I.** *Intellektual'nye sistemy i nechetkaja logika*. Moscow: KURS, NIC INFRA-M, 2017. 352 p. (in Russian).

2. **Vasil'ev E. P., Oreshkov V. I.** Modelirovanie urozhajnosti na osnove dannyh agro-himicheskogo obsledovanija pochv s pomoshh'ju metoda associativnogo analiza. *Vestnik RGATU*. 2012, no. 4 (16), pp. 8-13 (in Russian).

3. **Paklin N. B., Oreshkov V. I.** *Biznes-analitika: ot dannyh k znanijam. 2-e izd., pererab. i dop*. SPb.: Piter, 2010. 704 p.: il (in Russian).

4. **Baoli Li, Shiwen Yu, Qin Lu** *An Improved k-Nearest Neighbor Algorithm for Text Categorization Expert Systems with Applications* 39(1):1503-1509.

5. **Luo X, Yu J.X., Li Z. (Eds.):** kNN Algorithm with Data-Driven *k* Value. *International Conference on Advanced Data Mining and Applications – ADMA 2014*. Springer, Cham, pp. 499-512. ISBN 978-3-319-14716-1.

6. **Lall U., Sharma A.:** *A nearest neighbor bootstrap for resampling hydrologic time series Water Resources Research*. vol. 32(Issue 3), pp. 679-693 (1996). ISSN:1944-7973.