НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
март–апрель 2024 Том 24 № 2 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
March–April 2024 Vol. 24 No 2 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2024-24-2-322-329
УДК 004.8

Цензурирование обучающих выборок с использованием регуляризации отношений связанности объектов классов

Николай Александрович Игнатьев[1]✉, Даврбек Худаёрович Турсунмуротов[2]

[1,2] Национальный университет Узбекистана имени Мирзо Улугбека, Ташкент, 100174, Узбекистан

[1] n_ignatev@rambler.ru✉, https://orcid.org/0000-0002-7150-5837
[2] mr.davrbek@mail.ru, https://orcid.org/0009-0009-8664-9639

Аннотация

Введение. Рассмотрено цензурирование обучающих выборок с учетом специфики реализации алгоритмов метода ближайшего соседа. Процесс цензурирования связан с использованием множества граничных объектов классов по заданной метрике с целью: поиска и удаления шумовых объектов; анализа кластерной структуры обучающей выборки по отношению связанности. Исследуются специальные условия удаления шумовых объектов и формирования базы прецедентов для обучения алгоритмов. Распознавание объектов по такой базе должно обеспечивать более высокую точность с минимальными затратами вычислительных ресурсов относительно исходной выборки. **Метод.** Разработаны необходимые и достаточные условия для отбора шумовых объектов из множества граничных. Необходимое условие принадлежности граничного объекта к множеству шумовых задается в виде ограничения (порога) на отношение расстояний до ближайшего объекта из своего класса и его дополнения. Поиск минимального покрытия обучающей выборки эталонами производится на основе анализа кластерной структуры. Эталоны представлены объектами выборки. Структура отношений связанности объектов по системе гипершаров используется для их группировки. Состав групп формируется из центров (объектов выборки) для гипершаров, в пересечении которых содержатся граничные объекты. Значение меры компактности вычисляется как среднее число объектов обучающей выборки за вычетом шумовых, притягиваемое одним эталоном минимального покрытия. Выполняется анализ связи обобщающей способности алгоритмов при машинном обучении со значением меры компактности. Наличие связи обосновывается по критерию (регуляризатору) для отбора числа и состава множества шумовых объектов. Оптимальные коэффициенты регуляризации определяются как значения порогов для удаления шумовых объектов. **Основные результаты.** Показана связь между значением меры компактности обучающей выборки и обобщающей способностью алгоритмов распознавания. Связь выявлена по эталонам минимального покрытия выборки, из которых сформирована база прецедентов. Обнаружено, что точность распознавания по базе прецедентов выше, чем на исходной выборке. Минимальный состав базы прецедентов включает описания эталонов и параметры локальных метрик. При использовании процедур нормирования данных требуются дополнительные параметры. Анализ значений меры компактности востребован для обнаружения переобучения алгоритмов, связанного с размерностью признакового пространства. Распознавание по базе прецедентов минимизирует затраты вычислительных ресурсов с помощью алгоритмов метода ближайшего соседа. **Обсуждение.** Приводятся рекомендации по разработке моделей из области информационной безопасности, для обработки и интерпретации данных социологических исследований. Для использования в информационной безопасности формируется база прецедентов для идентификации DDOS-атак. Новые знания из области социологии предлагается получать через анализ значений показателей шумовых объектов и интерпретацию результатов разбиения респондентов на непересекающиеся группы по отношению к связанности объектов. Конфигурации групп по отношению связанности изначально не известны. Нет смысла вычислять их центры, которые могут размещаться за пределами конфигураций. Для объяснения содержимого групп предложено использовать эталоны минимального покрытия.

Ключевые слова

меры компактности, база прецедентов, коэффициенты регуляризации, минимальное покрытие эталонами, шумовые объекты

Благодарности

Работа выполнена в рамках плана научных исследований кафедры «Искусственный интеллект» Национального университета Узбекистана.

Ссылка для цитирования: Игнатьев Н.А., Турсунмуротов Д.Х. Цензурирование обучающих выборок с использованием регуляризации отношений связанности объектов классов // Научно-технический вестник информационных технологий, механики и оптики. 2024. Т. 24, № 2. С. 322–329. doi: 10.17586/2226-1494-2024-24-2-322-329

Censoring training samples using regularization of connectivity relations of class objects

Nikolay A. Ignatev[1]✉, Davrbek X. Tursunmurotov[2]

[1,2] National University of Uzbekistan named after Mirzo Ulugbek, Tashkent, 100174, Uzbekistan

[1] n_ignatev@rambler.ru✉, https://orcid.org/0000-0002-7150-5837
[2] mr.davrbek@mail.ru, https://orcid.org/0009-0009-8664-9639

Abstract

The censoring of training datasets is considered taking into account the specific implementation of the nearest neighbor method algorithms. The censoring process is associated with the use of a set of boundary objects of classes according to a given metric for the purpose of: searching and removing noise objects and analyzing the cluster structure of the training sample in relation to connectivity. Special conditions for removing noise objects and forming a precedent base for training algorithms are explored. Recognition of objects using such a database should provide higher accuracy with minimal computational resources relative to the original dataset. Necessary and sufficient conditions for selecting noise objects from a set of boundary ones have been developed. The necessary condition for a boundary object to belong to the noise set is specified in the form of a restriction (threshold) on the ratio of the distances to the nearest object from its class and its complement. The search for the minimum coverage of the training dataset with standards is carried out based on the analysis of the cluster structure. The standards are represented by sample objects. The structure of the connectivity relations of objects according to the hypersphere system is used to group them. The composition of the groups is formed from centers (dataset objects) for hyperspheres the intersection of which contains boundary objects. The value of the compactness measure is calculated as the average number of objects in the training dataset, excluding noise, pulled in by one standard of minimum coverage. An analysis is carried out of the connection between the generalizing ability of algorithms in machine learning and the value of the compactness measure. The presence of a connection is justified by a criterion (regularizer) for selecting the number and composition of a set of noise objects. Optimal regularization coefficients are defined as threshold values for removing noise objects. The relationship between the value of the training dataset compactness measure and the generalizing ability of recognition algorithms is shown. The connection was identified using the standards of minimum sample coverage from which the precedent base was formed. It was found that the recognition accuracy using the precedent base is higher than that using the original dataset. The minimum composition of the precedent base includes descriptions of standards and parameters of local metrics. When using data normalization procedures, additional parameters are required. Analysis of the values of the compactness measure is in demand to detect overfitting of algorithms associated with the dimension of the feature space. Recognition based on precedents minimizes the cost of computing resources using nearest neighbor algorithms. Recommendations are given for the development of models in the field of information security for processing and interpreting sociological research data. For use in information security, a precedent base is being formed to identify DDOS attacks. It is proposed to obtain new knowledge from the field of sociology through the analysis of the values of indicators of noise objects and the interpretation of the results of dividing respondents into non-overlapping groups in relation to the connectedness of objects. The configurations of groups in relation to connectivity are not initially known. There is no point in calculating their centers which can be located outside the configurations. To explain the contents of groups, it is proposed to use standards of minimum coverage.

Keywords

compactness measures, precedent base, regularization coefficients, minimum coverage with standards, noise objects

Acknowledgments

The work was carried out within the framework of the scientific research plan of the Department of Artificial Intelligence of the National University of Uzbekistan.

For citation: Ignatev N.A., Tursunmurotov D.X. Censoring training samples using regularization of connectivity relations of class objects. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2024, vol. 24, no. 2, pp. 322–329 (in Russian). doi: 10.17586/2226-1494-2024-24-2-322-329

Введение

Цензурирование обучающих выборок рассматривается как важная часть процесса машинного обучения. Основными целями цензурирования являются повышение обобщающей способности и снижение сложности алгоритмов. Реализация этих целей, как правило, связана с формированием обучающих выборок через поиск и удаление шумовых объектов и признаков [1–3]. Специфику цензурирования для метрических алгоритмов распознавания в работе [2] предложено рассматривать через меру компактности объектов классов и выборки в целом.

Процедура цензурирования в [4] предусматривала коррекцию диагностируемых объектов на примере выборок данных из области медицины. Смысл коррекции заключался в удалении таких объектов или исправлении ошибки в диагностируемом (целевом) признаке. Реализация процедуры основана на анализе изменений в оценке разделимости объектов классов, вычисляемой до и после внесения исправлений при использовании функции конкурентного сходства. Отказ от фильтрации

испорченных объектов мотивировался тем, что такие действия могут отрицательно отразиться на представительности обучающей выборки.

В работе [5] отмечено наличие влияния шумовых объектов на структуру отношений объектов обучающих выборок в метрических алгоритмах. Утверждалось, что множество шумовых объектов является подмножеством граничных по заданной метрике. Интерес к граничным объектам связан с использованием их для вычисления отношений связанности объектов классов по системе гипершаров. Отношение связанности применяется для разбиения объектов классов на непересекающиеся группы и вычисления по ним минимального покрытия обучающей выборки эталонами. Эталоны представлены объектами выборки. Предложены две меры компактности для оценки:

1) структуры отношений объектов в интервале (0; 1];
2) обобщающей способности алгоритмов распознавания.

Множество допустимых значений меры 1 в интервале (0; 1] зависит от количества групп в каждом классе и их мощности. Мера 2 определяется как среднее число объектов выборки за вычетом шумовых, притягиваемое одним эталоном минимального покрытия. Эта мера предлагается для оценки обобщающей способности алгоритмов в качестве альтернативы известному методу кросс-валидации. Число и состав шумовых объектов после их удаления меняет конфигурацию граничных и, как следствие, мощность множества эталонов покрытия. Число эталонов покрытия служат показателем представительности обучающей выборки.

Процесс поиска минимального покрытия в работе [5] реализован жадным алгоритмом. По этой причине шумовые объекты могут быть выбраны в качестве эталонов и повлиять на обобщающую способность алгоритмов распознавания в сторону ее уменьшения. Поскольку задачи распознавания являются некорректными, встает вопрос о наличии оптимального решения — получения максимума значения меры компактности. Поиск оптимального решения осуществлен через соотношение между числом эталонов минимального покрытия и числом определяемого состава удаляемых шумовых объектов. Для проверки принадлежности к множеству шумовых объектов предложено использовать дополнительный критерий — регуляризатор.

Меры компактности в отличии от метода кросс-валидации не являются средством для вычисления точности распознавания. Введение регуляризатора позволяет упорядочивать значения меры в зависимости от разных факторов, определяющих структуру отношений между объектами обучающей выборки и использовать их для анализа. Уменьшение значения меры при добавлении признаков в набор может рассматриваться как индикатор «проклятия размерности». Анализ порядка следования значений востребован при выборе метрики для вычисления расстояния между объектами и способов нормирования данных, отборе информативных наборов признаков.

Предмет исследования

В работе [6] описаны особенности использования эвристических метрик в методах типа «ближайших соседей». При реализации данных методов рассматриваются некоторые соотношения расстояний между различными парами объектов. Метрики, которые порождают совпадающие соотношения расстояний на множестве описаний объектов, оказываются эквивалентными при вычислении значений мер компактности. Изучение причин несовпадения значений компактности требует иных методов анализа с целью использования их результатов для цензурирования.

Существует потребность в проверке гипотезы о наличии последовательности признаков, удаление части из которых, согласно порядку их следования, обеспечивает монотонное неубывание значений меры компактности. Достоверность значений меры возрастает за счет использования регуляризаторов. Свойство монотонности снижает комбинаторную сложность процесса отбора информативных признаков.

Обобщающая способность алгоритма метода ближайшего соседа зависит от структуры отношений по множеству граничных объектов классов [5]. Как правило, эта структура меняется при удалении шумовых объектов из выборки. По сути дела, на обучающей выборке рассматриваются исходное и переопределенное множества граничных объектов, полученные до и после удаления шумовых объектов.

Удаление части граничных объектов (шумовых) рассматривается как способ поиска отступа между классами с целью повышения обобщающей способности алгоритмов распознавания. Отступ является относительной величиной, которая влияет на вычисление значений параметров локальных метрик и состав эталонов минимального покрытия выборки.

Для включения граничного объекта в состав шумовых используется отношение расстояний до двух ближайших объектов из своего класса и его дополнения. Утверждается, что для получения оптимального с точки зрения точности распознавания решения необходимо вводить ограничения (пороги) на значения оценок отношений. Выбор порогов (отступов) связан с решением задачи поиска максимального значения меры компактности на обучающей выборке по критерию регуляризации.

Численное решение задачи регуляризации структуры отношений объектов по мере компактности и формирование баз прецедентов через минимальное покрытие выборки эталонами в настоящей работе предложено впервые. Вероятным объяснением отсутствия аналогичных научных работ является то, что сложность решения задачи о минимальном покрытии оценивается как NP-полная. Практически получить решение поставленной задачи невозможно, так как требуется выполнить полный перебор всех вариантов. Для исключения полного перебора вариантов осуществлена предобработка данных через группировку объектов по отношению их связанности по системе пересекающихся гипершаров. Поиск эталонов после предобработки производится по каждой группе в отдельности.

Постановка задачи. Минимальное покрытие выборки эталонами

Рассмотрим задачу распознавания в стандартной постановке. Будем считать, что задано множество из m объектов $E = \{S_1, \ldots, S_m\}$, разделенное на l непересекающихся классов $K_1, \ldots, K_l$. Описание объектов производится с помощью n разнотипных признаков $X(n) = (x_1, \ldots, x_n)$, ξ из которых измеряются в интервальных шкалах, $n - \xi$ — в номинальной. На множестве объектов E_0 задана метрика $\rho(x, y)$.

Введем обозначения множеств: $B(E, \rho) = \left\{S \in E | \rho(S_i, S) = \min\limits_{S_i \in K_j,\, S_d \in CK_j} \rho(S_i, S_d)\right\}$ — граничных объектов классов; $T \subset B(E_0, \rho)$ — шумовых объектов, определяемых на E_0 по метрике $\rho(x, y)$; $E = E_0 \backslash T$.

Объекты $S_i, S_j \in K_t$, $t = 1, \ldots, l$ считаются связанными между собой $S_i \leftrightarrow S_j$, если $\{S \in B(E, \rho) | \rho(S, S_i) < r_i$ и $\rho(S, S_j) < r_j\} \neq \emptyset$, где $r_i(r_j)$ — расстояние до ближайшего от $S_i(S_j)$ объекта из дополнения $CK_t (CK_t = E \backslash K_t)$ к K_t по метрике $\rho(x, y)$. Множество $G_{tv} = \{S_{v1}, \ldots, S_{vc}\}$, $c \geq 2$, $G_{tv} \subset K_t$, $v < |K_t|$ представляет область (группу) со связанными объектами в классе K_t, если для любых S_{vi}, $S_{vj} \in G_{tv}$, существует путь $S_{vi}, \leftrightarrow S_{vk} \leftrightarrow \ldots \leftrightarrow S_{vj}$. Объект $S_i \in K_t$, $t = 1, \ldots, l$, принадлежит группе из одного элемента и считается несвязанным, если не существует пути $S_i \leftrightarrow S_j$ ни для одного объекта $S_j \neq S_i$ и $S_j \in K_t$.

Считается, что на множестве E определен жадный алгоритм формирования множества эталонов минимального покрытия E_{ob} и вычисления меры компактности

$$\mu(E, \rho) = |E|/|E_{ob}|. \quad (1)$$

Близость к эталону $S \in E_{ob} \cap K_t$ вычислим по локальной метрике $\rho_s(x, y) = \alpha_s \rho(x, y)$, где α_s — параметр, определяемый по граничным объектам из $E \cap CK_t$.

Требуется определить мощность множества шумовых объектов T и его состав, при котором

$$\mu(E, \rho) = \max_{T \subset E_0} \mu(E_0 \backslash T, \rho). \quad (2)$$

Процесс формирования минимального покрытия обучающей выборки эталонами [5] реализуется путем последовательного выполнения следующих этапов:

— выделение множества граничных объектов классов $B(E_0, \rho)$ по заданной метрике $\rho(x, y)$;
— поиск и удаление шумовых объектов $T \subset B(E_0, \rho)$ из множества граничных;
— разбиение объектов классов на непересекающиеся группы по отношению связанности по множеству граничных на $E = E_0 \backslash T$;
— формирование минимального покрытия из эталонов по каждой группе.

Вычисление минимального покрытия эталонами E_0 и значения меры компактности (1) служат основой для формирования многообразия баз прецедентов для реализации алгоритмов метода «ближайший сосед». Условия изменения отношений между объектами выборки, часть из которых перечислена в [5], объясняют причину появления такого многообразия.

Оптимизационная постановка задачи отбора эталонов, основанная на минимизации функционала полного скользящего контроля, рассмотрена в работе [7]. При разделении выборки на обучающую и контрольную предполагалось вхождение в них объектов из множества эталонов. Открытыми оставались проблемы численного решения оптимизационной задачи применительно к большим данным.

Для снижения вычислительной сложности задачи отбора эталонов в настоящей работе использована предобработка данных. Процесс предобработки заключается в разбиении объектов выборки на непересекающиеся группы по каждому классу. Считается, что с ростом объема выборки и удаления из нее выбросов снижается вариабельность конфигурации выборки по множеству граничных объектов классов. На снижение вариабельности должны указывать уменьшение разброса значений: числа непересекающихся групп объектов классов, определяемых по отношению связанности по системе гипершаров; числа эталонов минимального покрытия, вычисляемого по группам объектов классов; величины зазора (отступа) между объектами классов.

Проверка состоятельности данных значений необходима для оценки обобщающей способности алгоритмов. Использование асимптотических способов оценки для этих целей, как правило, представляет лишь теоретический интерес. Рекомендуемая асимптотическими способами завышенная длина обучающих выборок для практической реализации неприменима.

Выбор латентных признаков в пространстве большой размерности можно рассматривать как способ решить проблему проклятия размерности для метрических алгоритмов. Каждый латентный признак, синтезированный по методу обобщенных оценок [8], представляется как ансамбль из элементарных классификаторов по технологии стекинга. Доказано, что метод кросс-валидации для проверки точности распознавания ансамблем на контрольных выборках неприменим. По этой причине значения меры компактности (2) можно задействовать для формирования баз прецедентов на разных наборах латентных признаков.

Поиск шумовых объектов из множества граничных

Определение типичности объекта из E_0 (близок к своему классу, является граничным, близок к дополнению класса) по значению функции конкурентного сходства в алгоритме FRIS–STOLP [4] применяется при формировании множества из шумовых и эталонных объектов. Выполним оценку типичности граничного объекта по отношению двух ближайших от него расстояний до объектов из своего класса и его дополнения. Решение о включении (не включении) граничного объекта в множество шумовых принимается на основе анализа этого отношения. Для анализа требуется определить пороговое значение λ и условия, на основе которых принимается решение.

На множестве граничных объектов $B = B(E_0, \rho)$ сформируем множество пар $BG = \{(S_i, S_j)\}$, $S_i \in K_t \cap B$, $t \geq 2$, $S_j \in CK_t \cap B$, $\rho(S_i, S_j) = \min\limits_{S_v \in B \cap CK_t} \rho(S_i, S_v)$. Для

$(S_i, S_j) \in BG$ введем обозначения $r(S_i) = \rho(S_i, S_j)$, $d(S_i) = \rho(S_i, S_v)$, где $S_v = \arg \min_{S_a \in E_0 \cap K_t \setminus \{S_i\}} \rho(S_i, S_a)$. Аналогично для $S_j \in CK_t \cap B$ определим $r(S_j) = \rho(S_k, S_j) = \min_{S_v \in B \cap K_t} \rho(S_v, S_j)$, $d(S_j) = \rho(S_j, S_\mu)$, где $S_\mu = \arg \min_{S_a \in E_0 \cap CK_t \setminus \{S_j\}} \rho(S_k, S_a)$. Отношение $\frac{r(S_i)}{d(S_i)} < \lambda$, $0 < \lambda < 1$ рассматривается как необходимое условие отнесение объекта $S_i \in K_t \cap B$ к множеству шумовых. Достаточным условием является

$$\frac{r(S_i)}{d(S_i)} < \lambda \text{ и } \frac{r(S_j)}{d(S_j)} \geq \lambda . \quad (3)$$

Иллюстрация определения принадлежности граничного объекта $S_i \in K_1$ к множеству шумовых на выборке $E_0 = K_1 \cup K_2$ по отношениям расстояний $r(S_i)$, $d(S_j)$, $d(S_i)$ показана на рисунке.

Значение λ, определяемое по (3) в качестве параметра (коэффициента) регуляризатора, применяется для поиска экстремального значения меры компактности (2) при фиксированных факторах. Решение об эффективности выбора факторов (мера расстояния между объектами, способ нормирования, состав набора признаков и т. д.), изменяющих структуру отношений объектов, как правило, принимается по результатам вычислительного эксперимента.

При разработке программного обеспечения для распознавания произвольного допустимого объекта в приложении к базе прецедентов как минимум необходимо хранить сведения об используемой метрике и значения весов локальных метрик эталонов.

О регуляризации отношений связанности объектов классов при моделировании

Результаты, полученные при регуляризации отношений связанности объектов классов, востребованы при построении моделей, основанных на знаниях в слабоструктурированных предметных областях. Рекомендуются следующие варианты применения:

1) распознавание объектов с минимальными затратами вычислительных ресурсов;
2) анализ кластерной структуры объектов и свойств эталонов минимального покрытия;
3) исследование причин появления шумовых объектов.

Реализация варианта 1 связана с формированием базы прецедентов из эталонов минимального покрытия. Оптимальное значение меры компактности (2) позволяют отслеживать переобучение алгоритмов, связанное с проблемой проклятия размерности при машинном обучении. Индикатором переобучения служит уменьшение значения меры компактности и снижение обобщающей способности алгоритмов при росте числа признаков. Размерность пространства, выше которого фиксируется наличие переобучения, определяется при проведении вычислительного эксперимента.

Эффективность распознавания по базе прецедентов достигается за счет снижения затрат вычислительных ресурсов и возможности распараллеливания алгоритма при вычислении меры расстояния от произвольного допустимого объекта до прецедентов из базы. Минимизация ресурсов актуальна при разделении протоколов с нормальным трафиком и с DDOS-атаками (отказами в обслуживании) для борьбы с несанкционированным доступом в компьютерных сетях.

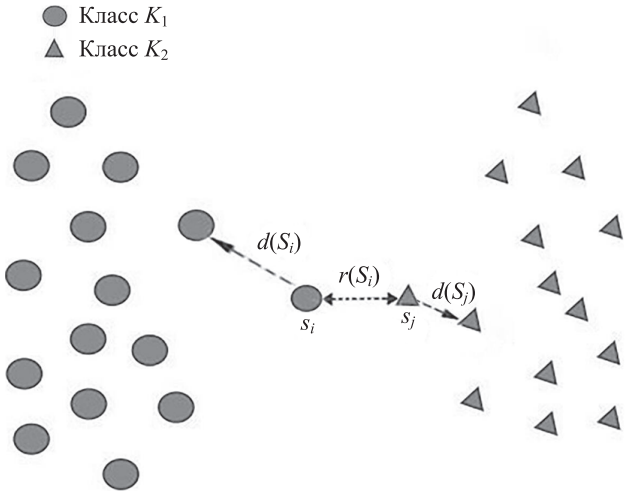

Рисунок. Отнесение граничного объекта $S_i \in K_1$ к множеству шумовых по отношениям расстояний $r(S_i)$, $d(S_j)$, $d(S_i)$

Figure. Assignment of the boundary object $S_i \in K_1$ to the noise set according to the distance relationship $r(S_i)$, $d(S_j)$, $d(S_i)$

Анализ кластерной структуры отношений объектов (вариант 2) востребован при поиске скрытых закономерностей в данных путем проверки истинности гипотез, выдвигаемых экспертом. Считается, что эксперт задает классификацию объектов и использует ее для частичного обучения при группировке. Например, в социологии такая классификация применима при идентификации респондентов с низким, средним и высоким уровнями доходов. Частичное обучение необходимо в качестве условия, что состав каждой группы представлен респондентами одного класса. С учетом этого условия алгоритм метода гарантирует единственность числа групп и их состава, которое используется при поиске минимального покрытия выборки эталонами.

По отношению связанности объектов классов конфигурация (форма) групп может быть различной. Каждую группу идентифицирует как минимум один эталон. Нет необходимости (в случае разнотипности признаков и возможности) в вычислении центров групп, во введении ограничений на шкалы измерений признаков. В качестве альтернативы центрам групп при анализе предлагается использовать эталоны покрытия.

Эталонные объекты являются предметом отдельного исследования как типичные представители групп. Шумовые (нетипичные) объекты (вариант 3) рассматриваются как выбросы или отклонения от эмпирических закономерностей. Например, по социологическим данным уровень потребления респондента существенно различается от уровня заявленных им доходов.

Вычислительный эксперимент

Для демонстрации влияния коэффициентов регуляризации на значение меры компактности (1), связи коэффициентов с обобщающей способностью алгоритмов метода ближайшего соседа были использованы данные German[1] и Spambase[2]. Связь коэффициентов регуляризации с результатами отбора шумовых и эталонных объектов на данных German по метрике Журавлёва показана в табл. 1. Описание каждого из 1000 объектов выборки при соотношении классов $|K_1|:|K_2| = 700:300$ производилось 7-ю количественными и 13-ю номинальными признаками. Значения количественных признаков в данных дробно-линейным преобразованием отображались в диапазоне [0; 1].

В качестве оптимального решения для данных German (табл. 1) рекомендуется удаление 42 шумовых объектов и отбор 260 эталонов при коэффициенте регуляризации 0,8.

При эксперименте на выборке данных Spambase количество объектов с исходных 4601 было уменьшено до 4204, из них 2528 представителей 1-го класса и 1676 — 2-го. Удалены пересекающиеся объекты из двух классов и из сходных по описанию объектов в каждом классе оставлено по одному представителю. Одна из целей эксперимента — демонстрация возможностей анализа многообразий отношений объектов как при нормировании данных в [0; 1] так и при использовании комбинаций вида $\rho(x, y) = \dfrac{\rho_1(x, y)}{1 + \rho_2(x, y)}$ по базовым метрикам Евклида и Чебышева. Оптимальные значения коэффициентов регуляризации на разных метриках по нормированным в диапазоне [0; 1] данным Spambase представлены в табл. 2.

Сильная корреляционная зависимость по (2) с базовыми метриками Евклида и Чебышева (табл. 2) получена по комбинации Евклида/(1 + Чебышева).

Для проверки эффективности отбора эталонных объектов в качестве прецедентов для обучения было произведено разбиение 4204 объектов Spambase на две равные по мощности выборки. При этом использован порядок следования четных и нечетных номеров индексов объектов в каждом классе. Каждая выборка (Chet и Nechet) применялась для обучения и контроля. Результаты отбора прецедентов по двум выборкам представлены в табл. 3. Прецедентами считаются эталоны минимального покрытия, при формировании которого использовались локальные метрики для вычисления расстояния по описаниям данных в [0; 1].

Минимальные покрытия равных по мощности выборок (табл. 3), полученные при разных значениях коэффициента регуляризации, отличаются разнообразием как по числу эталонов, так и их распределению по классам. Объясняется это способностью алгоритма обучения адаптироваться к конфигурации граничных объектов после удаления шумовых за счет выбора параметров локальных метрик эталонов.

Для проверки обобщающей способности алгоритмов распознавания в качестве прецедентов использованы эталоны минимального покрытия выборок Chet

[1] [Электронный ресурс]. Режим доступа: https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data, свободный. Яз. англ. (дата обращения: 19.01.2024).

[2] [Электронный ресурс]. Режим доступа: https://archive.ics.uci.edu/dataset/94/spambase, свободный. Яз. англ. (дата обращения: 19.01.2024).

Таблица 1. Отбор шумовых и эталонных объектов на данных German в зависимости от значений коэффициентов регуляризации по метрике Журавлёва

Table 1. Selection of noise and reference objects on German data depending on the values of regularization coefficients using the Zhuravlev metric

Коэффициент регуляризации	Число объектов		Компактность по (2)
	шумовых	эталонов	
0,5	42	267 (126, 141)	3,4373
0,6	60	259 (120, 139)	3,4116
0,7	54	259 (112, 147)	3,4553
0,8	42	260 (114, 146)	3,5299
0,9	27	277 (127, 150)	3,4178

Примечание. В скобках (табл. 1–3) указано число эталонов из 1-го и 2-го классов.

Таблица 2. Оптимальные значения коэффициентов регуляризации на данных Spambase

Table 2. Optimal values of regularization coefficients based on Spambase data

Метрика	Коэффициент регуляризации	Число объектов		Компактность по (2)
		шумовых	эталонов	
Евклида	0,7	70	395 (196, 199)	10,2915
Чебышева	0,7	115	370 (233, 137)	10,7490
Чебышева/(1 + Евклида)	0,6	82	498 (239, 259)	8,1157
Евклида/(1 + Чебышева)	0,8	83	400 (201, 199)	10,0991

Таблица 3. Результаты отбора прецедентов по выборкам Chet и Nechet
Table 3. Results of selection of precedents from the Chet and Nechet samples

Метрики		Евклида		Чебышева		Чебышева/ (1 + Евклида)		Евклида/ (1 + Чебышева)	
Выборки		Chet	Nechet	Chet	Nechet	Chet	Nechet	Chet	Nechet
Коэффициент регуляризации		0,7	0,5	0,9	0,8	0,6	0,5	0,8	0,9
Число объектов	шумовых	41	15	55	65	41	17	45	43
	эталонов	223 (113, 110)	246 (122, 124)	176 (159, 17)	210 (137, 73)	299 (154, 145)	320 (148, 172)	225 (119, 106)	239 (125, 114)
Компактность по (2)		9,0619	8,4232	11,3264	9,4000	6,7585	6,4629	8,9465	8,4388

Таблица 4. Точность распознавания по метрике Евклида
Table 4. Recognition accuracy using the Euclidean metric

Прецеденты по выборке	Контрольная выборка	
	Chet	Nechet
Chet	—	88,20 (87,01)
Nechet	88,73 (88,63)	—

Таблица 5. Точность распознавания по метрикам Евклида, Чебышева и их комбинации
Table 5. Recognition accuracy using Euclidean, Chebyshev metrics and their combinations

Метрика	Обучение	Контроль	Обучение	Контроль
	Chet	Nechet	Nechet	Chet
Евклида	Не определялось	78,35	Не определялось	77,55
Чебышева	Не определялось	71,27	Не определялось	74,02
Евклида/(1 + Чебышева)	Не определялось	83,30	Не определялось	84,11

и Nechet, число которых приведено в табл. 3. Точность распознавания при выборе ближайшего соседа на основе меры расстояния Евклида показана в табл. 4. В скобках указаны результаты по обучающим выборкам без удаления шумовых объектов и отбора эталонов. Нормирование данных на контроле проводилось по параметрам обучающей выборки.

В первой строке табл. 4 база прецедентов из 246 эталонов (табл. 3) для выборки Chet используется для тестирования 2102 объектов из Nechet. Аналогично во второй строке база прецедентов из 223 эталонов (табл. 3) для выборки Nechet применена для тестирования 2102 объектов из Chet. Точности 88,63 и 87,01, указанные в скобках, ниже результатов распознавания по базам прецедентов с применением локальных метрик. В разы снижена сложность вычислений алгоритмом по эталонам минимального покрытия.

Уменьшить ошибки распознавания через изменение структуры отношений объектов также можно по мере расстояния, вычисляемой по комбинациям метрик. В табл. 5 точность распознавания в пространстве из 57 исходных признаков демонстрируется по метрикам Евклида, Чебышева и их комбинации.

Результаты применения комбинации из метрик в табл. 5 показывают перспективность поиска различных путей для повышения эффективности систем распознавания. Эффективность может выражаться в отсутствии необходимости хранения специальных параметров и алгоритмов для предобработки данных при формировании и использовании баз прецедентов, например, для различных способов нормирования.

Заключение

Разработана новая методика формирования баз прецедентов для алгоритмов распознавания по методу ближайшего соседа. С целью повышения обобщающей способности алгоритмов предложен дополнительный критерий-регуляризатор для отбора шумовых объектов по заданному отступу между классами. Отступ определяет ограничение на отношение расстояний между граничными объектами и их ближайшими соседями. Регуляризатор использовался при вычислении максимума меры компактности по минимальному покрытию обучающей выборки эталонами после удаления шумовых объектов. Эффективность применения эталонов в качестве базы прецедентов выразилась в повышении точности распознавания при снижении затрат вычислительных ресурсов.

Литература

1. Борисова И.А., Кутненко О.А. Цензурирование ошибочно классифицированных объектов выборки // Машинное обучение и анализ данных. 2015. Т. 1. № 11. С. 1632–1641.
2. Загоруйко Н.Г., Кутненко О.А. Цензурирование обучающей выборки // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. 2013. № 1(22). С. 66–73.
3. Кутненко О.А., Плясунов А.В. NP-трудность некоторой задачи цензурирования данных // Дискретный анализ и исследование операций. 2021. Т. 28. № 2(148). С. 60–73. https://doi.org/10.33048/daio.2021.28.692
4. Борисова И.А., Кутненко О.А. Исправление диагностических ошибок в целевом признаке с помощью функции конкурентного сходства // Математическая биология и биоинформатика. 2018. Т. 13. № 1. С. 38–49. https://doi.org/10.17537/2018.13.38
5. Ignatyev N.A. Structure choice for relations between objects in metric classification algorithms // Pattern Recognition and Image Analysis. 2018. V. 28. N 4. P. 695–702. https://doi.org/10.1134/s1054661818040132
6. Рудаков К.В. О некоторых факторизациях полуметрических конусов и оценках качества эвристических метрик в задачах анализа данных // Доклады Российской Академии наук. Математика, Информатика, Процессы Управления. 2020. Т. 492. № 1. С. 101–103. https://doi.org/10.31857/S2686954320030236
7. Зухба А.В. Оценка вычислительной сложности задач отбора эталонных объектов и признаков: диссертация на соискание ученой степени кандидата физико-математических наук наук. М., 2018. 113 с.
8. Ignatev N.A., Rahimova M.A. Formation and analysis of sets of informative features of objects by pairs of classes // Scientific and Technical Information Processing. 2022. V. 49. N 6. P. 439–445. https://doi.org/10.3103/S0147688222060053

References

1. Borisova I.A., Kutnenko O.A. Outliers detection in datasets with misclassified objects. *Machine Learning and Data Analysis*, 2015, vol. 1, no. 11, pp. 1632–1641. (in Russian)
2. Zagoruiko N.G., Kutnenko O.A. Training dataset censoring. *Tomsk State University Journal of Control and Computer Science*, 2013, no. 1(22), pp. 66–73. (in Russian)
3. Kutnenko O.A., Plyasunov A.V. NP-hardness of some data cleaning problem. *Journal of Applied and Industrial Mathematics*, 2021, vol. 15, no. 2, pp. 285–291. https://doi.org/10.1134/S1990478921020095
4. Borisova I.A., Kutnenko O.A. The problem of correction diagnostic errors in the target attribute with the function of rival similarity. *Mathematical Biology and Bioinformatics*, 2018, vol. 13, no. 1, pp. 38–49. (in Russian). https://doi.org/10.17537/2018.13.38
5. Ignatyev N.A. Structure choice for relations between objects in metric classification algorithms. *Pattern Recognition and Image Analysis*, 2018, vol. 28, no. 4, pp. 695–702. https://doi.org/10.1134/s1054661818040132
6. Rudakov K.V. On some factorizations of semi-metric cones and quality estimates of heuristic metrics in data analysis problems. *Doklady Mathematics*, 2020, vol. 101, no. 3, pp. 257–258. https://doi.org/10.1134/S1064562420030230
7. Zukhba A.V. *Computational complexity estimation of the problems of selecting reference objects and features*. Dissertation for the degree of candidate of physical and mathematical sciences. Moscow, 2018, 113 p. (in Russian)
8. Ignatev N.A., Rahimova M.A. Formation and analysis of sets of informative features of objects by pairs of classes. *Scientific and Technical Information Processing*, 2022, vol. 49, no. 6, pp. 439–445. https://doi.org/10.3103/S0147688222060053

Авторы

Игнатьев Николай Александрович — доктор физико-математических наук, профессор, профессор, Национальный университет Узбекистана имени Мирзо Улугбека, Ташкент, 100174, Узбекистан, SC 39361638900, https://orcid.org/0000-0002-7150-5837, n_ignatev@rambler.ru

Турсунмуротов Даврбек Худаёрович — докторант, Национальный университет Узбекистана имени Мирзо Улугбека, Ташкент, 100174, Узбекистан, https://orcid.org/0009-0009-8664-9639, mr.davrbek@mail.ru

Authors

Nikolay A. Ignatev — D.Sc. (Physics & Mathematics), Full Professor, National University of Uzbekistan named after Mirzo Ulugbek, Tashkent, 100174, Uzbekistan, SC 39361638900, https://orcid.org/0000-0002-7150-5837, n_ignatev@rambler.ru

Davrbek X. Tursunmurotov — Doctoral Student, National University of Uzbekistan named after Mirzo Ulugbek, Tashkent, 100174, Uzbekistan, https://orcid.org/0009-0009-8664-9639, mr.davrbek@mail.ru

Статья поступила в редакцию 24.01.2024
Одобрена после рецензирования 05.03.2024
Принята к печати 27.03.2024

Received 24.01.2024
Approved after reviewing 05.03.2024
Accepted 27.03.2024