ВЕСТНИК ТОМСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА

2025 Управление, вычислительная техника и информатика № 73
Tomsk State University Journal of Control and Computer Science

Научная статья
УДК 519.95
doi: 10.17223/19988605/73/11

# Регуляризаторы по наборам обобщенных оценок

**Николай Александрович Игнатьев**

*Национальный университет Узбекистана им. Мирзо Улугбека, Ташкент, Узбекистан, n_ignatev@rambler.ru*

**Аннотация.** Предложен новый метод формирования ансамблей алгоритмов распознавания на основе технологии стекинга с использованием регуляризаторов для повышения обобщающей способности моделей. Основное внимание уделено предотвращению переобучения за счет мажорирующих функций, корректирующих отступы объектов от границы между классами. Исследованы условия корректного разделения объектов при обучении базовых алгоритмов и метаалгоритма. Предложен иерархический агломеративный алгоритм группировки признаков, формирующий латентные признаки с учетом внутриклассового сходства и межклассового различия. Показано, что регуляризация отступов и преобразование количественных признаков в номинальные повышают точность распознавания. Установлено, что выбор параметров мажорирующих функций минимизирует расхождение по точности между базовыми и метаалгоритмами.

**Ключевые слова:** ансамбли алгоритмов; стекинг; регуляризация; мажорирующие функции.

***Для цитирования*:** Игнатьев Н.А. Регуляризаторы по наборам обобщённых оценок // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. 2025. № 73. С. 90–99. doi: 10.17223/19988605/73/11

Original article
doi: 10.17223/19988605/73/11

# Regularizers on sets of generalized estimates

**Nikolay A. Ignatev**

*National University of Uzbekistan named after Mirzo Ulugbek, Tashkent, Uzbekistan, n_ignatev@rambler.ru*

**Abstract.** This paper proposes a novel method for constructing ensembles of recognition algorithms based on stacking technology, incorporating regularizers to enhance the generalization capability of models. The primary focus is on preventing overfitting through the use of majorizing functions that adjust the margins (offsets) of objects from the class boundary. The study investigates the conditions necessary for the correct separation of objects during the training of both base algorithms and the meta-algorithm. A hierarchical agglomerative feature grouping algorithm is proposed, which forms latent features based on intra-class similarity and inter-class differences. It is demonstrated that margin regularization and the transformation of quantitative features into nominal ones improve recognition accuracy. The results show that choosing appropriate parameters for the majorizing functions minimizes the accuracy gap between the base and meta-algorithms. Key advantages of the proposed method:

– flexible feature selection for the meta-algorithm based on a greedy strategy;
– unification of measurement scales through feature transformation;
– robustness to overfitting due to margin regularization.

**Keywords:** algorithm ensembles; stacking; regularization; majorizing functions.

***For citation*:** Ignatev, N.A. (2025) Regularizers on sets of generalized estimates. *Vestnik Tomskogo gosudarstvennogo universiteta. Upravlenie, vychislitelnaja tehnika i informatika – Tomsk State University Journal of Control and Computer Science*. 73. pp. 90–99. doi: 10.17223/19988605/73/11

## Введение

Повышение обобщающей способности при распознавании является основной целью формирования ансамблей алгоритмов. Как правило, в процессе формирования используют одну из технологий [1] машинного обучения. Чтобы предотвратить переобучение (overfitting), широко используются регуляризаторы. Регуляризаторы являются средством обобщения знания на новые данные, накладывая штрафы на сложность модели.

Смысл термина «обобщенные оценки» [2] ассоциируется с вычислением значений латентных признаков для описания объектов в двухклассовой задаче распознавания. Реализация методов вычисления обобщенных оценок основана на идее противопоставления описаний объектов двух классов как оппозиции друг другу.

Формализация описания многообразия пространств из латентных признаков, формируемых разными методами, сделана в [3]. Методы делятся на линейные и нелинейные, для реализации которых используются или не используются критерии оптимизации.

В статье [2] был предложен алгоритм вычисления обобщенных оценок по комбинации из базовых или элементарных классификаторов. Элементарные классификаторы из одного признака формировали обобщенные оценки объектов для разделения их на два класса. Для каждого признака вне зависимости от шкалы измерений (номинальной или количественной) в представлении прецедентов производилось нелинейное преобразование посредством значений функции принадлежности. Обобщенные оценки, вычисленные по набору элементарных классификаторов, рассматривались как проекции описаний объектов на числовую ось, которая использовалась для отнесения объекта к классу по пороговому решающему правилу.

В [4] описан многокритериальный метод формирования наборов латентных признаков из исходных. Согласно жадной стратегии иерархического агломеративного алгоритма группировки происходит формирование наборов латентных признаков. Множество исходных признаков, входящих в состав латентного с наибольшей дискриминантной способностью для определения принадлежности объектов к классам, считалось информативным набором. Показано, что существует латентный признак, сформированный из подмножества исходных, точность распознавания на котором выше, чем по аналогичному признаку для всего множества. Включение очередного признака в группу при синтезе основано на минимизации отношения внутриклассового сходства и межклассового различия. Если минимизации не происходит, то согласно правила иерархической агломеративной группировки процесс формирования латентного признака завершается. Нет гарантии, что при этом по значениям латентного признака объекты двух классов корректно разделяются на числовой оси.

Согласно технологии стекинга [5], один или несколько латентных признаков, полученных на исходных данных, могут использоваться в качестве входных данных для метаалгоритма, который делает финальное предсказание. Обучение метаалгоритма на входных данных базовых алгоритмов заключается в оптимальном их комбинировании.

Интерес для исследования представляет повышение точности с использованием регуляризаторов. Необходимо определить и обосновать условия, обеспечивающие существование (отсутствие) корректного разделения объектов обучения на классы, не приводящего к переобучению. Обобщающую способность алгоритмов имеет смысл оценивать как частоту ошибок на конечной выборке, так как вероятность ошибки является величиной ненаблюдаемой, которую невозможно вычислить точно [6].

Применение регуляризаторов для повышения точности основывается:

– на увеличении значений отступов между ближайшими объектами из разных классов;

– удалении аномальных объектов(выбросов) из обучающих выборок;

– достижении максимума значения меры компактности классов и выборки в целом.

Отступ – это расстояние между ближайшими объектами двух классов, которые не пересекаются, может представляться и как абсолютная, и как относительная величина. Относительный отступ применялся в регуляризаторах метода «ближайший сосед» (NN) [7]. Классическим пониманием отступа как абсолютной величины пользуются при обучении алгоритмов распознавания с разделяющимися

поверхностями (например, в методе опорных векторов – SVM). Значение отступа является аргументом для мажорирующих функций по технологии бустинга.

Бустинг для задач классификации объектов на два класса $K_1$ и $K_2$ строится как серия алгоритмов регрессии $a_1(\cdot)$, …, $a_k(\cdot)$, $k \geq 2$, со значениями целевого признака из {–1, 1} и правилом:

– если $a_1(S) + \ldots + a_k(S) > 0$, то объект $S$ из класса $K_1$;

– если $a_1(S) + \ldots + a_k(S) < 0$, то объект $S$ из класса $K_2$.

Решение по бустингу заключается в вычислении отступов и выбранной маожорирующей функции. Для практической реализации, как правило, выбор производился из следующего набора функций: квадратичная, кусочно-линейная, сигмоидная, логистическая, экспоненциальная. Итерации бустинга можно повторять. Такой вид бустинга называется градиентным, так как фактически вычисляется градиент функции ошибок (потерь), и новый алгоритм получается, как шаг против градиента функции ошибок. Интерес представляет применимость к значениям латентного признака (обобщенным оценкам объектов) мажорирующих функций для формирования ансамблевых алгоритмов.

## 1. Предмет исследования

Технология стекинга является методом формирования ансамблей алгоритмов в машинном обучении [7]. Различают базовые модели и модели метауровня, или метамодели. Базовые модели обучаются на исходных данных, затем их результаты используются в качестве входных данных для метамодели. В качестве преимуществ технологии указывалось на возможность комбинирования разных базовых моделей (решающих деревьев, логистической регрессии, градиентного бустинга и т.д.). Комбинирование позволяет использовать сильные и слабые стороны разных моделей с целью повышения точности. Недостатки технологии выражались в более высокой вычислительной сложности и рисках переобучения.

Для устранения недостатков вводились ограничения на применение технологии стекинга. Запрещалось обучать метаалгоритм на данных, по которым обучались базовые алгоритмы. Считалось, что игнорирование запрета приводит к переобучению и недостоверным результатам на новых данных.

Интерес для исследования представляет процесс формирования латентного признакового пространства в качестве входных данных для метаалгоритма. Реализация процесса происходит по результатам иерархической агломеративной группировки с учетом следующих условий:

– вхождение всех исходных признаков в состав латентных не является обязательным;

– максимальная точность распознавания при обучении достигается по базовым алгоритмам;

– базовый алгоритм необходим для оценки объекта лишь по части признакового пространства.

Пусть $a_1(\cdot)$, …, $a_k(\cdot)$ – набор из последовательности базовых алгоритмов. Действия базового алгоритма:

– выбор и включение исходного признака в состав латентного;

– корректировка значений латентного признака (оценок) объектов по отступу от границы между классами с помощью мажорирующих функций;

– включение оценок объектов в обучающую выборку как дополнительного признака для обучения метаалгоритма.

Каждый базовый алгоритм, включаемый в набор, не уменьшает точность распознавания предыдущего. Эффективность распознавания при обучении растет, так как известны отступы объектов от границы между классами, а также в каком направлении от нее нужно производить коррекцию (регуляризацию) оценок с помощью мажорирующих функций. Высокая точность распознавания, полученная на последнем базовом алгоритме из набора, не является основанием для выводов о высокой обобщающей способности для метаалгоритма.

Требуется изучение влияния мажорирующих функций на результаты метаалгоритма. Проблема качества обучения имеет скорее комбинаторную, нежели вероятностную природу [5]. Применительно к рассматриваемым ансамблевым алгоритмам речь идет о комбинациях латентных и исходных признаков в зависимости от параметров мажорирующих функций.

Эффект от применения мажорирующих функций проявляется в повышении точности распознавания на обучающей выборке по базовым алгоритмам. Процедуры обучения на основе жадных стратегий являются причиной порождения эффекта переобучения. Включение дополнительных признаков в описание объектов обучающей выборки проводится с целью сохранения в них частичной информации о скрытых закономерностях в данных. На использование этих закономерностей адаптирован метаалгоритм при разделении объектов на классы.

С целью унификации шкал измерений производится преобразование значений количественных признаков в градации номинальных. Для преобразования используется разбиение на непересекающиеся интервалы, оптимальные значения границ которых определяются по специальному критерию.

Есть предположение, что переобучение по технологии стекинга связано с выбором параметров базовых алгоритмов. Как правило, число базовых алгоритмов изначально не фиксировано, так как оно зависит от использования правил иерархической агломеративной группировки и настраиваемых параметров мажорирующей функции. Правила группировки реализованы на основе выбора первого исходного признака для организации ансамбля, результатов анализа отношения внутриклассового сходства и межклассового различия, условия останова процесса формирования значений латентного признака.

Для выбора параметров мажорирующих функций необходим анализ сходимости процесса обучения к состоянию корректного разделения обучающей выборки на классы и максимальному показателю обобщающей способности по метаалгоритму. Показателем для контроля сходимости к корректному разделению является отношение внутриклассового сходства и межклассового различия. Анализ вариабельности отношений востребован для исследования устойчивости метаалгоритма от переобучения.

Для вычисления выходных данных (значений дополнительных признаков) по базовым алгоритмам помимо отступов объектов от границы требовался идентификатор класса. Вычисление оценок для произвольного объекта по набору базовых алгоритмов связывалось с необходимостью использования функции потерь. Вид функции и множество ее допустимых значений в общем-то неизвестны. Для базового алгоритма нет информации, в каком направлении от границы между классами производить корректировку значений отступа.

Утверждается, что для произвольного объекта:

– базовые алгоритмы не вычисляют значения оценок;

– метаалгоритм может производить распознавание лишь с учетом привнесения в состав таблицы обучения дополнительных признаков с помощью базовых алгоритмов.

Следовало обосновать применимость предлагаемого метода на данных, которые не участвовали в процессе обучения. Обоснование строится на доказательстве утверждения, что для распознавания объекта не требуется вычислять значения его оценок по базовым алгоритмам. Метаалгоритм для распознавания использует исходные данные объекта и значения дополнительных признаков, полученных на объектах обучающей выборки.

Информацию о вариациях отношений значений внутриклассового сходства к межклассовому различию при добавлении исходного признака в латентный можно получать в процессе реализации метода агломеративной иерархической группировки [4]. Ценность (востребованность) информации заключается в ее использовании в процессе принятия решений как при включении, так и при отказе от включения очередного исходного признака в качестве кандидата в состав латентного.

## 2. Постановка задачи

Рассматривается стандартная постановка задачи распознавания для объектов из двух непересекающихся классов $K_1$ и $K_2$. Описание объектов в обучающей выборке $E_0 = \{S_1, \ldots, S_m\}$ представлено набором разнотипных признаков $X(n) = (x_1, \ldots, x_n)$, $\sigma$ из которых являются номинальными, $(n - \sigma)$ – количественными.

Считается, что набор базовых алгоритмов формируется с использованием параметров мажорирующих функций по значениям отступов объектов от границы между классами. Для предобработки данных используются процедуры:

– преобразования значений количественных признаков в градации номинальных;

– вычисления значения весов признаков и их вкладов в распознавание объектов классов;

– оптимизации критериев для синтеза латентных признаков из исходных.

Требуется:

– определить параметры для вычисления величин штрафов по мажорирующим функциям;

– построить набор базовых алгоритмов с использованием мажорирующих функций;

– разработать метаалгоритм для распознавания принадлежности объектов к классам с использованием дополнительных признаков, определяемых по базовым алгоритмам.

Пусть для значений признака $x_c \in X(n)$ в описании объектов $E_0 = K_1 \cup K_2$ построена упорядоченная по неубыванию последовательность

$$r_1, \ldots, r_j, \ldots, r_m. \tag{1}$$

В качестве границ двух непересекающихся интервалов $[\pi_1; \pi_2]$, $(\pi_2; \pi_3]$, определяемых по (1), используются $\pi_1 = r_1$, $\pi_2 = r_j$, $1 < j < m$, $\pi_3 = r_m$. Интервалы $[\pi_1; \pi_2]$ и $(\pi_2; \pi_3]$ идентифицируются соответственно как первый и второй. Вес признака у объектов классов по (1) вычисляется как максимум произведения внутриклассового сходства и межклассового различия по критерию из [8]:

$$\left( \frac{\sum\limits_{d=1}^{2}\sum\limits_{i=1}^{2}\left(u_i^d - 1\right)u_i^d}{\sum\limits_{i=1}^{2}|K_i|\left(|K_i|-1\right)} \right) \left( \frac{\sum\limits_{d=1}^{2}\sum\limits_{i=1}^{2}u_i^d\left(|K_{3-i}| - u_{3-i}^d\right)}{2|K_1||K_2|} \right) \to \max\limits_{\pi_1<\pi_2<\pi_3}, \tag{2}$$

где $u_i^d\left(u_{3-i}^d\right)$ – количество значений признака $x_c$ у объектов из класса $K_i$ ($K_{3-i}$) в $d$-м интервале. Множество допустимых значений критерия (2) принадлежит (0; 1] и используется для оценки объектов классов на числовой оси. Если в каждом интервале содержатся все значения признака объектов из одного класса, то его вес равен 1.

Граница между классами (порог) для количественного признака $x_c$ вычисляется как

$$\Gamma_c = \frac{\pi_2 + b}{2}, \tag{3}$$

где $b$ – ближайшее к $\pi_2$ значение из интервала $(\pi_2; \pi_3]$, определяемого по (2). В данной работе граница (3) используется для преобразования значений признака $x_c \in X(n)$ в виде двух градаций (по факту принадлежности к одному из интервалов) в номинальной шкале измерений. Далее будем считать, что выборка $E_0$ представлена значениями номинальных признаков.

Обозначим через $g_{1c}^j, g_{2c}^j$ – количество значений градации $j \in \{1, \ldots, \mu\}$ признака $x_c \in X(n)$ в описании объектов соответственно класса $K_1$ и $K_2$. Межклассовое различие по признаку $x_c$ определяется как величина

$$\lambda_c = 1 - \frac{\sum\limits_{t=1}^{2}\sum\limits_{d=1}^{\mu} g_{tc}^d g_{3-t,c}^d}{2|K_1||K_2|}.$$

Степень однородности (мера внутриклассового сходства) $\beta_c$ значений градаций признака при $\mu \geq 2$ по классам $K_1$, $K_2$ вычисляется по формулам

$$D_{dc} = \begin{cases} \left(|K_d| - l_{dc} + 1\right)\left(|K_d| - l_{dc}\right), & p_c > 2, \\ |K_d|\left(|K_d| - 1\right), & p_c \leq 2, \end{cases}$$

$$\beta_c = \begin{cases} \dfrac{\sum\limits_{j=1}^{p_c} g_{1c}^j\left(g_{1c}^j - 1\right) + g_{2c}^j\left(g_{2c}^j - 1\right)}{D_{1c} + D_{2c}}, & D_{1c} + D_{2c} > 0, \\ 0, & D_{1c} + D_{2c} = 0, \end{cases}$$

где $l_{dc}$ – число градаций признака $x_c$ в описании объектов из $K_d$, $d = 1, 2$.

Вес признаку $x_c \in X(n)$ определяется как

$$\omega_c = \beta_c \lambda_c. \quad (4)$$

Множество допустимых значений весов признаков, вычисленных по (4), лежит в интервале [0; 1]. Для получения обобщенных оценок объектов [2] на $E_0$ используются вклады градаций признаков. Вклад градации $j \in \{1, \ldots, \mu\}$ признака $x_c \in X(n)$ вычисляется как

$$\eta_c(j) = \omega_c \left( \frac{\alpha_{cj}^1}{|K_1|} - \frac{\alpha_{cj}^2}{|K_2|} \right), \quad (5)$$

где $\alpha_{cj}^1, \alpha_{cj}^2$ – количество значений градации $j$ признака $x_c$ соответственно в классах $K_1$ и $K_2$; $\omega_c$ – вес признака $x_c$ по (4). Обобщенная оценка объекта $S_r \in E_0$, $S_r = \{x_{ri}\}$, по описанию на наборе $X(d) \subset X(n)$ без использования мажорирующих функций вычисляется как

$$R(S_r) = \sum_{x_i \in X(d)} \eta_i(x_{ri}).$$

### *2.1. Формирование дополнительных признаков для метаалгоритма*

Согласно технологии стекинга вычисление каждого дополнительного признака реализуется отдельным базовым алгоритмом. Базовый алгоритм явно не участвует в вычислении оценок произвольного допустимого объекта. Считается, что для каждого $x_i \in X(n)$ определены вес $\omega_i$ по (4) и значения вкладов $\eta_i(j)$, $j \in \{1, \ldots, \mu\}$, по (5). Особенности реализации вычисления дополнительных признаков в качестве входных данных для метаалгоритма заключаются:

– в выборе первого признака из $X(n)$ для вычисления обобщенных оценок (значений дополнительных признаков) объектов $E_0$ базовыми алгоритмами;

– наборе правил для включения (не включения) признака в группу;

– вычислении значений обобщенных оценок объектов $E_0$ по вкладам признаков (5) и отступу между классами по мажорирующей функции.

Обозначим через $P$, *TUPLAM* – множество индексов признаков соответственно из $X(n)$ и формируемых алгоритмом группировки, $f(\cdot)$ – мажорирующая функция, $\alpha$ – параметр для регуляризации отступа, $0 < \alpha < 1$, $\delta$ ($0 < \delta < 0{,}5$) – порог для отношения внутриклассового сходства $\theta$ и межклассового различия по латентному признаку $\gamma$, $\varkappa$ – максимальное число дополнительных признаков, $\varkappa \le n - 1$. Реализация алгоритма по шагам будет следующей:

Шаг 1. $P = \{i \mid x_i \in X(n)\}$.

Шаг 2. Вычислить $u = \arg\max_{j \in P} \omega_j$. $TUPLAM = \{u\}$.

**Цикл** по $t \in \{1, \ldots, m\}$ $R(S_t) = \eta_u(a_{tu})$. Конец **цикла**; $cr1 = 10$. $P = P/\{u\}$.

Шаг 3. **Цикл** по $u \in P$. **Цикл** по $t \in \{1, \ldots, m\}$. $b_t = R(S_t) + \eta_u(a_{tu})$.

Если $S_t \in K_1$, то $b_t = b_t + \alpha f(-b_t)$ иначе $b_t = b_t - \alpha f(-b_t)$. Конец **цикла**;

$M_1 = \sum_{S_t \in K_1} b_t \cdot M_2 = \sum_{S_t \in K_2} b_t$. $M_1 = M_1 / |K_1|$. $M_2 = M_2 / |K_2|$. $\Theta = 0$. $\gamma = 0$.

**Цикл** по $t \in \{1, \ldots, m\}$. Если $S_t \in K_1$, то $\theta = \theta + |b_t - M_1|$, $\gamma = \gamma + |b_t - M_2|$. Иначе $\theta = \theta + |b_t - M_2|$, $\gamma = \gamma + |b_t - M_1|$. Конец **цикла;**

Если $\theta/\gamma < cr1$, то $cr1 = \theta/\gamma$, $q = u$. Конец **цикла;**

Шаг 4. $crit = cr1$. $P = P/\{q\}$. $TUPLAM = TUPLAM \cup \{q\}$. $cr1 = 10$.

**Цикл** по $t \in \{1, \ldots, m\}$. $R(S_t) = R(S_t) + \eta_q(a_{tq})$.

Если $S_t \in K_1$, то $R(S_t) = R(S_t) + \alpha f(-R(S_t))$. Иначе $R(S_t) = R(S_t) - \alpha f(-R(S_t))$.

Конец **цикла;** Вывод $\{R(S_t)\}_{t \in \{1,\ldots,m\}}$. Если $|TUPLAM| < \varkappa$ *and* $crit > \delta$, то идти 3.

Вывод *TUPLAM*.

Шаг 5. Конец.

Множество значений $\{R(S_t)\}_{t \in \{1,\ldots,m\}}$, полученное на шаге 4 алгоритма, формируют дополнительные (латентные) признаки в описания объектов $K_1$ и $K_2$.

Выводы, которые можно сделать по результатам иерархического агломеративного алгоритма:

– мощность множества исходных признаков для обучения метаалгоритма $|TUPLAM| \le n$;

– число дополнительных (латентных) признаков $p = |TUPLAM| - 1$.

Поставим в соответствие каждому исходному признаку индексы согласно порядку их следования в *TUPLAM*. Обозначим набор исходных и дополнительных признаков объектов $E_0$ для реализации метаалгоритма как $Y(2p-1) = (y_0, \ldots, y_p, r_1, \ldots, r_p)$. С учетом такого обозначения произвольный допустимый объект $S$ будет представлен признаками из *TUPLAM* как $S = (a_0, \ldots, a_p)$. В описание по $Y(2p-1)$ объекта обучающей выборки $S_i \in E_0$, $S_i = (a_{i0}, \ldots, a_{ip}, d_{i1}, \ldots, d_{ip})$ включены дополнительные признаки $d_{i1}, \ldots, d_{ip}$. Реализация метаалгоритма по шагам для распознавания объекта $S$ будет такой:

Шаг 1. $B1(a_0) = \{S_i \in K_1 |\ a_0 = a_{i0}\}$, $B2(a_0) = \{S_i \in K_2 |\ a_0 = a_{i0}\}$, $j = 0$.

Шаг 2. $j = j + 1$. $B1(a_j) = \{S_i \in B1(a_{j-1}) |\ a_j = a_{ij}, d_{ij} > 0\}$, $B2(a_j) = \{S_i \in B2(a_{j-1}) |\ a_j = a_{ij}, d_{ij} < 0\}$.

Шаг 3. Если $j < p$, то идти 2.

Шаг 4.
$$\begin{cases} S \in K_1, |B1(a_j)|/|K_1| > |B2(a_j)|/|K_2|, \\ S \in K_2, |B2(a_j)|/|K_2| > |B1(a_j)|/|K_1|, \\ 0, |B1(a_j)|/|K_1| = |B2(a_j)|/|K_2|. \end{cases}$$

Шаг 5. Конец.

### *2.2. О точности ансамблевых алгоритмов*

Для вычисления точности ансамблевых алгоритмов распознавания методы кросс-валидации неприменимы. Доказательство этого утверждения относительно ансамблевых алгоритмов, формируемых по методу вычисления обобщенных оценок, приводится в [2]. Возникла необходимость поиска альтернативных способов оценки точности. Предлагается исследование связи между точностью на обучении по базовым алгоритмам и метаалгоритму.

Для обозначения отношения внутриклассового сходства к межклассовому различию, так же как в описании алгоритма иерархической агломеративной группировки, будем использовать $\theta/\gamma$. Значение данного отношения рассматривается как средство для проверки истинности гипотезы о компактности объектов классов на многообразии латентных признаков. Интерес для исследования при формирования ансамблевых алгоритмов представляют:

– размеры отступов между классами в зависимости от параметров мажорирующей функции;

– наличие (отсутствие) равномерной сходимости значений $\theta/\gamma \to \min$ при иерархической агломеративной группировке по последовательности латентных признаков базовых алгоритмов;

– условия отсутствия корректного разделения объектов обучающей выборки по базовому и метаалгоритму.

Результаты исследования свойств сходимости $\theta/\gamma \to \min$ могут быть использованы при:

– выборе параметров мажорирующей функции;

– обосновании сходства точности по базовому и метаалгоритму.

Преобразование значений количественных признаков в градации номинальных увеличивает вероятность появления совпадающих описаний объектов обучения из двух классов. Граница между классами по (3), используемая для такого преобразования, определяется на основе реальной (не гипотетической) плотности распределения значений признака.

Проблема появления объектов из разных классов с совпадающими описаниями решается за счет использования мажорирующих функций. С помощью этих функций формируются несовпадающие значения латентных или дополнительных признаков для метаалгоритма. Наличие набора дополнительных признаков уменьшает вероятность отказа от распознавания по метаалгоритму.

## 3. Вычислительный эксперимент

Формирование ансамблей алгоритмов зависит от свойств распределений признаков в описании прецедентов для обучения. Учет этих свойств позволяет гибко подстраивать систему выбора как исходных, так и дополнительных признаков для метаалгоритма. В качестве мажорирующей используется сигмоидная функция с параметром $\alpha \in (0; 1)$. Применение других мажорирующих функций и их влияние на отступы с целью сравнительного анализа в данном исследовании не рассматриваются. Считается, что для количественных признаков произведено преобразование в градации номинальных по (1). В табл. 1 представлен пример построения ансамбля при наличии равномерной сходимости $\theta / \gamma \to \min$ на данных Molecular-biology [9]. Значение отношения $\theta/\gamma$ получено на последнем в последовательности дополнительном признаке, сформированным базовым алгоритмом. Условием останова процесса формирования было $\theta/\gamma < 0{,}1$. В скобках указана точность по метаалгоритму.

Таблица 1

**Результаты распознавания при наличии равномерной сходимости**

| α | Число дополнительных признаков | Значение отношения θ/γ | Точность, % |
|---|---|---|---|
| 0,10 | 9 | 0,0982 | 100 (100) |
| 0,20 | 6 | 0,0885 | 100 (100) |
| 0,30 | 5 | 0,0818 | 100 (100) |

Сходимость по $\theta/\gamma$ при разных значениях параметра α (см. табл. 1) к относительно малому значению $\delta$ ($\delta = 0{,}1$) при корректном разделении объектов на классы по базовому и метаалгоритму увеличивает возможности выбора набора из исходных и дополнительных признаков. При $\alpha = 0{,}3$ набор был представлен шестью из 56 исходных и пятью дополнительными признаками.

На данных Heart + Disease [10] показана важность отбора значения параметров α и $\varkappa$ для установления максимального соответствия результатов между базовыми алгоритмами и метаалгоритмом при отсутствии равномерной сходимости по отношению $\theta/\gamma$ (табл. 2).

Таблица 2

**Результаты распознавания при отсутствии равномерной сходимости**

| α | Число дополнительных признаков $\varkappa$ | |
|---|---|---|
| | 5 | 12 |
| 0,05 | 91,11 (87,04) | 92,96 (92,96) |
| 0,10 | 94,44 (86,67) | 98,15 (98,15) |
| 0,20 | 99,63 (97,41) | 99,63 (99,26) |
| 0,30 | 100,00 (97,41) | 100,00 (99,26) |

Компромисс или малое расхождение при высокой точности распознавания между результатами по базовым и метаалгоритму достигнут на $\alpha = 0{,}2$ и $\varkappa = 12$. Максимум по базовому алгоритму составил 99,63%, и 99,26% – по метаалгоритму (см. табл. 2).

Для демонстрации влияния коэффициентов регуляризации на значение меры компактности обучающей выборки и связи коэффициентов с обобщающей способностью метрического алгоритма «ближайший сосед» в [6] были использованы данные Spambase [11]. Так же, как и в [6], для тестирования предлагаемого ансамбля алгоритмов было произведено разбиение 4 204 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 тестирования.

По причине отсутствия на данных Spambase равномерной сходимости $\theta / \gamma \to \min$ для анализа состава исходных и дополнительных признаков использовалось значение параметра $\varkappa$. Результаты распознавания по составам исходных и дополнительных признаков при коэффициенте $\alpha = 0{,}2$ приводятся в табл. 3.

Таблица 3

**Результаты распознавания на данных Chet и Nechet**

| Выборка | Число дополнительных признаков $\varkappa$ | | |
|---|---|---|---|
| | 10 | 20 | 56 |
| Chet | 100,00 (94,96) | 100,00 (97,95) | 100,00 (99,24) |
| Nechet | 100,00 (93,34) | 100,00 (96,34) | 100,00 (99,48) |

Как видно из табл. 3, точность распознавания по метаалгоритму не превосходит точность по базовому алгоритму.

Для проверки обобщающей способности ансамбля алгоритмов распознавания в качестве прецедентов использованы выборки Chet и Nechet, результаты которых приведены в табл. 4. В скобках указана точность, полученная на регуляризаторах для метрического алгоритма NN из [7].

Таблица 4

**Точность распознавания на тестовых выборках, %**

| Обучающая выборка | Тестовая выборка | |
|---|---|---|
| | Chet | Nechet |
| Chet | – | 99,48 (88,20) |
| Nechet | 99,24(88,73) | – |

В табл. 4 демонстрируется превосходство по точности при использовании регуляризаторов по обобщенным оценкам относительно регуляризаторов для метрического алгоритма NN.

## Заключение

Разработан новый метод формирования ансамблей алгоритмов распознавания по технологии стекинга. Отметим изменения, которые привнесены в эту технологию:

– число базовых алгоритмов связано с выбором параметров модели;

– есть два ограничения на число признаков, определяемых явно или по специальному условию; выполнение условия зависит от значения отношения внутриклассового сходства и межклассового различия;

– роль базовых алгоритмов сводится к вычислению дополнительных признаков для обучающей выборки;

– высокая обобщающая способность на обучающих и тестовых выборках по метаалгоритму объясняется размерами отступов между классами.

Метод рекомендуется для использования в моделях, основанных на знаниях. Совершенствование метода связано с решением проблемы выбора прецедентов для обучающей выборки и разработкой новых способов формирования множества дополнительных признаков.

**Список источников**

1. Zhou Z.H. Ensemble learning: foundations and algorithms. Chapman & Hall/CRC, 2021. 394 p.
2. Ignatev N.A. On Nonlinear Transformations of Features Based on the Functions of Objects Belonging to Classes // Pattern Recognition and Image Analysis. 2021. V. 31 (2). P. 197–204.
3. Игнатьев Н.А., Акбаров Б.Х. Оценка близости структур отношений объектов обучающей выборки на многообразиях наборов латентных признаков // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. 2023. № 65. С. 69–78. doi: 10.17223/19988605/65/7
4. 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 (6). P. 439–445.
5. Hastie T., Tibshirani R., Friedman J. The elements of statistical learning: data mining, inference and prediction. 2nd ed. Springer, 2009. 767 p. (Springer Series in Statistics).
6. Воронцов К.В. Комбинаторный подход к оценке качества обучаемых алгоритмов // Математические вопросы кибернетики / под ред. О.Б. Лупанов. М. : Физматлит, 2004. Т. 13. С. 5–36.

7. 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. 2226–1494. doi: 10.17586/2226-1494-2024-24-2-322-329
8. Згуральская Е.Н. Алгоритм выбора оптимальных границ интервалов разбиения значений признаков при классификации // Известия Самарского научного центра Российской академии наук. 2012. № 4-3. С. 826–829.
9. UCI repository of machine learning databases/molecular-biology/promoter-gene-sequences. URL: https://archive.ics.uci.edu/dataset/67/molecular+biology+promoter+gene+sequences (accessed: 02.07.2025).
10. UCI repository of machine learning databases. Ionosphere. URL: http://archive.ics.uci.edu/ml/ datasets/Heart+Disease (accessed: 02.07.2025).
11. UCI repository of machine learning databases. spambase. URL: https://archive.ics.uci.edu/dataset/94/spambase (accessed: 02.07.2025).

**References**

1. Zhou, Z.H. (2021) *Ensemble Learning: Foundations and Algorithms.* Chapman & Hall/CRC.
2. Ignatev, N.A. (2021) On Nonlinear Transformations of Features Based on the Functions of Objects Belonging to Classes. *Pattern Recognition and Image Analysis*. 31(2). pp. 197–204.
3. Ignatev, N.A. & Akbarov, B.Kh. (2023) Estimation of the proximity of structures of relations of objects of the training sample on manifolds of sets of latent features. *Vestnik Tomskogo gosudarstvennogo universiteta. Upravlenie, vychislitelnaya tekhnika i informatika – Tomsk State University Journal of Control and Computer Science*. 65. pp. 69–78. doi: 10.17223/19988605/65/7
4. Ignatev, N.A. & Rahimova, M.A. (2022) Formation and Analysis of Sets of Informative Features of Objects by Pairs of Classes. *Scientific and Technical Information Processing*. 49(6). pp. 439–445.
5. Hastie, T., Tibshirani, R. & Friedman, J. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. Springer.
6. Vorontsov, K.V. (2004) Combinatorial approach to assessing the quality of learning algorithms. *Matematicheskie voprosy kibernetiki.* 13. pp. 5–36.
7. Ignatev, N.A. & Tursunmurotov, D.Kh. (2024) Censoring training samples using regularization of relatedness relations of class objects. *Nauchno-tekhnicheskiy vestnik informatsionnykh tekhnologiy, mekhaniki i optiki*. 24(2). pp. 322–329. doi: 10.17586/2226-1494-2024-24-2-322-329
8. Zguralskaya, E.N. (2012) Algorithm for selecting optimal boundaries of intervals for partitioning feature values during classification. *Izvestiya Samarskogo nauchnogo tsentra Rossiyskoy akademii nauk.* 4–3. pp. 826–829.
9. *UCI Repository of Machine Learning Databases. Molecular Biology. Promoter Gene Sequences*. [Online] Available from: https://archive.ics.uci.edu/dataset/67/molecular+biology+promoter+gene+sequences (Accessed: 2nd July 2025).
10. *UCI Repository of Machine Learning Databases. Ionosphere*. [Online] Available from: http://archive.ics.uci.edu/ml/datasets/Heart+Disease (Accessed: 2nd July 2025).
11. *UCI Repository of Machine Learning Databases. Spambase*. [Online] Available from: https://archive.ics.uci.edu/dataset/94/spambase (Accessed: 2nd July 2025).

***Информация об авторе:***
**Игнатьев Николай Александрович** – доктор физико-математических наук, профессор кафедры программного инжиниринга и искусственного интеллекта Национального университета Узбекистана им. Мирзо Улугбека (Ташкент, Узбекистан). E-mail: n_ignatev@rambler.ru

***Автор заявляет об отсутствии конфликта интересов.***

***Information about the author:***
**Ignatev Nikolay A.** (Doctor of Physical and Mathematical Sciences, Professor of the National University of Uzbekistan named after Mirzo Ulugbek, Tashkent, Uzbekistan). E-mail: n_ignatev@rambler.ru

***The author declares no conflicts of interests.***

*Поступила в редакцию 19.07.2025; принята к публикации 02.12.2025*

*Received 19.07.2025; accepted for publication 02.12.2025*