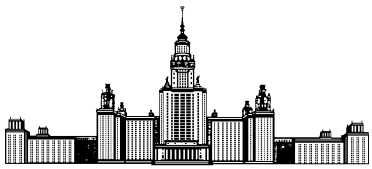
Московский государственный университет имени М. В. Ломоносова



Факультет Вычислительной Математики и Кибернетики

Кафедра Математических Методов Прогнозирования

ДИПЛОМНАЯ РАБОТА СТУДЕНТКИ 517 ГРУППЫ

«Логические корректоры в задачах распознавания»

Выполнила:

студенка 5 курса 517 группы

Любимцева Мария Михайловна

Научный руководитель:

д.ф-м.н., профессор

Дюкова Елена Всеволодовна

Москва, 2014

Содержание

1 Введение

В различных областях человеческой деятельности возникают задачи, в которых требуется найти решение на основе анализа большого объема накопленных знаний. Примерами являются задачи медицинской диагностики, обработки социологической информации, технического и геологического прогнозирования, анализа банковской деятельности, анализа пользовательской среды интернета и т. д. Для решения перечисленных задач успешно применяются методы распознавания образов, в частности методы, основанные на обучении по прецедентам.

Рассматривается задача распознавания по прецедентам с непересекающимися классами $K_1, \ldots, K_l$, набором целочисленных признаков $x_1, \ldots, x_n$ и обучающей выборкой $T = \{S_1, \ldots, S_m\}$, про каждый объект из которой известно, какому классу он принадлежит. Требуется по предъявленному признаковому описанию объекта S определить класс, к которому относится данный объект.

При решении прикладных задач классификации хорошо себя зарекомендовали методы, основанные на логическом (комбинаторном) анализе признаковых описаний объектов. В этих методах большое внимание уделяется вопросам синтеза корректных распознающих алгоритмов, т.е. алгоритмов, которые безошибочно классифицируют обучающие объекты.

В случае, когда данные представлены в целочисленном виде, используется понятие элементарного классификатора (эл.кл.). Эл.кл. — это элементарная конъюнкция, определенная на признаковых описаниях объектов. Если на описании некоторого объекта S элементарная конъюнкция обращается в единицу, то говорят, что объект S содержит данный эл.кл. В этом случае эл.кл. выделяет в описании S определенный фрагмент.

В классических дискретных процедурах распознавания ставится задача построения множества информативных фрагментов описаний обучающих объектов. Фрагмент называется информативным для класса K, если не существует пары обучающих объектов, в которой один объект принадлежит K, а второй — нет, и такой, что оба объекта содержат данный фрагмент. Эл.кл., порождающий информативный фрагмент, называется корректным. В классических логических алгоритмах в процессе

обучения строится некоторое семейство корректных эл.кл. На этапе классификации каждый из найденных эл.кл. участвует в процедуре голосования и формирования оценок принадлежности объектов к классам. В таких моделях корректность распознающего алгоритма обеспечивается корректностью каждого из найденных эл.кл.

Применение логических процедур распознавания наиболее эффективно в случае информации низкой значности (признаки могут принимать небольшое число значений). Для решения задач с целочисленной информацией высокой значности успешно используются различные способы перекодировки данных, которые приводят к понижению исходной значности. Однако, существуют прикладные задачи, на которых классические логические алгоритмы показывают плохие результаты даже после перекодировки. Например, такая ситуация возникает, если после перекодировки большинство информативных фрагментов редко встречается в описаниях обучающих объектов. При этом могут существовать фрагменты, которые часто встречаются в одном классе и редко в других. Таким фрагментам соответствуют некорректные эл.кл.

В [1] предложен подход к конструированию корректных логических алгоритмов распознавания на базе некорректных эл.кл. Этот подход сочетает логические методы и идеи алгебраического подхода. В качестве базовых алгоритмов выступают некорректные эл.кл., в качестве корректирующей функции берется булева функция. Основным понятием является классифицирующий набор — корректный набор из необязательно корректных эл.кл. Ставится задача построения семейства классифицирующих наборов. Распознающие алгоритмы, основанные на построении таких семейств называют логическими корректорами.

В работах [3] и [4] рассмотрены вопросы практического применения различных моделей логических корректоров.

В [3] построен логический корректор MON, который отличается следующими свойствами. Во–первых, используются монотонные классифицирующие наборы, когда в качестве корректирующей функции выступает монотонная булева функция. Во–вторых, классифицирующие наборы конструируются из эл.кл. ранга 1. В–третьих, итоговое семейство формируется при помощи генетического алгоритма из классифицирующих наборов с хорошей распознающей способностью.

Использование эл.кл. ранга 1 — довольно жесткое условие, однако снятие этого ограничения приводит к дополнительным временным затратам. Для решения данной проблемы в [4] предложено для составления классифицирующих наборов использовать не все множество эл.кл., порождаемых описаниями обучающих объектов, а некоторое его подмножество. Это подмножество названо локальным базисом.

По результатам тестирования построенный в [4] логический корректор LOBAGA превзошел корректор MON по качеству распознавания. В LOBAGA использован бустинг над классифицирующими наборами и основанная на бустинге процедура формирования локального базиса. В связи с этим сложно оценить, как повлияло на качество распознавания снятие ограничения на ранг эл.кл.

Цель дипломной работы — построение новых моделей логических корректоров и исследование эффективности использования корректоров с эл.кл. произвольного ранга.

В дипломной работе на базе эл.кл. произвольного ранга построен стохастический монотонный логический корректор MONS. На этапе обучения выделяется валидационная выборка и далее применяется итеративная процедура, на каждой итерации которой происходит следующее. Случайным образом формируется локальный базис определенной мощности. Затем из эл.кл. локального базиса при помощи генетического алгоритма формируется семейство классифицирующих наборов. После этого для каждого объекта валидационной выборки по найденному семейству вычисляется отступ — разность между оценкой за свой класс и максимальной оценкой за чужой класс. Обучение заканчивается, когда для всех объектов валидационной выборки усредненные по итерациям значения отступов перестают существенно изменяться. Итоговое семейство классифицирующих наборов получается объединением семейств, полученных на всех итерациях.

Описанный корректор протестирован на ряде прикладных задач. На основании полученных результатов и применении статистического критерия Уилкоксона можно утверждать, что корректор MONS по точности значимо превзошел корректор MON. Следовательно, снятие ограничения на ранг эл.кл. дает значимое повышение качества распознавания.

Показано превосходство корректора MONS по точности распознавания над рядом логических алгоритмов, таких как ABO, алгоритм голосования по тупиковым тестам, Логические Закономерности, RIPPER, различные алгоритмы на основе решающих деревьев.

По результатам экспериментов корректор MONS в среднем по качеству распознавания незначительно уступил алгоритму голосования по представительным наборам. При этом MONS значимо превзошел его на задачах с малым числом часто встречающихся корректных эл.кл., где алгоритм голосования по представительным наборам показал результаты, сопоставимые с бросанием монетки.

Дополнительно построена новая модель антимонотонного логического корректора AMON. Он отличается от корректора MON видом корректирующей функции, которая как и в случае MON выбрана монотонной булевой. Доказано утверждение об эквивалентности корректоров MON и AMON в случае двух классов. Проведен ряд экспериментов на задачах с более, чем двумя классами. Практически на всех задачах корректор модель с монотонными классифицирующими наборами показала себя лучше, чем с антимонотонными.

Полученные в дипломной работе результаты частично опубликованы в [5], [6], [7].

2 Основные определения

Определение. $T \cap K$ — множество обучающих объектов из T из класса K;
$T \cap \overline{K}$ — множество обучающих объектов из T, не принадлежащих классу K.

Определение. Пусть
$H = \{x_{j_1}, \ldots, x_{j_r}\}$ — набор различных целочисленных признаков,
$\sigma = \{\sigma_1, \ldots, \sigma_r\}$ — набор некоторых допустимых значений этих признаков.
Тогда эл.кл. — конъюнкция вида $B_{(H,\sigma)} = x_{j_1}^{\sigma_1} \cdots x_{j_r}^{\sigma_r}$, число r — ранг эл.кл.

Определение. Эл.кл. считается корректным, если
$\nexists\, S' \in T \cap K,\, S'' \in T \cap \overline{K}$: $B_{(H,\sigma)}(S') = B_{(H,\sigma)}(S'') = 1$.

Определение. Объект S содержит эл.кл. $B_{(H,\sigma)}$, если $B_{(H,\sigma)}(S) = 1$.

Определение. $U = \{B_{(x_{j_1},\sigma_{j_1})}, \ldots, B_{(x_{j_q},\sigma_{j_q})}\}$ — набор эл.кл.

Определение. Набор эл.кл. U называется корректным для класса K, $K \in \{K_1, \ldots, K_l\}$, если $\forall (S', S'')$ — пара обучающих объектов, $S' \in K$, $S'' \notin K$ существует монотонная функция алгебры логики $F_{U,K}$:

$$F_{U,K}(U(S')) \neq F_{U,K}(U(S'')).$$

Определение. Корректный для класса K набор из эл.кл. U называется классифицирующим набором для класса K.

Определение. Классифицирующий набор U для класса K называется монотонным, если существует монотонная функция алгебры логики $F_{U,K}$:

$$F_{U,K}(U(S)) = \begin{cases} 1, & \text{если } S \in K \cap T, \\ 0, & \text{если } S \notin K \cap T. \end{cases}$$

Определение. Классифицирующий набор U для класса K называется антимонотонным, если существует монотонная функция алгебры логики $F_{U,K}$:

$$F_{U,K}(U(S)) = \begin{cases} 0, & \text{если } S \in K \cap T, \\ 1, & \text{если } S \notin K \cap T. \end{cases}$$

3 Предыдущие результаты

В классических алгоритмах распознавания, основанных на поиске эл.кл., в процессе обучения стоится некоторое семейство корректных эл.кл. На этапе классификации каждый из найденных эл.кл. участвует в процедуре голосования и формирования оценок принадлежности объектов к классам. В таких моделях корректность распознающего алгоритма обеспечивается корректностью каждого из найденных эл.кл.

Применение логических процедур распознавания наиболее эффективно в случае информации низкой значности (признаки могут принимать небольшое число значений). Для решения задач с целочисленной информацией высокой значности успешно используются различные способы перекодировки данных, которые приводят к понижению исходной значности. Однако, существуют прикладные задачи, на которых классические логические алгоритмы показывают плохие результаты даже после перекодировки. Например, такая ситуация возникает, если после перекодировки большинство информативных фрагментов редко встречается в описаниях обучающих объектов. При этом могут существовать фрагменты, которые часто встречаются в одном классе и редко в других. Таким фрагментам соответствуют некорректные эл.кл.

Современный подход к конструированию корректных логических алгоритмов распознавания на базе эл.кл., предложенный в [1], сочетает алгебраические и описанные логические методы построения корректных распознающих процедур. Ключевая идея алгебраического подхода, предложенная в [2], заключается в построении корректных алгоритмов с использованием, вообще говоря, некорректных базовых алгоритмов и обеспечении корректности итогового распознающего алгоритма применением корректирующей функции.

В терминах алгебраического подхода в классических логических методах используются корректные эл.кл. в качестве базовых алгоритмов и простейшая корректирующая функция. Эта функция лишь формально называется таковой, так как корректность итогового алгоритма достигается за счет корректности базовых алгоритмов.

В [1] предлагается ослабить условие на базовые алгоритмы, позволив им быть некорректными, и при этом использовать более сложную корректирующую функ-

цию. Ставится задача построения семейства корректных наборов из некорректных эл.кл., в котором каждый набор обладает хорошей распознающей способностью. Итоговый алгоритм корректен за счет корректности каждого построенного набора из эл.кл. Получаемые распознающие алгоритмы, называют логическими корректорами.

В [1] также показано, что задача построения (монотонных) корректных наборов эл. кл. для класса K сводится к задаче поиска покрытий булевой матрицы L_K, специальным образом построенной по обучающей выборке. Число столбцов в указанной матрице равно числу используемых эл.кл. В случае бинарной информации и использовании эл.кл. только ранга 1 показано, что поиск монотонных корректных наборов эл.кл. можно свести к поиску специальных наборов столбцов матрицы сравнения. Удалось добиться сокращения перебора, так как число столбцов этой матрицы равно размерности признакового пространства, а число строк такое же, как у L_K. Для рассмотренного случая в указанной работе предложен алгоритм для перечисления всех (монотонных) корректных наборов.

С распространением генетических алгоритмов стало возможным эффективно строить решения для задач поиска неприводимых покрытий булевых матриц, которые возникают в задачах построения корректных наборов эл. кл. Стало возможным исследовать вопросы практического применения различных моделей логических корректоров, и продолжить развитие алгебро–логического подхода, что и было сделано в работах [3] и [4].

В [3] на основе генетического подхода проведено исследование двух моделей: модели, основанной на построении корректного набора по мощности близкого к минимальному, и модели, основанной на построении коллектива корректных наборов, каждый из которых обладает хорошей распознающей способностью. В качестве особей генетического алгоритма выступали тупиковые корректные наборы простейших эл.кл., порождаемых элементарными конъюнкциями ранга 1. Дополнительно исследован случай использования только монотонных корректных наборов из эл.кл., когда в качестве корректирующей функции выступает монотонная булева функция. Наилучшие результаты получены при использовании модели MON, основанной на построении коллектива монотонных корректных наборов из эл.кл.

Использование эл.кл. ранга 1 — довольно жесткое условие. Снятие этого ограничения приводит к дополнительным временным затратам. Для решения данной проблемы в [4] для построения (монотонных) корректных наборов из эл.кл. предложено использовать не множество эл.кл., порождаемых подописаниями всех обучающих объектов, а в некоторое его подмножество — локальный базис.

В работе рассмотрено несколько подходов для формирования локального базиса, среди которых бустинг-метод, метод построения бинарного дерева, метод монотонной коррекции и собственный метод. Наилучших результатов удалось добиться, используя метод, основанный на бустинге.

В основе обучения построенного в [4] логического корректора LOBAGA лежит итеративный алгоритм, использующий бустинг–метод. При инициализации берутся пустые семейства корректных наборов из эл.кл. для каждого класса. На каждой итерации сначала выбирается один из классов. На первых итерациях поочередно выбирается каждый из имеющихся классов, затем на выбор класса влияют результаты, полученные на предыдущих итерациях. После выбора класса для него строится локальный базис по объектам подвыборки, специальным образом выделенной из множества обучающих объектов. В построенном локальном базисе запускается генетический алгоритм. Найденные корректные наборы из эл.кл. добавляются в семейство рассматриваемого класса. Для каждого добавляемого набора вычисляется вес, используемый при распознавании объектов. Алгоритм завершает работу, когда число найденных корректоров превосходит некоторый заданный наперед порог.

На этапе классификации применяется взвешенное голосование корректных наборов из эл.кл., веса которых настроены при помощи бустинга. По результатам тестирования построенный алгоритм превзошел по качеству распознавания алгоритм MON. Однако, из–за применения бустинга — мощного средства для построения композиции алгоритмов — сложно оценить вклад снятия ограничения на ранг эл.кл. в повышение качества распознавания. Таким образом, актуальными остаются вопросы использования в логических корректорах эл.кл. произвольного ранга.

4 Общая схема работы логического корректора

Общая схема работы логического корректора состоит из следующих этапов:

1. Обучение

   (a) Для каждого класса K, $K \in \{K_1, \ldots, K_l\}$, распознающий алгоритм A конструирует семейство классифицирующих наборов $W_A(K)$ — некоторое подмножество множества всех классифицирующих наборов для класса K.

   (b) Каждому классифицирующему набору U приписывается вес α_U.

2. Распознавание

   (a) Для каждого распознаваемого объекта S вычисляется оценка принадлежности этого объекта к каждому из классов K по формуле:

   $$\Gamma_K^{LC}(S) = \sum_{U \in W_A(K)} \alpha_U F_{U,K}(U(S)),$$

   где $F_{U,K}$ —корректирующая булева функция для набора U.

   В простейшем случае все наборы имеют одинаковые веса $\alpha_U = |W_A(K)|^{-1}$, тогда формула выше преобразуется к виду

   $$\Gamma_K^{LC}(S) = \frac{1}{|W_A(K)|} \sum_{U \in W_A(K)} F_{U,K}(U(S)).$$

   (b) Классификация

   Алгоритм относит неизвестный объект S к классу, за принадлежность к которому была получена наибольшая оценка. Если таких классов несколько, то алгоритм отказывается от классификации.

5 Логические корректоры MON и AMON

В данном разделе рассматриваются построенный в [3] логический корректор MON и построенный по аналогии с ним в дипломной работе логический корректор AMON.

MON — корректор на базе монотонных корректных наборов из эл.кл. ранга 1.
AMON — корректор на базе антимонотонных корректных наборов из эл.кл. ранга 1.

Для корректора MON итоговое семейство классифицирующих наборов является подмножеством множества всех монотонных классифицирующих наборов. Аналогично для корректора AMON итоговое семейство является подмножеством множества всех антимонотонных классифицирующих наборов.

Также стоит отметить, что в обоих алгортимов классифицирующим наборам одного класса приписываются одинаковые веса, то есть, как указано в предыдущем разделе $\alpha_U = |W_A(K)|^{-1}$.

Прочие детали схемы работы корректоров описаны ниже.

5.1 Вычисление оценок принадлежности

Определение Функция голосования классифицирующего набора U по обучающему объекту S' для объекта S имеет вид:

$$\delta_U(S, S') = \prod_{B_{(H,\sigma)} \in U} \left[B_{(H,\sigma)}(S) \geqslant B_{(H,\sigma)}(S') \right].$$

В модели MON для монотонного классифицирующего набора U для класса K корректирующая функция имеет вид:

$$F^{MON}_{U,K}(S) = \frac{1}{|T \cap K|} \sum_{S' \in T \cap K} \delta_U(S, S').$$

В модели AMON для антимонотонного классифицирующего набора U для класса K корректирующая функция имеет вид:

$$F^{AMON}_{U,K}(S) = \frac{1}{|T \cap \overline{K}|} \sum_{S'' \in T \cap \overline{K}} \left(1 - \delta_U(S, S'')\right).$$

5.2 Построение семейства наборов эл.кл.

Построение (анти)монотонных классифицирующих наборов сводится к построению покрытий булевых матриц L'_K (L''_K), которые сконструированы по обучающей выборке описанным ниже способом.

Пусть $C^*(K) = \{B_{(x_{j_1},\sigma_{j_1})}, \ldots, B_{(x_{j_{N(K)}},\sigma_{j_{N(K)}})}\}$ — множество эл.кл. для класса K, на основе которых производится построение (анти)монотонных классифицирующих наборов. В случае корректора MON $C^*(K)$ — множество всех эл.кл., порождаемых описаниями обучающих объектов из класса K, в случае корректора AMON — порождаемых описаниями обучающих объектов не из класса K.

В случае корректора MON паре объектов S' и S'' ставится в соответствие строка $D'(S', S'') = (d'_1, \ldots, d'_{N(K)})$, в которой

$$d'_i = \begin{cases} 1, & \text{если } B_{(x_{j_i},\sigma_{j_i})}(S') = 1 \text{ и } B_{(x_{j_i},\sigma_{j_i})}(S'') = 0, \\ 0, & \text{иначе.} \end{cases}$$

Для каждого класса K конструируется булева матрица L'_K из всех строк $D'(S', S'')$ таких, что S', S'' — обучающие объекты, $S' \in K$, $S'' \in \overline{K}$.

В случае корректора AMON паре объектов S' и S'' ставится в соответствие строка $D''(S', S'') = (d''_1, \ldots, d''_{N(K)})$, в которой

$$d''_i = \begin{cases} 1, & \text{если } B_{(x_{j_i},\sigma_{j_i})}(S') = 0 \text{ и } B_{(x_{j_i},\sigma_{j_i})}(S'') = 1, \\ 0, & \text{иначе.} \end{cases}$$

Аналогично для каждого класса K конструируется булева матрица L''_K из всех строк $D''(S', S'')$ таких, что S', S'' — обучающие объекты, $S' \in K$, $S'' \in \overline{K}$.

Для сокращения перебора при поиске классифицирующих наборов с хорошей распознающей способностью использован генетический алгоритм. В качестве особей популяции в алгоритмах используются неприводимые покрытия матриц L'_K, L''_K, в качестве оценки степени приспособленности — оценка качества распознавания $\tau_{(A,K)}(U)$:

$$\tau_{(A,K)}(U) = \frac{1}{|T_1 \cap K|} \sum_{(S \in T_0 \cap K)} \sum_{S' \in T_1 \cap K} \delta_U(S, S') - \frac{1}{|T_1 \cap \overline{K}|} \sum_{S \in T_0 \cap K} \sum_{S' \in T_1 \cap \overline{K}} \delta_U(S, S'),$$

где T_1 — заранее выделенная, не использовавшаяся для построения матриц L'_K, L''_K валидационная выборка.

5.3 Эквивалентность алгоритмов MON и AMON в случае двух классов

Рассматривается задача классификации на два класса K_1 и K_2. В этом случае булева матрица L'_{K_1} для поиска всех монотонных классифицирующих наборов для класса K_1 совпадает с матрицей L''_{K_2} для поиска всех антимонотонных классифицирующих наборов для класса K_2. Следовательно, множество $W_{K_1}^{MON}$ всех монотонных классифицирующих наборов для класса K_1 совпадает с множеством $W_{K_2}^{AMON}$ всех антимонотонных классифицирующих наборов для класса K_2. Аналогично $W_{K_2}^{MON} = W_{K_1}^{AMON}$.

Пусть U — монотонный классифицирующий набор для K_1. Он же будет являться антимонотонным набором для K_2.

Оценка объекта S за принадлежность классу K_1 по монотонному для K_1 набору U записывается как

$$\Gamma_U^{MON}(S) = \frac{1}{|K_1|} \sum_{S' \in K_1} \delta_U(S, S').$$

Оценка объекта S за принадлежность классу K_2 по антимонотонному для K_2 набору U записывается как

$$\Gamma_U^{AMON}(S) = \frac{1}{|K_1|} \sum_{S' \in K_1} (1 - \delta_U(S, S')) = 1 - \frac{1}{|K_1|} \sum_{S' \in K_1} \delta_U(S, S') = 1 - \Gamma_U^{AMON}(S).$$

Обозначим $W_1 = W_{K_1}^{MON}$, $W_2 = W_{K_2}^{MON}$. Пусть выполнено соотношение

$$\Gamma_{W_1}^{MON}(S) \quad > \quad \Gamma_{W_2}^{MON}(S).$$

Тогда корректор MON отнесет объект S к классу K_1. С учетом расписанных выше соотношений рассмотрим, к какому классу отнесет объект S корректор AMON.

$$\Gamma_{W_2}^{AMON}(S) \quad ? \quad \Gamma_{W_1}^{AMON}(S)$$

$$1 - \Gamma_{W_2}^{MON}(S) \quad > \quad 1 - \Gamma_{W_1}^{MON}(S),$$

то есть корректор AMON также отнесет объект S к классу K_1.

Таким образом, получаем, что при использовании семейств из всех возможных классифицирующих наборов в случае двух классов корректоры MON и AMON всегда относят неизвестные объекты к одному классу.

6 Логический корректор MONS

MONS (MON Stochastic) — монотонный логический корректор из эл.кл. произвольного ранга со стохастической процедурой формирования локального базиса.

Для обучения применена итеративная процедура с использованием общей для всех итераций валидационной выборки. На каждой итерации сначала случайным образом формируется локальный базис мощности $2]log_2 m[+3$, m — число строк расширенной булевой матрицы L'_K [9]. Затем при помощи генетического алгоритма из эл.кл. локального базиса формируется семейство классифицирующих наборов. После этого для каждого объекта валидационной выборки по найденному семейству вычисляется отступ — разность между оценкой за свой класс и максимальной оценкой за чужой класс. Обучение заканчивается, когда для всех объектов валидационной выборки усредненные по итерациям значения отступов перестают существенно изменяться. Итоговое семейство классифицирующих наборов получается объединением семейств, полученных на всех итерациях.

Algorithm 6.1 Процедура обучения корректора MONS

Вход: $Train$ — обучающая выборка, $maxi$ — максимальное число итераций, ε;

Выход: W_k, $k = 1..l$ — семейства монотонных классифицирующих наборов;

случайным образом выделить $Val \subset Train$ // *валидационная выборка*

$Train := Train \setminus Val$;

для всех $i = 1..maxi$

  для всех $k = 1..l$

    случайным образом сформировать LB; // *локальный базис*

    $W_k^i := GA(Train, k, LB)$; // *семейство мон. наборов из эл.кл. LB*

    $W_k := W_k \cup W_k^i$;

  для всех $S \in Val$

    $M_i(S) := \Gamma^{MON}_{y(S)}(W^i_{y(S)}, S) - \max\limits_{k \in \{1,\ldots,l\} \setminus y(S)} \Gamma^{MON}_k(W^i_k, S)$; // *отступ*

    $M_i^{avg}(S) := \frac{1}{i} \sum\limits_{j=i}^{i} M_j(S)$; // *средний по итерациям отступ*

  если $\forall S \in Val \quad M_i^{avg}(S) - M_{i-1}^{avg}(S) < \varepsilon$ **то**

    выход

7 Связь логических корректоров с классическими логическими алгоритмами распознавания

Для того чтобы показать, что некоторый алгоритм A представим в виде логического корректора LC , необходимо:

1. Задать вид корректных наборов из эл.кл.
2. Задать вид корректирующей функции
3. Доказать $\forall K \in \{K_1, \ldots, K_l\}$, $\forall S \quad \Gamma_K^A(S) = \Gamma_K^{LC}(S)$

7.1 Связь логических корректоров с алгоритмами голосования по представительным наборам

Утверждение 1

Алгоритм голосования по представительным наборам можно привести к виду логического корректора.

Доказательство

Каждому представительному набору $B_{(H,\sigma)}$ ставится в соответствие набор $U = \{B_{(H,\sigma)}\}$, содержащий всего один эл.кл. — сам представительный набор.

Данный набор будет корректным, так как эл.кл., который он содержит, корректен. Следовательно, достаточно воспользоваться простейшей корректирующей функцией:

$$F_{U,K}(U(S)) = \prod_{B_{(H_i,\sigma_i)} \in U} B_{(H_i,\sigma_i)}(S).$$

Для оценок принадлежности неизвестного объекта S к классу K получаем

$$\Gamma_K^{RS}(S) = \frac{1}{|W_K^{RS}|} \sum_{B_{(H,\sigma)} \in W_K^{RS}} B_{(H,\sigma)}(S) = \frac{1}{|W_K^{LC}|} \sum_{U \in W_K^{LC}} F_{U,K}(U(S)) = \Gamma_K^{LC}(S),$$

W_K^R — множество представительных наборов для класса K,
W_K^{LC} — множество корректных для класса K наборов из эл.кл., построенных логическим корректором LC; $|W_K^R| = |W_K^{LC}|$.

7.2 Связь логических корректоров с алгоритмом голосования по тупиковым тестам

Далее для всех объектов S их признаковое описание обозначается как $(a_1, \ldots, a_n)$.

Утверждение 2

Тестовый алгоритм можно привести к виду логического корректора.

Доказательство

Пусть $H = \{x_{j_1}, \ldots, x_{j_r}\}$ — тест, $a_H^i = \{a_{j_1}^i, \ldots, a_{j_r}^i\}$ — фрагменты описаний обучающих объектов из класса K, которые выделяет тест H, $i = 1..m_k$, $m_k = |K|$.

Тестовый алгоритм можно рассматривать как алгоритм голосования по представительным наборам, так как каждый тест H выделяет множество представительных наборов $\{B_{(H,a_H^1)}, \ldots, B_{(H,a_H^{m_k})}\}$, где $B_{(H,a_H^i)}$ — представительный набор, порождаемый тестом H и обучающим объектом S_i.

Из утверждения 1 следует, что и тестовый алгоритм можно привести к виду логического корректора. Для этого каждому тесту H ставится в соответствие множество наборов $\{U_1, \ldots, U_{m_k}\}$,
где $U_i = \{B_{(H,a_H^i)}\}$.

Данный набор будет корректным, так как эл.кл., который он содержит, корректен. Как и в утверждении 1 воспользуемся корректирующей функцией вида

$$F_{U,K}(U(S)) = \prod_{B_{(H_i,\sigma_i)} \in U} B_{(H_i,\sigma_i)}(S).$$

Для оценок принадлежности неизвестного объекта S к классу K получаем

$$\Gamma_K^T(S) = \frac{1}{|W^T|}\frac{1}{|K|}\sum_{H \in W^T}\sum_{i=1}^{m_k} B_{(H,a_H^i)}(S) = \frac{1}{|W_K^{LC}|}\sum_{U \in W_K^{LC}} F_{U,K}(U(S)) = \Gamma_K^{LC}(S),$$

W^T — множество тупиковых тестов, построенных тестовым алгоритмом,
W_K^{LC} — множество корректных для класса K наборов из эл.кл., построенных логическим корректором LC; $|W_K^{LC}| = |W^T||K|$.

Утверждение 3

Для любых исходных данных можно построить алгоритм MON, который будет эквивалентен тестовому алгоритму.

Доказательство

Для доказательства этого утверждения достаточно показать, что из любого теста H можно построить монотонный корректный набор эл.кл. U, который на всех обучающих объектах голосует так же, как и исходный тест, то есть

$$\forall\, S, \forall\, S^i \in K \cap T \quad \delta_U(S, S^i) = \Gamma_H(S, S^i).$$

Пусть $H = \{x_{j_1}, \ldots, x_{j_r}\}$ — тест, $a_i = \{a^i_{j_1}, \ldots, a^i_{j_r}\}$ — фрагменты описаний обучающих объектов из класса K, которые выделяет тест H, $i = 1..m_k$, $m_k = |K|$.

Покажем, что $U = \{B_{(H,a_1)}, \ldots, B_{(H,a_{m_k})}\}$ — искомый монотонный корректный набор.

Для $i = 1..m_k$ верно, что $B_{(H,a^i)}(S^i) = 1$. Так как во всех эл.кл. из U берутся признаки H, то значение

$$\delta_U(S, S^i) = \prod_{t=1}^{q} \left[B_{(H,a_t)}(S) \geqslant B_{(H,a_t)}(S^i)\right]$$

$$B_{(H,a_t)}(S^i) \quad = \quad \begin{cases} 0, & a_t \neq a_i \\ B_{(H,a_i)}(S^i), & a_t = a_i \end{cases} \quad = \quad \begin{cases} 0, & a_t \neq a_i \\ 1, & a_t = a_i \end{cases}$$

Подставляя данные значения в формулу для функции голосования, получаем, что значение $\delta_U(S, S^i)$ определяется только значением $B_{(H,a^i)}(S)$. Далее при помощи несложных преобразований завершаем доказательство утверждения:

$$\delta_U(S, S^i) = \left[B_{(H,a_i)}(S) \geqslant B_{(H,a_i)}(S^i)\right] = \left[B_{(H,a_i)}(S) \geqslant 1\right] = \left[B_{(H,a_i)}(S) = 1\right] = \Gamma_H(S, S^i).$$

Замечание

Получается, что тестовый алгоритм — частный случай алгоритма MON.

Утверждение 4

$U = \{(x_{j_1}, \sigma_1), \ldots, (x_{j_r}, \sigma_r)\}$ — корректный для класса K набор эл. кл.

$H = \bigcup\limits_{i=1}^{r} \{x_{j_i}\}$ — множество признаков, участвующих в образовании набора U.

$\Rightarrow H$ — тест.

Доказательство

Корректность набора U означает, что

$$\forall\ S' \in K \cap T,\ S'' \in \overline{K} \cap T \quad \exists (x_{j_i}, \sigma_i) \in U : \quad B_{(x_{j_i}, \sigma_i)}(S') \neq B_{(x_{j_i}, \sigma_i)}(S'').$$

Это равносильно тому, что

$$\forall\ S' \in K \cap T,\ S'' \in \overline{K} \cap T \quad \exists x_{j_i} \in H : a'_{j_i} \neq a''_{j_i}.$$

А это по определению означает, что H — тест.

Утверждение 5

$U = \{(H_1, \Sigma_1), \ldots, (H_r, \Sigma_r)\}$ — корректный для класса K набор эл. кл.

$H_p = \bigcup\limits_{i=1}^{r_p} \{x^p_{j_i}\}$, $H = \bigcup\limits_{p=1}^{r} H_p$ — множество признаков, участвующих в образовании набора U.

$\Rightarrow H$ — тест.

Замечание

Таким образом, каждый корректный набор эл. кл. порождает тест.

Доказательство

Корректность набора U означает, что

$$\forall\ S' \in K \cap T,\ S'' \in \overline{K} \cap T \quad \exists (H_p, \Sigma_p) \in U : \quad B_{(H_p, \Sigma_p)}(S') \neq B_{(H_p, \Sigma_p)}(S'')$$

$$\Leftrightarrow \exists (x^p_{j_i}, \sigma^p_i) \in U : \quad B_{(x^p_{j_i}, \sigma^p_i)}(S') \neq B_{(x^p_{j_i}, \sigma^p_i)}(S'').$$

Это равносильно тому, что

$$\forall\ S' \in K \cap T,\ S'' \in \overline{K} \cap T \quad \exists x_{j_i} \in H : a'_{j_i} \neq a''_{j_i}.$$

А это по определению означает, что H — тест.

8 Тестирование

Тестирование проводилось по методу leave–one–out на 25 прикладных задачах, собранных в отделе Математических проблем разпознавания и методов комбинаторного анализа ВЦ РАН. Так как логические корректоры нацелены на работу с информацией низкой значности, все задачи были предварительно перекодированы с помощью метода из [13]. Характеристики всех перекодированных задач представлены в таблице 4. Для каждой задачи указаны число объектов m, число признаков n, число классов k, распределение объектов по классам, среднее по признакам значение величины $\frac{\text{число значений признака}}{m} * 100\%$ (доля уникальных значений), размер расширенной матрицы L'_{K_1} для алгоритма MON.

Для оценки качества распознавания использован следующий функционал:

$$R = \frac{1}{l} \sum_{i=1}^{l} \frac{tp_i}{|K_i|},$$

где tp_i — число правильно распознанных объектов из класса K_i. Он имеет смысл средней по всем классам точности распознавания.

8.1 Эксперимент 1

Цель эксперимента — сравнение алгоритмов MON и AMON. В разделе 5.3 показано, что в случае двух классов данные алгоритмы эквивалентны, поэтому запуски проводились только на задачах с числом классов более двух. Результаты тестирования приведены в таблице 1. Почти на всех задачах алгоритм MON превосходит алгоритм AMON.

8.2 Эксперимент 2

В таблице 5 приведены результаты тестирования алгоритмов MON и MONS на 24 задачах. На большинстве задач алгоритм MONS превосходит алгоритм MON. Для определения, является ли разница в результатах счета алгоритмов значимой, воспользуемся ранговым критерием Уилкоксона для связных выборок.

Пусть x — разности результатов MONS и MON. Нулевая гипотеза критерия Уилкоксона предполагает, что вектор x пришел из распределения с нулевой медианой.

Название	MON	AMON
eco_l	91.15%	88.86%
Hea_r	38.68%	31.05%
melanoma	83.58%	77.64%
Pnevmo_r	78.02%	64.20%
TLMEL	87.52%	77.09%
TRMEL	83.78%	87.33%
TUML	64.90%	60.44%

Таблица 1: Качество распознавания R для корректоров MON и AMON.

Положим, что альтернативной гипотезе будет соответствовать утверждение, что x пришел из распределения с медианой больше нуля.

При выбранной альтернативе нулевая гипотеза отвергается на уровне значимости 0.05 с p–value = 0.012. То есть по тесту Уилкоксона результаты алгоритма MONS значимо лучше результатов алгоритма MON. Получается, что снятие ограничения на ранг эл.кл. дает значимое повышение качества распознавания.

8.3 Эксперимент 3

В ходе данного эксперимента проведено сравнение корректора MONS со следующими логическими алгоритмами:

- AVO — алгоритм вычисления оценок
- TV — алгоритм голосования по тупиковым тестам
- RS — алгоритм голосования по представительным наборам длины 3
- LR — логические закономерности классов
- JRip — реализация алгоритма RIPPER (Repeated Incremental Pruning to Produce Error Reduction)
- DT — решающие деревья

- J48 — реализация алгоритма C4.5 для построения решающих деревьев с отсечением ветвей
- LADT — Logitboost Alternating Decision Tree — бустинг над решающими деревьми
- SC — Classification and Regression Tree
- RF — Random Forest — баггинг над решающими деревьями

Для чистоты эксперимента для всех используемых алгоритмов, проведено сравнение качества распознавания на задачах до и после перекодировки. Результаты представлены в таблице 2, где для каждого алгоритма указан в процентах средний по задачам прирост качества после перекодирования. Для всех алгоритмов прирост положителен, поэтому дальнейшее сравнение полученных результатов с результатами построенных в дипломной работе корректоров правомерны.

AVO	TV	RS	LR	JRip	DT	J48	LADT	SC	RF
3,70%	5,97%	25,60%	5,14%	4,62%	0,17%	4,73%	3,43%	3,51%	3,35%

Таблица 2: Прирост качества распознавания после перекодирования данных.

В таблицах 6 и 7 представлены результаты тестирования десяти различных логических алгоритмов распознавания на 24 задачах.

В таблице 3 представлены результаты сравнения с корректором MONS. Для каждого алгоритма A в первой строке указано, на скольких задачах корректор MONS превзошел A по R, во второй — средняя по всем задачам разность оценок R для MONS и для A.

AVO	TV	RS	LR	JRip	DT	J48	LADT	SC	RF
20	19	6	24	22	21	21	21	21	22
12,62%	7,32%	-1,32%	24,57%	11,54%	26,51%	11,09%	10,35%	14,96%	10,09%

Таблица 3: Сравнение результатов MONS с результатами других логических алгоритмов.

9 Заключение

В дипломной работе получены следующие результаты:

- Предложена модель логического корректора MONS на базе эл.кл. произвольного ранга со стохастической процедурой формирования локального базиса.
- Проведены вычислительные эксперименты, которые подтвердили предположение, что снятие ограничения на ранг эл.кл. приводит к повышению качества распознавания.
- Показано превосходство корректора MONS по точности распознавания над рядом логических алгоритмов, таких как ABO, алгоритм голосования по тупиковым тестам, Логические Закономерности, RIPPER, различные алгоритмы на основе решающих деревьев.
- По результатам экспериментов корректор MONS в среднем по качеству распознавания незначительно уступил алгоритму голосования по представительным наборам. При этом MONS значимо превзошел его на задачах с малым числом часто встречающихся корректных эл.кл., где алгоритм голосования по представительным наборам показал результаты, сопоставимые с бросанием монетки.
- Построена новая модель логического корректора AMON на базе антимонотонных наборов из эл.кл. ранга 1.
- Доказано утверждение об эквивалентности корректоров MON и AMON в случае двух классов.
- Проведен ряд экспериментов с MON и AMON на задачах с более, чем двумя классами. Практически на всех задачах корректор модель с монотонными классифицирующими наборами показала себя лучше, чем с антимонотонными.

10 Список литературы

[1] *Дюкова Е. В., Журавлев Ю. И., Рудаков К. В.* Об алгебро-логическом синтезе корректных процедур распознавания на базе элементарных алгоритмов // Ж. вычисл. матем. и матем. физ., 2000. Т.40. №8. — С. 1264–1278.

[2] *Журавлев Ю. И.* Об алгебраическом подходе к решению задач распознавания или классификации // Проблемы кибернетики. №33, М.: Наука, 1978. — С.5–66.

[3] *Dyukova E. V., Zhuravlev Yu. I., Sotnezov M. R.* Construction of an Ensemble of Logical Correctors on the Basis of Elementary Classifiers // Pattern Recognition and Image Analysis, 2011, Vol. 21, №4, pp. 599—605.

[4] *Dyukova E. V., Prokofjev P. A.* Models of Recognition Procedures with Logical Correctors // Pattern Recognition and Image Analysis, 2013, Vol. 23, №2, pp. 235—244

[5] *Дюкова Е. В., Любимцева М. М., Прокофьев П. А.* Об алгебро–логической коррекции в задачах распознавания по прецедентам // Машинное обучение и анализ данных, 2013. Т.41. № 6. — С. 705–713.

[6] *Djukova E. V., Lyubimtseva M. M., Prokofjev P. A.* Logical correctors in recognition problems // 11th International Conference «Pattern Recognition and Image Analysis: New Information Technologies (PRIA–11–2013)», —- Samara, September 23–28, 2013. Conference Proceedings, Vol. I. Samara: IPSI RAS. P. 82–83.

[7] *Дюкова Е. В., Любимцева М. М., Прокофьев П. А.* Логические корректоры в задачах классификации по прецедентам // Тезисы докладов Всероссийской конференции «Математические методы распознавания образов (ММРО–16)», Казань, 6-–12 октября 2013. — М.: Торус Пресс. С. 7.

[8] *Дюкова Е. В.* Асимптотические оптимальные тестовые алгоритмы в задачах распознавания // Проблемы кибернетики. Вып. 39. М.: Наука, 1982. — С. 165–199.

[9] *Кузнецов В. Е.* Об одном стохастическом алгоритме вычисления информационных характеристик таблиц по методу тестов

[10] *Genrikhov I. E.* Synthesis and analysis of recognizing procedures on the basis of full decision trees // Pattern Recognition and Image Analysis, 2011, Vol. 21, no. 1., pp. 45-–51.

[11] *Воронцов К. В.* Лекции по логическим алгоритмам классификации — 2010. — http://www.machinelearning.ru.

[12] *Дюкова Е. В., Песков Н. В.* Поиск информативных фрагментов описаний объектов в дискретных процедурах распознавания // Ж. вычисл. матем. и матем. физ. 2002. Т. 42, № 5. С. 741-753.

[13] *Cheng–Jung Tsai, Chien–I. Lee, Wei–Pang Yang* A discretization algorithm based on Class-Attribute Contingency Coeficient // Information Sciences, 178, pp. 714-–731.

[14] *Yildiz O. T., Aslan O., Alpaydin E.* Multivariate statistical tests for comparing classification algorithms // Learning and Intelligent Optimization, 2011 — Springer.

Таблица 4: Характеристики используемых задач.

Название	m	n	k	Распределение объектов по классам	Доля уникальных значений	Размер булевой матрицы
botwinklbl	196	9	2	23/173	2.95%	3979×42
botwinSt	196	17	2	23/173	2.49%	3979×60
crest	100	2	2	50/50	30.00%	2500×36
dorovskih	33	12	2	16/17	25.76%	272×62
ech_l	60	8	2	40/20	10.42%	800×33
ech_r	71	8	2	48/23	7.39%	1104×32
echu	131	9	2	89/42	5.35%	3738×43
eco_l	144	7	4	76/33/24/11	2.28%	5168×16
Hea_r	136	13	5	78/24/14/15/5	8.37%	4524×95
Hep_l	72	19	2	15/57	5.26%	855×55
Hep_r	83	19	2	17/66	4.06%	1122×56
manelis1	145	35	2	38/107	1.91%	4066×92
manelis2	107	35	2	35/72	2.38%	2520×83
manelis3	73	35	2	38/35	5.83%	1330×129
manelis4	110	35	2	38/72	2.29%	2736×82
matchak2	132	24	2	30/102	5.05%	3060×111
melanoma	80	33	3	29/30/21	35.87%	1479×401
patomorfoz	77	7	2	47/30	7.42%	1410×24
Pnevmo_r	57	41	4	17/18/12/10	11.60%	680×123
SARComa	80	18	2	40/40	18.19%	1600×160
sigapur	58	15	2	17472	3.45%	517×30
stupenexper	61	18	2	39/22	16.67%	858×127
surv	77	8	2	52/25	16.72%	1300×67
TLMEL	48	33	3	17/20/11	37.88%	527×237
TRMEL	32	33	3	40522	38.92%	240×186
TUML	114	5	3	60/15/39	2.63%	3240×14

Таблица 5: Качество распознавания R для корректоров MON и MONS

	MON	MONS
botwinklbl	69,00%	**73,93%**
botwinSt	**63,56%**	61,90%
dorovskih	87,87%	**90,99%**
ech_l	72,50%	**86,25%**
ech_r	81,84%	**82,97%**
echu	84,37%	**87,73%**
eco_l	93,94%	**94,51%**
Hea_r	36,19%	**41,81%**
Hep_l	67,93%	**68,77%**
Hep_r	79,32%	**79,99%**
manelis1	**80,98%**	77,12%
manelis2	71,17%	**74,52%**
manelis3	**80,93%**	79,47%
manelis4	**81,79%**	78,66%
matchak2	65,71%	**69,22%**
melanoma	89,45%	**90,48%**
patomorfoz	91,21%	**91,96%**
Pnevmo_r	76,26%	**78,70%**
SARComa	**95,00%**	**95,00%**
sigapur	93,33%	**94,39%**
stupenexper	89,62%	**93,18%**
surv	75,12%	**78,15%**
TLMEL	**88,05%**	84,71%
TUML	63,85%	**76,84%**

Таблица 6: Качество распознавания R для ряда логических алгоритмов

	AVO	TstAlg	Repr Sets	Loreg	JRip
botwinklbl	50,00%	62,20%	**74,50%**	57,85%	60,00%
botwinSt	50,00%	62,05%	**64,64%**	54,20%	47,11%
dorovskih	32,70%	87,85%	**97,06%**	48,35%	57,90%
ech_l	60,00%	60,00%	**85,00%**	52,50%	67,50%
ech_r	74,25%	71,95%	**81,75%**	58,10%	77,49%
echu	74,75%	77,40%	**89,49%**	76,65%	76,95%
eco_l	87,25%	84,73%	94,27%	90,35%	90,63%
Hea_r	39,24%	**57,00%**	49,14%	27,36%	26,34%
Hep_l	73,15%	71,60%	70,70%	47,70%	**74,74%**
Hep_r	66,15%	74,75%	**84,54%**	71,15%	78,57%
manelis1	71,10%	67,00%	**87,90%**	73,00%	69,10%
manelis2	73,60%	72,25%	**75,91%**	61,10%	68,73%
manelis3	70,80%	70,50%	**92,97%**	59,40%	72,44%
manelis4	83,45%	80,90%	**87,79%**	75,25%	77,41%
matchak2	50,00%	66,25%	**72,55%**	47,10%	64,71%
melanoma	65,70%	70,77%	**90,52%**	36,53%	69,02%
patomorfoz	94,50%	88,95%	88,48%	86,80%	**94,54%**
Pnevmo_r	78,25%	69,83%	**85,00%**	29,70%	60,92%
SARComa	78,75%	80,00%	**96,25%**	68,75%	86,25%
sigapur	89,85%	93,30%	49,71%	37,40%	92,26%
stupenexper	67,60%	80,50%	**90,91%**	61,80%	74,13%
surv	50,00%	63,75%	**80,15%**	35,75%	46,23%
TLMEL	68,07%	64,43%	**95,30%**	25,70%	45,66%
TUML	79,13%	77,70%	**78,50%**	59,00%	75,64%

Таблица 7: Качество распознавания R для ряда логических алгоритмов

	DecTree	J48	LAD Tree	Simple Cart	Random Forest
botwinklbl	50,00%	53,77%	61,31%	62,47%	61,89%
botwinSt	50,00%	47,69%	62,91%	51,31%	54,21%
dorovskih	0,00%	60,29%	72,61%	39,34%	73,16%
ech_l	50,00%	63,75%	61,25%	65,00%	68,75%
ech_r	45,85%	65,58%	70,97%	73,05%	62,18%
echu	78,55%	82,91%	78,91%	78,01%	81,15%
eco_l	96,35%	95,31%	94,51%	**96,36%**	93,47%
Hea_r	24,44%	32,62%	28,15%	27,45%	29,16%
Hep_l	46,50%	69,65%	57,89%	58,95%	62,28%
Hep_r	44,70%	76,38%	78,57%	56,46%	72,68%
manelis1	72,35%	71,26%	75,59%	71,43%	75,68%
manelis2	50,00%	68,77%	66,65%	64,40%	70,89%
manelis3	72,35%	73,98%	80,79%	77,29%	87,48%
manelis4	79,55%	74,63%	83,30%	80,88%	77,34%
matchak2	50,00%	65,39%	65,88%	64,02%	63,43%
melanoma	57,20%	47,16%	59,14%	60,56%	59,78%
patomorfoz	94,50%	**94,54%**	87,87%	**94,54%**	92,87%
Pnevmo_r	17,40%	61,45%	44,90%	32,43%	61,18%
SARComa	0,00%	90,00%	91,25%	80,00%	85,00%
sigapur	92,25%	92,26%	93,33%	92,26%	**94,39%**
stupenexper	80,50%	87,06%	77,97%	79,25%	81,53%
surv	50,00%	65,58%	64,54%	50,08%	57,27%
TLMEL	30,87%	50,53%	48,10%	49,47%	48,98%
TUML	61,67%	74,57%	76,45%	67,26%	74,36%