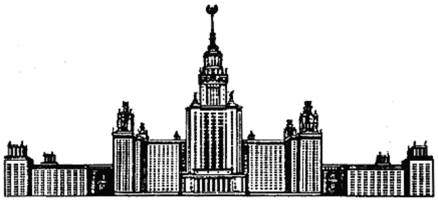
Московский государственный университет имени М. В. Ломоносова



Факультет Вычислительной Математики и Кибернетики

Кафедра Математических Методов Прогнозирования

КУРСОВАЯ РАБОТА СТУДЕНТА 317 ГРУППЫ

«Классификация по прецедентам: процедуры корректного голосования и анализ формальных понятий»

«Supervised classification: correct voting procedures and formal concept analysis»

Выполнил:

студент 3 курса 317 группы

Янаков Дмитрий Спартакович

Научный руководитель:

д.ф.-м.н., доцент

Дюкова Елена Всеволодовна

Москва, 2022

Содержание

Введение

Одной из основных задач машинного обучения является классификация по прецедентам. В общем случае эта задача ставится следующим образом.

Исследуется некоторое множество объектов M. Известно, что M представимо в виде объединения l непересекающихся подмножеств $K_1, \dots, K_l$, называемых классами. Объекты множества M описываются признаками $x_1, \dots, x_n$. Имеется конечный набор объектов $S_1, \dots, S_m \in M$, о которых известно, каким классам они принадлежат. Эти объекты называются прецедентами или обучающими объектами, а их описания имеют вид $S_i = (a_{i1}, \dots, a_{in})$, где a_{ij} – значение признака x_j для объекта S_i. Требуется по предъявленному набору значений признаков $(b_1, \dots, b_n)$, описывающему некоторый объект из M, о котором, вообще говоря, неизвестно какому классу он принадлежит, определить (распознать) этот класс.

Для решения сформулированной задачи успешно используется аппарат дискретной математики. Главным достоинством данного подхода, известного как дискретный или логический, является получение результата при отсутствии дополнительных предположений вероятностного характера и при небольшом числе прецедентов. Рассматриваемый подход включает три основных направления: Correct Voting Procedures (CVP), Formal Concept Analysis (FCA) и Logical Analysis of Data (LAD).

Центральными в CVP [3, 6] являются вопросы построения моделей классификаторов, которые безошибочно распознают обучающие объекты. Фундаментальную роль в создании отечественных процедур корректного голосования сыграли работы С.В. Яблонского, в которых введено хорошо известное в дискретной математике понятие теста [11], и работы Ю.И. Журавлева, опубликованные в 70-х и 80-х годах прошлого века. Основы проблематики заложены также в статьях российских ученых М.М. Бонгарда и М.Н. Вайнцвайга, в которых описывался распознающий алгоритм «Кора» [1]. В дальнейшем это направление в основном развивалось в работах Ю.И.

Журавлева, Е.В. Дюковой, Н.В. Пескова, П.А. Прокофьева, Г.О. Маслякова и др. [5, 12].

На этапе обучения классификаторов используются методы построения покрытий булевых и целочисленных матриц и преобразования нормальных форм логических функций. Основными понятиями для CVP являются понятие (корректного) элементарного классификатора и понятие корректного набора элементарных классификаторов.

Основополагающие идеи FCA предложены в начале 1980-х годов Рудольфом Вилле [20]. Данное направление изучает, как объекты могут быть иерархически сгруппированы вместе с их общими признаками. В качестве математического аппарата, FCA использует аппарат теории множеств и теории решеток. В 1976 г. В.К. Финном предложен так называемый метод автоматического порождения гипотез или ДСМ-метод (как средство для формализации схемы правдоподобного и достоверного вывода) [10]. Использование аппарата FCA позволило в дальнейшем применять ДСМ-метод для решения задач классификации с бинарными данными, представленными положительными и отрицательными примерами.

Первоначальная идея LAD предложена Питером Хаммером в 1986 году [14]. Построение алгоритмов классификации в данном направлении базируется на поиске так называемых логических закономерностей в описаниях прецедентов, для нахождения которых в основном применяются оптимизационные методы. Для описания моделей, как правило, используются понятия из теории логических функций.

В России методы LAD и FCA развиты в работах И.С. Масича, В.В. Рязанова [8, 9] и С.О. Кузнецова, Д.И. Игнатова, М.И. Забежайло [7, 16] соответственно. За рубежом идеи LAD излагаются в работах Хаммера и Когана [15], а FCA – в работах Гантера [13].

Все три направления логического анализа данных (CVP, FCA и LAD) имеют много общего. В частности, в случае бинарных данных, логическая закономерность – это элементарный классификатор специального вида. Есть схожесть и между основными понятиями из FCA и CVP. С другой стороны, каждый из подходов демонстрирует некоторую оригинальность.

В настоящей работе приведён сравнительный анализ методов CVP и FCA в случае бинарных данных (каждый признак имеет два допустимых значения). С использованием алгоритма из [19] реализован на C++ ДСМ-классификатор. Тестирование ДСМ-классификатора проведено на модельных и реальных данных, бинаризация которых выполнена с помощью OneHot-кодирования.

1. Корректное голосование (Correct Voting Procedures)

Пусть H – набор из r различных признаков вида $H = \{x_{j_1}, \dots, x_{j_r}\}$, $\sigma = (\sigma_1, \dots, \sigma_r)$, σ_i – допустимое значение признака x_{j_i}, $i = \overline{1, r}$. Пару (σ, H) назовем *элементарным классификатором (эл.кл.).*

Близость объекта $S = (a_1, \dots, a_r) \in M$ и эл.кл. (σ, H), $\sigma = (\sigma_1, \dots, \sigma_r), H = \{x_{j_1}, \dots, x_{j_r}\}$, будем оценивать величиной $B(\sigma, S, H)$ равной 1, если $a_{j_t} = \sigma_t$ при $t = \overline{1, r}$, и равной 0 в противном случае.

Если $B(\sigma, S, H) = 1$, то будем говорить, что объект S содержит эл.кл. (σ, H).

Эл.кл. (σ, H) является *корректным для класса* K, если нельзя указать пару обучающих объектов S' и S'': $S' \in K$, $S'' \notin K$ и $B(\sigma, S', H) = B(\sigma, S'', H) = 1$.

Пусть $\overline{K} = \{K_1, \dots, K_l\} \backslash K$.

Эл.кл. (σ, H) называется *представительным для класса* K, если ни один обучающий объект из $\overline{K}$ не содержит (σ, H) и хотя бы один обучающий объект из K содержит (σ, H).

Представительный эл.кл. (σ, H) для класса K называется *тупиковым*, если не является представительным для K любой эл.кл. вида (σ', H'), где $\sigma' \subset \sigma$, $H' \subset H$.

В классической модели алгоритма голосования по тупиковым представительным эл.кл. для каждого класса K строится множество тупиковых представительных эл.кл., обозначаемое далее через $\mathcal{T}(K)$. Для нахождения $\mathcal{T}(K)$ используются алгоритмы монотонной дуализации, например, асимптотически оптимальные алгоритмы, впервые предложенные Е. В. Дюковой [4]. Эти алгоритмы на сегодняшний день являются лидерами по скорости счета. Задача монотонной дуализации относится к числу труднорешаемых задач дискретной математики и формулируется как построение сокращенной дизъюнктивной нормальной формы монотонной булевой функции, заданной конъюнктивной нормальной формой. Труднорешаемость дуализации обусловлена экспоненциальным ростом числа решений с ростом размера задачи и сложностью нахождения каждого нового решения.

Распознавание объекта S осуществляется на основе процедуры голосования. Для этого вычисляется оценка принадлежности объекта S классу K по следующей формуле:

$$\Gamma(S,K) = \frac{1}{|\mathcal{T}(K)|} \sum_{(\sigma,H)\in\mathcal{T}(K)} P_{(\sigma,H)} B(\sigma,S,H),$$

где $P_{(\sigma,H)}$ – вес эл.кл. (σ,H). В качестве $P_{(\sigma,H)}$ обычно берется число обучающих объектов из K, содержащих (σ,H).

Объект S относится к классу с наибольшей оценкой. Если таких классов несколько, то происходит отказ от классификации.

2. Анализ формальных понятий (Formal Concept Analysis)

Формальным контекстом называется тройка вида $C = (G,X,I)$, где G – множество объектов, X – множество признаков, I – бинарное отношение между множествами G и X. Запись вида $(S,x) \in I$ означает, что объект $S \in G$ обладает признаком $x \in X$.

Иначе говоря, формальный контекст – это булева матрица L, строками которой являются признаковые описания объектов. Если объект S обладает признаком x, то соответствующий элемент матрицы L равен 1, в противном случае равен 0.

Для множества $A \subseteq G$, положим $A' = \{x \in X \,|(S, x) \in I \;\; \forall S \in A\}$. Аналогично, для множества $B \subseteq X$, положим $B' = \{S \in G \,|(S, x) \in I \;\; \forall x \in B\}$. Полученные множества A' и B' называются *операторами вывода* для формального контекста $C = (G, X, I)$.

То есть, A' – это оператор, возвращающий столбцы матрицы L, которые в пересечении с заданными строками образуют «максимальную» подматрицу, все элементы которой равны 1. Аналогично, B' – это оператор, возвращающий строки матрицы L, которые в пересечении с заданными столбцами образуют «максимальную» подматрицу, все элементы которой равны 1.

На рисунке 1 приведен пример формального контекста с четырьмя геометрическими фигурами и четырьмя признаками.

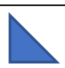
№	G \ X	x_1	x_2	x_3	x_4
S_1		1	0	0	1
S_2		1	0	1	0
S_3		0	1	1	0
S_4		0	1	1	1

Рис. 1

В случае использования операторов вывода в данном контексте, можно получить, к примеру, следующие результаты:

1) $\{S_1, S_4\}' = \{x_4\}$;
2) $\{x_2, x_3\}' = \{S_3, S_4\}$;
3) $\{S_1, S_2, S_3, S_4\}' = \emptyset$.

Формальным понятием формального контекста $C = (G, X, I)$ называется пара вида (A, B), $A \subseteq G, B \subseteq X$, такая, что $A' = B$, $B' = A$.

Заметим, что пара (A, B) образует «максимальную» подматрицу матрицы L, все элементы которой равны 1. Причём число строк или столбцов в полученной подматрице нельзя увеличить.

В случае рассмотренного выше примера, формальными понятиями являются:

1) $(\{S_1\},\ \{x_1, x_4\})$;
2) $(\{S_3, S_4\},\ \{x_2, x_3\})$;
3) $(\{S_2, S_3, S_4\}, \{x_3\})$.

3. ДСМ-метод в задаче классификации

ДСМ-метод изначально был сформулирован в терминах математической логики, однако позже была установлена эквивалентность между ДСМ-гипотезами и формальными понятиями [17].

Будем называть эл.кл. (σ, H) *формальным*, если все элементы σ равны 1.

Формальное понятие (A, B) порождает формальный эл.кл. вида (σ, B), который называется *положительной ДСМ-гипотезой*, если (σ, B) – представительный эл.кл. для класса K, и *отрицательной ДСМ-гипотезой*, если эл.кл. (σ, B) не содержится ни в одном обучающем объекте из K, и содержится хотя бы в одном обучающем объекте из $\overline{K}$.

Обозначим через $R(K)$ – множество прецедентов класса K, а через $R(\overline{K})$ – множество прецедентов, не принадлежащих классу K. То есть $R(\overline{K}) = \{S_1, \dots, S_m\} \backslash R(K)$.

На этапе обучения ДСМ-метода необходимо найти все формальные понятия формального контекста $(R(K), X, I)$, которые порождают положительные ДСМ-гипотезы, а также все формальные понятия формального контекста $(R(\overline{K}), X, I)$, которые порождают отрицательные ДСМ-гипотезы.

Существует немало алгоритмов для нахождения формальных понятий. Один из них – *Close By One* был предложен Кузнецовым С.О. в [18]. С помощью него будут проведены дальнейшие эксперименты.

Один шаг алгоритма *Close By One* соответствует операции «*Замыкай-по-одному-вниз*» ($CbODown$), определяемой следующим образом:

$$CbODown((A,B),S) \ = ((A \cup \{S\})'', B \cap \{S\}')$$

Корректность алгоритма обеспечивается тем, что для любого формального понятия (A,B) и любого объекта $S \in G$, пара $CbODown\big((A,B),S\big)$ также является формальным понятием [2].

Псевдокод алгоритма приведён ниже.

Алгоритм. Close By One

Вход: $C = (G,X,I)$ – формальный контекст

Выход: L – множество формальных понятий

1. $L \coloneqq \emptyset$
2. *Для каждого* $S \in G$
3. $\quad$ *Process* $(\{S\},\ S,\ (\{S\}'',\ \{S\}'))$
4. *Вернуть* L

Процедура. Process $\big(A, S, (C, D)\big)$, причем $C = A'', D = A'$

1. *Если* $\{H\ | H\ \in \mathrm{C} \backslash A$ и $H < S\} = \emptyset$
2. $\quad$ $L \coloneqq \ L \cup \{(C,D)\}$
3. $\quad$ *Для каждого* $F\ \in \{H\ |H\ \in G$ и $S < H\}$
4. $\quad\quad$ $Z \coloneqq \ C \cup \{F\}$
5. $\quad\quad$ $Y \coloneqq \ D \cap \{F\}'$
6. $\quad\quad$ $X \coloneqq Y'$
7. $\quad\quad$ *Process* $(Z,\ F,\ (X, Y))$

Считается, что на множестве объектов задан лексикографический порядок. В данном случае, это означает, что объект $S_i < S_j$, если $(S_i)_{10} < (S_j)_{10}$, где $(S)_{10}$ – это десятичное представление двоичного числа S. К примеру, в случае контекста с геометрическими фигурами (см. Рис.1), объект S_3 будет меньше объекта S_4, так как $0110_2 = 6_{10} < 7_{10} = 0111_2$. Формальные понятия перечисляются в лексикографическом порядке.

Классификация распознаваемого объекта S происходит при помощи построенных гипотез. Для этого в ДСМ-методе используется более строгая процедура, чем процедура голосования:

- если S содержит хотя бы одну положительную гипотезу и не содержит ни одну отрицательную, то S **относится к классу** $\boldsymbol{K}$;
- если S содержит хотя бы одну отрицательную гипотезу и не содержит ни одну положительную, то S **не относится к классу** $\boldsymbol{K}$;
- если S содержит как положительную, так и отрицательную гипотезу, или если S не содержит ни одну гипотезу, то происходит **отказ от классификации**.

Стоит отметить, что на практике данная процедура приводит к большему числу отказов, по сравнению со стандартной процедурой корректного голосования, что снижает качество классификации.

4. Экспериментальное исследование

Изначально была рассмотрена задача нахождения всех формальных понятий формального контекста, которая решалась, как уже было отмечено, с помощью алгоритма *Close By One*, реализованного на языке программирования C++.

Целью экспериментов было выявить зависимость времени работы алгоритма от числа строк/столбцов булевой матрицы из дискретного равномерного распределения. Результаты вычислений, усредненные по десяти запускам, изображены на графиках ниже.

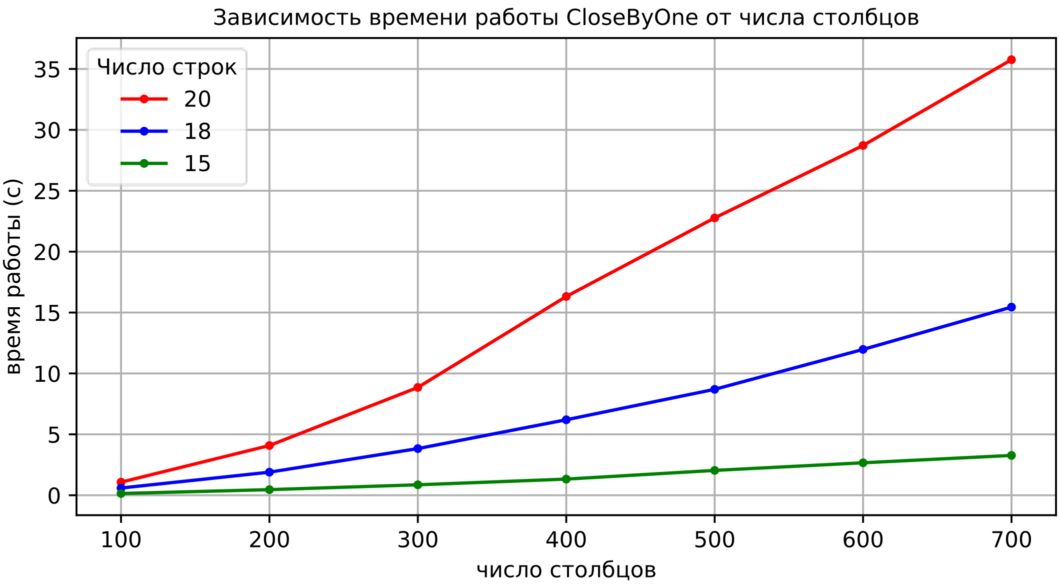


Рис. 2

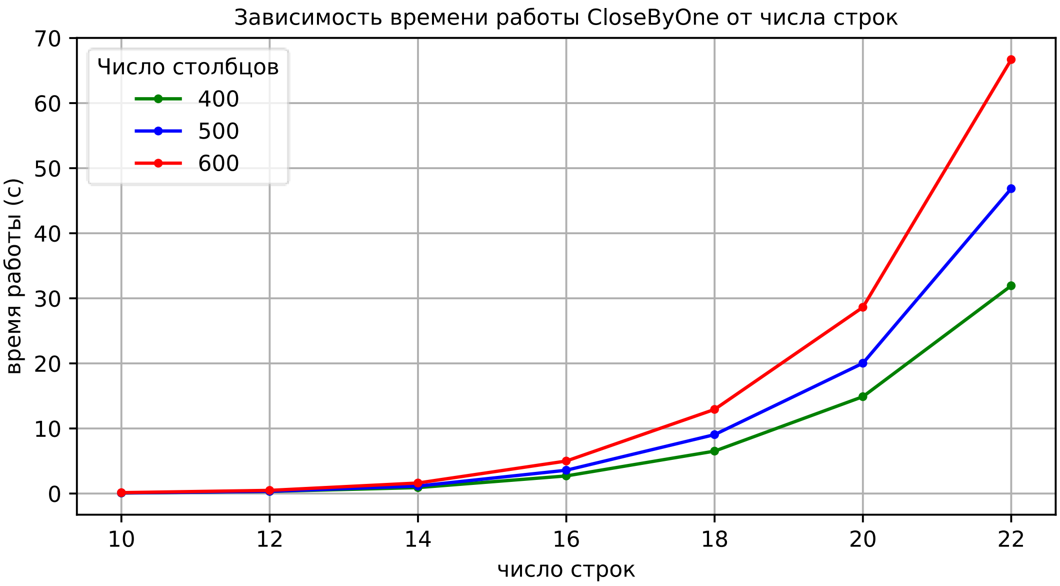


Рис. 3

Из графиков можно сделать следующие выводы:

1) рост числа строк сильнее увеличивает время работы алгоритма, чем рост числа столбцов;
2) время работы алгоритма зависит от числа столбцов линейно;
3) время работы алгоритма зависит от числа строк кубически.

Логично предположить, что ДСМ-классификатор, использующий данный алгоритм для нахождения положительных и отрицательных гипотез, будет хорошо работать при небольшом числе строк (как показали эксперименты, при количестве строк > 35, время работы могло достигать нескольких часов, или алгоритм мог вовсе не завершать работу). Однако в задачах машинного обучения построение классификатора при количестве строк меньшем, чем количество столбцов обычно не даёт хорошего результата. Поэтому была предложена модификация для ускорения работы алгоритма: для нахождения формальных понятий (A, B) исходной матрицы (у которой число строк больше, чем число столбцов), искались формальные понятия вида (B, A) транспонированной матрицы, после чего в них менялись местами множества A и B .

Далее была рассмотрена задача классификации, которая решалась при помощи ДСМ-метода (с предложенной модификацией). Эксперименты проводились на реальных данных из репозитория UCI: датасетах «Молекулярная биология» (МБ), «Крестики-нолики» (КН) и «Машины» (М).

Размеры задач приведены в таблице ниже.

Датасет	Обучающая выборка	Тестовая выборка	Количество признаков
МБ	1228	307	284
КН	765	192	27
М	1274	319	21

Рис. 4

Основной целью было сравнить ДСМ-метод с алгоритмом голосования по тупиковым представительным эл.кл., поиск которых был произведен с помощью алгоритма из CVP (обозначим через CVP). Результаты по счету для алгоритма из CVP предоставил Н.А. Драгунов, аспирант ФИЦ ИУ РАН.

Помимо классической процедуры распознавания из ДСМ-метода (обозначим через ДСМ), была рассмотрена также процедура голосования (обозначим через ДСМ-Г), в которой объект из тестовой выборки относился к тому классу, чьих гипотез он содержал больше. При равном количестве происходил отказ от классификации.

Результаты экспериментов, усреднённые по десяти запускам при случайных разбиениях датасетов на обучающую и тестовую выборки, приведены в таблице ниже.

	Время			Точность		
Датасет	ДСМ	ДСМ-Г	CVP	ДСМ	ДСМ-Г	CVP
МБ	—		50 мин	—		0.963
КН	43 мин		95 сек	0.448	0.986	0.968
М	75 сек		170 сек	0.684	0.943	0.841

Рис. 5

Из таблицы можно сделать следующие выводы:

1) классический ДСМ-метод показал результат хуже, чем алгоритм голосования по тупиковым представительным эл.кл.: как по времени, так и по качеству классификации;
2) алгоритм ДСМ-Г не отличился по времени работы от ДСМ, однако достиг наилучшей точности среди всех трёх алгоритмов на тех задачах, в которых был получен результат;
3) на датасете «Молекулярная биология» ДСМ-метод не отработал. Произошло это потому, что размерность задачи достаточно велика для алгоритма *Close By One*.

Как отмечалось ранее, низкое качество классификации классического ДСМ-метода связано с тем, что происходит достаточно много отказов от распознавания. Однако стоит отметить тот факт, что если убрать объекты, на которых произошли отказы, то качество классификации будет достаточно высоким: 1.0 для «Крестиков-ноликов» и 0.986 для «Машин».

Таким образом, в результате проведенных экспериментов было установлено, что ДСМ-метод в его исходном виде достиг весьма хорошей точности на тех объектах, которые удалось классифицировать. Модификация ДСМ-Г показала наилучшее качество на всех объектах, однако оба эти алгоритма, в среднем, работали гораздо дольше, чем алгоритм голосования по тупиковым представительным эл.кл., который, к тому же, отработал на матрице большой размерности.

5. Заключение

В настоящей работе рассмотрены подходы FCA и CVP к решению задачи классификации по прецедентам в случае бинарных данных. Реализован ДСМ-метод, а также предложена его модификация. Проведено экспериментальное исследование на случайных данных различной размерности, а также сравнение на реальных данных с алгоритмом голосования по тупиковым представительным эл.кл., используемому в CVP. Показано, что модификация ДСМ-метода существенно увеличивает точность классификации, но все равно проигрывает по времени алгоритму из CVP.

Список литературы

1. *Вайнцвайг М.Н.* Алгоритм обучения распознаванию образов «Кора» // Алгоритмы обучения распознаванию образов. М.: Сов. Радио. 1973. С. 82–91.
2. *Виноградов Д.В.* Вероятностно-комбинаторный подход к автоматическому порождению гипотез // В кн.: Гуманитарные чтения. — М.: РГГУ, 2015. С. 771–775
3. *Дюкова Е.В., Журавлёв Ю.И., Рудаков К.В.* Об алгебраическом синтезе корректирующих процедур распознавания на базе элементарных алгоритмов // Ж. вычисл. матем. и матем. физ. 1996. Т. 36, №8. С. 215–223.
4. *Дюкова Е. В., Инякин А. С.* Об асимптотически оптимальном построении тупиковых покрытий целочисленной матрицы // Математические вопросы кибернетики. Вып. 17 — М.: Физматлит, 2008. С. 247–262.
5. *Дюкова Е.В., Масляков Г.О., Прокофьев П.А.* О логическом анализе данных с частичными порядками в задаче классификации по прецедентам // Ж. вычисл. матем. и матем. физ., 2019. Т. 59, № 9. С. 1605–1616
6. *Дюкова Е. В., Песков Н. В.* Построение распознающих процедур на базе элементарных классификаторов // Математические вопросы кибернетики. Вып. 14. — М.: Физматлит, 2005. С. 57–92.
7. *Забежайло М.И., Трунин Ю.Ю.* К проблеме доказательности медицинского диагноза: интеллектуальный анализ эмпирических данных о пациентах в выборках ограниченного размера // Ж. Научно-техническая информация. Серия 2: Информационные процессы и системы. — М.: ВИНИТИ, 2019. С. 12–18
8. *Масич И.С.* Поисковые алгоритмы решения задач условной псевдобулевой оптимизации // Ж. системы управления, связи и безопасности, 2016. №1, С. 1605–1616

9. *Рязанов В. В.* Логические закономерности в задачах распознавания (параметрический подход) // Ж. вычисл. матем. и матем. физ. 2007. Т. 47, №10. С. 1793–1808.
10. *Финн В. К.* О возможности формализации правдоподобных рассуждений средствами многозначных логик // Всесоюзн. симп. по логике и методологии науки. — Киев: Наукова думка, 1976. С. 82–83.
11. *Чегис И. А., Яблонский С. В.* Логические способы контроля электрических схем // Труды математического института им. В. А. Стеклова АН СССР. 1958. Т. 51.С. 270–360.
12. *Djukova E.V, Masliakov G.O.,* Correct classification over a product of partial orders // IEEE Proceedings of the VII International Conference on Information Technology and Nanotechnology, Samara, Russia, 2021. P. 1–5.
13. *Ganter, B., Wille, R.,* Formal Concept Analysis: Mathematical Foundations (Heidelberg: Springer), 1999
14. *Hammer P.L.* Partially defined boolean functions and cause-effect relationships. // International Conference on Multi-attribute Decision Making Via OR-based Expert Systems. University of Passau. Passau. Germany, April, 1986.
15. *Hammer P., Bonates T., Kogan A.,* Maximum patterns in datasets. Discrete Applied Mathematics. 2008, Vol. 156(6), P. 846–861.
16. *Karpov N., Ignatov D.I., Braslavski P.*, Information Retrieval // 8th Russian Summer School, RuSSIR 2014, Nizhniy, Novgorod, Russia, August 18–22, 2014, Revised Selected Papers
17. *Kuznetsov S.O.,* Mathematical aspects of concept analysis // Journal of Mathematical Science, 1996, Vol. 80, Issue 2, pp. 1654–1698.
18. *Kuznetsov S.O.,* A fast algorithm for computing all intersections of objects from an arbitrary semilattice // Nauchno-Tekhnicheskaya Informatsiya. Seriya 2 — Informatsionnye protsessy i sistemy, 1993, No. 1, pp.17–20.

19. *Kuznetsov S.O. and Obiedkov S.A.*, Comparing Performance of Algorithms for Generating Concept Lattices // Journal of Experimental and Theoretical Artificial Intelligence, 2002, Vol. 14, no. 2–3, pp. 189–216.
20. *R. Wille,* Restructuring lattice theory: An approach based on hierarchies of concepts // Rival, I., Ed., Ordered Sets: Proceedings. NATO Advanced Studies Institute, 83, Reidel, Dordrecht. 1982, 445–470.