19-я Всероссийская конференция с международным участием «Математические методы распознавания образов»

26-29 ноября 2019, Москва, Россия

Высокопроизводительный метод средних решающих правил для решения больших двухклассовых задач SVM в пространстве признаков

Курбаков М. Ю., Макарова А. И., Сулимова В. В.

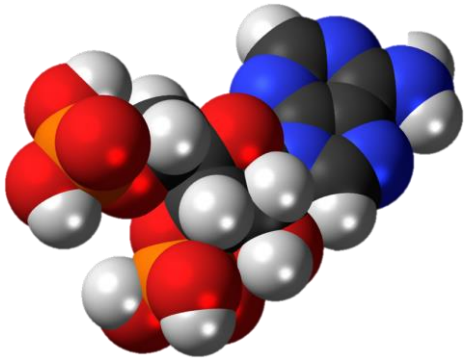
muwsik@mail.ru, aleksarova@gmail.ru, vsulimova@yandex.ru

Тульский государственный университет

Лаборатория анализа данных

Области применения двухклассового распознавания

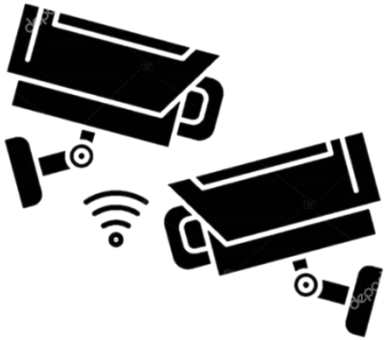
Молекулярная биология



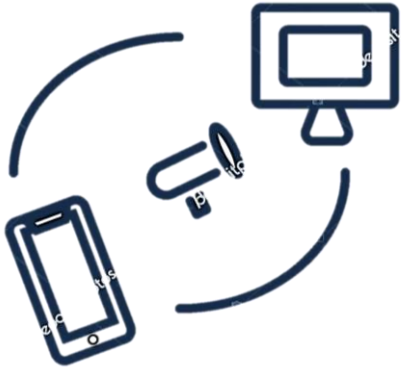
Горнодобывающая и нефтяная промышленности

Медицинские системы

Системы видеонаблюдения



Маркетинговые системы



Анализ текстовых данных

Общая особенность: большой объём исходных данных для обучения.

Постановка задачи двухклассового распознавания

Исходные данные:

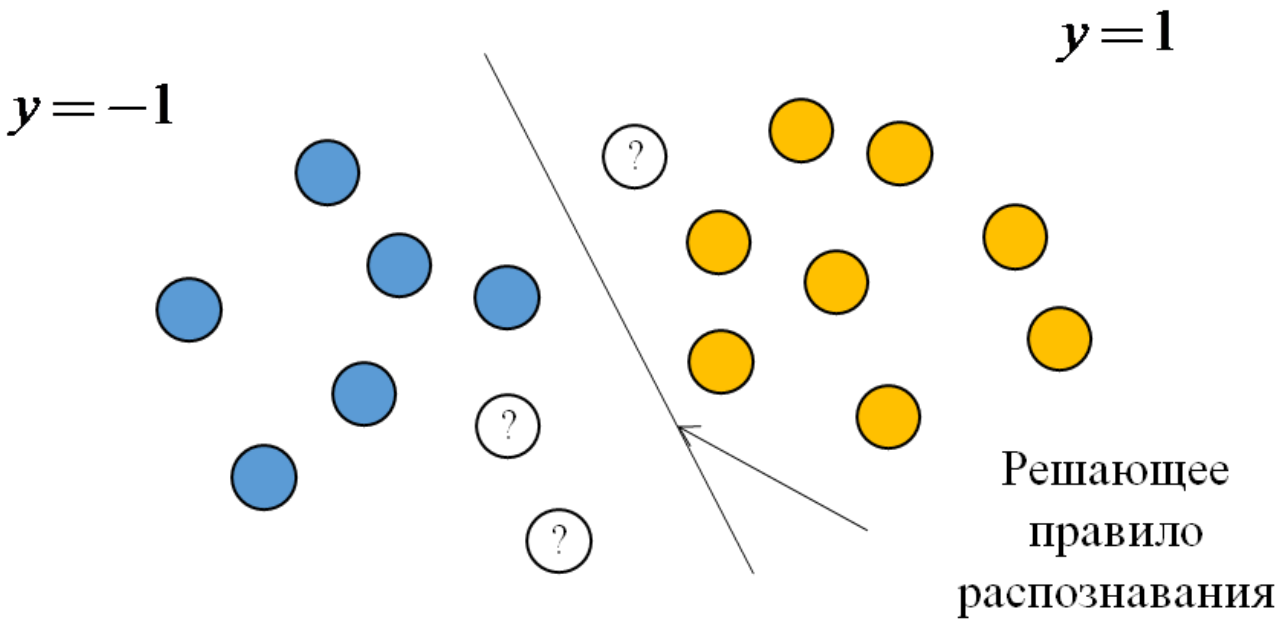
Ω - совокупность объектов, где $\omega \in \Omega$ - объект произвольного вида;

Неизвестная функция принадлежности объектов к классу: $y(\omega) : \Omega \to \{+1; -1\}$;

Обучающая совокупность объектов: $\{(\omega_j, y_j), j = 1,...,N\}, \quad \omega_j \in \Omega^* \subset \Omega, \quad y_j = y(\omega_j) \in \{+1; -1\}$;

Требуется:

Найти решающее правило распознавания всевозможных объектов ω: $\hat{y}(\omega) = \pm 1$



Метод опорных векторов (SVM). Обучение в линейном признаковом пространстве

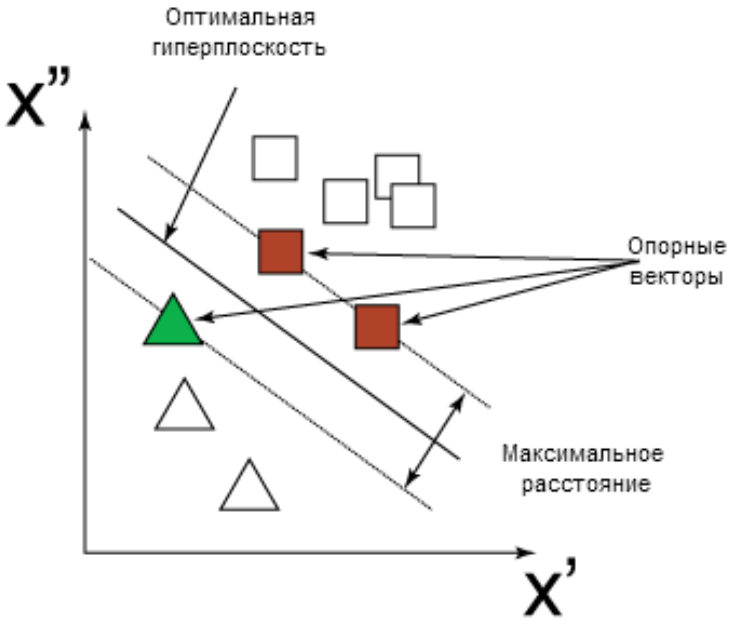
Представление объектов в виде точек в m-мерном признаковом пространстве: $\mathbf{x}(\omega) \in R^m$

Обучающее множество: $\{\mathbf{x}_j, y_j\}, \quad \mathbf{x}_j = \mathbf{x}_j(\omega), \quad j = 1,\ldots,N$

Решающее правило в виде линейной разделяющей гиперплоскости:

$$d(\mathbf{x};\mathbf{a},b) = \mathbf{a}^T\mathbf{x} + b \qquad \begin{matrix} \geq 0 \Rightarrow \hat{y}(\mathbf{x}) = +1, \\ < 0 \Rightarrow \hat{y}(\mathbf{x}) = -1, \end{matrix}$$

$\mathbf{a} \in R^m$ - направляющий вектор;
b - смещение вдоль направляющего вектора;



Недостаток:

- Высокая вычислительная сложность в условиях больших объёмов данных.

Метод средних решающих правил*

Исходная обучающая совокупность: $[X,Y]$, $X=[\mathbf{x}_j, j=1,...,N]$, $\mathbf{x}_j \in R^m$

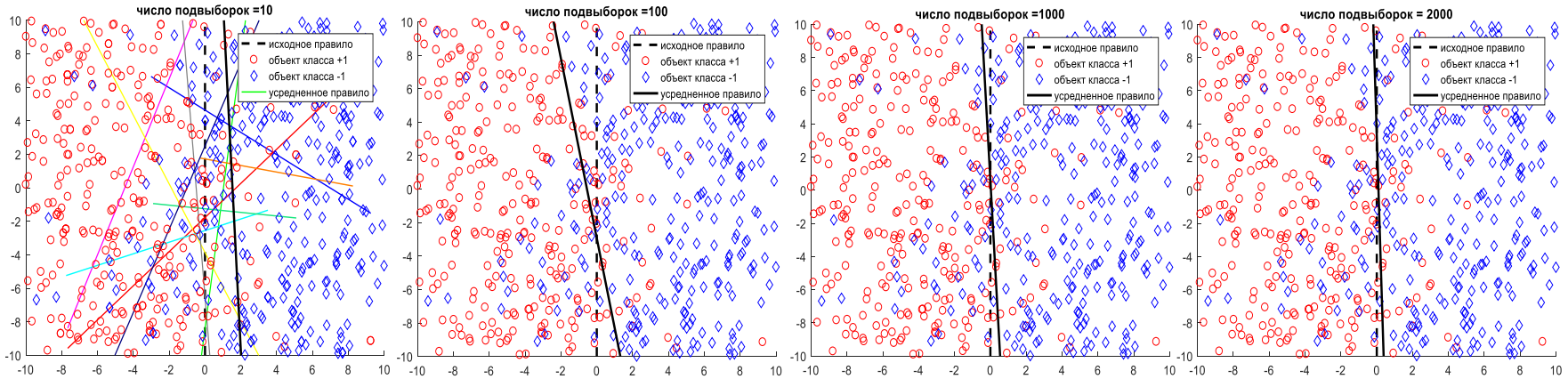
$Y=[y_j, j=1,...,N]$, $y_j \in \{-1;1\}$

Набор случайных подвыборок: $[X,Y]^{(i)} \in [X,Y]$, $i=1,...,k$

Результатом обучения по i-й подвыборке $[X,Y]^{(i)}$ является частное решающее правило с параметрами $[\mathbf{a}^{(i)}, b^{(i)}]$, $i=1,...,k$

Усредненное решающее правило: $[\mathbf{a},b]$, $\mathbf{a}=\frac{1}{k}\sum_{i=1}^{k}\mathbf{a}^{(i)}$, $b=\frac{1}{k}\sum_{i=1}^{k}b^{(i)}$

Изменение положения усредненной гиперплоскости при изменении числа подвыборок



*Макарова А.И., Сулимова В.В. Быстрое приближенное решение задачи SVM для больших обучающих совокупностей // Труды международной конференции ITNT-2019

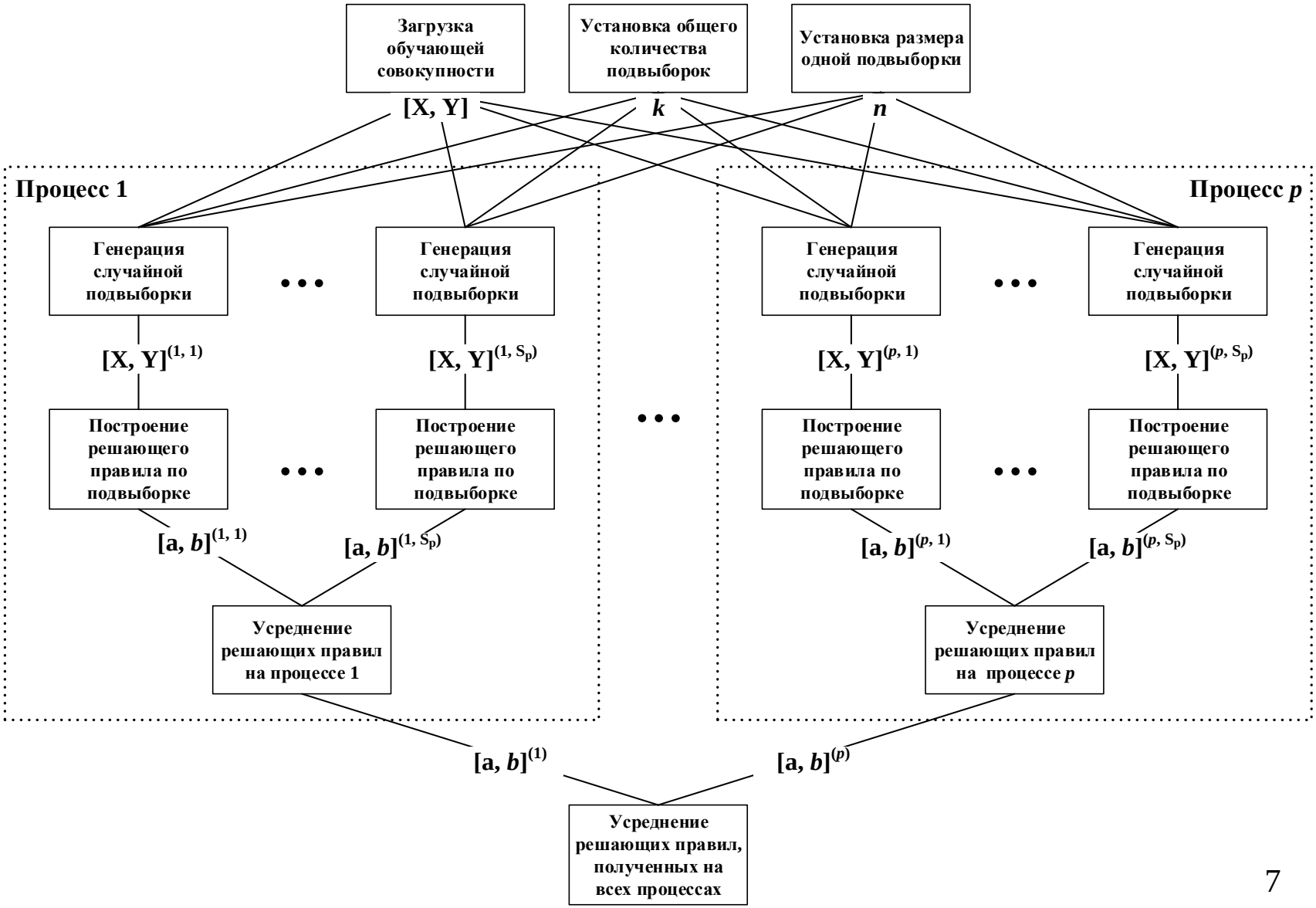
Важные достоинства метода средних решающих правил

- Позволяет быстро найти приближенное, но при этом достаточно близкое к точному, решение задачи SVM даже на одной вычислительной машине;

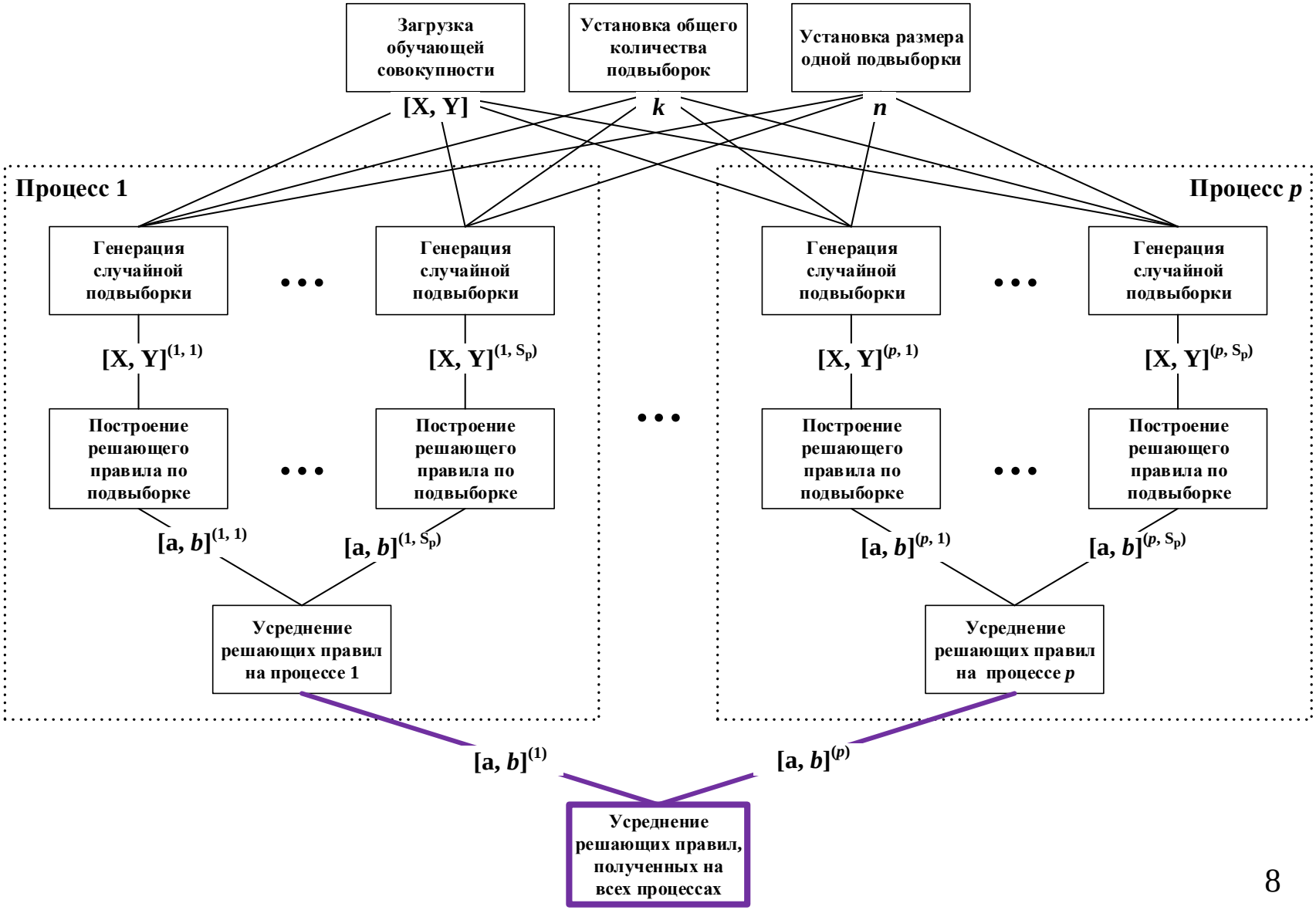
- Не имеет теоретического ограничения размера обучающей совокупности, поскольку не требует одновременного хранения всех объектов в оперативной памяти;

- Обладает высокой степенью параллелизма по данным.

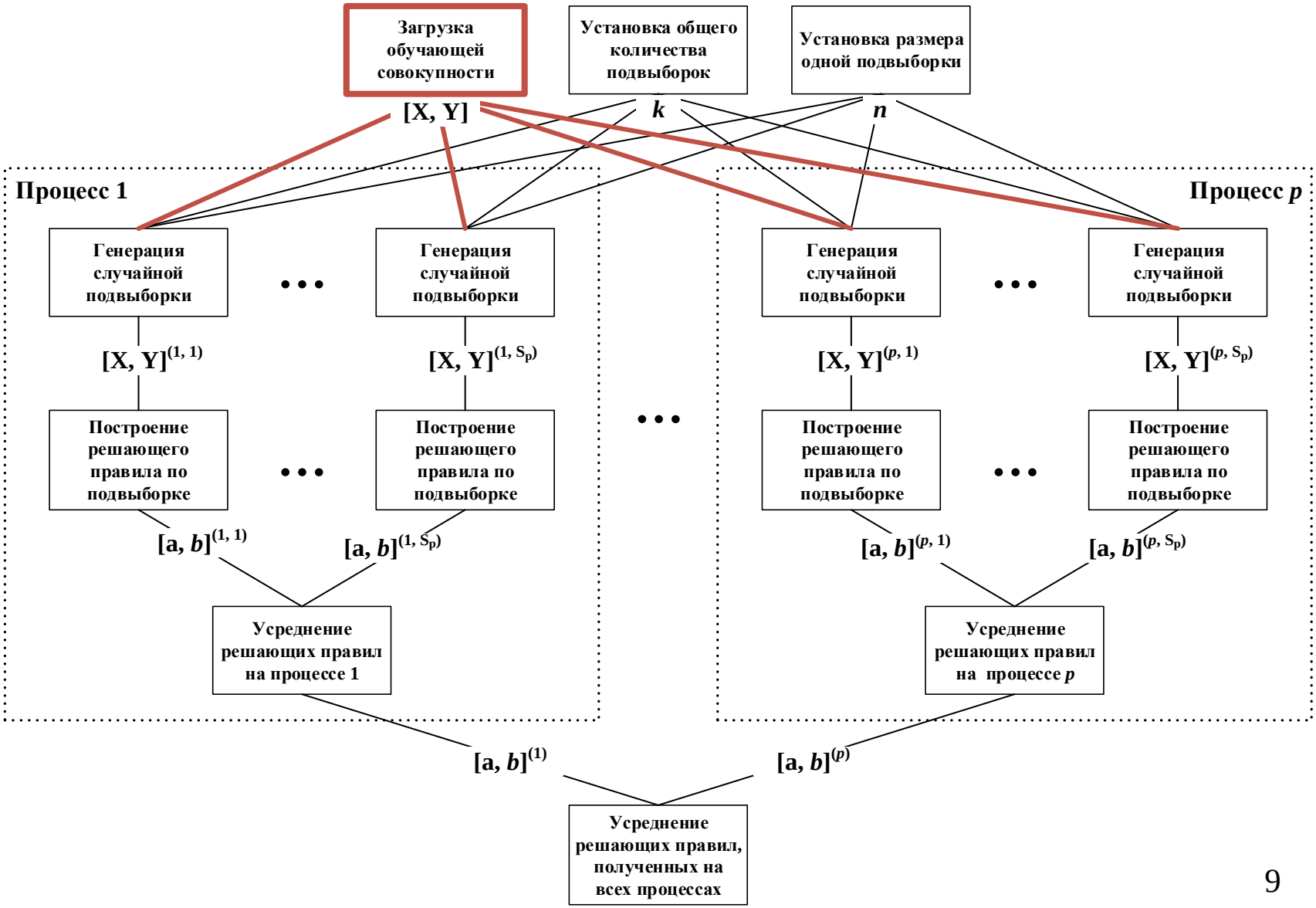
Модель параллельных вычислений в виде графа операции-операнды для метода средних решающих правил



Модель параллельных вычислений в виде графа операции-операнды для метода средних решающих правил



Модель параллельных вычислений в виде графа операции-операнды для метода средних решающих правил



Ограничения применения метода средних решающих правил

Традиционный способ работы с данными предполагает **единовременное чтение всей обучающей совокупности** в оперативную память.

Следствия:

- С ростом числа объектов время их чтения увеличивается, снижая эффективность работы метода;
- Обучающая совокупность может не поместиться целиком в оперативную память, ограничивая применение метода.

Вариант решения:

Разработка специальной стратегии, которая позволила бы оптимизировать работу с данными.

Формат файла исходных данных libsvm

Каждая строка соответствует одному объекту и имеет следующую структуру:

<метка класса> <номер признака>:<значение> ... <номер признака>:<значение>

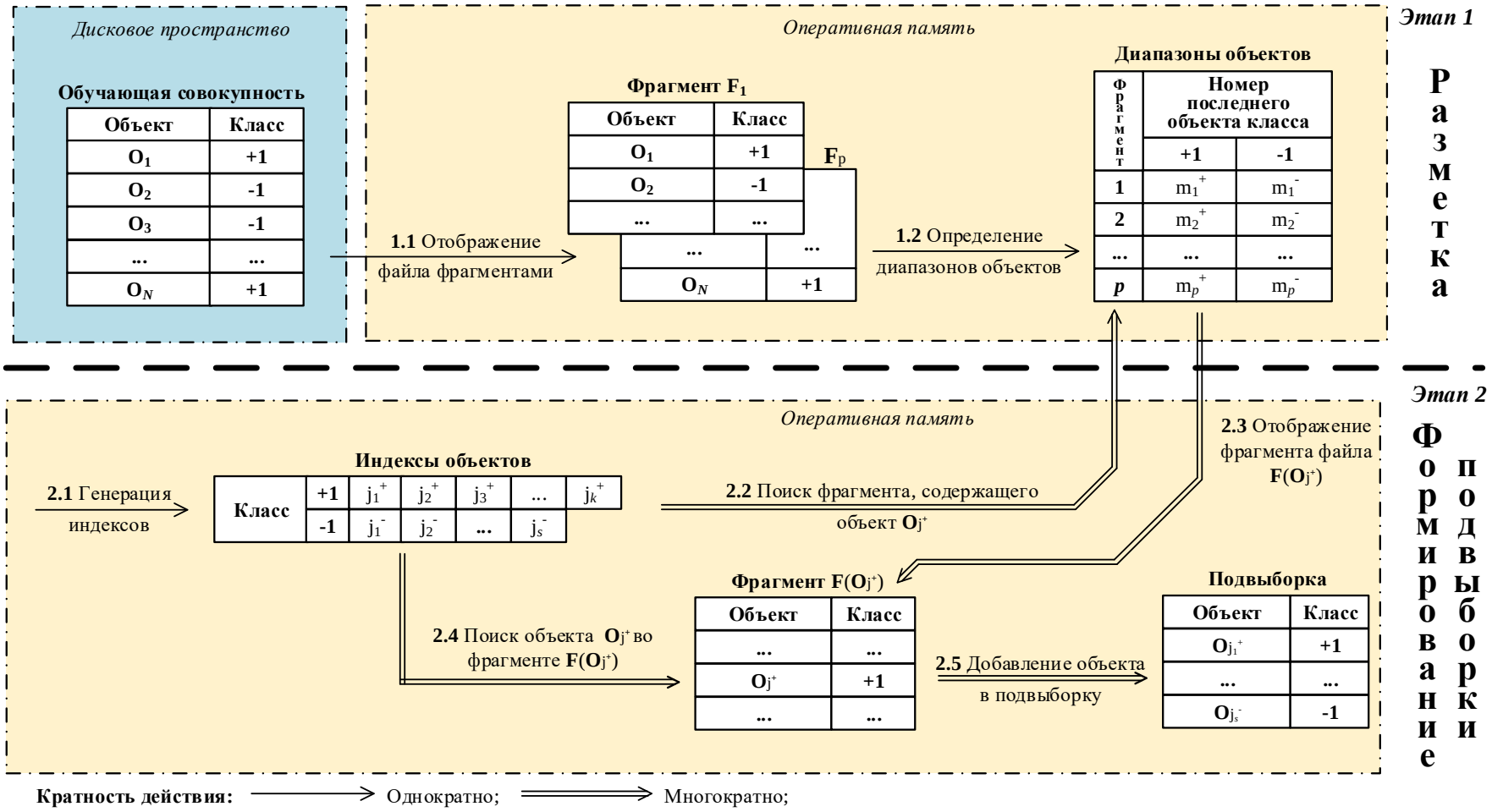
Главное достоинство:

- Позволяет хранить разреженные данные (с большим числом нулевых признаков) в сжатой форме.

Недостаток с точки зрения метода средних решающих правил:

- Невозможно по порядковому номеру объекта вычислить его местоположение в файле.

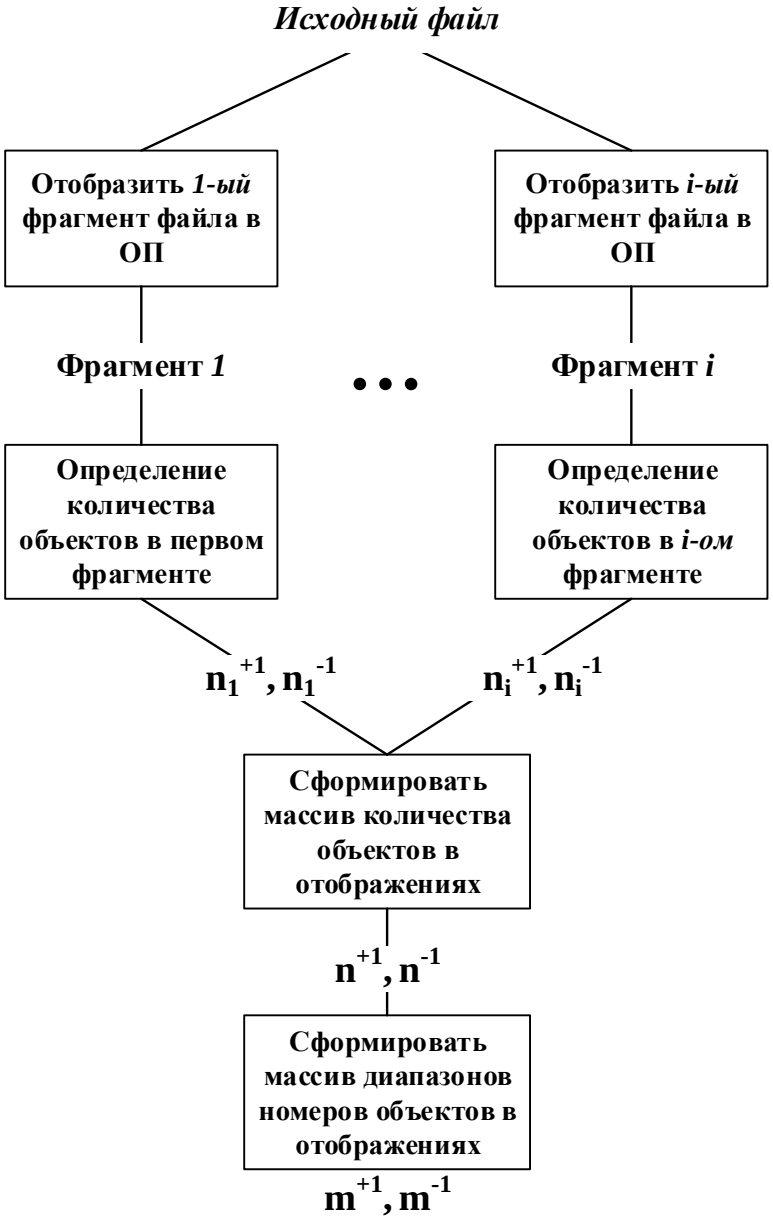
Стратегия работы с данными



Достоинства:

- Нет ограничения на размер обучающей совокупности;
- Малый объём оперативной памяти для хранения разметки;

Модель параллельного алгоритма разметки данных



Модель параллельных вычислений **для бесконечного числа процессов** в виде графа операции-операнды, где $n_i^{+1/-1}$ – кол-во объектов класса +1/-1 в *i*-ом фрагменте файла.

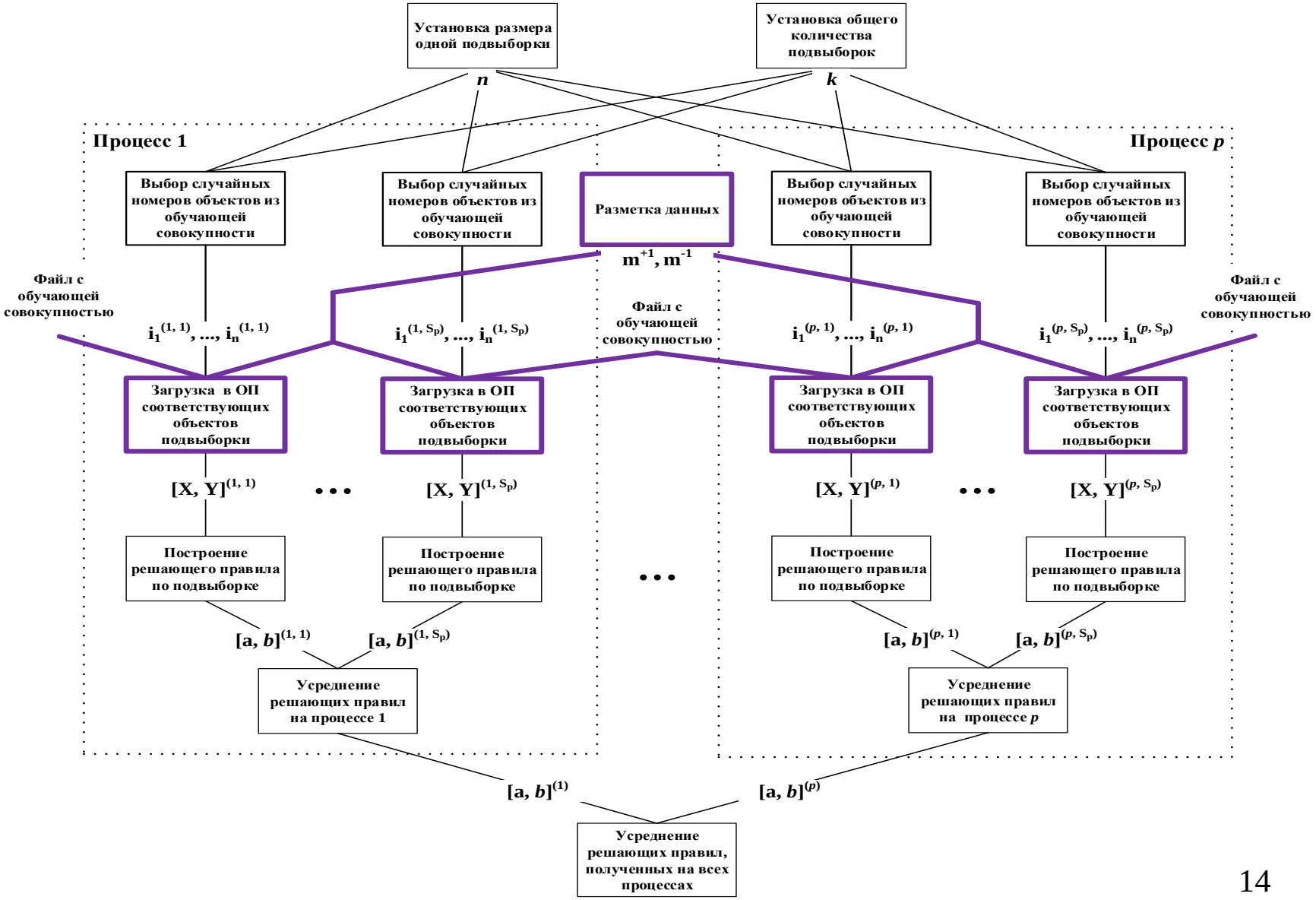
Достоинства:

- Высокая степень параллелизма по данным.

Недостаток:

- Требуется дополнительное время на поиск и чтение объектов при формировании случайных подвыборок в процессе обучения.

Вычислительная схема параллельного алгоритма MDR с использование предложенной стратегии обработки данных



Экспериментальное исследование. Описание данных и платформы тестирования

Характеристики набора данных:

Название набора данных	Объектов на обучении	Объектов на контроле	Число признаков	Доля ненулевых признаков
kddcup122	4 898 430	145253	122	10,23%

Работа выполнена с использованием оборудования Центра коллективного пользования сверхвысокопроизводительными вычислительными ресурсами МГУ имени М.В. Ломоносова.

Конфигурация одного узла суперкомпьютера НИИ ВЦ МГУ «Ломоносов»:

процессор – Intel Xeon X5570 (2.9 ГГц), 8 ядер;

оперативная память – 8 ГБ.

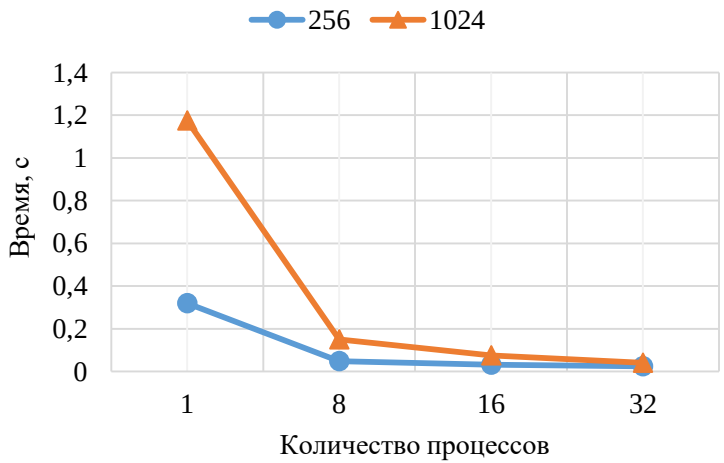
Количество узлов суперкомпьютера НИИ ВЦ МГУ «Ломоносов»: 5104.

Экспериментальное исследование. Результаты тестирования стратегий

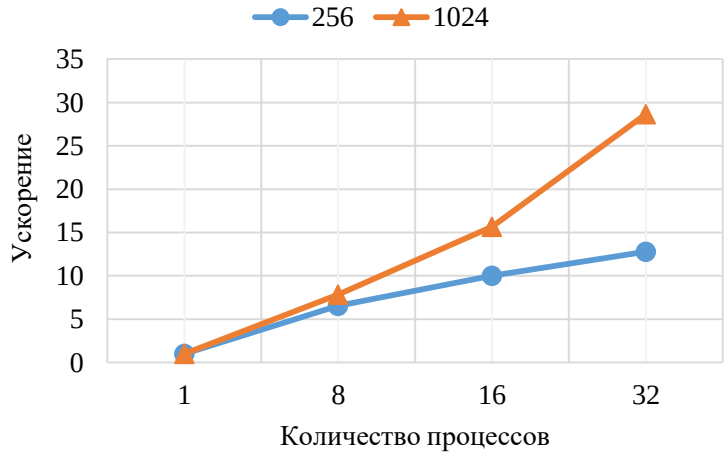
Среднее время работы отдельных этапов параллельной реализации метода средних решающих правил с **традиционной стратегией работы с данными**.

Этапы работы с данными	Количество объектов в одной подвыборке							
	256				1024			
	Количество процессов							
	1	8	16	32	1	8	16	32
1. Чтение данных	12,559	12,791	12,831	13,345	12,580	12,778	12,810	13,591
2. Формирование подвыборок	0,166	0,022	0,011	0,006	0,599	0,077	0,039	0,022
3. Обучение по подвыборкам	0,154	0,027	0,021	0,019	0,576	0,073	0,036	0,019
Общее время работы	12,878	12,840	12,863	13,370	13,761	12,936	12,896	13,645

a) Время обучения



b) Ускорение обучения

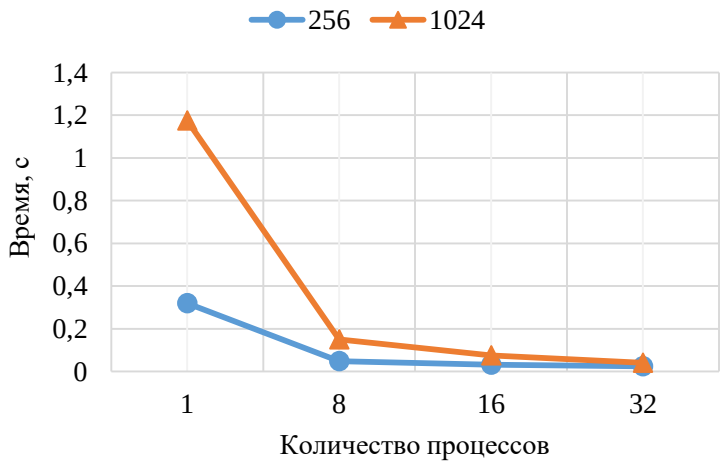


Экспериментальное исследование. Результаты тестирования стратегий

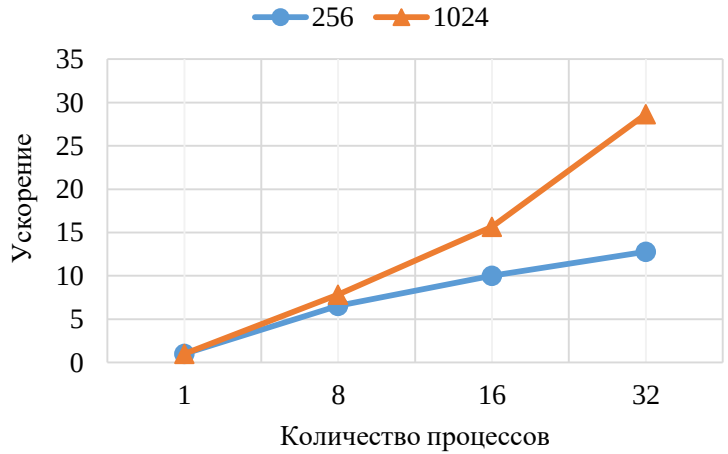
Среднее время работы отдельных этапов параллельной реализации метода средних решающих правил с **традиционной стратегией работы с данными**.

Этапы работы с данными	Количество объектов в одной подвыборке							
	256				**1024**			
	Количество процессов							
	1	8	16	32	**1**	8	16	**32**
1. Чтение данных	*12,559*	12,791	*12,831*	13,345	***12,580***	12,778	12,810	***13,591***
2. Формирование подвыборок	0,166	0,022	0,011	0,006	**0,599**	0,077	0,039	**0,022**
3. Обучение по подвыборкам	0,154	0,027	0,021	0,019	**0,576**	0,073	0,036	**0,019**
Общее время работы	*12,878*	12,840	*12,863*	13,370	***13,761***	12,936	12,896	***13,645***

a) Время обучения



b) Ускорение обучения

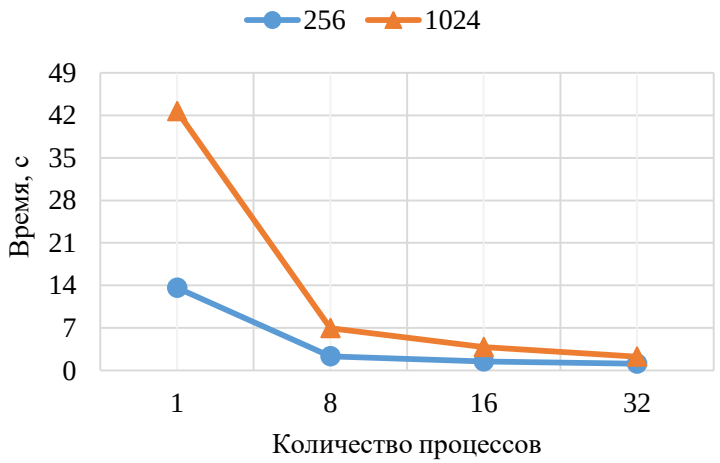


Экспериментальное исследование. Результаты тестирования стратегий

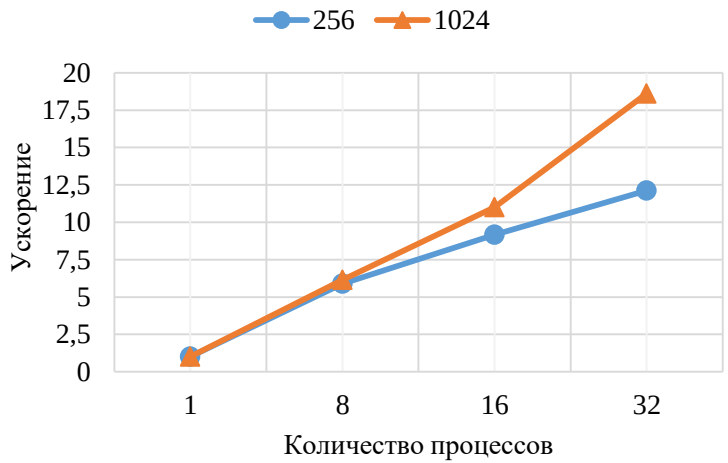
Среднее время работы отдельных этапов параллельной реализации метода средних решающих правил с **предложенной стратегией работы с данными.**

Этапы работы с данными	Количество объектов в одной подвыборке							
	256				**1024**			
	Количество процессов							
	1	8	16	32	1	8	16	32
1. Подготовка данных	3,757	0,770	0,418	0,211	***3,663***	0,766	0,381	***0,206***
2. Формирование подвыборок	9,733	1,523	1,056	0,902	**38,520**	6,109	3,457	**2,064**
3. Обучение по подвыборкам	0,134	0,019	0,014	0,012	**0,502**	0,071	0,041	**0,024**
Общее время работы	13,625	2,312	1,488	1,125	***42,680***	6,946	3,879	***2,294***

a) Общее время работы



b) Ускорение общего времени работы



Экспериментальное исследование. Результаты тестирования параллельной реализации метода средних решающих правил

Средние значения времени полного обучения и точности для различных параметров параллельной реализации метода средних решающих правил.

Кол-во фрагментов*	Кол-во подвыборок		Количество процессов				
			1	8	16	32	64
2560	1024	acc.	91,70±0,04	91,68±0,03	91,69±0,04	91,68±0,04	91,70±0,03
		time	127,01	31,277	15,465	8,469	4,018
25600	256	acc.	91,69±0,10	91,68±0,10	91,66±0,13	91,69±0,17	91,68±0,09
		time	13,62	2,312	1,488	1,125	1,235
	512	acc.	91,70±0,08	91,69±0,09	91,67±0,08	91,67±0,07	91,68±0,08
		time	27,29	4,421	2,270	1,528	1,458
	1024	acc.	91,69±0,05	91,71±0,04	91,70±0,04	91,69±0,03	91,71±0,04
		time	42,68	6,946	3,879	2,294	1,975
	2048	acc.	91,68±0,04	91,69±0,04	91,69±0,03	91,69±0,03	91,69±0,03
		time	97,54	15,878	8,808	3,841	3,087

(*****) параметр параллельной стратегии

Экспериментальное исследование. Результаты тестирования параллельной реализации метода средних решающих правил

Средние значения времени полного обучения и точности для различных параметров параллельной реализации метода средних решающих правил.

Кол-во фрагментов*	Кол-во подвыборок		Количество процессов				
			1	8	16	32	**64**
2560	**1024**	acc.	***91,70±0,04***	91,68±0,03	91,69±0,04	91,68±0,04	***91,70±0,03***
		time	**127,01**	31,277	15,465	8,469	**4,018**
25600	256	acc.	91,69±0,10	91,68±0,10	91,66±0,13	91,69±0,17	91,68±0,09
		time	13,62	2,312	1,488	1,125	1,235
	512	acc.	91,70±0,08	91,69±0,09	91,67±0,08	91,67±0,07	91,68±0,08
		time	27,29	4,421	2,270	1,528	1,458
	1024	acc.	***91,69±0,05***	91,71±0,04	91,70±0,04	91,69±0,03	***91,71±0,04***
		time	**42,68**	6,946	3,879	2,294	**1,975**
	2048	acc.	91,68±0,04	91,69±0,04	91,69±0,03	91,69±0,03	91,69±0,03
		time	97,54	15,878	8,808	3,841	3,087

(*****) параметр параллельной стратегии

Экспериментальное исследование. Результаты тестирования параллельной реализации метода средних решающих правил

Средние значения времени полного обучения и точности для различных параметров параллельной реализации метода средних решающих правил.

Кол-во фрагментов*	Кол-во подвыборок		Количество процессов				
			1	8	**16**	32	**64**
2560	1024	acc.	91,70±0,04	91,68±0,03	91,69±0,04	91,68±0,04	91,70±0,03
		time	127,01	31,277	15,465	8,469	4,018
25600	**256**	acc.	91,69±**0,10**	91,68±0,10	91,66±**0,13**	91,69±0,17	91,68±**0,09**
		time	***13,62***	2,312	***1,488***	1,125	***1,235***
	512	acc.	91,70±**0,08**	91,69±0,09	91,67±**0,08**	91,67±0,07	91,68±**0,08**
		time	***27,29***	4,421	***2,270***	1,528	***1,458***
	1024	acc.	91,69±**0,05**	91,71±0,04	91,70±**0,04**	91,69±0,03	91,71±**0,04**
		time	***42,68***	6,946	***3,879***	2,294	***1,975***
	2048	acc.	91,68±**0,04**	91,69±0,04	91,69±**0,03**	91,69±0,03	91,69±**0,03**
		time	***97,54***	15,878	***8,808***	3,841	***3,087***

(*****) параметр параллельной стратегии

Экспериментальное исследование. Результаты тестирования параллельной реализации метода средних решающих правил

Средние значения времени полного обучения и точности для различных параметров параллельной реализации метода средних решающих правил.

Кол-во фрагментов*	Кол-во подвыборок		Количество процессов				
			128				
			Количество объектов в одной подвыборке				
			300	500	**700**	1000	**2000**
25600	**1024**	acc.	**91,68±0,06**	91,72±0,05	**91,72±0,03**	91,74±0,02	**91,77±0,01**
		time	***2,13***	2,829	***3,098***	3,239	***4,676***

(*) параметр параллельной стратегии

Благодарю за внимание!