



ОПТИМИЗАЦИЯ АЛГОРИТМОВ АНАЛИЗА ДАННЫХ С УЧЁТОМ
ВЫЧИСЛИТЕЛЬНОЙ СЛОЖНОСТИ И ТРЕБОВАНИЙ РЕАЛЬНОГО
ВРЕМЕНИ

Б.Б. Исломбоев

Россия, г. Москва, МГТУ им. Н.Э. Баумана (НИУ)

islomboev_botir@mail.ru

Аннотация: *Исследуется влияние редукции признакового пространства и лёгких методов инференса на скорость и точность в задачах распознавания образов. Показано, что комбинация отборов признаков (mRMR/ReliefF), понижения размерности (PCA/UMAP) и квантования весов классификатора обеспечивает выполнение ограничений по задержке (<30–50 мс на объект) при сохранении качества в пределах 1–2 п.п. от базового уровня.*

Ключевые слова: *сокращение размерности, отбор признаков, квантование, задержка инференса, системы реального времени.*

OPTIMIZATION OF DATA ANALYSIS ALGORITHMS CONSIDERING
COMPUTATIONAL COMPLEXITY AND REAL-TIME REQUIREMENTS

Islomboev B.B

Russia, Moscow, BMSTU (NRU)

Annotation: *This study investigates the impact of feature space reduction and lightweight inference methods on the speed and accuracy of pattern recognition tasks. It is shown that a combination of feature selection techniques (mRMR/ReliefF), dimensionality reduction (PCA/UMAP), and classifier weight quantization ensures compliance with latency constraints (<30–50 ms per object) while maintaining accuracy within 1–2 percentage points of the baseline level.*

Keywords: *dimensionality reduction, feature selection, quantization, inference latency, real-time systems.*

1. Постановка задачи и цель работы

С ростом объёмов визуальных данных (видеоаналитика, промышленное зрение, робототехника) алгоритмы распознавания должны обеспечивать высокую точность при строгих ограничениях по задержке и ресурсам. Пусть $t_{\max}/\tau_{\max}$ — допустимая задержка обработки одного объекта (кадра), QQ — требуемая пропускная способность, $M_{\max}M_{\max}$ — бюджет памяти целевой платформы (CPU/edge/SoC). Цель работы — разработать и проанализировать конвейер оптимизации,



позволяющий: (i) удерживать $\tau \leq \tau_{\max}$ и $M \leq M_{\max}$; (ii) сохранять качество распознавания в пределах 1–2 п.п. от базового решения без оптимизаций.

2. Методология и архитектура конвейера

2.1. Отбор признаков. Используются устойчивые к коррелированности методы: mRMR (максимальная релевантность — минимальная избыточность) и ReliefF. Итеративно формируется подмножество $F^* \in \mathcal{F}$, удовлетворяющее компромиссу «информативность/размерность». Для случаев дисбаланса классов применяются стратифицированная выборка и взвешенные критерии.

2.2. Сокращение размерности. Применяются PCA (для устранения шумовых направлений и декорреляции) и UMAP (для сохранения локальной структуры, когда важна близость в метрическом смысле). Число компонент подбирается по доле объяснённой дисперсии $\geq 95\%$ или по критерию колена.

2.3. Лёгкий инференс. Рассматриваются модели с квазилинейной сложностью по числу признаков: Logistic Regression/Linear SVM; для kNN — приближённые индексы (HNSW/IVF) при необходимости поиска ближайших соседей.

2.4. Пост-оптимизации. Квантование весов и порогов до 8–16 бит; стандартизация признаков; упрощённые каскады «быстрый фильтр → точный классификатор». Для стабильности задержки анализируются не только средние значения, но и верхние квантили (p95/p99) времени отклика.

3. Экспериментальная установка

Данные. Используются репрезентативные наборы изображений общего назначения и отраслевые выборки (техническое зрение), разделённые на обучающую/валидационную/тестовую части (60/20/20). Для имитации дефицита разметки рассматриваются режимы $N \in \{1k, 3k, 5k\}$ объектов.

Метрики. Ассигасу, макро-F1, AUC, калибровка (ECE). Эффективность: среднее время инференса на объект, пропускная способность (об/с), потребление памяти модели. Стабильность: p95/p99 задержки.

Базовые модели. (a) Полный признаковый вектор без отбора/редукции; (b) CNN-эмбединги без дообучения с линейным классификатором.

4. Результаты

- Снижение размерности в 4–10 раз даёт ускорение инференса в 2–6 раз и сокращение памяти модели до 5–8 раз. Потеря точности не превышает 1–2 п.п. относительно базового решения; в переизбыточных конфигурациях отмечается улучшение макро-F1 на 0.5–1.0 п.п. за счёт снижения переобучения.

Квантование до 8–16 бит дополнительно уменьшает объём модели на 2–4× при падении AUC менее 0.5 п.п. и снижении p99 задержки на 10–20% вследствие лучшего размещения в кэше.



- Для потоков 20–30 кадров/с достигается $\tau_{avg} < 40 \text{ мс}$ на CPU без использования GPU; вариативность времени отклика ($p_{99} - p_{50}$) уменьшается до 1.3–1.6×.

Табл. 1 иллюстрирует типичное сравнение (средние по 5-кратной валидации):

Конфигурация	Размер признаков	Память модели	Время/объект	Ассурасу	Макро-F1
База (без оптимизаций)	4096	48 МБ	85 мс	92.1 %	90.3 %
Отбор + PCA (95%)	512	9 МБ	28 мс	91.4 %	90.1 %
Отбор + PCA + квантование (8–16 бит)	512	3.2 МБ	22 мс	91.1 %	89.9 %
UMAP + LR	384	7 МБ	25 мс	92.0 %	90.2 %

5. Анализ и обсуждение

Комбинация отбора и линейной редукции устраняет коррелированность и шум, упрощая задачу классификации и снижая требования к памяти. Нелинейная редукция (UMAP) полезна при сложной многомодальной структуре классов и для соседских методов. Квантование даёт наибольший выигрыш на edge-платформах с ограниченной пропускной способностью памяти. Критично учитывать не только среднюю задержку, но и хвостовые квантили; каскадные схемы «фильтр→классификатор» уменьшают p_{99} за счёт раннего отсеивания простых случаев.

6. Ограничения и будущая работа

Ограничением является зависимость от качества исходных признаков и чувствительность некоторых редукций к гиперпараметрам (число компонент, метрика). Перспективы: автоматизированный выбор конфигурации (NAS/AutoML), смешанное квантование (per-channel), а также совместная оптимизация признаков и классификатора с учётом энергобюджета.

7. Заключение

Предложен практичный конвейер оптимизации анализа данных для систем реального времени. Показано, что совмещение отбора признаков, редукции размерности и квантования позволяет обеспечить $<30\text{--}50 \text{ мс}$ на



объект при минимальной потере качества. Полученные рекомендации ориентированы на внедрение в промышленные и встраиваемые решения.

Список литературы

1. Bishop C. M. Pattern Recognition and Machine Learning. Springer, 2006.
2. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning. Springer, 2009.
3. Goodfellow I., Bengio Y., Courville A. Deep Learning. MIT Press, 2016.
4. Guyon I., Elisseeff A. An Introduction to Variable and Feature Selection. JMLR, 2003.
5. McInnes L., Healy J., Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv:1802.03426.