Структуры и алгоритмы для повышения производительности классификации данных на основе байесовского подхода

А. С. Писарев, И. А. Писарев
Санкт-Петербургский государственный электротехнический университет
«ЛЭТИ» им. В.И. Ульянова (Ленина)
a_pisarev@mail.ru, pisarevivan@yandex.ru

***Аннотация*. Разработаны модификации структур и алгоритмов классификации данных на основе Байесовского подхода. Структура модели обеспечивает минимизацию операций алгоритма классификации. Исследован вариант реализации алгоритма при выполнении условия, когда в цикле тестирования на каждом шаге изменяется только часть признаков. Рассмотрены примеры классификации паттернов в изображениях и документов в корпусах тематических текстов. Алгоритмы реализованы для случаев хранения модели в оперативной и внешней памяти (базе данных). Приведены результаты тестирования версии программ с использованием графического процессора (Graphics Processing Unit – GPU). Применение разработанных алгоритмов позволит повысить производительность классификации данных.**

Ключевые слова: автоматизированная система научных исследований; байесовский подход; база данных; классификация данных; графический процессор

I. Введение

Важным этапом выполнения научных исследований является анализ полученных экспериментальных данных. Актуальной задачей является повышение производительности интеллектуального анализа данных в составе автоматизированной системы управления научными исследованиями (АСНИ).

Вопросам создания эффективных АСНИ в передовых областях науки с использованием современных информационных технологий посвящены труды В.В. Александрова, В.И. Городецкого, В.В. Иванищева, В.М. Пономарева, Г.С. Поспелова, Д.А. Поспелова, Б.Я. Советова, Р.М. Юсупова и др.

Целью исследований в области нового направления автоматизации интеллектуального анализа данных на основе методов машинного обучения (Automated Machine Learning – AutoML), является извлечение новых знаний из разнородных данных большого объема [1-6].

В СПбГЭТУ «ЛЭТИ» разработан сетевой программный комплекс ОнтоМАСТЕР-Ресурс [7], который позволяет проводить совместные научные исследования с использованием веб-интерфейсов и сценариев анализа разнородных экспериментальных данных (текстов, видео и изображений, сигналов).

Результаты автоматизированного выполнения сценариев сохраняются в базе данных, что позволяет проводить всесторонний анализ и поиск наилучших моделей на основе многокритериального подхода.

Отличительной особенностью разработанного сетевого программного комплекса является реализация свойств адаптации, масштабирования, функционального расширения существующей базы моделей и методов машинного обучения.

В работе представлены результаты разработки и исследования алгоритмов и структур для повышения производительности интеллектуального анализа данных на основе Байесовского подхода.

II. Алгоритмы и структуры данных

Предварительный анализ моделей и результатов интеллектуального анализа в среде ОнтоМАСТЕР-Ресурс [8] показал высокую точность тематической классификации текстов документов с применением методов, основанных на Байесовском подходе, деревьях решений, нейронных сетях и др.

Для повышения точности классификации применяется расширение числа атрибутов, заданных в виде однословных ключевых терминов, их колокациями: многословными терминами предметных областей. Замена последовательности терминов многословными терминами так же позволяет более полно выполнить условие взаимной независимости признаков-терминов [9–10].

На точность и производительность классификации влияют число используемых атрибутов и их частота в корпусе текстов. При подготовке данных используются частотные словари терминов, ранжированных в порядке убывания частоты. На рис. 1 представлен пример графика кумулятивной частоты терминов tfc в зависимости от ранга в частотном словаре для корпуса текстов по тематической области знаний «Методы обработки изображений и сигналов в гидроакустике». Приблизительно 25% наиболее часто встречающихся

Работа частично выполнена при финансовой поддержке РНФ, проект № 17-71-20077.

терминов покрывают 95% данного корпуса текстов. Для повышения производительности классификации может быть сокращено число редко используемых терминов. Для поиска минимального числа атрибутов применяется итерационный алгоритм решения задачи классификации при выполнении ограничения на допустимое снижение значения информационного критерия (например, точности, F_1, χ^2) по сравнению с полным набором атрибутов.

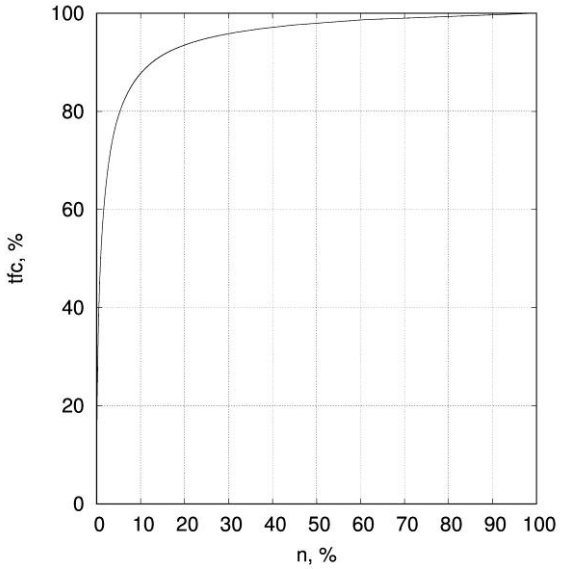

Рис. 1. Пример графика кумулятивной частоты терминов из корпуса текстов «Методы обработки изображений и сигналов в гидроакустике»

Мультиноминальный метод Байеса (Multinominal Naive Bayes – NB) является одним из видов машинного обучения с учителем (Supervised Machine Learning – SML) [11].

Мультиноминальный метод Байеса при больших размерах словарей обеспечивает уменьшение средней ошибки на 27% и превосходит по точности многовариантную модель Бернулли (multi-variate Bernoulli model) при любом размере словаря [12].

Классификация документов на классы осуществляется на основе модели языка корпуса текстов, содержащихся в документах.

В модели классификации используются априорные вероятности (частоты) встречаемости терминов в документах к которым применяется сглаживание Лапласа (Laplace smoothing). Использование информации о частотах встречаемости терминов позволяет повысить точность в задачах классификации тестов [13].

В качестве атрибута (тематического признака) классификации применяется частота встречаемости словоформы (термина) $\mathrm{tf}_{k,d}$ в тексте документа.

Оценки $P(c \mid d)$ – априорных вероятностей принадлежности документа d к классу c используются для определения наиболее вероятного класса с максимальной апостериорной вероятностью (Maximum A Posteriori – MAP) [11]:

$$c_{map} = \arg\max_{c \in C} P(c) \prod_{k=1}^{n} P(t_k \mid c)^{tf_{k,d}}$$

Для снижения влияния погрешностей вычислений применяется логарифмическая шкала для представления априорных вероятностей:

$$c_{map} = \arg\max_{c \in C} \left[\log P(c) + \sum_{k=1}^{n} \mathrm{tf}_{k,d} \log P(t_k \mid c) \right]$$

При использовании правила Байеса в логарифмической шкале используется матрица значений $\log P(t_k \mid c)$, к которой добавляется столбец со значениями из вектора вероятностей $\log P(c)$. К матрице $S_{\mathrm{l,n+1}}$ со значениями $\mathrm{tf}_{k,d}$ признаков тестовой выборки добавляется единичный столбец.

Тогда процедура классификации может быть представлена в виде умножения матриц:

$$C_{m,l} = A_{m,n+1} S_{n+1,l}$$

где $m = |C|$ – число классов; n – число признаков; l – размер тестовой выборки.

Модифицированная матричная структура позволяет реализовать алгоритмы параллельной классификации на основе Байесовского подхода.

III. Результаты

Разработанные алгоритмы реализованы с применением технологий параллельных вычислений. Тестирование разработанных алгоритмов производилось на нескольких задачах классификации текстовых документов и изображений.

Классификация с использованием нескольких вычислительных потоков. В многопроцессорных и многоядерных вычислительных системах производительность классификации может быть повышена за счет реального одновременного выполнения умножения матриц и векторов в отдельных потоках.

Классификация с использованием нескольких вычислительных потоков показала повышение производительности приблизительно в n раз, где n - число ядер в многопроцессорной системе.

Классификация с использованием графического процессора (graphics processing unit – GPU) с интерфейсом CUDA. Повышение производительности классификации на один или два порядка зависит от типов используемых графических процессоров и наиболее существенно при возрастании объемов обрабатываемых данных. Для дальнейшего повышения производительности необходима минимизация операций по обмену данными с GPU, например, предварительная загрузка модели и обновление только матрицы (вектора) признаков тестовой выборки.

Классификация больших данных (Big Data) с использованием баз данных (БД). Для интеллектуального анализа больших данных, превышающих по объему размеры требуемой оперативной памяти компьютера, хорошие результаты демонстрируются при использовании технологии MapReduce и БД, в которых алгоритмы машинного обучения реализуются с помощью определенных пользователями функций (user defined functions – UDF) [14–15]. Некоторые исследования демонстрируют более высокую производительность БД по сравнению с MapReduce при решении задач классификации [16].

Разработаны структуры и реализованы алгоритмы классификации средствами языка SQL запросов с применением присоединенных процедур и триггеров БД. Разработана модификация алгоритма Байесовской классификации для случаев, когда обновление признаков классификации осуществляется асинхронно из нескольких источников (сенсоров) данных. Алгоритм предусматривает пересчет вероятностей классов при возникновении изменений индивидуальных признаков.

Результаты разработки и исследования алгоритмов и структур классификации могут быть использованы как в виде автономных программных модулей, так и в составе автоматизированной подсистемы интеллектуального анализа данных АСНИ, которая состоит из веб-интерфейсов исследователей, программных агентов, гибридной базы данных, базы знаний, онтологии методов и моделей.

IV. Заключение

Разработаны алгоритмы и структуры для повышения производительности классификации данных на основе Байесовского подхода.

Исследованы реализации алгоритмов при использовании нескольких потоков в многопроцессорных и многоядерных вычислительных системах, графических процессоров с интерфейсом CUDA, баз данных.

Точность тематической классификации текстовых документов может быть повышена при использовании алгоритма анализа ключевых терминов предметных областей [9]. При этом происходит сокращение числа используемых атрибутов классификации, что способствует повышению производительности.

Дальнейшее повышение точности обеспечивается расширением числа атрибутов классификации, заданных в виде однословных ключевых терминов, их колокациями: многословными терминами предметных областей. Замена последовательности терминов многословными терминами позволяет более полно выполнить условие взаимной независимости признаков-терминов.

Разработанные алгоритмы применены в адаптивной сетевой среде ОнтоМАСТЕР-Ресурс для автоматизированного анализа данных при решении задач классификации. Данный подход позволяет повысить производителmность при решении задач классификации в многопользовательском режиме и применять различные алгоритмы с учетом характеристик вычислительных узлов и обрабатываемых объемов и типов данных для задач обработки разнородных ресурсов (например, при классификации гидроакустических данных).

Результаты работы могут быть использованы в научно-исследовательской, образовательной и инженерной деятельности для повышения эффективности научно-исследовательских, опытно-конструкторских работ и учебного процесса.

Список литературы

[1] Cvetković L., Milašinović B., Fertalj K. A tool for simplifying automatic categorization of scientific paper using Watson API //Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2017 40th International Convention on. IEEE, 2017. pp. 1501-1505.

[2] Witten I. H., Frank E., Hall M. A., Pal C. J. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2016. 629 p.

[3] Hanif M. H. M., Adewole K. S., Anuar N. B., Kamsin A. Performance Evaluation of Machine Learning Algorithms for Spam Profile Detection on Twitter Using WEKA and RapidMiner //Advanced Science Letters. – 2018. V. 24. No 2. Pp. 1043-1046.

[4] Fillbrunn A., Dietz C., Pfeuffer J., Rahn R., Landrum G. A., Berthold M. R. KNIME for reproducible cross-domain analysis of life science data //Journal of Biotechnology. 2017. V. 261. Pp. 149-156

[5] Triguero I., González S., Moyano J. M., García S., Alcalá-Fdez J., Luengo J., Herrera F. KEEL 3.0: an open source software for multi-stage analysis in data mining //International Journal of Computational Intelligence Systems. 2017. V. 10. No 1. Pp. 1238-1249.

[6] Abadi M., Barham P., Chen J., Chen Z., Davis A., Dean J., Kudlur, M. TensorFlow: A System for Large-Scale Machine Learning //OSDI. 2016. V. 16. Pp. 265-283.

[7] Котова Е.Е., Писарев А.С., Писарев И.А. Программный комплекс анализа информационных ресурсов ОнтоМАСТЕР-Ресурс. Свидетельство о государственно регистрации программы для ЭВМ № 2018611107 от 24 января 2018 г.

[8] Pisarev I.A., Kotova E.E., Pisarev A.S., Stash N.V. Structure of knowledge base of methods for processing hydroacoustic signals //Young Researchers in Electrical and Electronic Engineering (EIConRus), 2018 IEEE Conference of Russian. IEEE, 2018. Pp. 1132-1135.

[9] Котова Е.Е., Писарев И.А. Построение тематических онтологий с применением метода автоматизированной разработки тезаурусов. // Известия СПбГЭТУ «ЛЭТИ». 2016. № 3. С. 37-47.

[10] Писарев И.А., Котова Е.Е., Писарев А.С. Анализ ключевых понятий областей знаний для поддержки научных исследований. Качество. Инновации. Образование. 2017. № 7 (146). С. 38-50..

[11] Schütze H., Manning C. D., Raghavan P. Introduction to information retrieval. Cambridge University Press, 2008. 544 P.

[12] McCallum A., Nigam K. A comparison of event models for naive bayes text classification //AAAI-98 workshop on learning for text categorization. 1998. V. 752. No. 1. Pp. 41-48.

[13] Rosario R. R. A data augmentation approach to short text classification. Dissertation. – University of California, Los Angeles, 2017. 210 P.

[14] Ordonez C., Pitchaimalai S. K. Bayesian classifiers programmed in SQL //IEEE Transactions on Knowledge and Data Engineering. 2010. V. 22. No. 1. Pp. 139-144.

[15] Ordonez C., Pitchaimalai S. K. One-pass data mining algorithms in a DBMS with UDFs //Proceedings of the 2011 ACM SIGMOD International Conference on Management of data. ACM, 2011. Pp. 1217-1220.

[16] Pitchaimalai S. K., Ordonez C., Garcia-Alvarado C. Comparing SQL and MapReduce to compute Naive Bayes in a single table scan //Proceedings of the second international workshop on Cloud data management. ACM, 2010. Pp. 9-16.