Быстрая одноклассовая SVM классификация для большой обучающей совокупности

М.Ю. Курбаков
Тульский государственный университет,
Лаборатория когнитивных технологий и симуляционных систем,
Тула, Россия
muwsik@mail.ru

В. В. Сулимова
Тульский государственный университет,
Лаборатория когнитивных технологий и симуляционных систем,
Тула, Россия
vsulimova@yandex.ru

***Аннотация*—В основу данной работы положен популярный метод одноклассовой классификации OCSVM. Мы предлагаем усовершенствованный вариант данного метода, целью создания которого является обеспечение возможности работы с большими обучающими совокупностями, что является проблематичным для OCSVM из-за высокой трудоемкости обучения. Основная идея предлагаемого подхода заключается в применении OCSVM к независимым случайным подвыборкам из исходной обучающей совокупности с последующим объединением результатов в единое решение, совпадающее по виду с решающим правилом OCSVM. Экспериментальное исследование показало, что предложенный подход позволяет существенно ускорить решение задачи, получая при этом точное (или близкое к точному) решение.**

Ключевые слова — одноклассовая классификация, OCSVM, большие задачи, повышение производительности.

1. Введение

Одноклассовая классификация является частным случаем многоклассовой классификации, когда обучающая совокупность состоит только из объектов одного класса (как правило, целевого). При этом требуется на основе анализа этой обучающей совокупности построить решающее правило, позволяющее определить, относится ли новый объект к присутствующему на обучении классу или нет [1].

Одним из наиболее популярных методов решения задачи одноклассовой классификации является метод OCSVM [2], который активно используется для решения различных прикладных задач, в частности, для классификации текстов [3], обнаружения мошеннических транзакций по кредитным картам [4], обнаружения вторжений [5], в системах видеонаблюдений [6], медицинских системах [7] и т.д.

В случае небольшого размера обучающей совокупности задача построения оптимального решающего правила OCSVM может быть достаточно быстро решена при помощи классических методов оптимизации. Однако в случае большого количества объектов, что характерно для многих современных задач анализа данных, построение решающего правила оказывается очень трудоемким по времени и памяти, что существенно затрудняет, а в ряде случаев и делает невозможным его непосредственное применение на практике.

Несмотря на массовые исследования в данной области, до сих пор не удалось найти решение, которое бы полностью удовлетворяло практические нужды. В частности, работы, направленные на повышение скорости обучения, обычно не решают проблему нехватки оперативной памяти и обеспечения быстрого доступа к объектам [8], методы, ориентированные на более экономичное использование памяти, часто оказываются существенно менее точными или имеют низкую скорость сходимости [9]. Применение технологий параллельных и распределенных вычислений [10] смягчает проблему большого количества данных, но не решает ее, поскольку большинство методов имеют итерационную природу с зависимостями по данным, что существенно ограничивает возможности повышения производительности данным путем.

2. Предлагаемый подход

В данной работе мы предлагаем добиться повышения производительности решения задачи одноклассовой классификации по методу OCSVM путем замены одной большой исходной задачи на серию существенно более мелких задач, исходными данными для каждой из которых является случайная выборка объектов из полной обучающей совокупности, сформированная независимо от остальных. Результаты обучения по частным подвыборкам предлагается объединять в единое решающее правило согласно принципу усреднения, предложенному нами ранее для двухклассовых задач SVM [11], в результате чего итоговое решение имеет ту же структуру, что и традиционное решение задачи по методу OCSVM. В результате такой подход имеет существенное преимущество перед бэггингом, который, предполагает построение ансамбля классификаторов для частных подвыборок, а итоговое решение получает путем голосования или какой-либо другой схемы построения ансамблей, что требует хранения всех частных классификаторов и отдельного применения каждого из них на этапе распознавания.

Следует отметить, что предложенный подход снимает теоретическое ограничение на размер обучающей совокупности, поскольку, как и бэггинг [12], фактически, не требует единовременной загрузки всех объектов в память. Однако, даже при использовании выборочных методов, на практике в большинстве случаев по-прежнему в память загружается полная обучающая совокупность (например, в рамках пакета scikit-learn python). Это обусловлено тем фактом, что традиционный формат хранения данных libsvm не позволяет вычислить позицию начала объекта с заданным номером, что сильно снижает эффективность произвольного доступа к объектам в файле при его традиционном чтении.

Для снятия практического ограничения на объем обучающей совокупности при решении задачи одноклассовой классификации мы предлагаем использовать принцип оптимальной работы с данными [13], разработанный нами ранее для двухклассового распознавания, но который может быть почти без изменений применен и для одноклассовой ситуации. Его основная идея заключается в осуществлении предварительной разметки данных с сохранением областей файла, содержащих диапазоны объектов обучающей совокупности, и последующем использовании этой разметки для быстрого формирования подвыборок за счет предварительной примерной локализации объектов, а также за счет привлечения предоставляемого операционной системой механизма отображения файлов в память, который позволяет работать с данными на диске как с обычными данными в памяти.

3. Результаты экспериментов

В таблице 1 приведено время обучения и AUC (на тестовой выборке) для libsvm и для предложенного метода (для различного размера случайных подвыборок SRS и числа случайных подвыборок *NRS* для двух модельных наборов данных с разным числом объектов и выбросов (аномальных объектов). Модельные данные генерировались по следующей схеме. Положительный класс (нормальные объекты) генерировались в соответствии с нормальным распределением, выбросы – равномерно вокруг положительного класса. Для параметров libsvm (как отдельно, так и для обучения по подвыборкам в составе предложенного подхода) были установлены значения: $nu = 0.001$, $\gamma = 0.01$.

Таблица I. Результаты работы методов

Метод	**Параметры**		**Набор данных (объектов/выбросов)**			
			100000 / 100		**500000 / 500**	
	NRS	***SRS***	***время***	***AUC***	***время***	***AUC***
Libsvm	-	-	0,2314	0,9750	6,2159	0,9920
Предло-женный подход	10	300	0,0063	0,9966	0,0199	0,9975
	10	1000	0,0047	0,9968	0,0103	0,9996
	100	300	0,0196	0,9973	0,0296	0,9975
	100	1000	0,0362	0,9999	0,0468	0,9999

Как видно из таблицы I, предложенный подход при различных значениях параметров позволяет получить решение, превосходящее по качеству решение, полученное при помощи популярной библиотеки libsvm. При этом время, затраченное на построение решающего правила, оказывается на 1-2 порядка.

Существенное повышение производительности (даже в условиях последовательной реализации предложенного метода) обусловлено нелинейной вычислительной сложностью решения задачи SVM относительно числа объектов, в результате чего оказывается вычислительно более выгодно решать серию небольших задач, чем одну более крупную. Повышение качества обусловлено снижением чувствительности к наличию выбросов в обучающей совокупности за счет усреднения частных решающих правил, построенных по небольшим подвыборкам.

Заключение

В работе предложен подход к решению больших одноклассовых задач SVM, основанный на обучении на подвыборках и последующем усреднении результатов с получением решения, совпадающего по структуре с традиционным решением задачи SVM. Эксперименты на модельных данных показали существенное преимущество предложенного подхода по сравнению с наиболее популярной библиотекой для решения задач SVM - libsvm.

Благодарности

Работа выполнена при финансовой поддержке Министерства науки и высшего образования РФ в рамках государственного задания FEWG-2021-0012.

Литература

[1] Perera, P. One-Class Classification: A Survey / P. Perera, P. Oza, V. Patel // Computer Vision and Pattern Recognition. – 2021. – P. 19. doi:10.48550/arXiv.2101.03064.

[2] Scholkopf, B. Estimating the support of a high-dimensional distribution / B. Scholkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola. R. C. Williamson // Neural computation. – 2001. – Vol. 13(7). – P. 1443-1471. doi: 10.1162/089976601750264965.

[3] Shravan, K. One-Class Text Document Classification with OCSVM and LSI / K. Shravan, V. Ravi // Artificial Intelligence and Evolutionary Computations in Engineering Systems. – 2017. – Vol. 517. doi:10.1007/978-981-10-3174-8_50.

[4] Wu, T. Locally Interpretable One-Class Anomaly Detection for Credit Card Fraud Detection / T. Wu, Y. Wang // 2021 International Conference on Technologies and Applications of Artificial Intelligence. – 2021. – P. 25-30. doi:10.48550/arXiv.2108.02501.

[5] Krishnaveni, S. Anomaly-Based Intrusion Detection System Using Support Vector Machine / S. Krishnaveni, P. Vigneshwar, S. Kishore, B. Jothi, S. Sivamohan // Artificial Intelligence and Evolutionary Computations in Engineering Sys. – 2020. – Vol. 1056. doi:10.1007/978-981-15-0199-9_62.

[6] Seredin, O. S. A Skeleton Features-Based Fall Detection Using Microsoft Kinect v2 with One Class-Classifier Outlier Removal / O. S. Seredin, A. V. Kopylov, S. C. Huang, D. S. Rodionov // ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. – 2019. – Vol. 4212. – P. 189-195.

[7] Xiao-Kang, W. KDE-OCSVM model using Kullback-Leibler divergence to detect anomalies in medical claims / W. Xiao-Kang, H. Wen-Hui, Z. Hong-Yu, W. Jian-Qiang, G. Mark, T. Zhang-Peng, S. Kai-Wen // Expert Systems with Applications. – 2022. – Vol. 200. doi:10.1016/j.eswa.2022.117056.

[8] Zeyi, W. ThunderSVM: A Fast SVM Library on GPUs and CPUs / W. Zeyi, S. Jiashuai, L. Qinbin, H. Bingsheng, C. Jian // Journal of Machine Learning Research. – 2018. – Vol.19(21). – P. 1-5.

[9] Lee, Y.-J. RSVM: Reduced Support Vector Machines / Y.-J. Lee, O. L. Mangasarian // In Proceedings of the SIAM International Conference on Data Mining. –2021. – Vol. 1. – P. 325-361.

[10] Stolpe, M. Distributed Support Vector Machines: An Overview / M. Stolpe, K. Bhaduri, K. Das // Solving Large Scale Learning Tasks. Challenges and Algorithms. – 2016. – Vol. 9580. – P. 109-138. doi:10.1007/978-3-319-41706-6_5.

[11] Makarova, A. Mean Decision Rule Method for Consrtucting Nonlinear Boundaries in Large Binary SVM Problems / A. Makarova, M. Kurbakov, V. Sulimova // Inf. Technology and Nanotechnology. – 2020. – P. 1-6. doi:10.1109/ITNT49337.2020.9253181.

[12] Shieh, A. D. Ensembles of One Class Support Vector Machines / A. D. Shieh, D. F. Kamm // Multiple Classifier Systems. – 2009. – Vol. 5519. – P. 181-190. doi:10.1007/978-3-642-02326-2_19.

[13] Курбаков, М.Ю. Оптимизация загрузки данных в формате libsvm при решении двухклассовой задачи SVM методом усреднения решающих правил в условиях большой обучающей совокупности / М.Ю. Курбаков, А.И. Макарова, В.В. Сулимова // Информационные технологии и нанотехнологии. – 2019. – Т. 4. – С. 53-60.