Быстрое решение больших задач SVM-регрессии

А.И. Макарова[1], А.В. Копылов[1], С.Д. Двоенко[1], В.В. Сулимова[1]

[1]*Тульский государственный университет, проспект Ленина 92, Тула, Россия, 300012*

Аннотация

В работе предложен подход, позволяющий быстро найти приближенное решение задачи SVM-регрессии в случае большого числа объектов. Проведенное экспериментальное исследование подтверждает эффективность предложенного подхода и показывает его преимущества при восстановлении регрессионных зависимостей по сравнению с другими методами, находящимися в открытом доступе.

Ключевые слова

SVM-регрессия, большие задачи, сэмплирование, потенциальные функции

1. Введение

Задача SVM-регрессии [1] является одной из распространенных задач анализа данных. В случае небольшого числа объектов ее решение может быть легко найдено классическими методами оптимизации. Если же объектов много решению данной задачи препятствуют две существенные проблемы: высокая вычислительная сложность построения регрессионной модели и низкая производительность при работе с данными.

Однако существующие методы, направленные на повышение скорости решения задачи SVM-регрессии [2-4], как правило, ориентированы на решение только одной из указанных проблем, а часто имеют и другие недостатки, наиболее существенными из которых являются отсутствие возможности восстановления нелинейных зависимостей, а также итерационный характер вычислений с зависимостями по данным, существенно снижающими эффективность применения технологий параллельного и распределенного программирования.

В данной работе предлагается простой подход, лишенный всех указанных выше недостатков.

2. Метод средних функций (MF) для больших задач SVM-регрессии

Основная идея заключается в формировании множества небольших выборок из обучающей совокупности, независимом восстановлении регрессионной зависимости по каждой из них и последующем объединении полученных частных функций в одну путем усреднения.

Предлагаемый подход эксплуатирует идею бэггинга [5], но отличается от него способом формирования выборок (сэмплирования) и наличием возможности обучения в пространстве, порожденном потенциальной функцией, позволяя восстанавливать нелинейные зависимости.

В случае признакового пространства усреднение производится непосредственно в терминах параметров восстанавливаемой функции. А для случая потенциальных функцией, ввиду невозможности явного вычисления параметров, предложен специальный способ усреднения, выраженный в терминах множителей Лагранжа, являющихся решением частных задач восстановления регрессионных зависимостей.

В данной работе предложен новый принцип формирования выборок (названных умными), заключающийся в выборе объектов, наиболее влияющих на результат решения исходной задачи. А именно, предлагается формировать выборку из объектов, имеющих ненулевые множители Лагранжа, полученным в результате обучения по небольшим случайным выборкам из исходной обучающей совокупности. Предлагаемый подход позволяет существенно повысить скорость сходимости метода, являясь в случае большого числа объектов менее трудозатратным по сравнению с альтернативными подходами [6, 7], поскольку обучение по небольшим выборкам может быть осуществлено достаточно быстро.

В таблице 1 приведены время работы и квадрат коэффициента корреляции ($r2$) истинного и оцененного вектора значений функции для предложенного метода с умными (SmartMF) и случайными выборками (MF), библиотеки LibSVM (являющейся эталоном качества), а также python-реализациями бэггинга с ансамблем из 1 и 2 моделей и стохастического градиента (SGD).

Таблица 1

Результаты работы методов для квадратичной зависимости с нормальным шумом

Метод	Набор данных (число объектов / число признаков)								
	1 (10 000 / 10)			2 (100 000 / 10)			3 (100 000 / 100)		
	время (сек)		$r2$	время (сек)		$r2$	время (сек)		$r2$
	обуч.	контр.		обуч.	контр.		обуч.	контр.	
LibSVM	7.538	0.72	0.992	5178	29.53	0.996	>5000	-	-
Бэггинг-1	5.028	0.58	0.990	4662	22.9	0.996	>5000	-	-
Бэггинг-2	11.07	1.21	0.991	>5000	-	-	>5000	-	-
SGD	0.32	0.0005	0	0.39	0.004	0	0.52	0.024	0.002
MF	1.46	0.97	0.971	1.74	15.16	0.971	1.45	25.61	0.22
SmartMF	4.91	0.38	0.990	10.65	4.61	0.993	15.14	24.54	0.901

3. Заключение

Как видно из таблицы 1, методы LibSVM и бэггинг позволяют достичь наиболее высокого качества (идеальное значение $r2 = 1$). Однако для данных методов характерно большое время обучения, которое для второго и третьего наборов данных составляет более часа. Метод стохастического градиента (SGD) очень быстро находит решение для всех наборов данных, но он позволяет восстанавливать только линейные зависимости, поэтому качество результата для него является очень низким ($r2 = 0$ - минимально возможное значение критерия качества).

Предложенные методы MF и SmartMF позволяют за небольшое время найти приближенное, но достаточно близкое к точному решение задачи SVM-регрессии. При этом преимущество по времени перед другими методами становится все более ощутимым с ростом числа объектов. Не трудно заметить, что SmartMF имеет очевидное преимущество по качеству перед MF, которое обусловлено применением предложенного способа формирования выборок. Особенно ярко этот эффект выражен для 3-го набора данных, имеющего большое число признаков.

4. Благодарности

Работа выполнена при поддержке грантов РФФИ № 20-07-00055, № 20-07-00441.

5. Литература

[1] Smola, A.J. A tutorial on support vector regression, statist / A.J. Smola, B. Schölkopf // Comput. – 2004. – Vol. 14. – P. 199-222.

[2] Rivas-Perea, P. An Algorithm for Training a Large Scale Support Vector Machine for Regression Based on Linear Programming and Decomposition Methods / P. Rivas-Perea, J. Cota-Ruiz // Pattern Recognition Letters. – 2013. – Vol. 34(4). – P. 439-451. DOI: 10.1016/j.patrec.2012.10.026.

[3] Rizzi, A.M. Support vector regression model for BigData syst. // ArXiv: 1612.01458 [cs.DC]. – 2016.

[4] Hoi, S. Online Learning: A Comprehensive Survey / S. Hoi, D. Sahoo, J. Lu, P. Zhao // Technical report. – 2018. – 100 p.

[5] L. Breiman. Bagging predictors // ML. – 1996. – Vol. 24(2). – P. 123-140.

[6] Csiba, D. Importance sampling for minibatches / D. Csiba, P. Richtárik // Machine Learning Research. – 2018. – Vol. 19(1). – P. 962-982.

[7] Sadrfaridpour, E. Engineering fast multilevel support vector machines / E. Sadrfaridpour, T. Razzaghi, I. Safro // Machine Learning. – 2019. – Vol. 108(11). – P. 1879-1917.