Метод оценки эффективности сжатия матриц данных в процедурах рандомизированного машинного обучения[1]

Ю.С. Попков

Аннотация. Предложен метод оценки эффективности уменьшения размерности пространства признаков, ориентированный на использование в процедурах рандомизированнго машинного обучения. Эффективность измеряется в терминах функции Кульбака-Ляйблера, интерпретируемой как информационное расстояние между энтропийно-оптимальными функциями плотности распределения вероятностей для исходной и редуцированной входной обучающей коллекции.

Ключевые слова: рандомизированное машинное обучение, Энтропийный функционал, информационное расстояние Кульбака-Ляйблера.

Введение

Существенное расширение возможностей и ресурсов современных средств вычислительной техники позволило реализовать многочисленные ранее созданные методы машинного обучения (МО) и на основе их синтеза построить эффективные прикладные системные платформы для решения задач классификации, кластеризации и динамической регрессии. Их эксплуатация проявила некоторые особенности процедур МО, наиболее неприятная из которых есть следствие неопределенности в данных и моделях соответствующих объектов.

В [1] было предложено применить оптимизированную рандомизацию моделей для снижения влияния неопределенности. В результате появилась новая ветвь в МО - рандомизированное машинное обучение (РМО), которое для некоторых задач классификации, регрессии и кластеризации оказалось достаточно эффективным средством их решения в условиях неопределенности [2, 3].

Важной компонентой РМО являются данные, входящие в состав обучающей коллекции. Обучающая коллекция включают два блока данных. Входные данные, характеризуемые матрицами определенных размеров $(m \times s)$, где m - количество объектов в обучающей коллекции и s - размерность пространства признаков, в терминах которых характеризуются объекты. Выходные данные характеризуются m -мерным вектором, компонентами которого являются параметры принадлежности классу, либо параметры состояния объектов.

По разным причинам в перечисленных выше задачах возникает необходимость <<сжать>> входную компоненту обучающей коллекции, т.е. описывать ее $(m \times r)$-матрицей вместо исходной $(m \times s)$-матрицы. Содержательно это сводится к редукции пространства признаков, на массиве которых проводится обучение модели. Существует довольно много методов редукции, среди которых наиболее популярен

[1]Работа поддержана РФФИ (гранты 17-07-00286, 17-29-03119).

метод главных компонент (МГК) [4-7]. В [8] предложен метод энтропийной редукции матрицы (*DIP-метод*) данных со случайными элементами до заданного размера r признакового пространства.

Однако практическое применение всех методов редукции связано с количественной оценкой последствий редукции. Причем, такая оценка либо связана только с матрицей данных и не зависит от ее применения в различных задачах МО, либо зависит от класса задач МО. Так, оценивание эффекта от применения МГК зависит только от матрицы данных. К тому же, оно весьма чувствительно к вариациям значений ее элементов.

Данная статья посвящена процедуре оценки эффективности DIP-метода, ориентированного на решение задач РМО с линейными стохастическими моделями. Решением задач РМО являются функции плотности распределения вероятностей (ПРВ) параметров модели и шумов измерений [1]. Поэтому оценивание эффективности редукции сводится к оцениванию информационного расстояния между ПРВ выхода модели с исходной и редуцированной матрицей данных. Развивается метод оценивания с использованием расстояния Кульбака-Ляйблера [9, 10].

1. Характеризация линейных статических моделей с исходной и редуцированной входной обучающей коллекцией

Линейные статические модели являются довольно распространенным средством описания объекта в задачах РМО. Рассмотрим более детально это описание для случая исходной и редуцированной входной обучающей коллекции.

Пусть имеется линейный статический объект с одним выходом z и s входами $u \in R^s$, состояние которого характеризуется s признаками с весами-параметрами a. Математическая модель объекта - рандомизированная, т.е. параметры a случайные, интервальные:

$$a \in \mathcal{A} = \{a^- \le a \le a^+\}. \qquad (1)$$

Вероятностные свойства параметров характеризуются функцией плотности распределения вероятности (ПРВ) $P(a)$. Поскольку объект линейный и статический, выход z его модели связан со входами следующим соотношением:

$$z = \langle u, a \rangle, \qquad (2)$$

где $\langle , \rangle$ - скалярное произведение соответствующих векторов.

Допустим, что имеется входная обучающая коллекция $\mathcal{U}$ из m таких объектов, состояние которых характеризуются векторами $u^{(1)}, \dots, u^{(m)}$ из пространства признаков R^s, а выход характеризуется вектором $y = \{y_1, \dots, y_m\}$. В силу линейности объекта выход модели будет иметь следующий вид:

$$z = Ua, \qquad z = \{z_1, \dots, z_m\}, \qquad (3)$$

где $U = U_{(m \times s)}$ - $(m \times s)$-матрица с неотрицательными элементами, вектор

$$z \in \mathcal{Z} = [z^-, z^+], \qquad z^- = Ua^-; \; z^+ = Ua^+. \qquad (4)$$

Обычно предполагается, что выходы объектов содержат измерительные ошибки, которые имитируются случайным вектором $\bar{\xi} = \{\xi_1, \dots, \xi_m\}$ интервального типа, т.е.

$$\bar{\xi} \in \mathcal{K} = \{\bar{\xi}^- \le \bar{\xi} \le \bar{\xi}^+\}, \qquad (5)$$

с функцией плотности $L(\bar{\xi})$. Предполагается, что ПРВ параметров и шумов - непрерывно-дифференцируемые функции.

Наблюдаемый выход модели искажен шумом и имеет вид:

$$v = z + \bar{\xi} = U_{(m \times s)}\, a + \bar{\xi}. \qquad (6)$$

Вектор v принадлежит m-мерному параллелепипеду

$$\mathcal{V} = [v^-, v^+], \quad v^- = z^- + \bar{\xi}^-; \; v^+ = z^+ + \bar{\xi}^+. \qquad (7)$$

В процедурах РМО строятся оценки $\hat{P}(a)$ и $\hat{L}(\bar{\xi})$ соответствующих ПРВ. При этом используются обучающая коллекция $(\mathcal{U}, y)$ и рандомизированная модель (6). Полученные оценки позволяют реализовать эту модель, т.е. сгенерировать ансамбль $\mathcal{V}$ случайных векторов $v \in R^m$. Это - одна из задач рандомизированного машинного обучения [1], которую будем обозначать как *s-задачу* РМО.

Часто размерность s признакового пространства оказывается слишком большой для возможностей компьютера, на котором реализуется тот или иной алгоритм рандомизирован-

ного машинного обучения. Поэтому имеет смысл его редуцировать до некоторой заданной размерности $r < s$.

Введем матрицу входных данных $W_{(m\times r)}$, которая характеризует коллекцию $\mathcal{W}$ из m объектов в редуцированном признаковом пространстве R^r, и рандомизированную модель с параметрами $b = \{b_1, \dots, b_r\} \in R^r$:

$$w = W_{(m\times r)}b, \qquad b \in R^r,\ w \in R^m. \qquad (8)$$

Вектор параметров b имеет случайные компоненты интервального типа, т.е.

$$b \in \mathcal{B} = [b^-, b^+], \qquad (9)$$

с ПРВ $B(b)$. Наблюдаемый вектор

$$t = w + \bar{\zeta}, \qquad \bar{\zeta} \in R^m. \qquad (10)$$

Случайный вектор $\bar{\zeta}$, имитирующий ошибки измерений, интервального типа, т.е.

$$\bar{\zeta} \in \mathcal{J} = [\bar{\zeta}^-, \bar{\zeta}^+], \qquad (11)$$

с ПРВ $Q(\bar{\zeta})$. Векторы

$$t \in \mathcal{T} = [t^-, t^+],$$
$$t^- = Wb^- + \bar{\zeta}^-, \qquad t^+ = Wb^+ + \bar{\zeta}^+. \qquad (12)$$

Все функции плотности, характеризующие параметры модели и измерительные шумы, предполагаются непрерывно-дифференцируемыми функциями. Задача оценивания ПРВ $B(b)$ и генерация соответствующего ансамбля $\mathcal{T}$ есть *r-задача* РМО.

2. Метод сравнения *s*- и *r*-задач РМО

Развиваемый метод сравнения состоит из следующих этапов. На первом, с помощью процедуры РМО решаются s- и r-задачи с использованием обучающих коллекций $(\mathcal{U}, y)$ и $(\mathcal{W}, y)$ соответственно. Результатом этого этапа являются энтропийно-оптимальные ПРВ параметров и шумов: $P^*(a), L^*(\bar{\xi})$ - для s-задачи, и $B^*(b), Q^*(\bar{\zeta})$ - для r-задачи. На втором этапе, используя линейные модели (6, 10), вычисляются ПРВ $F_s(v)$ и $F_r(t)$, нормированные на множестве $\mathcal{C} = \mathcal{V} \cap \mathcal{T}$. Наконец, на третьем этапе вычисляется функция Кульбака-Ляйблера [9], значение которой характеризует абсолютное расхождение между ПРВ $F_s(v)$ и $F_r(t)$. Будем обозначать данный метод сравнения - <<$s\&r$ comparison (SRC)>>.

Рассмотрим первый этап и напомним процедуру РМО для решения s-задачи, модель в которой характеризуется равенством (6).

Критерий качества обучения, который позволяет получать наилучшие решения при максимальной неопределенности, формулируется как задача максимизации энтропийного функционала, определенного на функциях ПРВ $P(a), L(\bar{\xi})$

$$\mathcal{H}\left[P(a), L(\bar{\xi})\right] = -\int_{\mathcal{A}} P(a)\ln P(a)da - \\ -\int_{\mathcal{K}} L(\bar{\xi})\ln L(\bar{\xi})d\bar{\xi} \Rightarrow \max, \qquad (13)$$

при ограничениях:

$$\int_{\mathcal{A}} P(a)da = 1, \qquad \int_{\mathcal{K}} L(\bar{\xi})d\bar{\xi} = 1, \qquad (14)$$

$$\mathcal{M}\{z\} = \int_{\mathcal{A}} UP(a)da + \int_{\mathcal{K}} \bar{\xi}L(\bar{\xi})d\bar{\xi} = y \qquad (15)$$

Эта задача энтропийно-линейного функционального программирования [1]. Ее решение имеет вид:

$$P^*(a) = \pi^{-1}(\bar{\theta})\exp\langle -\bar{\theta}, U\,a\rangle,$$
$$L^*(\bar{\xi}) = \varpi^{-1}(\bar{\theta})\exp\langle -\bar{\theta}, \bar{\xi}\rangle, \qquad (16)$$

где

$$\pi(\bar{\theta}) = \int_{\mathcal{A}} \exp\langle -\bar{\theta}, U\,a\rangle da,$$
$$\varpi(\bar{\theta}) = \int_{\mathcal{K}} \exp\langle -\bar{\theta}, \bar{\xi}\rangle d\bar{\xi}. \qquad (17)$$

Множители Лагранжа $\bar{\theta}$ определяются решением балансовых уравнений (15):

$$\int_{\mathcal{A}} UP^*(a)da + \int_{\mathcal{K}} \bar{\xi}L^*(\bar{\xi})d\bar{\xi} = y. \qquad (18)$$

Итак, в результате решения s-задачи РМО получаем две энтропийно-оптимальных ПРВ параметров $P^*(a)$ и шумов $L^*(\bar{\xi})$.

На втором этапе используем линейную модель (6), параметры и шумы в которой - независимы. Функция плотности распределения вероятностей $F(v)$ наблюдаемого вектора v имеет вид:

$$F(v) = \int_{\mathcal{K}} G(v - \bar{\xi})L^*(\bar{\xi})d\bar{\xi}, \qquad (19)$$

где $G(\bullet)$ -плотность вектора z (6). Согласно этому равенству имеем:

$$a = (U^TU)^{-1}U^Tz. \qquad (20)$$

Поэтому

$$P^*((U^TU)^{-1}U^Tz) = \eta(z), \qquad (21)$$

где вектор z определен на множестве $\mathcal{Z}$ (4). Нормируя эту функцию, получим функцию ПРВ вектора z:

$$G(z) = \frac{\eta(z)}{\int_{\mathcal{Z}} \eta(z)} dz, \quad (22)$$

нормированную на множестве $\mathcal{Z}$ (4).

Итак, используя равенство (19), получим ПРВ $F(v), v \in \mathcal{V}$ (7) для рандомизированной модели *s* -задачи. Обозначим ее $F_s(v)$.

Аналогичную процедуру нужно проделать для *r*-задачи (8 - 10) с редуцированной матрицей данных W (1) и получить ПРВ $F_r(t)$ нормированную на множестве $\mathcal{T}$ (12).

Для сравнения ПРВ $F_s(v)$ и $F_r(t)$ введем множество

$$\mathcal{C} = \mathcal{V} \cap \mathcal{T}, \quad (23)$$

и пронормируем указанные функции на этом множестве. Получим следующие функции ПРВ:

$$\tilde{F}_s(c) = \frac{F_s(c)}{\int_{\mathcal{C}} F_s(c)dc}, \tilde{F}_r(c) = \frac{F_r(c)}{\int_{\mathcal{C}} F_r(c)dc}.$$

На третьем этапе, сравнение полученных функций ПРВ будем проводить, используя функцию Кульбака-Ляйблера в качестве меры *абсолютной информационной ошибки* Δ между ПРВ $\tilde{F}_r(c)$ и $\tilde{F}_s(c)$:

$$KL(\tilde{F}_s, \tilde{F}_r) = \int_{\mathcal{C}} \tilde{F}_r(c) \ln \frac{\tilde{F}_r(c)}{\tilde{F}_s(c)} = \Delta \geq 0, \quad (24)$$

Заметим, что минимум $\Delta = 0$ достигается при $\tilde{F}_r(c) = \tilde{F}_s(c)$.

Введем также *относительную информационную ошибку* δ в виде:

$$\delta = \frac{\Delta}{H_s + H_r}, \quad (25)$$

где

$$H_s = \int_{\mathcal{C}} \tilde{F}_s(c) \ln \tilde{F}_s(c) dc,$$
$$H_r = \int_{\mathcal{C}} \tilde{F}_r(c) \ln \tilde{F}_r(c) dc. \quad (26)$$

Последние равенства представляют собой информационные энтропии функций ПРВ $\tilde{F}_s$ и $\tilde{F}_r$, определенные на носителе $\mathcal{C}$ (23).

Заключение

В статье предложен метод оценки эффективности уменьшения размерности пространства признаков, ориентированный на использование в процедурах рандомизированного машинного обучения. Эффективность измеряется в терминах функции Кульбака-Ляйблера, интерпретируемой как информационное расстояние между энтропийно-оптимальными функциями ПРВ для исходной и редуцированное входной обучающей коллекции.

Литература

1. Yu. S. Popkov, Yu. A. Dubnov, A. Yu. Popkov. Randomized Machine Learning: Statement, Solution, Applications // Proceedings of 2016 IEEE 8-th International Conference on Intelligent Systems (IS16). September 4-6, 2016. Sofia, Bulgaria, P.27-39.
2. Yuri S. Popkov, Zeev Volkovich, Yuri A. Dubnov, Renata Avros and Elena Ravve. // Entropy '2'-Soft Classification of Objects // Entropy, 2017, Vol. 19, Iss. 4, No.178.
3. Popkov Y.S., Dubnov Y.A. Entropy-robust randomized forecasting under small sets of retrospective data. Automation and Remote Control, 2016,v.77, No.5, p.839-854.
4. Bruckstein A.M., Donoho D.L., Elad M. From Sparse Solutions of Systems of Equations to Sparse Modeling of Signals and Images. SIAM Rev. 2009, v.51, No.1, p.34-81.
5. Кендал М.Дж., Стюарт А. Статистические выводы и связи./ Пер. с англ. М.. Наука, 1973.
6. Jollife I.T. Principal Component Analysis. N.Y. Springer-Verlag, 2002.
7. Поляк Б.Т., Хлебников М.В. Метод главных компонент: робастные версии. Автоматика и Телемеханика, 2017, №3, с.130-148.
8. Попков Ю.С. Энтропийный метод сжатия матриц со случайными значениями элементов. ИТВС, 2018, №1, с.
9. Kullback S., Leibler R.A. On information and Sufficiency. Ann. of Math. Statistics, 1951, v.22(1), p. 79-86.
10. Zhang Y., Li S., Wang T., Zhang Z. Divergence-based feature selection for separate classes. Neurocomputing, 2013, v.101, p. 32-42.

Попков Юрий Соломонович. Директор ИСА ФИЦ ИУ РАН. Окончил МЭИ в 1960 году. Академик, доктор технических наук, профессор. Количество печатных работ: 200. Область научных интересов: математическое программирование, динамические макросистемы. E-mail: popkov.yuri@gmail.com

Effectiveness estimation of matrices compression in the procedures of randomized machine learning

Yu.S. Popkov

The metod of estimation effectiveness of the matrices compressions? That oriented to the procedures ranomized machine learning. It is proposed to measure of effectiveness in the term of the Kullback-Leibler function.

Keywords: randomized machine learning, entropy, KL-distance.

References

1. Yu. S. Popkov, Yu. A. Dubnov, A. Yu. Popkov. Randomized Machine Learning: Statement, Solution, Applications // Proceedings of 2016 IEEE 8-th International Conference on Intelligent Systems (IS16). September 4-6, 2016. Sofia, Bulgaria, P.27-39.
2. Yuri S. Popkov, Zeev Volkovich, Yuri A. Dubnov, Renata Avros and Elena Ravve. // Entropy '2'-Soft Classification of Objects // Entropy, 2017, Vol. 19, Iss. 4, No.178.
3. Popkov Y.S., Dubnov Y.A. Entropy-robust randomized forecasting under small sets of retrospective data. Automation and Remote Control, 2016,v.77, No.5, p.839-854.
4. Bruckstein A.M., Donoho D.L., Elad M. From Sparse Solutions of Systems of Equations to Sparse Modeling of Signals and Images. SIAM Rev. 2009, v.51, No.1, p.34-81.
5. Kendall M, Stewart A. Statisticheskie vivodi I sviazi, Nauka , 1973.
6. Jollife I.T. Principal Component Analysis. N.Y. Springer-Verlag, 2002.
7. Polyak B.T., Hlebnikov M.V. Metod glavnih komponent: robastnie versii // Avtomatika I telemehanika, 2017, №3, c.130-148.
8. Popkov Y.S. Entropiinii metod szhatia matric so sluchainimi znacheniami elemntov // ITVS, 2018, №1, c.
9. Kullback S., Leibler R.A. On information and Sufficiency. Ann. of Math. Statistics, 1951, v.22(1), p. 79-86.
10. Zhang Y., Li S., Wang T., Zhang Z. Divergence-based feature selection for separate classes. Neurocomputing, 2013, v.101, p. 32-42.

Popkov Yuri Solomonovich. Director of ISA FRC CSC RAS. D.Sc, Professor. The number of publications : 200 publications. Research interests: mathematial programming, dynamic macrosystems. E-mail: popkov.yuri@gmail.com