Модификация метода t-SNE для задачи классификации

Молибог Игорь
Московский физико-технический институт, ЗАО Анти-Плагиат
molybog.igor@phystech.edu

Мотренко Анастасия
Московский физико-технический институт, ЗАО Анти-Плагиат
anastasiya.motrenko@phystech.edu

Аннотация

В работе исследуется задача классификации объектов в многомерных пространствах. Для повышения качества классификации предлагается модификация алгоритма снижения размерности t-SNE. В предлагаемой модификации при обучении используется информация о разметке, не возникает необходимость заново обучать алгоритм при добавлении новых данных, а также предусмотрена параллельная реализация.

Ключевые слова: понижение размерности, Stochastic Neighbor Embedding, классификация.

1. Введение

В работе рассматривается задача классификации многомерных объектов, признаковое описание которых имеет в себе скрытые функциональные зависимости. Предполагается, что объекты содержатся вблизи многообразия много меньшей размерности, чем размерность исходного пространства. Назовем это предположение гипотезой многообразия [3]. Данные многих практических задач, включая задачи анализа генома, анализа текста и распознавания изображений подчиняются этой гипотезе [8]. Методы проверки этой гипотезы описаны в [9]. Практической задачей, рассматриваемой в данной работе, является задача обнаружения внутреннего плагиата [10], [5].

Для решения задачи снижения размерности предлагается модификация метода визуализации данных t-distributed Stochastic Neighbor Embedding (t-NSE) [8]. Этот метод сохраняет структуру близости между ближайшими объектами. Преимуществом метода t-SNE является склонность к локализации изолированных плотных пространственных структур произвольной геометрии. Предполагается, что именно в такие структуры в признаковом пространстве формируются схожие по смыслу и стилю части текста.

Предлагаемая модификация вкладывает данные в пространство низкой размерности, в котором после этого строится классификатор. Она предусматривает вложение тестовых данных без повторного вложения обучающих, однако может учитывать разметку обучающих данных, а также имеет параллельную реализацию.

2. Снижение размерности методом t-SNE

Обозначим через $\mathscr{X} \subset \mathbb{R}^n$ множество всех возможных векторов признаков изучаемых объектов. Предполагается, что объекты $\mathscr{X}$ подчиняются *гипотезе многообразия*: $\exists f : \mathbb{R}^d \to \mathbb{R}^n$ — такой гладкий гомоморфизм, что

$$\forall \mathbf{x} \in \mathscr{X} \ \exists \, \mathbf{z}^* \in \mathbb{R}^d : \mathbf{x} = f(\mathbf{z}^*) + \varepsilon(\mathbf{z}^*),$$

где $\varepsilon(\mathbf{z}^*)$ — случайный вектор с нулевым матожиданием и конечной матрицей корреляции. Будем называть d эффективной размерностью исходного пространства $\mathscr{X}$. Она определяется природой признакового пространства. Поскольку d заранее неизвестно, введем понятие результирующего пространства $\mathbb{R}^k$, в котором выполняется поиск решения. В общем случае $k \neq d$. Процесс поиска образов объектов выборки в результирующем пространстве назовем вложением в него.

Рассмотрим выборку из m объектов $X = [\mathbf{x}_1 \ \ldots \ \mathbf{x}_m]^T \subset \mathscr{X}$. Пусть $p_{ij} = P(\mathbf{x}_i, \mathbf{x}_j)$ и $q_{ij} = Q(\mathbf{z}_i, \mathbf{z}_j)$ — вероятностные меры сходства объектов в $\mathbb{R}^n$ и $\mathbb{R}^k$ соответственно, задаваемые по формулам

$$p_{i|j} = \frac{\exp\left(-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{2\sigma_i^2}\right)}{\sum\limits_{k \neq i} \exp\left(-\frac{||\mathbf{x}_i - \mathbf{x}_k||^2}{2\sigma_i^2}\right)}, \quad p_{ij} = \frac{p_{j|i} + p_{i|j}}{2m}.$$

$$q_{ij} = \frac{(1 + ||\mathbf{z}_i - \mathbf{z}_j||^2)^{-1}}{\sum\limits_{k \neq i} (1 + ||\mathbf{z}_i - \mathbf{z}_k||^2)^{-1}}, \quad q_{ii} = 0.$$

Расположение $Z = [\mathbf{z}_1 \ \dots \ \mathbf{z}_m]^T \subset \mathbb{R}^k$ как образов X в результирующем пространстве $\mathbb{R}^k$ находится путем минимизации дивергенции Кульбака-Лейбера

$$Z_{min} = \underset{Z \in \mathbb{R}^{m \times k}}{\operatorname{argmin}} C(X, Z), \tag{1}$$

$$C(X, Z) = KL(P||Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}. \tag{2}$$

Задача решается градиентными методами [8].

Рассмотрим задачу классификации с обучающей выборкой X и тестовой выборкой из m' объектов $X' = [\mathbf{x}_{m+1} \ \dots \ \mathbf{x}_{m+m'}]^T \subset \mathscr{X}$. Соответственно, метки классов $y_i \in \{0, 1\}$, $i = 1, \dots, m$ известны, а $\hat{y}_i$, $i = m+1, \dots, m+m'$ необходимо оценить.

3. Предлагаемая модификация t-SNE

Так как на этапе обучения данные X' могут быть недоступны, метод непараметрического t-SNE не применим для снижения размерности в задачах классификации. Будем называть это проблемой не просмотренных объектов ("out-of-sample problem"). Для решения этой проблемы предлагается минимизировать (2) независимо по различным подмножествам переменных.

Для повышения качества классификации в результирующем пространстве предлагается добавить метки как признаки в обучающую выборку и улучшить начальное приближение градиентного метода.

Использование исходной разметки выборки при вложении для обучения классификатора. Для учета разметки обучающей выборки признаковая матрица X расширяется дополнительным столбцом признаков: $\tilde{X} = \left(\begin{array}{c|c} X & \mu \mathbf{y} \end{array} \right)$, где μ — вес меток как признаков. В модифицированном алгоритме на основе расширенной матрицы $\tilde{X}$ выполняется поиск образов Z, на котором обучается классификатор. Таким образом, при построении вложения обучающей выборки решается задача

$$Z_{min} = \underset{Z \in \mathbb{R}^{m \times k}}{\operatorname{argmin}} C\left(\left(\begin{array}{c|c} X & \mu \mathbf{y} \end{array} \right), Z \right).$$

Вложение новых объектов в пространство со сниженной размерностью для классификации. Обозначим через $Z' = [\mathbf{z}_{m+1} \ \dots \ \mathbf{z}_{m+m'}]^T$ образы X' в результирующем пространстве. Аналогично (1), сформулируем задачу поиска Z' в виде m' задач k-мерной минимизации, которые могут быть решены независимо:

$$\mathbf{z}_i^{min} = \underset{z_i \in \mathbb{R}^{m'}}{\operatorname{argmin}} C\left(\left(\frac{X}{\mathbf{x}_i^T} \right), \left(\frac{Z}{\mathbf{z}_i^T} \right) \right),$$

$$i = m+1, \dots, m+m',$$

где матрицы $\left(\frac{X}{\mathbf{x}_i^T} \right)$ и $\left(\frac{Z}{\mathbf{z}_i^T} \right)$ получены из X и Z добавлением строк $\mathbf{x}_i^T$ и $\mathbf{z}_i^T$ соответственно. При использовании такого подхода предполагается, что в обучающая выборка достаточно репрезентативна.

Для инициализации образов $z_{i'}$ классифицируемых объектов предлагается использовать метод среднего по образам соседей:

$$\mathbf{z}_{i'}^{(0)} = \sum_{i=1}^{m} \mathbf{z}_i w_{ii'}, \quad \sum_{i=1}^{m} w_{ii'} = 1, \quad i' = m+1, \dots, m+m'$$

где $w_{ii'}$ — веса образов объектов $\mathbf{x}_i$, $i = 1, \dots, m$. В работе рассмотрены два способа задания весов:

$$w_{ii'}^{softmax} = \frac{\exp(-\|\mathbf{x}_i - \mathbf{x}_{i'}\|)}{\sum\limits_{k=1}^{m} \exp(-\|\mathbf{x}_k - \mathbf{x}_{i'}\|)}, \text{ или} \tag{3}$$

$$w_{ii'}^{stud} = \frac{(1 + \|\mathbf{x}_k - \mathbf{x}_{i'}\|^2)^{-1}}{\sum\limits_{k=1}^{m} (1 + \|\mathbf{x}_k - \mathbf{x}_{i'}\|^2)^{-1}}. \tag{4}$$

4. Исследование свойств алгоритма на синтетических данных

Для эмпирического исследования свойств предлагаемого алгоритма использовались синтетические выборки, заведомо подчиняющиеся гипотезе многообразия. Использование синтетических выборок при тестировании алгоритма гарантирует, что наблюдаемые результаты связаны со свойствами алгоритма, а не качеством исходных данных. Использованная выборка представляла собой многомерный аналог выборки “swissroll dataset” [11]. Генерировалось одинаковое количество объектов разных классов, а на обучение и контроль выборка разбивалась в соотношении 1/4.

Ускорение и распараллеливание. Для ускорения процедуры вложения при работе с большими данными предлагается процедура поэтапного вложения объектов из X блоками, размер которых, S_s для первого по очереди и S_b для всех остальных, много меньше размера m всей выборки.

Была исследована зависимость качества классификации от величины отношения размера стартовой части к размеру выборки S_s/m. Согласно результатам эксперимента, зависимость от этих параметров незначительна. При этом скорость работы алгоритма увеличивается при наличии разбиений на стартовую и дополняющую часть. Таким образом показано, что предложенная модификация алгоритма позволяет значительно ускорить его работу без существенного понижения качества.

Было проведено исследование зависимости значения функции качества классификации от S_b. Обнаружено, что предлагаемый метод устойчив относительно параметра S_b.

Сравнение с альтернативными методами снижения размерности. Для сравнения предлагаемого алгоритма и его исследования рассматривается классификация в комбинации с другими методами снижения размерности: Principal Component Analysis (PCA) [4], Local Linear Embedding (LLE) [6], Isometric Mapping (ISOMAP) [7], а также без применения снижения размерности. В качестве классификатора использовался алгоритм SVM [2]. На рис. 1 изображены результаты при размерности выборки $n = 600$.

Было обнаружено, что качество классификации F_1 повышается с ростом μ. Разработанный метод при достаточно больших значениях μ показывает в среднем лучшие результаты среди всех рассмотренных методов снижения размерности, а также превосходит по качеству классификацию в исходном пространстве. Эксперименты проводились при постоянных $m = 500, k = 3, S_b = 100, S_s = 400$.

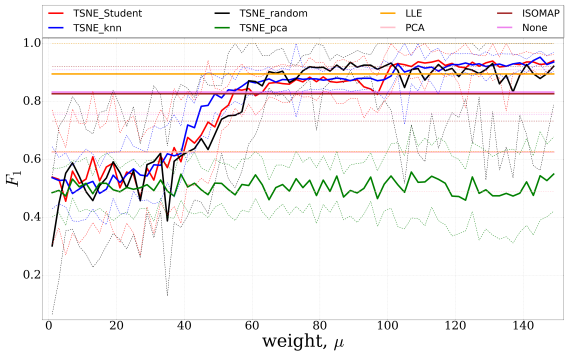
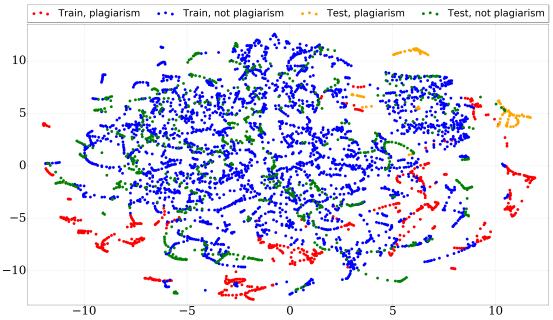


Рис. 1: Зависимость F_1 от μ при размерности выборки $n = 600$

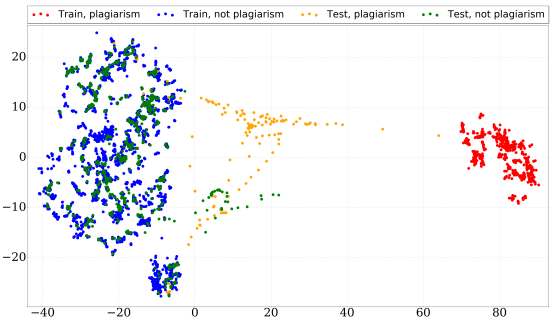
5. Демонстрация работы алгоритма на реальных данных

Предполагается, что гипотеза многообразия справедлива для данных задачи внутреннего плагиата. Эта задача состоит в поиске отклонений стиля части текста от стиля текста в целом.

На рис. 2 приведено сравнение результатов работы исходного и модифицированного алгоритмов в применении к одному и тому же документу из коллекции PAN. Исходный алгоритм вкладывал обучающую и тестовую выборки одновременно. Из рисунка видно, что предложенный метод привел данные документа к виду, облегчающему построение классификатора.



(a) Original t-SNE



(b) Modified t-SNE

Рис. 2: Визуализация реальных данных оригинальным и модифицированным t-SNE

Авторы планируют проведение дальнейших исследований в области проверки гипотезы многообразия в задачах, связанных с анализом текстов.

6. Заключение

В работе была предложена модификация непараметрического метода снижения размерности t-SNE. Алгоритм был распараллелен, решена проблема не просмотренных объектов и внедрена возможность учета разметки при вложении для классификации. Был проведен вычислительный эксперимент на синтетических данных, показывающий эффективность предложенного метода в применении к задаче классификации. Была определена зависимость качества классификации с применением описанного

метода от его параметров, экспериментально обосновано использование поэтапного обучающего вложения. Было проведено сравнение полученного значения качества с результатами классификации с применением других методов снижения размерности, а также без их применения.

Список литературы

[1] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. *MIT Press*, 2001.

[2] L´eon Bottou. Stochastic gradient descent tricks. *Microsoft Research, Redmond, WA*, 2012.

[3] Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. 2013.

[4] Hyunsoo Kim, Haesun Park, and Hongyuan Zha. Distance preserving dimension reduction for manifold learning. *Proceedings of the 2007 SIAM International Conference on Data Mining*, 2007.

[5] Mikhail Kuznetsov, Anastasia Motrenko, Rita Kuznetsova, and Vadim Strijov. Methods for intrinsic plagiarism detection and author diarization. *Notebook for PAN at CLEF 2016*, 2016.

[6] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 2000.

[7] J. Tenenbaum, V. Silva, and J. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 2000.

[8] Laurens van der Maaten. Visualizing data using t-sne. *Journal of Machine Learning Research*, 2008.

[9] Hariharan Narayanan, Sanjoy Mitter. Sample complexity of testing the manifold hypothesis. 2010.

[10] Sven Meyer zu Eissen and Benno Stein. Intrinsic plagiarism detection. *Proceedings of the 28th European Conference on IR Research, ECIR*, 2006.

[11] Xiaohui Wang and J. S. Marron. A scale-based approach to finding effective dimensionality in manifold learning. *Electronic Journal of Statistics*, 2008.