UDC 004.8

Ярослав Подзірей, Павло Регіда

КЛАСТЕРИЗАЦІЯ. ЗМЕНШЕННЯ РОЗМІРНОСТІ ОЗНАК ОБ'ЄКТІВ В МАШИННОМУ НАВЧАННІ ТА ВІЗУАЛІЗАЦІЯ ДАНИХ. МЕТОД t-SNE.

CLUSTERING. DECREASE OF DIMENSION OBJECT DESCRIPTION IN MASTER EDUCATION AND VISUALIZATION OF DATA. T-SNE METHOD

У даній статті розглядається питання кластеризації, зменшення розмірності ознак вибірки вхідних об'єктів аналізу і візуалізації даних за допомогою алгоритму t-SNE. Пропонується варіант можливого покращення даного алгоритму. Для тестування використовується мова Python та популярна бібліотека scikit-learn.

Ключові слова: машинне навчання, кластеризація, зменшення розмірності простору вибірки ознак, візуалізація даних, алгоритм t-SNE.

Рис.: 1.. Бібл.: 5.

In this article, the question of clustering, reducing the size of the characteristics of the sample of input objects of analysis and data visualization using the t-SNE algorithm. A variant of possible improvement of this algorithm is proposed. For testing, Python language and the popular scikit-learn library are used.

Key words: machine learning, clustering, diminishing of the dimension of the sample of traits, data visualization, t-SNE algorithm.

Fig .: 1 .. Bible: 5.

Актуальність теми дослідження. Машинне навчання розвивається дуже швидкими темпами і кількість даних для аналізу росте з кожним днем. Така тенденція, в свою чергу, призводить до поступового ускладнення аналізу даних і розуміння взаємозв'язків між ними, що в машинному навчанні є ключовою особливістю. Саме тому люди в першу чергу шукають можливості для спрощення представлення даних без втрати деякої важливої інформації. Дана стаття присвячена одному з відносно молодих алгоритмів, який досить добре справляється з задачами подібного типу.

Постановка проблеми. Не дивлячись на те, що алгоритм t-SNEдосить добре себе зарекомендував в сфері машинного навчання та аналізу даних, він все ж має певні недоліки, які намагаються бути вирішеними в даній статті.

Аналіз останніх досліджень і публікацій. Протягом останніх років з'являється все більше статей присвячених різним модифікаціям алгоритму t-SNE, проте більшість цих досліджень направлені на підвищення швидкості даного алгоритму, оскільки він має одну з найбільших складностей серед конкурентів. Більшість інших недоліків даного алгоритму не є досить детально вивченими на даний час.

Виділення недосліджених частин загальної проблеми. Дана стаття присвячена дослідженню потенційності покращення алгоритму t-SNE за рахунок можливості використання інформації про розмітку та склад вибірки, а також можливості додавати нові точки в вибірку для покращення візуального результату без перерахунку всіх координат.

Постановка завдання. Провести модифікацію алгоритму t-SNEта аналіз його тестування.

Викладення основного матеріалу. Зазвичай в машинному навчанні ми маємо справу з досить багатовимірними вибірками даних, оскільки дуже часто аналіз даних проводиться одночасно відносно багатьох параметрів і уникнути цього в еру технічного розвитку неможливо. Але при цьому ми хочемо якось дивитися на ці дані, розуміти, як вони влаштовані, які там взаємозв'язки, які ознаки важливі, а які - ні, як співвідносяться класи між собою. Існує досить велика кількість алгоритмів які допомагають провести певну характеристику даних без пониження розмірності, але такі данні досить важко сприймаються людиною, оскільки ми не можемо мислити більш ніж в 3 вимірах.

Хотілося б відобразити всю вибірку в двовимірний або тривимірний простір так, щоб відразу було видно всі закономірності в даних, вся їх структура. Наприклад, якщо класи сильно перемішані між собою, якісь класи виділяються і їх можна добре відокремити від решти, тобто провести процес кластеризації. Власне, саме так ми приходимо до задачі візуалізації даних: це окремий випадок нелінійного зниження розмірності, коли розмірність простору, в яке ми намагаємося спроектувати нашу вибірку, це 2 або 3.

На сьогоднішній день, одним з найбільш популярних підходів до вирішення такого класу задач є використання методу t-SNE(t-distributed stochastic neighbor embedding).

Суть алгоритму t-SNE полягає в розрахунку двох матриць схожості: перша - для вихідних точок в багатовимірному просторі, друга - для кінцевих точок в двовимірному. Друга матриця підбирається таким чином, щоб якомога краще відображати властивості першої.

На відміну від алгоритмів конкурентів, метод t-SNEне потребує максимальної близькості розміщення об’єктів у результуючому просторі відносно заданого. Даний алгоритм потребує лише збереженню пропорцій об’єктів, що дає досить великі переваги, оскільки при досить великому скупченні об’єктів в багатовимірному просторі, досить важно відобразити дані закономірності в маломірному, тобто данні виходять більше чіткими і зрозумілими для аналізу.

Алгоритм створює статистичну модель різних особливостей даних, які формують вірогідну картину будь-якого заданого значення. Він випадковим чином розсіює точки даних і починає переміщати ці точки крок за кроком в процесі, який намагається знайти золоту середину, де точки збалансовані статистично. Тобто, точки тягнуть і натискають одна на одну, ґрунтуючись на тому, як незбалансовані вони з усіма іншими точками, вони самоорганізуються за алгоритмом. Рух кожної характеристики точок в кожному напрямку визначається впливом інших точок, які повинні бути розміщені на відстані за їх особливою ймовірністю.

Загальні формули для обчислення ймовірностей в заданому та результуючому просторах:

$$p(x_j|x_i) = \frac{\exp\left(\| x_i - x_j\|^2/2\sigma^2\right)}{\sum_{k \neq i} \exp\left(\|x_i - x_k\|^2/2\sigma^2\right)} \quad (1)$$

$$q(\tilde{x}_j|\tilde{x}_i) = \frac{\left(1 - \|\tilde{x}_i - \tilde{x}_j\|^2\right)^{-1}}{\sum_{k \neq i} \left(1 + \|\tilde{x}_i - \tilde{x}_k\|^2\right)^{-1}} \quad (2)$$

Ці формули показують на скільки j-та точка близька до і-тої з заданим відхиленням σ. Сигма відрізняється для кожної точки. Воно підбирається так, щоб виконувалась оцінка перплексії:

(3)

де H(Pi) – ентропія Шеннона в бітах

(4)

Для розуміння фізичних властивостей алоритму, перплексію можна вважати пом'якшеною оцінкою ефективної кількості сусідніх точок для заданої. Зазвичаний беруться значення від 5 до 50.

Для порівняння розподілень ймовірностей заданого і результуючого простору з точки зору математики дуже добре підходить дивергенція Кульбака-Лейблера:

$$p(x_j|x_i)\log\frac{p(x_j|x_i)}{q(\tilde{x}_j|\tilde{x}_i)} \to \min_{\tilde{x}_1,\ldots,\tilde{x}_z} \tag{5}$$

Загальне розподілення ймовірностей:

$$p_{ij} = \frac{p_{i|j} + p_{j|i}}{m}, \tag{6}$$

де m – кількість точок в наборі даних.

Заданий алгоритм мінімізує суму всіх ймовірних відстаней за допомогою алгоритму градієнтного спуску з наступним градієнтом:

$$\frac{\partial C\,cost}{\partial y_i} = 4\sum_j (p_{ij} - y_{ij})(y_i - y_j)(1 - \| y_i - y_j \|^2)^{-1} \tag{7}$$

Для розуміння фізичних властивостей даного процесу варто зазначити, що алгоритм буде певним чином притягувати точки простору відображення для близьких між собою точок заданого простору, та відштовхувати для точок, які розміщені далеко в заданому багатовимірному просторі.

Спроба модифікації заданого алгоритму проводилась з урахуванням вже існуючих рішень, які зазвичай направлені на зменшення швидкості виконання або зменшення складності алгоритму. Ціллю модифікації алгоритму є можливість використовувати інформацію про розмітку вибірки та додавати нові точки без перерахунку всіх координат.

Для початку пропонується взяти до уваги вибірку $X^{m0} = \{\mathbf{x}_1,\ldots,\mathbf{x}_m 0\}$. Для неї оцінимо $P^{m0\times m0}$ і знайдемо його відображення Z^{m0} и $Q^{m0\times m0}$. Додамо нову множину об'єктів $X^{m-m0} = \{\mathbf{x}_m 0_{+1},\ldots,\mathbf{x}_m\}$. Не міняючи положень $\mathbf{z}_i$, $i = m^0+1,\ldots,m$ початкових точок знайдемо відображення Z^{m-m0} и $Q^{(m-m0)\times m}$, зміщуючи точки в напрямку градієнта.(7)

Для підвищення надійності Z^{m0} використаємо інформацію про розмітку y_i об'єктів $\mathbf{x}_i$, $i = 1,\ldots,m^0$. Пропонується використовувати інформацію про розмітку при обчисленні $P^{m0\times m0}$, наприклад:

$$p_{j|i} = \frac{\exp(-||\mathbf{x}_i - \mathbf{x}_j||^2/\sigma_{ij}^2)}{\sum_{k\neq i}\exp(-||\mathbf{x}_i - \mathbf{x}_k||^2/\sigma_{ij}^2)}, \quad \sigma_{ij} = \begin{cases}\sigma^2, y_i = y_j,\\ \varepsilon \to 0, y_i \neq y_j\end{cases} \tag{8}$$

Для тестування була використана задача:

Для існуючого документу C —послідовність символів $c_1,\ldots,c_{Ld}$. В даному документі присутні зашифровані блоки інформації. Необхідно для кожного символу c_l документу провести відповідну кластеризацію за характеристикою приналежності заданого символу до зашифрованого блоку інформації.

Результати виконання модифікації:

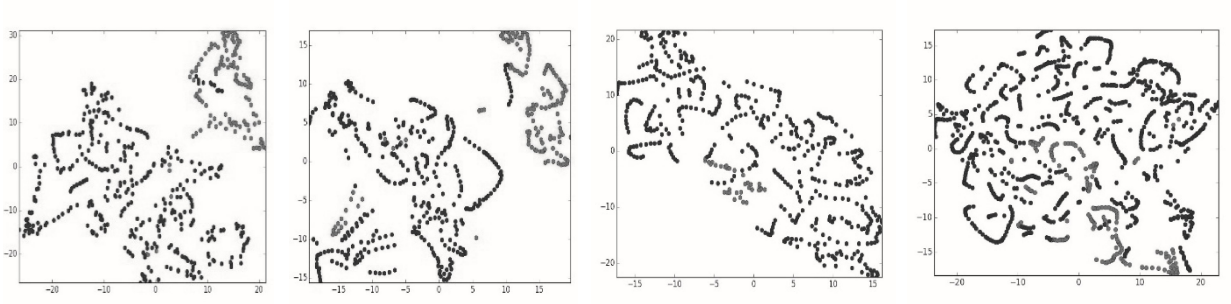

Рис. 1. Результати виконання для,
відповідно, 500, 500, 1000, 1000 речень в файлі

Висновки. Проаналізувавши результати можна зробити висновок, що задана модифікація алгоритму, на даний момент, має досить непоганий результат лише для малого об'єму даних. При вхідних параметрах великого розміру, результат виходить не дуже чітким, не можна відразу побачити чітку кластеризацію об'єктів, а відповідно, не можна провести якісний аналіз даних. Даний метод модифікації потребує більш детального вивчення.

Список використаних джерел

1. L.J.P. van der Maaten and G.E. Hinton. Visualizing High-Dimensional Data Using t-SNE. Journal of Machine Learning Research 9(Nov):2579-2605, 2008. PDF

2. G.E. Hinton and S.T. Roweis. Stochastic Neighbor Embedding. In Advances in Neural Information. Processing Systems, volume 15, pages 833–840, Cambridge, MA, USA, 2002. The MIT Press. PDF

3. L.J.P. van der Maaten. Accelerating t-SNE using Tree-Based Algorithms. Journal of Machine Learning Research 15(Oct):3221-3245, 2014. PDF

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

5. Wes McKinney. Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython. O'Reilly, 2012.

ДОВІДКА ПРО АВТОРІВ

Подзірей Ярослав Іванович – студент, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

Yaroslav Podzirei – student, Department of Computer Engineering,
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute".
E-mail: ginofa00@gmail.com

Регіда Павло Геннадійович – аспірант, кафедра обчислювальної техніки, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

Rehida Pavlo – PhD student, Department of Computer Engineering, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute".
E-mail: pavel.regida@gmail.com

Yaroslav Podzirei, Pavlo Rehida

CLUSTERING. DECREASE OF DIMENSION OBJECT DESCRIPTION IN MASTER EDUCATION AND VISUALIZATION OF DATA. T-SNE METHOD

Target setting. People are primarily looking for opportunities to simplify the presentation of data without losing some important information. This article is devoted to one of the relatively young algorithms, which quite well copes with problems of this type.

Formulation of the problem. Despite the fact that the t-SNE algorithm has proven itself well in the field of machine learning and data analysis, it nevertheless has certain disadvantages that are trying to be resolved in this article.

Actual scientific researches and issues analysis. In recent years, there are more articles devoted to various modifications of the t-SNE algorithm, but most of these studies are aimed at increasing the speed of this algorithm, since it has one of the greatest difficulties among competitors. Most of the other disadvantages of this algorithm are not sufficiently studied at present.

Uninvestigated parts of general matters defining. This article is devoted to the study of the potential of improving the t-SNE algorithm due to the possibility of using information about the markup and composition of the sample, as well as the ability to add new points in the sample to improve the visual result without recalculating all coordinates.

The research objective. Modify the t-SNE algorithm and analyze its testing.

General model structure. Usually in machine learning, we deal with rather high-dimensional data samples, because very often the analysis of data is carried out simultaneously with respect to many parameters and it is impossible to avoid this in the era of technical development. But at the same time, we want to somehow look at these data, understand how they are arranged, what are the interrelationships there, which features are important, and which - no, how the classes relate to each other. There is a fairly large number of algorithms that help to carry out a certain characteristic of the data without diminishing the dimension, but such data are hardly perceived by a person, since we can not think in more than 3 dimensions. I would like to display the entire sample in two-dimensional or three-dimensional space so that all the regularities in the data, all their structure, were immediately visible. For example, if the classes are strongly mixed with each other, some classes are allocated and they can be well separated from the rest, that is to hold the process of clusterization. Actually, this is the way we come to the task of data visualization: this is a special case of a nonlinear decrease in dimensionality, when the dimension of the space in which we try to design our sample is 2 or 3.To date, one of the most popular approaches to solving such a task class is the use of the t-SNE method (t-distributed stochastic neighbor embedding).

Conclusions. After analyzing the results it can be concluded that the given modification of the algorithm, at the moment, has a fairly good result for only a small amount of data. When input parameters are large, the result is not very clear, it is impossible to immediately see a clear clustering of objects, and accordingly, it is impossible to conduct a qualitative analysis of data.

Key words: machine learning, clustering, diminishing of the dimension of the sample of traits, data visualization, t-SNE algorithm.