Improving classification quality for intrinsic plagiarism problem*

I O. Molybog,[1] A .P. Motrenko,[2] V. V. Strijov[3]

The paper addresses the classification problem in multidimensional spaces. The authors propose a supervised modification of t-distributed Stochastic Neighbor Embedding algorithm. Additional features of the proposed modification are that, unlike the original algorithm, it does not require retraining if new data is added to the training set and can be easily parallelized. The novel method was applied to detect intrinsic plagiarism in a collection of documents. The authors also test the performance of their algorithm using synthetic data and show that the quality of classification is higher with the algorithm than without or with other algorithms for dimension reduction.

Keywords: *data analysis, dimension reduction, nonlinear dimension reduction, manifold learning, intrinsic plagiarism detection.*

References

[1] Fefferman, Charles and Mitter, Sanjoy and Narayanan, Hariharan. 2016. Testing the manifold hypothesis. *Journal of the American Mathematical Society.* 29(4):983–1049.

[2] Maaten, Laurens van der and Hinton, Geoffrey. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research.* 9(Nov):2579–2605.

[3] Narayanan, Hariharan and Mitter, Sanjoy. 2010. Sample complexity of testing the manifold hypothesis. *Advances in Neural Information Processing Systems.* 1786–1794.

*This publication is funded by the Russian Foundation for Basic Research, award number 16-07-01155

[1]Center for Energy Systems, Skolkovo institute of science and technology; Moscow institute of physics and technology; Antiplagiat, i.molybog@skoltech.ru

[2]Moscow institute of physics and technology; Antiplagiat, anastasiya.motrenko@phystech.edu

[3]Dorodnicyn Computing Centre, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, strijov@ccas.ru

[4] Zu Eissen, Sven Meyer and Stein, Benno. 2006. Intrinsic plagiarism detection. *European Conference on Information Retrieval, Springer.* 565–569.

[5] Kuznetsov, Mikhail and Motrenko, Anastasia and Kuznetsova, Rita and Strijov, Vadim. 2016. Methods for intrinsic plagiarism detection and author diarization. *Working Notes Papers of the CLEF.*

[6] Muhr, Markus and Kern, Roman and Zechner, Mario and Granitzer, Michael. 2010. External and intrinsic plagiarism detection using a cross-lingual retrieval and segmentation system. *Notebook Papers of CLEF 2010 LABs and Workshops.*

[7] Stamatatos, Efstathios. 2009. Intrinsic plagiarism detection using character n-gram profiles. *SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 2009).* 38–46.

[8] Kestemont, Mike and Luyckx, Kim and Daelemans, Walter. 2011. Intrinsic plagiarism detection using character trigram distance scores. *Proceedings of the PAN.*

[9] Potthast, Martin and Eiselt, Andreas and Cedeño, Luis Alberto Barrón and Stein, Benno and Rosso, Paolo. 2011. Overview of the 3rd international competition on plagiarism detection. *CEUR Workshop Proceedings.* 1177.

[10] Fodor, Imola K. 2002. A survey of dimension reduction techniques. *Center for Applied Scientific Computing, Lawrence Livermore National Laboratory.* 9:1–18.

[11] Brooke, Julian and Hirst, Graeme. 2012. Paragraph Clustering for Intrinsic Plagiarism Detection using a Stylistic Vector Space Model with Extrinsic Features. *CLEF (Online Working Notes/Labs/Workshop).*

[12] Brooke, Julian and Hammond, Adam and Hirst, Graeme. 2012. Unsupervised stylistic segmentation of poetry with change curves and extrinsic features. *Proceedings of the 1st NAACL-HLT Workshop on Computational Linguistics for Literature. Association for Computational Linguistics, Stroudsburg, PA, USA.* 26–35.

[13] Gorban, Alexander N and Kégl, Balázs and Wunsch, Donald C and Zinovyev, Andrei Y and others. 2008. Principal manifolds for data visualization and dimension reduction. *Springer.* 58.

[14] Tenenbaum, Joshua B and De Silva, Vin and Langford, John C. 2000. A global geometric framework for nonlinear dimensionality reduction. *Science.* 290(5500):2319–2323.

[15] Belkin, Mikhail and Niyogi, Partha. 2001. Laplacian eigenmaps and spectral techniques for embedding and clustering. *NIPS.* 14(14):585–591.

[16] Roweis, Sam T and Saul, Lawrence K. 2000. Nonlinear dimensionality reduction by locally linear embedding. *Science.* 290(5500):2323–2326.

[17] Donoho, David L and Grimes, Carrie. 2003. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences.* 100(10):5591–5596.

[18] Zhang, Zhen-yue and Zha, Hong-yuan. 2004. Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. *Journal of Shanghai University (English Edition), Springer.* 8(4):406–424.

[19] Weinberger, Kilian Q and Saul, Lawrence K. 2006. Unsupervised learning of image manifolds by semidefinite programming. *International Journal of Computer Vision, Springer.* 70(1):77–90.

[20] Chen, Changyou and Zhang, Junping and Fleischer, Rudolf. 2010. Distance approximating dimension reduction of Riemannian manifolds. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics).* 40(1):208–217.

[21] van der Maaten, Laurens. 2009. Learning a parametric embedding by preserving local structure, *RBM.* 500(500):26.

[22] Van Der Maaten, Laurens. 2014. Accelerating t-SNE using tree-based algorithms. *Journal of machine learning research.* 15(1):3221–3245.

[23] Kim, Hyunsoo and Park, Haesun and Zha, Hongyuan. 2007. Distance preserving dimension reduction for manifold learning. *Proceedings of the 2007 SIAM International Conference on Data Mining.* 527–532.

[24] Bottou, Léon. 2012. Stochastic gradient descent tricks. *Neural networks: Tricks of the trade, Springer.* 421–436.

Повышение качества классификации в задаче обнаружения внутреннего плагиата*

И. О. Молибог,[1] А. П. Мотренко,[2] В. В. Стрижов[3]

Аннотация: В работе исследуется задача классификации объектов в многомерных пространствах. Для снижения размерности задачи предлагается модификация алгоритма t-SNE, в которой при обучении используется информация о разметке, не возникает необходимости заново обучать алгоритм при добавлении новых данных, а также предусмотрена параллельная реализация. Предлагаемый алгоритм решает задачу внутреннего плагиата, в которой признаками являются частотные словесные профили сегментов текста. Показано, что качество классификации после применения алгоритма выше, чем без него или с другими алгоритмами.

Ключевые слова: анализ данных; снижение размерности; нелинейные методы снижения размерности; обучение многообразий; обнаружение внутреннего плагиата

1 Введение

В работе рассматривается задача классификации объектов в пространствах большой размерности, 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 [1]. Данные ряда практических задач, включая задачи анализа генома, анализа текста и распознавания изображений, не противоречат этой гипотезе [2]. В [3] было дано ее формальное определение и перечислены идеи

*Работа выполнена при финансовой поддержке РФФИ (проект 16-07-01155).

[1]Центр энергетических систем, Сколковский институт науки и технологий; Московский физико-технический институт; ЗАО Анти-Плагиат, i.molybog@skoltech.ru

[2]Московский физико-технический институт; ЗАО Анти-Плагиат, anastasiya.motrenko@phystech.edu

[3]Вычислительный центр им. А. А. Дородницына Федерального исследовательского центра «Информатика и управление» Российской академии наук, strijov@phystech.edu

методов, которыми ее можно проверить. Практической задачей, рассматриваемой в данной работе, является задача обнаружения внутреннего плагиата [4, 5].

Задача обнаружения внутреннего плагиата состоит в поиске заимствованных частей документа без использования внешних источников. При решении задачи исследуемый текст некоторым образом разбивается на сегменты. Каждому сегменту соответствует его вектор признаков. Сегмент считается минимальной единицей заимствования. Он считает либо полностью заимствованным, либо полностью оригинальным. Тогда задача обнаружения внутреннего плагиата является задачей классификации, где объектами являются векторы признаков сегментов, а классами — метки заимствования или оригинальности.

Способы разбиения на сегменты, как и способы вычисления вектора признаков, являются предметом отдельного исследования. Подходы [4, 6–8], продемонстрировали на конкурсе PAN-2011 [9] наилучшее качество решения задачи обнаружения внутреннего плагиата. Они включают разбиение документа на абзацы, предложения, блоки слов или символов. В них используются признаки, основанные на частотных профилях сегментов. Такие признаки имеют размерность, пропорциональную числу слов в документе, сильно разрежены и не всегда информативны.

В данной работе предполагается, что объекты с таким признаковым описанием подчиняются гипотезе многообразия. Это означает, что метрически близкие объекты могут быть геодезически далекими, и дает возможность применить методы снижения размерности для улучшения качества классификации.

В задаче понижения размерности требуется построить гладкое отображение множества $\mathbf{X}$ в пространстве исходных данных в некоторое множество $\mathbf{Z}$ в пространстве меньшей размерности. Будем называть элементы $\mathbf{Z}$ образами элементов $\mathbf{X}$. Пространство образов будем называть результирующим. В конкретных алгоритмах на это отображение накладывают необходимые ограничения, исходя из специфики задачи [10]. Приведем некоторые из них.

Для снижения размерности широко применяются линейные методы, основанные на анализе дисперсии: латентно-семантический анализ [11, 12], анализ главных компонент [13]. Однако они могут не сохранять кластерную структуру исходных данных, и потому не применимы для решения задач вложений из нелинейных многообразий.

Для выполнения вложений из нелинейных многообразий были разработаны алгоритмы, использующие изометрические отображения. Алгоритмы ISOMAP [14] и Laplacian Eigenmap [15] приближают геодезическое расстояние с помощью графа k ближайших соседей. Алгоритмы Local Linear Embedding [16] и Hessian-based Locally Linear Embedding [17] основаны на предположении, что многообразие аппроксимируется кусочно-линейной функцией. Для каждого объекта исходного пространства строится его линейное приближенное описание через соседние объекты, после чего по этим описаниям строятся образы в результирующем пространстве. Метод [17] использует для описания объектов специальную квадратичную форму, что гарантирует асимптотическую оптимальность метода даже в случае невыпуклых множеств.

Алгоритм Local Tangent Space Alignment Algorithm [18] также использует кусочно-линейную аппроксимацию. Многообразие приближается гиперплоскостью в окрестно-

сти каждой точки, после чего полученные приближения сглаживаются между собой. При помощи Semidefinite Embedding [19] можно получить вложение, в котором сохранены точные расстояния между ближайшими объектами. Для этого метод максимизирует след матрицы Грама для образов при ограничениях, накладываемых отношением соседства объектов исходного пространства и их матрицей Грама.

Все перечисленные методы нацелены на наиболее точное сохранение расстояний между объектами при снижении размерности. Это может привести к неустойчивости решения, связанной с тем, что изменения расстояния между далекими и близкими объектами штрафуются одинаково. Кроме того, они не приспособлены для решения задачи классификации, поскольку не учитывают разметку при выполнении вложения, хотя существуют их модификации, обладающие этим свойством. В [20] метод аппроксимации расстояний, используемый в ISOMAP, модифицирован в методе оптимизации целевого функционала. Полученный метод получил название TRIMAP. В нем при обучении используется разметка обучающей выборки.

В данной работе применяется метод t-NSE (англ. t-distributed Stochastic Neighbor Embedding) [2]. Выгодной особенностью метода t-SNE является склонность к локализации изолированных плотных пространственных структур произвольной геометрии. Под изолированной плотной структурой подразумевается множество точек, имеющих близких соседей из той же структуры, но сравнительно удаленных от всех точек не из нее. Такой эффект достигается тем, что близким и далеким объектам назначаются разные приоритеты.

Недостатком метода t-SNE в отношении задачи классификации является то, что в нем не предусмотрено функции вложения объектов, не участвовавших в построении уже существующего вложения. В работе [21] описана параметрическая модификация t-SNE, которая частично избавлена от этой особенности, однако в данной работе она не использовалась.

Дополнительным ограничением применимости метода t-SNE является высокая по сравнению с другими методами вложений вычислительная сложность. Хотя в [22] предлагается два способа вычисления градиента, при использовании которых сложность непараметрического t-SNE составляет $O(k \cdot m \log(m))$, где m — размер выборки, а k — размерность результирующего пространства, этого ускорения недостаточно для обеспечения комфортной работы даже с выборками длиной порядка 10^3.

Основным вкладом данной статьи в теорию распознавания образов является предложенная модификация метода t-SNE, позволяющая строить классификаторы в результирующем пространстве. Преимуществом предлагаемого метода является то, что он расширяет границы применимости оригинального метода t-SNE. Разработанная модификация предусматривает вложение тестовых данных без повторного вложения обучающих, а также может учитывать разметку обучающих данных и имеет параллельную реализацию.

2 Постановка задачи

Обозначим $\mathbb{X} \subset \mathbb{R}^n$ множество всех возможных векторов $\mathbf{x}$ признаков изучаемых объектов. Предполагается, что объекты $\mathbb{X}$ подчиняются *гипотезе многообразия*: найдется гладкое отображение $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^n$ такое, что

$$\text{для } \mathbf{x} \in \mathbf{X} \text{ существует } \mathbf{z}^* \in \mathbb{R}^d : \mathbf{x} = \mathbf{f}(\mathbf{z}^*) + \boldsymbol{\varepsilon},$$

где $\boldsymbol{\varepsilon}$ — случайный вектор с нулевым матожиданием и конечной матрицей корреляций. Будем называть d эффективной размерностью исходного пространства $\mathbb{X}$. Она определяется природой признакового пространства. Поскольку d заранее не известно, 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, заданную матрицей

$$\mathbf{X} = [\mathbf{x}_1 \ \dots \ \mathbf{x}_m]^\mathsf{T}, \quad \mathbf{x}_i \in \mathbb{X}, \quad i = 1, \dots, m. \tag{1}$$

Пусть $p_{ij} = P(\mathbf{x}_i, \mathbf{x}_j)$ и $q_{ij} = Q(\mathbf{z}_i, \mathbf{z}_j)$ — расстояния между объектами в $\mathbb{R}^n$ и $\mathbb{R}^k$ соответственно:

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

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

Параметр σ_i в условном распределении p_{ij} задан для каждого i и зависит от расположения $\mathbf{x}_i$ относительно других объектов в исходном пространстве. Если он расположен в области высокой концентрации исходных данных, то коэффициент σ_i имеет меньшие значения, чем если бы концентрация была низкой.

Расположение

$$\mathbf{Z} = [\mathbf{z}_1 \ \dots \ \mathbf{z}_m]^\mathsf{T} \subset \mathbb{R}^k \tag{2}$$

как образов $\mathbf{X}$ в результирующем пространстве $\mathbb{R}^k$ находится путем минимизации дивергенции Кульбака-Лейблера

$$\mathbf{Z}_{\min} = \operatorname*{argmin}_{\mathbf{Z} \in \mathbb{R}^{m \times k}} C(\mathbf{X}, \mathbf{Z}), \tag{3}$$

$$C(\mathbf{X}, \mathbf{Z}) = \mathrm{KL}(P||Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}. \tag{4}$$

Заметим, что минимизация происходит только по координатам объектов $\mathbf{z}_1, \ \dots, \ \mathbf{z}_m$ как по переменным, а координаты $\mathbf{x}_1, \ \dots, \ \mathbf{x}_m$ считаются известными константами.

Задача решается градиентными методами [2]. Для инициализации начальных точек $\mathbf{Z}^{(0)} = [\mathbf{z}_1^{(0)} \ \ldots \ \mathbf{z}_m^{(0)}]^\mathsf{T}$ градиентного спуска в стандартной реализации было предложено [2] два метода: инициализировать случайными точками, либо использовать для задания начальной инициализации метод Principal Components Analysis. От качества начальной инициализации, в случае с невыпуклой задачей оптимизации, зависит не только скорость сходимости к оптимуму, но и локальный минимум, к которому будет сходиться градиентный метод.

3 Предлагаемая модификация t-SNE

Рассмотрим задачу классификации с обучающей выборкой $\mathbf{X}$ (1) и тестовой выборкой из m' объектов $\mathbf{X}' = [\mathbf{x}_{m+1} \ \ldots \ \mathbf{x}_{m+m'}]^\mathsf{T} \subset \mathbb{X}$. Соответственно, метки классов $y_i \in \{0, 1\}, \quad i = 1, \ldots, m$, известны, а $\hat{y}_i, \quad i = m+1, \ldots, m+m'$, необходимо оценить. Так как на этапе обучения данные $\mathbf{X}'$ могут быть недоступны, метод непараметрического t-SNE не применим для снижения размерности в задачах классификации. Назовем это проблемой непросмотренных объектов (out-of-sample problem). Для ее решения предлагается минимизировать (4) независимо по различным подмножествам объектов.

Для повышения качества классификатора в результирующем пространстве предлагается перед вложением обучающей выборки добавить в ней метки классов в качестве признаков и улучшить таким образом начальное приближение градиентного метода. Идея такого подхода заключается в том, что, поскольку t-SNE сохраняет только локальную структуру схожести между объектами, после проведения процедуры понижения размерности классифицируемые объекты отображаются в кластеры, предварительно разнесенные учетом меток. При этом используется предположение, что объекты из $\mathbf{X}'$ больше схожи с объектами $\mathbf{X}$ того же класса, чем с объектами противоположного. Таким образом удается увеличить расстояние между образами классифицируемых объектов из различных классов, что упрощает их классификацию. На диаграммах (5) и (6) показаны основные отображения оригинального непараметрического t-SNE

$$\mathbf{X} \in \mathbb{R}^{m \times n} \longrightarrow \mathbf{Z} \in \mathbb{R}^{m \times k} \tag{5}$$

$$\mathbf{X}' \in \mathbb{R}^{m' \times n} \longrightarrow \mathbf{Z}' \in \mathbb{R}^{m' \times k}$$

и предложенной модификации

$$\begin{array}{ccccc} \mathbf{X}|\mu\mathbf{y} \in \mathbb{R}^{m \times (n+1)} & \xrightarrow[\text{Дополнительные партии}]{\text{Начальная партия}} & \mathbf{Z} \in \mathbb{R}^{m \times k} & & \\ & \searrow & & \swarrow & \\ \mathbf{X}' \in \mathbb{R}^{m' \times n} & \longrightarrow & \mathbf{Z}'^{(0)} \in \mathbb{R}^{m' \times k} & \longrightarrow & \mathbf{Z}' \in \mathbb{R}^{m' \times k} \end{array} \tag{6}$$

Использование исходной разметки выборки при вложении для обучения классификатора. Для учета разметки обучающей выборки признаковая матрица $\mathbf{X}$ расширяется дополнительным столбцом признаков

$$\tilde{\mathbf{X}} = \left(\begin{array}{c|c} \mathbf{X} & \mu\mathbf{y} \end{array} \right),$$

где μ — вес меток как признаков. В модифицированном алгоритме на основе расширенной матрицы $\tilde{X}$ выполняется поиск образов $\mathbf{Z}$ (4), на которых обучается классификатор. Таким образом, при построении вложения обучающей выборки решается задача

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

Вложение новых объектов в пространство со сниженной размерностью для классификации. Обозначим через $\mathbf{Z}' = [\mathbf{z}_{m+1} \ \ldots \ \mathbf{z}_{m+m'}]^\mathsf{T}$ образы $\mathbf{X}'$ в результирующем пространстве. Аналогично (3), сформулируем задачу поиска $\mathbf{Z}'$ в виде m' задач k-мерной минимизации, которые могут быть решены независимо:

$$\mathbf{z}_i^{\min} = \underset{\mathbf{z}_i\in\mathbb{R}^{m'}}{\operatorname{argmin}} C\left(\left[\frac{\mathbf{X}}{\mathbf{x}_i^\mathsf{T}}\right], \left[\frac{\mathbf{Z}}{\mathbf{z}_i^\mathsf{T}}\right]\right), \qquad i = m+1, \ldots, m+m',$$

где матрицы $\left[\frac{\mathbf{X}}{\mathbf{x}_i^\mathsf{T}}\right]$ и $\left[\frac{\mathbf{Z}}{\mathbf{z}_i^\mathsf{T}}\right]$ получены из $\mathbf{X}$ и $\mathbf{Z}$ добавлением строк $\mathbf{x}_i^\mathsf{T}$ и $\mathbf{z}_i^\mathsf{T}$ соответственно. При использовании такого подхода предполагается, что обучающая выборка $\mathbf{X}$ (1) достаточно репрезентативна.

Для инициализации образов $\mathbf{z}_{i'}$ классифицируемых объектов предлагается использовать метод взвешенного среднего по образам соседей:

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

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

$$w_{ii'}^{\text{softmax}} = \frac{\exp(-\|\mathbf{x}_i - \mathbf{x}_{i'}\|)}{\sum\limits_{k=1}^{m} \exp(-\|\mathbf{x}_k - \mathbf{x}_{i'}\|)} \text{ или} \tag{7}$$

$$w_{ii'}^{\text{stud}} = \frac{(1 + \|\mathbf{x}_k - \mathbf{x}_{i'}\|^2)^{-1}}{\sum\limits_{k=1}^{m} (1 + \|\mathbf{x}_k - \mathbf{x}_{i'}\|^2)^{-1}}. \tag{8}$$

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

Псевдокод предложенного метода приведен в алгоритмах 1 и 2.

4 Вычислительный эксперимент

Вычислительный эксперимент состоит из двух частей: исследование разработанного алгоритма на синтетических данных и применение разработанного алгоритма для решения задачи внутреннего плагиата.

Для инициализации вложения тестовых данных использовались четыре различных подхода: случайный — инициализация случайным образом, PCA — инициализация образами при снижении размерности методом главных компонент, Softmax и Student — задаваемые по формулам (7) и (8). На рис. 1 — 4, 6, 7 представлены результаты экспериментов, проведенных при использовании всех этих способов инициализации.

Все инициализированные таким образом объекты далее преобразуются, минимизируя (4) при фиксированных образах объектов обучающей выборки. После этого происходит классификация полученных образов.

Методы инициализации (7) и (8) были предложены так, чтоб инициализированные данные обладали свойством сохранения локальной структуры исходной выборки. Предполагалось, что это улучшит сходимость градиентного метода, используемого для минимизации (4), по сравнению с инициализациями PCA и random.

Алгоритм 1: Вложение выборки с известным вектором ответов классификации $\mathbf{y}$

Data: $\mathbf{X}, \mathbf{y}, \mu, S_s, S_b$
Result: $\mathbf{Z}$

$\tilde{\mathbf{X}} = (\mathbf{X}|\mu\mathbf{y})$
Инициализировать $\mathbf{Z}$ (2) случайно или при помощи PCA($\tilde{\mathbf{X}}$). Положить инициализацию начальной точкой градиентного метода: $\mathbf{Z}^{(0)}$.
if $m > S_s$ **then**
    Разбить $\mathbf{Z}$ на партии: начальная партия $\mathbf{Z}_0$ размером S_s и $B = \left\lceil \frac{m - S_s}{S_b} \right\rceil$ дополнительных партий $\mathbf{Z}_1, \ldots, \mathbf{Z}_B$ размером не больше чем S_b каждая.
    Оптимизировать (4) по $\mathbf{Z}_0$, зафиксировав координаты остальных объектов из $\mathbf{Z}$, известные из предыдущего шага.
    for $\mathbf{Z}_i \in \{\mathbf{Z}_1, \ldots, \mathbf{Z}_B\}$ **do**
        Оптимизировать (4) по $\mathbf{Z}_i$, зафиксировав координаты остальных объектов из $\mathbf{Z}$, известные из предыдущего шага.
    end
else
    Оптимизировать (4)
end

Алгоритм 2: Вложение выборки без известного вектора ответов классификации

Data: $\left[\frac{\mathbf{X}}{\mathbf{X}'}\right], \mathbf{Z}$

Result: $\mathbf{Z}'$

1 Инициализировать $\mathbf{Z}$ (2) случайно, при помощи PCA($\tilde{\mathbf{X}}$), либо используя (8) или (7) для расчета $\mathbf{W}$ и считать $\mathbf{Z}'^{(0)} = \mathbf{Z}^\mathsf{T}\mathbf{W}$.

2 **for** $i \in \{m+1, \ldots, m+m'\}$ **do**

3 Оптимизировать (4) по $\mathbf{z}_i$, зафиксировав координаты остальных объектов из $\left[\frac{\mathbf{Z}}{\mathbf{Z}'}\right]$, известные из предыдущего шага.

4 **end**

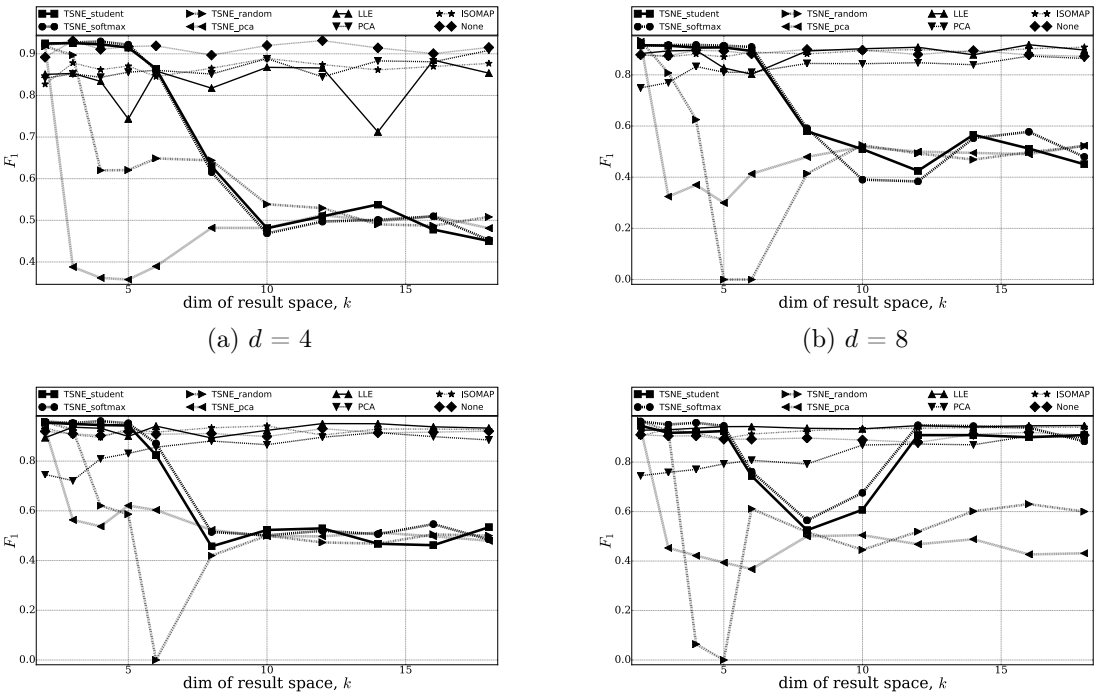
4.1 Исследование свойств алгоритма на синтетических данных

В данном разделе для эмпирического исследования свойств предлагаемого алгоритма использовались синтетические выборки $\mathbf{X} = [\mathbf{x}_1 \ \ldots \ \mathbf{x}_m]^\mathsf{T}$. Для любого вектора $\mathbf{x}_i$ компоненты были сгенерированы как стандартные нормальные распределения на гранях гиперкуба. При этом эффективная размерность выборки составляла d, а оставшиеся признаки были шумовыми. Далее выборка сворачивалась в спираль по одной из размерностей. Это делалось для того, чтобы реализовать предположение о существовании многообразия меньшей размерности, в котором содержится выборка. Генерировалось одинаковое количество объектов разных классов, а на обучение и контроль выборка разбивалась в соотношении $1:4$.

В этом разделе описывается исследование качества классификации с применением предлагаемого алгоритма в зависимости от основных его параметров и специфики выборки. Для сравнения предлагаемого алгоритма и его исследования рассматривается классификация в комбинации с другими методами снижения размерности: Principal Component Analysis (PCA) [23], Local Linear Embedding (LLE) [16], Isometric Mapping (ISOMAP) [14], а также без применения снижения размерности. Для построения классификатора использовался метод логистической регрессии на основе Stochastic Gradient Descent (SGD) [24].

На рис. 1 изображена зависимость меры качества F_1 от размерности вложения k при различных значениях эффективной размерности d. Пунктиром отмечено стандартное отклонение, срезанное по уровню единицы. На графиках видно, что качество значительно ухудшается при увеличении k независимо от соотношения k и d. Эксперимент проведен при постоянных $m = 500, n = 20, S_b = 100, S_s = 400, \mu = 150$.

На рис. 2 изображена зависимость F_1 от веса μ меток класса y в стартовой выборке при различных значениях размерности выборки n. Из них можно сделать вывод, что качество классификации повышается с ростом μ, при этом скорость роста падает с ростом n. Также видно, что разработанный метод при достаточно больших значениях μ показывает в среднем лучшие результаты среди всех рассмотренных методов сниже-



(a) $d = 4$ (b) $d = 8$

(c) $d = 12$ (d) $d = 16$

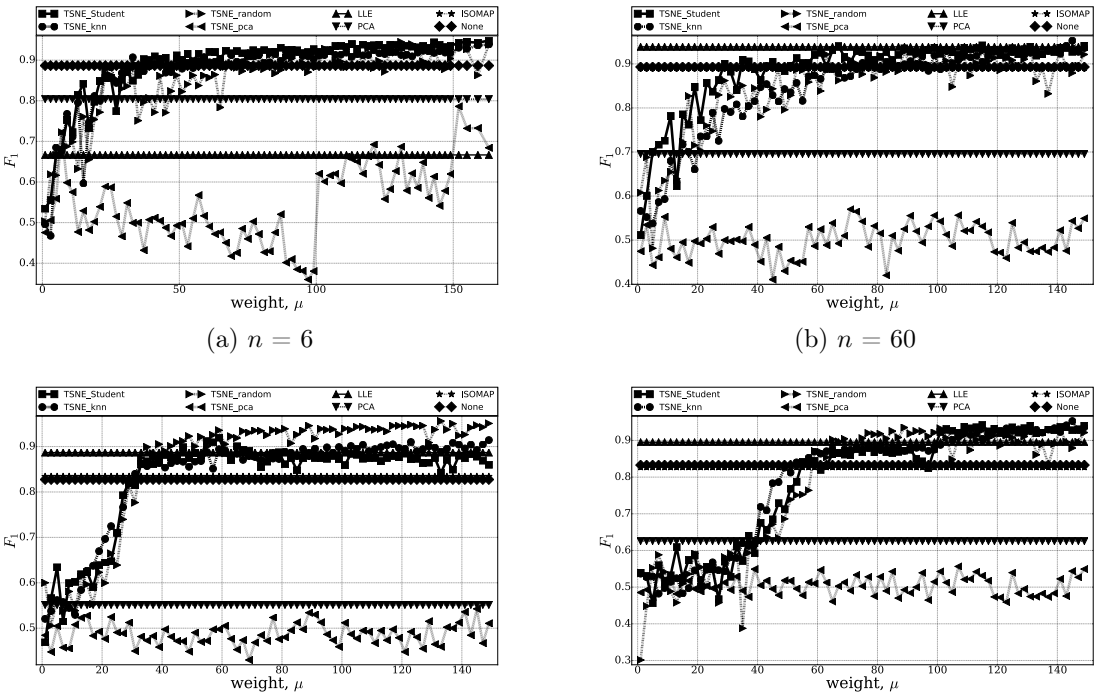
Рис. 1: Зависимость F_1 от k при различных эффективных размерностях выборки d

ния размерности, а также превосходит по качеству классификацию в исходном пространстве. Эксперимент проведен при постоянных $m = 500, k = 3, S_b = 100, S_s = 400$. В этом эксперименте все исходные признаки были информативными.

Для исследования зависимости качества классификации от величины отношения размера стартовой части к размеру выборки S_s/m был поставлен эксперимент, где при постоянных $n = 6, k = d = 3, \mu = 150$ исследовалась зависимость меры качества F_1 от размера выборки m и размера начального вложения S_s. При этом размер дополнительно вкладываемых блоков S_b принимался заведомо бóльшим размера выборки m, так что дополняющая часть не разбивалась на блоки. На рис. 3 выведены результаты. Можно видеть, что 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 снижения качества.

На графиках на рис. 3 также видно, что методы инициализации с помощью (8) и случайной инициализации дают лучшие результаты, в то время как метод инициализации PCA показал результаты порядка $0,5$, по причине чего было принято решение не выносить его на рисунок.

Целью эксперимента, результаты которого приведены на рис. 4, было исследование



(a) $n = 6$ (b) $n = 60$

(c) $n = 300$ (d) $n = 600$

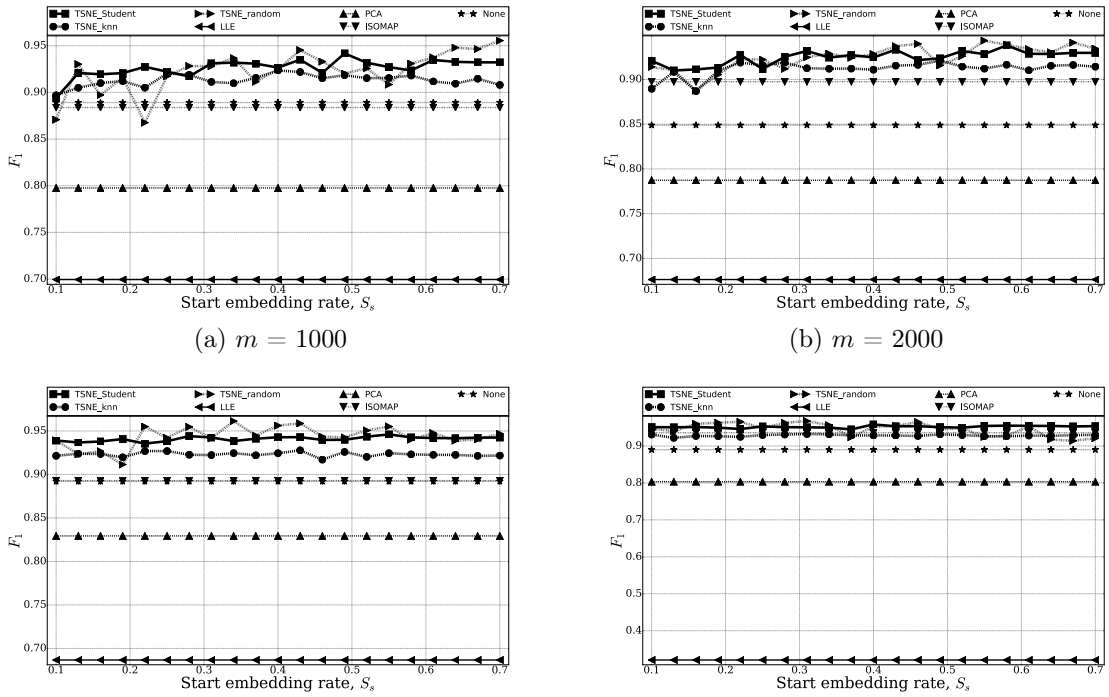
Рис. 2: Зависимость F_1 от μ при различных размерностях выборки n

зависимости значения функции качества классификации от S_b. Он был проведен при постоянных $n = 6, k = d = 3, \mu = 150, S_s = 200$. В результате было обнаружено, что предлагаемый метод устойчив относительно параметра S_b.

4.2 Задача обнаружения внутреннего плагиата

Целью данной части эксперимента был анализ предложенного метода снижения размерности в применении к реальным данным задачи внутреннего плагиата. Рассматривается набор документов. Каждый документ рассматривается как последовательность сегментов s_i, каждый из которых описывается вектором признаков $\mathbf{x}$. В данной работе в качестве сегментов рассматриваются предложения. Каждому s_i поставлена в соответствие метка класса $y_i \in \{0, 1\} : y_i = 1$, если s_i — заимствованный сегмент, иначе $y_i = 0$. Задача распознавания внутреннего плагиата ставится как задача восстановления меток y_i по документу.

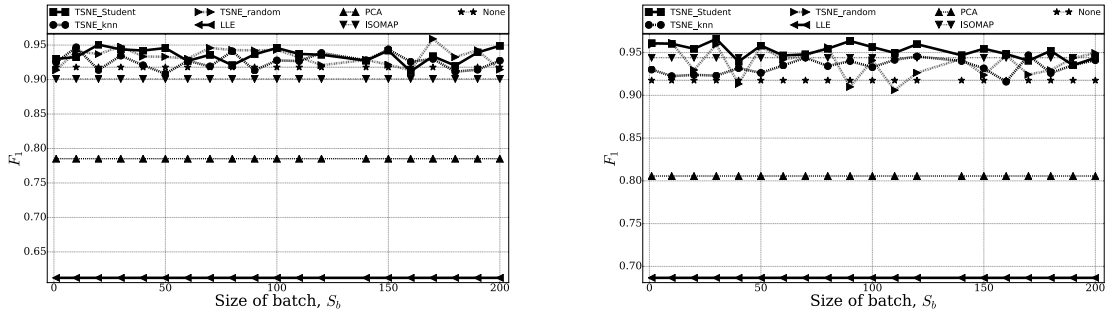
Иллюстрация вложения реальных данных. Для демонстрации работы алгоритма на реальных данных из предоставленного корпуса [25] `part1` выделен один из документов. Выделенные из него объекты были разделены на обучающую и тестовую выборки. Каждой из них соответствуют непрерывные части текста. Это разделение



(a) $m = 1000$ (b) $m = 2000$

(c) $m = 3000$ (d) $m = 4000$

Рис. 3: Зависимость F_1 от $\frac{S_s}{m}$ при различных размерах выборки m



(a) $m = 500$ (b) $m = 1000$

Рис. 4: Зависимость F_1 от S_s при различных размерах выборки m

необходимо для демонстрации работы предложенной модификации и не учитывается при применении оригинального t-SNE. На рис. 5 приведен результат применения оригинального непараметрического метода t-SNE к объектам, выделенным из выбранного документа. На нем видно, что объекты, соответствующие заимствованным частям текста, имеют очаги концентрации в исходном пространстве, что свидетельствует об информативности выбранных признаков.

На рис. 6, 7 представлены результаты вложения данных выбранного документа

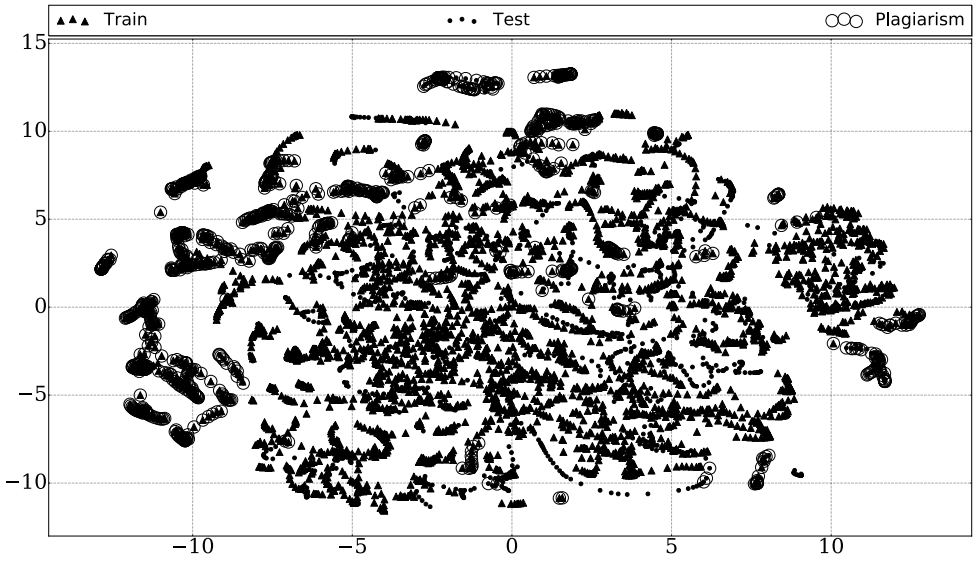
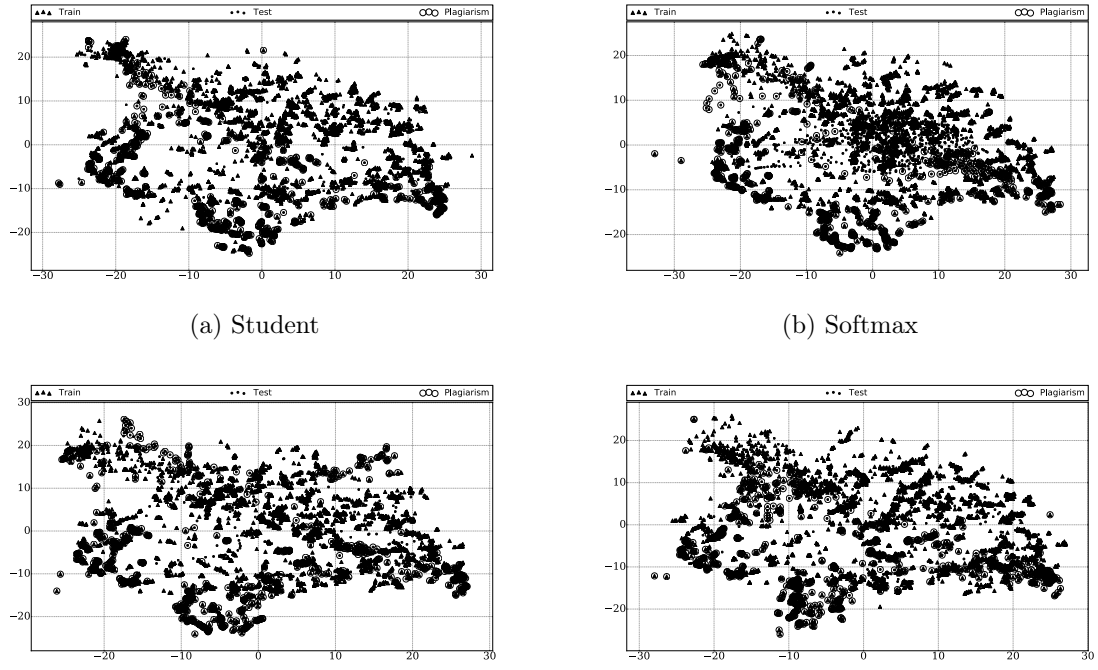


Рис. 5: Визуализация документа с использованием оригинального алгоритма t-SNE

с использованием предложенного алгоритма при различных методах начальной инициализации и при различных значениях веса μ. При выполнении вложений были зафиксированы параметры $S_s = 500$ и $S_b = 200$. В эксперименте данные из выбранного документа были разделены на обучающую и тестовую выборки. В тестовую часть попали образы предложений, которые образовывали в исходном тексте непрерывную цепочку. Обучающая часть вкладывалась с учетом ее разметки, а тестовая — без учета.

Результаты. Из полученных графиков можно сделать вывод, что предложенная модификация при больших значениях веса μ принимает на себя часть ответственности за классификацию. Она склонна разделять и кластеризовать тестовую выборку по целевому признаку. Таким образом, исходя из описанных выше свойств t-SNE, любой построенный в результирующем пространстве классификатор получает свойство классификатора ближайших соседей с адаптивной константой, подстраиваемой под локальную геометрию выборки. Следует отметить также, что при больших значениях μ минимизация целевой функции (4) требует больше шагов градиентного алгоритма. Таким образом, этот параметр следует выбирать с оглядкой на время работы программы. Авторы рекомендуют значение порядка характерной величины координат векторов обучающей выборки.



(a) Student

(b) Softmax

(c) PCA

(d) Random

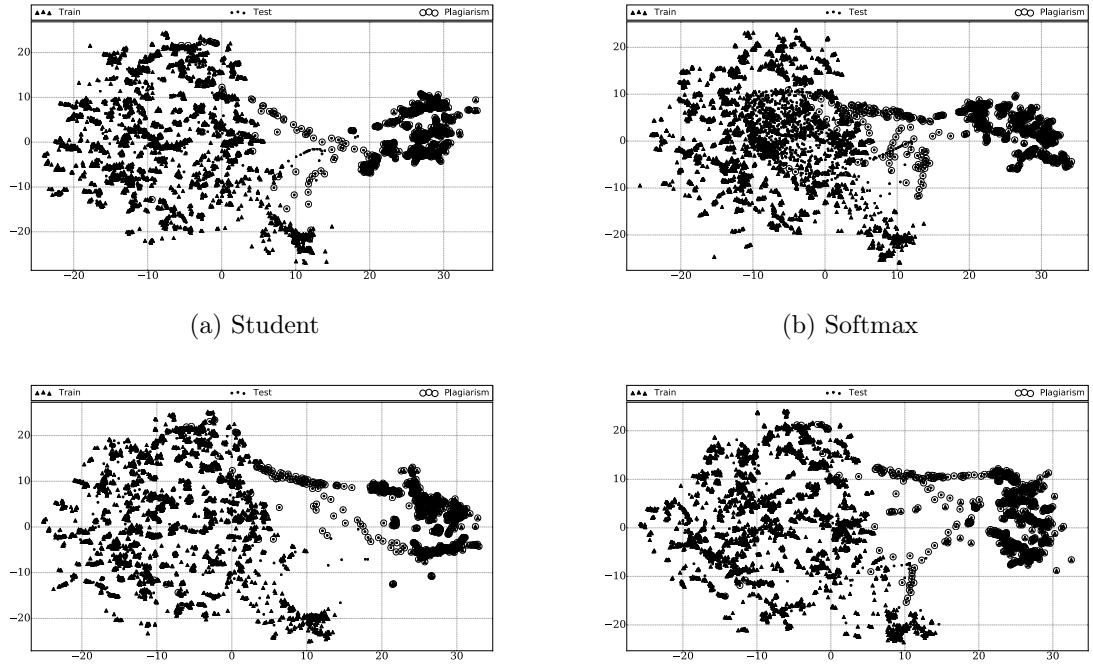
Рис. 6: Демонстрация вложения, выполненного предлагаемым методом при параметре $\mu = 0$

5 Заключение

В работе была предложена модификация непараметрического метода снижения размерности t-SNE, состоящая в воплощении возможности выполнения вложения поэтапно, решении проблемы непросмотренных объектов и внедрении возможности учета разметки при выполнении вложения 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. Полученные значения качества сравнивались с результатами классификации с применением других методов снижения размерности, а также без их применения.

Была показана устойчивость алгоритма к введенным параметрам размера начальной части S_s и максимального размера блоков S_b, что облегчает его использование на практике. Также явно продемонстрирована зависимость свойств метода от параметра веса разметки выборки μ.

Проанализировано признаковое пространство задачи внутреннего плагиата. Проиллюстрированы свойства предложенного алгоритма относительно данных задачи



(a) Student (b) Softmax

(c) PCA (d) Random

Рис. 7: Демонстрация вложения, выполненного предлагаемым методом при параметре $\mu = 10$

внутреннего плагиата. Продемонстрирована эффективность предложенных методов инициализации при вложении образов объектов, которые не были использованы при выполнении начального вложения.

Список литературы

[1] *Fefferman C., Mitter S., Narayanan H.* Testing the manifold hypothesis // Journal of the American Mathematical Society, 2016. Vol. 29 No. 4 P. 983–1049.

[2] *van der Maaten L., Hinton G.* Visualizing data using t-SNE // Journal of Machine Learning Research, 2008. Vol. 9 No. Nov P. 2579–2605.

[3] *Narayanan H., Mitter S.* Sample complexity of testing the manifold hypothesis // Advances in Neural Information Processing Systems, 2010. P. 1786–1794.

[4] *Zu Eissen S. M., Stein B.* Intrinsic plagiarism detection // European Conference on Information Retrieval, 2006. P. 565–569.

[5] *Kuznetsov M. P., Motrenko A. P., Kuznetsova M. V., Strijov V. V.* Methods for intrinsic plagiarism detection and author diarization // Working Notes Papers of the CLEF, 2016. P. 912–919.

[6] *Muhr M., Kern R., Zechner M., Granitzer M.* External and intrinsic plagiarism detection using a cross-lingual retrieval and segmentation system // Notebook Papers of CLEF 2010 LABs and Workshops, 2010.

[7] *Stamatatos E.* Intrinsic plagiarism detection using character n-gram profiles // SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse, 2009. P. 38–46.

[8] *Kestemont M., Luyckx K., Daelemans W.* Intrinsic plagiarism detection using character trigram distance scores // Notebook Papers of CLEF 2011 Labs and Workshops, 2011.

[9] *Potthast M., Eiselt A., Cedeño L. A., Stein B., Rosso P.* Overview of the 3rd international competition on plagiarism detection // CEUR Workshop Proceedings, 2011. P. 1177.

[10] *Fodor I. K.* A survey of dimension reduction techniques // Center for Applied Scientific Computing, Lawrence Livermore National Laboratory, 2002. Vol. 9 P. 1–18.

[11] *Brooke J., Hirst G.* Paragraph Clustering for Intrinsic Plagiarism Detection using a Stylistic Vector Space Model with Extrinsic Features // CLEF 2012 Evaluation Labs and Workshop – Working Notes Papers, 2012.

[12] *Brooke J., Hammond A., Hirst G.* Unsupervised stylistic segmentation of poetry with change curves and extrinsic features // Proceedings of the 1st NAACL-HLT Workshop on Computational Linguistics for Literature, 2012. P. 26–35.

[13] *Gorban A. N., Kégl B., Wunsch D. C., Zinovyev A. Y., et al* Principal manifolds for data visualization and dimension reduction – Springer, 2008. 58 p.

[14] *Tenenbaum J. B., De Silva V., Langford J. C.* A global geometric framework for nonlinear dimensionality reduction // Science, 2000. Vol. 290. No. 5500. P. 2319–2323.

[15] *Belkin M., Niyogi P.* Laplacian eigenmaps and spectral techniques for embedding and clustering // Proceedings of the 14th International Conference on Neural Information Processing Systems, 2001. Vol. 14. No. 14 P. 585–591.

[16] *Roweis S. T., Saul L. K.* Nonlinear dimensionality reduction by locally linear embedding // Science, 2000. Vol. 290. Vol. 5500. P. 2323–2326.

[17] *Donoho D. L., Grimes C.* Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data // Proceedings of the National Academy of Sciences, 2003. Vol. 100. No. 10 P. 5591–5596.

[18] *Zhang Z., Zha H.* Principal manifolds and nonlinear dimensionality reduction via tangent space alignment // Journal of Shanghai University (English Edition), 2004. Vol. 8. No. 4. P. 406–424.

[19] *Weinberger K. Q., Saul L. K.* Unsupervised learning of image manifolds by semidefinite programming // International Journal of Computer Vision, 2006. Vol. 70. No. 1. P. 77–90.

[20] *Chen C., Zhang J., Fleischer R.* Distance approximating dimension reduction of Riemannian manifolds // IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2010. Vol. 40. No. 1. P. 208–217.

[21] *van der Maaten L.* Learning a parametric embedding by preserving local structure // RBM, 2009. Vol. 500. No. 500 P. 26.

[22] *van der Maaten L.* Accelerating t-SNE using tree-based algorithms // Journal of machine learning research, 2014. Vol. 15 No. 1. P. 3221–3245.

[23] *Kim H., Park H., Zha H.* Distance preserving dimension reduction for manifold learning // Proceedings of the 2007 SIAM International Conference on Data Mining, 2007. P. 527–532.

[24] *Bottou L.* Stochastic gradient descent tricks // Neural networks: Tricks of the trade, 2012. P. 421–436.

[25] *Potthast M., Stein B., Barrón-Cedeño A., Rosso P.* An evaluation framework for plagiarism detection // Proceedings of the 23rd International Conference on Computational Linguistics: Posters, 2010. P. 997–1005.