Постановка задачи. Рассматривается документ d —последовательность символов $c_1, \ldots, c_{L_d}$. В документе возможно наличие чужеродных блоков. Требуется каждому символу c_l документа сопоставить метку класса

$$t_l = \begin{cases} 1, & \text{если символ принадлежит чужеродному блоку,} \\ 0, & \text{иначе,} \end{cases} \qquad l = 1, \ldots, L_d.$$

Стандартные подходы к решению задачи включают разбиение документа на сегменты s_i, $i = 1, \ldots, m$ (абзацы, предложения, блоки слов или символов), профилирование сегментов — выделение признаков $\mathbf{x}_i \in \mathbf{R}^n$, и выделение аномальных сегментов. На этом этапе построенное признаковое описание $\mathbf{x}_i$ используется для сравнения сегментов s_i и выделения сегментов, принадлежащих чужеродным блокам. Здесь используются методы классификации либо обнаружения выбросов.

Примеры признаков:

- вектор `w\_freq\_vec` частот слов w встречающихся в сегменте s_i. Размерность вектора равна числу уникальных слов в документе;
- вектор `ngr\_freq\_vec` частот 4-грам, встречающихся в сегменте;
- среднее число слов в предложении, среднее число предложений с сегменте, частота употреблений частей речи, знаков пунктуации, местоимений, стоп слов, доля различных типов символов в сегменте;
- признаки, связанные с читаемостью сегмента;
- правдоподобие сегмента, вычисляемое на основе марковской модели последовательности слов в тексте.

Для многомерных признаков выполняется нормализация, затем признаки объединяются в один вектор, для которого еще раз выполняется нормализация.

Так как построенное таким образом признаковое описание высокой размерности сильно разрежено и содержит много шума, предлагается перед обучением использовать методы снижения размерности. Также изучаются способы снижения размерности [1, 2]. Автры рассмотрели метод LSA.

Предлагается рассмотреть методы, сохраняющие метрику. На рисунках 1, 2, 3 сравниваются методы tNSE и PCA.

Преимущество метода tSNE [4] заключается в том, что данный метод сохраняет метрику. Недостаток в том, что, в отличие от PCA метод невоспроизводим, то есть его необходимо обучать заново для каждого нового документа. Существуют также модификации PCA, сохраняющие метрику [6]. Преимущество PCA – меньшая вычислительная сложность: $O(n^2m + n^3)$ по сравнению с $O(m^2n)$ tNSE, где m – число точек. В [5] предлагается два способа вычисления градиента, при использовании которого сложность по m опускается с квадратичной до $O(m\log(m))$.

Предлагаемый план действий. Предлагается модифицировать алгоритм, добавив возможность использовать информацию о разметке выборки и добавлять новые точки в выборку без пересчитывания всех координат.

Базовый алгоритм: для каждого объекта $\mathbf{x}_i \in \mathbb{R}^n$ выборки $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$, оценивается распределение вероятностей $p_{j|i}$, что $\mathbf{x}_j$ — ближайший сосед $\mathbf{x}_i$:

$$p_{j|i} = \frac{\exp(-||\mathbf{x}_i - \mathbf{x}_j||^2/\sigma^2)}{\sum_{k\neq i}\exp(-||\mathbf{x}_i - \mathbf{x}_k||^2/\sigma^2)}.$$

Совместное распределение вероятностей p_{ij} оценивается как

$$p_{ij} = \frac{p_{i|j} + p_{j|i}}{m}.$$

Задача метода состоит в нахождении координат $\mathbf{z}_i$ объектов $\mathbf{x}_i$ в пространстве $\mathbb{R}^d$ меньшей размерности $d < n$, в котором распределение q_{ij}

$$q_{i|j} = \frac{(1 + ||\mathbf{z}_i - \mathbf{z}_j||^2)^{-1}}{\sum_{k\neq l}(1 + ||\mathbf{z}_k - \mathbf{z}_l||^2)^{-1}}$$

ближайших соседей близко к исходному с точки зрения дивергенции Кульбака-Лейблера

$$KL(P||Q) = \sum_{i,j} p_{ij}\log\frac{p_{ij}}{q_{ij}}. \tag{1}$$

Шаг 1. Рассмотрим подвыборку $X^{m'} = \{\mathbf{x}_1, \ldots, \mathbf{x}_{m'}\}$. Для нее оценим $P^{m'\times m'}$ и найдем ее представление $Z^{m'}$ и $Q^{m'\times m'}$. Добавим новое множество объектов $X^{m-m'} = \{\mathbf{x}_{m'+1}, \ldots, \mathbf{x}_m\}$. Не меняя положений $\mathbf{z}_i$, $i =$

$m'+1,\ldots,m$ исходных точек, найдем представление $Z^{m-m'}$ и $Q^{(m-m')\times m}$, сдвигая точки в направлении градиента (1)

$$\frac{\partial KL}{\partial y_i} = 4\sum_{j=1}^{m}(p_{ij}-q_{ij})(y_i-y_j), \quad i = m'+1,\ldots,m.$$

Чтобы это работало, нужно чтобы исходный набор $X^{m'}$ был достаточно репрезентативен, а найденное представление $Z^{m'}$ было надежно.

Шаг 2. Для повышения надежности $Z^{m'}$ используем информацию о разметке y_i объектов $\mathbf{x}_i$, $i = 1,\ldots,m'$. Предлагается использовать знания о разметке при вычислении $P^{m'\times m'}$, например:

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

Список литературы

[1] Julian Brooke and Graeme Hirst. Paragraph Clustering for Intrinsic Plagiarism Detection using a Stylistic Vector-Space Model with Extrinsic Features, 2012.

[2] Julian Brooke et al. Unsupervised Stylistic Segmentation of Poetry with Change Curves and Extrinsic Features, 2012.

[3] Martin Potthast and Andreas Eiselt and Alberto Barrón-Cedeño and Benno Stein and Paolo Rosso. Overview of the 3rd International Competition on Plagiarism Detection // Working Notes Papers of the CLEF 2011 Evaluation Labs, 2011. http://www.clef-initiative.eu/publication/working-notes.

[4] Laurens van der Maaten. Visualizing Data using t-SNE
Journal of Machine Learning Research, 9 (2008) 2579-2605.

[5] Laurens van der Maaten. Accelerating t-SNE using Tree-Based Algorithms
Journal of Machine Learning Research, 15 (2014) 1-21.

[6] Hyunsoo Kim, Haesun Park, and Hongyuan Zha. Distance Preserving Dimension Reduction for Manifold Learning
Proceedings of the 2007 SIAM International Conference on Data Mining.

[7] [LinkReviw] Distance preserving dimensionality reduction and manifold learning. https://goo.gl/eoda3K

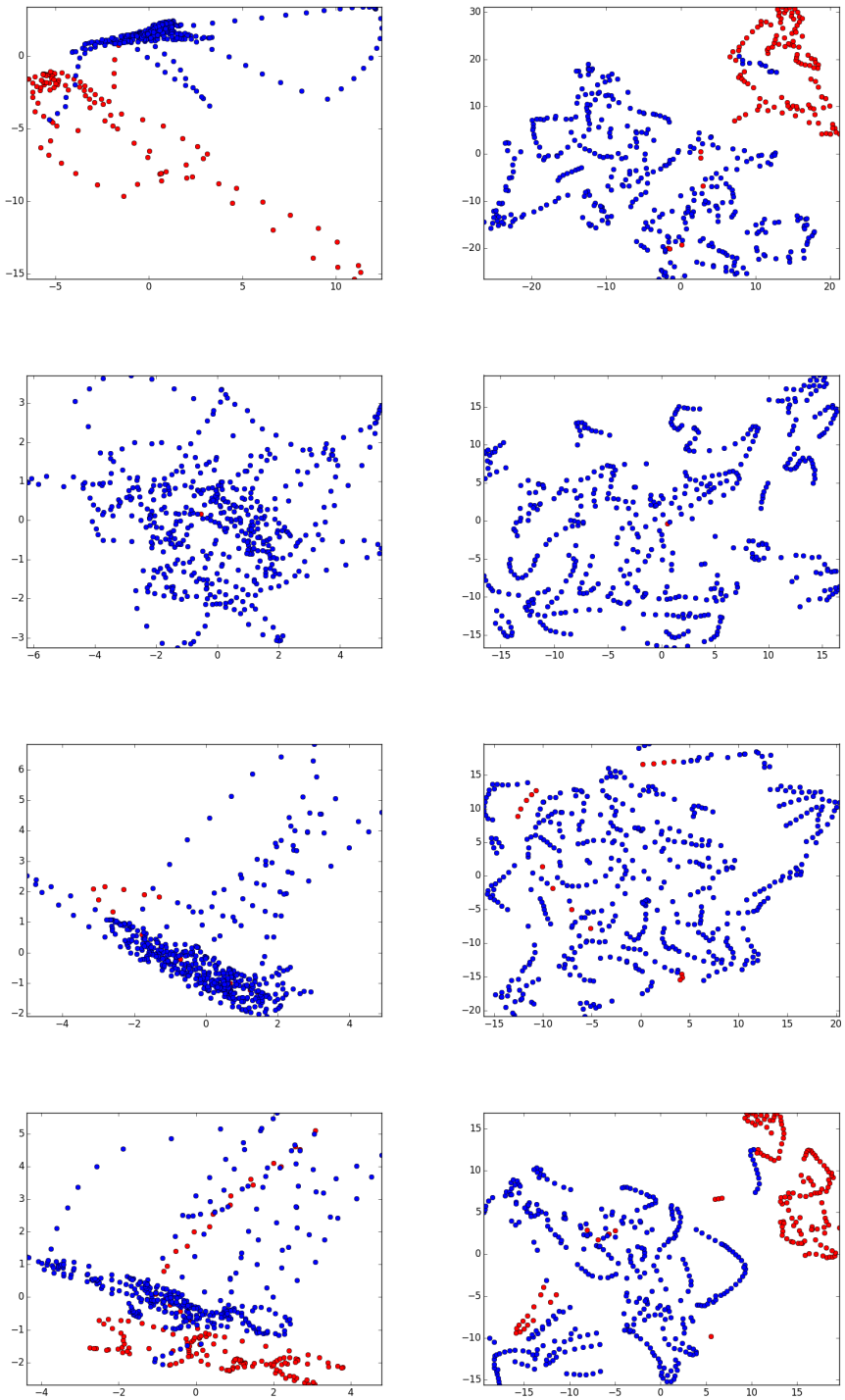


Рис. 1: Примеры работы алгоритма снижения размерности (PCA — слева, tSNe — справа), по 500 предложений документа suspicious-document00008.txt.

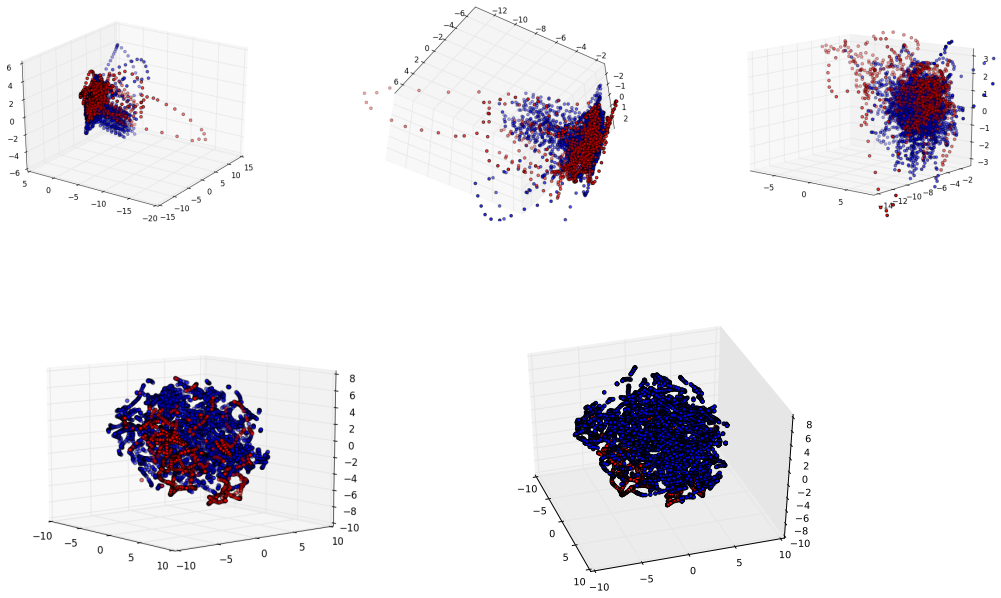
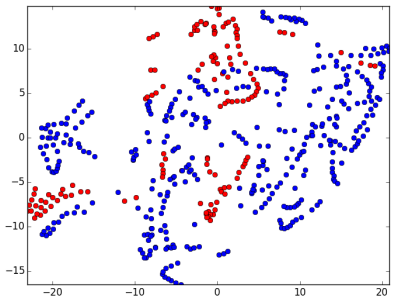
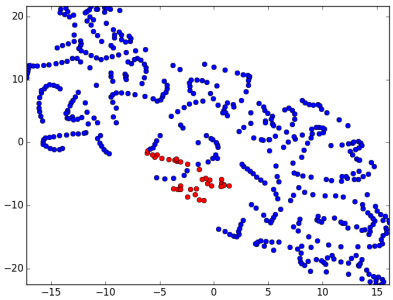


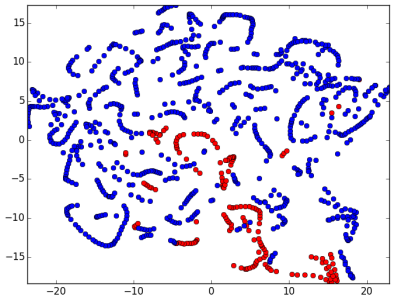
Рис. 2: Примеры работы алгоритмов снижения размерности PCA (сверху) tSNe (снизу) для документа suspicious-document00008.txt.



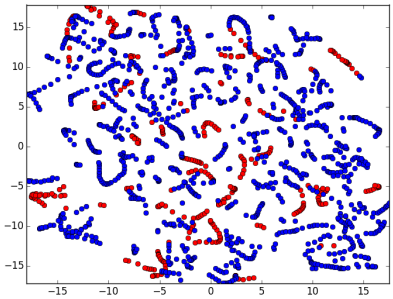
(a) 500



(b) 500



(c) 1000



(d) all

Рис. 3: Неуспех, tSNE.