Application of locally linear embedding algorithm on hotel data text classification

Jinming Huang

1College of mathematics and systems science, Shandong university of science and technology, Qingdao, Shandong, 266000, China

*Author’s e-mail: jinminghhh@163.com

Abstract. As a non-linear dimension reduction method, manifold learning algorithm projects high-dimensional input to a low-dimensional space by maintaining the local structure of the data, and discovers the inherent geometric structure hidden in the data. In this paper, we attempt to apply the manifold learning algorithm to the field of Chinese text classification, and use the locally linear embedding algorithm to reduce the dimension of the ctrip hotel review data set. Then, we utilize extreme gradient boosting (XGBoost) and logistic regression to classify the text. Experimental results show that it is effective and feasible to use manifold learning algorithm for text classification. Moreover, the classification effect of logistic regression is better than XGBoost in the text classification of hotel reviews.

1. Introduction

Text classification is a hot topic in the field of text mining. Most text classification methods, based on the assumption that features and document categories are closely related, use the vector space model to describe and represent documents. In the vector space model, every document can be represented as a vector in the feature space. Each element represents a feature of the document, usually a specific word, and its value corresponds to the weight of this word in the document. If there are n documents, each of which takes m words into consideration, it will become an n×m text matrix. Therefore, the feature dimension of a document is often relatively large. It is necessary to select features of the document and retain those features that can represent the category information of the document to improve the speed and accuracy of text classification. There are two types of feature selection methods[1]: one is based on frequency statistics, and the other is based on correlation statistics between features and text, features and category information. But each feature selection method has its shortcomings.

Another common method for dimension reduction of high-dimensional data is feature extraction. The feature extraction methods are divided into linear methods and nonlinear methods in the pattern recognition domain. Linear methods mainly include principal component analysis (PCA), linear discriminant analysis (LDA), etc. Nonlinear methods are mainly based on manifold learning algorithms, such as isometric feature mapping (ISOMAP), locally liner embedding (LLE).

In the field of text classification, there have been some studies about text classification based on manifold learning. Yang introduced ISOMAP algorithm into text mining domain[2]. In order to solve the problem of text classification, a text classification algorithm combined with manifold learning and support vector machine was proposed[3]. Wang improved the locally linear embedding algorithm and applied the new method to Chinese text classification, and proved the effectiveness of the improved LLE algorithm for Chinese text classification[4]. In order to solve the problem of dimensionality
disaster in text classification, Zhang proposed a text classification algorithm using nonlinear dimension reduction algorithm combined with k-nearest neighbor classifier, and proved that the algorithm can effectively improve the accuracy of text classification[5].

These studies all apply the manifold learning algorithm to the Chinese text classification research, and improve the accuracy of new algorithm from different aspects. It shows that it is feasible to apply the nonlinear feature extraction methods based on manifold learning to the Chinese text classification research. But different algorithms have various problems while improving the accuracy of text classification. Based on this, we preprocess the data set at first. Then, we apply locally linear embedding algorithm to select text features, and utilize logistic regression and XGBoost to classify the text. Finally, we use ctrip hotel review data to demonstrate its effectiveness by comparing with other dimension reduction algorithms.

2. Manifold learning

Three papers that proposed and discussed manifold learning from a cognitive perspective were published in Science in 2000. For the first time, the term "manifold learning" was used. Manifold learning gives a feasible and effective algorithm and provides a new research direction of dimension reduction methods[6]. Manifold learning can reduce the dimension of the high-dimensional data, reveal its manifold distribution, and find meaningful low-dimensional structures hidden in the high-dimensional observation data. Therefore, manifold learning has important applications in intelligent information processing and pattern recognition, and has been successfully used in face recognition and tumor classification. Classical manifold learning algorithms mainly include isometric feature mapping[7], locally linear embedding[8], laplacian eigenmaps and local preserving projection (LPP)[9], etc. LLE algorithm is the most representative manifold learning method.

3. Locally linear embedding algorithm

The main task of text classification is to train a high-performance classifier based on the existing text data information, so as to realize automatic classification of new unknown category text. Text feature extraction is the main step and method to achieve dimension reduction of high-dimensional text data. The quality of the text feature extraction algorithm directly affects the effect of text classification. The purpose of using feature extraction to reduce dimension is to project high-dimensional data into a low-dimensional feature space, so that different types of samples can be simply classified. In the manifold learning algorithm, LLE algorithm and some improved algorithms have been applied to the field of text classification.

LLE algorithm, an effective visualization dimension reduction method, can obtain better low-dimensional embedding for text data sampled in a single manifold. LLE algorithm believes that the structure of data is linear in the local space, so any point can be represented by a linear combination of its nearest neighbor points. Therefore, the main idea of LLE algorithm is to establish a local linear representation of the data in the high-dimensional space, and to achieve dimension reduction by keeping its local linear representation in the low-dimensional space as much as possible. The steps of the LLE algorithm can be summarized as follows:

Step one: select local neighbors. For a given data set \( X = \{x_1, x_2, \ldots, x_N\}, x_i \in \mathbb{R}^D \), we use the Euclidean distance to find \( k \) nearest neighbors of each sample point in the high-dimensional space.

Step two: calculate the local reconstruction weights of the sample points to minimize the reconstruction error of sample points, that is, to solve the following optimal problem:

\[
\begin{align*}
\min_{W} & \sum_{i=1}^{N} \left\| x_i - \sum_{j=1}^{k} w_{ij} x_j \right\|_2^2, \\
\text{s.t.} & \sum_{j=1}^{k} w_{ij} = 1.
\end{align*}
\]

(1)
is the jth nearest neighbor of \( x_i \), \( w_j \) represents contribution of the jth sample point to \( x_i \).

Step three: use the weight matrix \( W \) to find the low-dimensional embedding \( Y \). we minimize the reconstruction error and the function

\[
\min \phi(Y) = \sum_{i=1}^{N} \left\| y_i - \sum_{i=1}^{N} w_{ij}y_j \right\|_2^2.
\]

In order to fix \( Y \), \( Y \) can be simply restricted to

\[
\sum_{i=1}^{N} y_i = 0, \quad \frac{1}{N} \sum_{i=1}^{N} y_i y_i^T = 0,
\]

\( I \) represents the \( N \)-dimensional identity matrix. The corresponding optimization problem is transformed into the following constrained optimization problem:

\[
\begin{align*}
\min \phi(Y) &= \sum_{i=1}^{N} \left\| Y_i - YW_i \right\|_2^2 = \min tr(\hat{Y} - W)Y^T(Y - W)Y^T, \\
s.t. \quad YY^T &= I.
\end{align*}
\]

By using the Lagrange multiplier method, the solution is \( (I - W)^T(I - W)Y = \lambda Y^T \), \( Y \) is the eigenvector corresponding to the smallest \( d \) non-zero eigenvalues of \( (I - W)^T(I - W) \).

4. Experiments

4.1. Data selection and preprosessing

The ctrip hotel review data are used in this paper. This data set has a total of 7766 comments, including 5322 positive comments and 2444 negative comments. Obviously, this data set has the problem of sample imbalance. We use undersampling strategy to balance samples, that is, to use 2000 positive samples and 2000 negative samples.

4.2. Experiments based on the logistic regression

The text classifiers mainly include logistic regression[10] and XGBoost[11, 12]. In this experiment, we first choose logistic regression. In order to make the obtained experimental results credible, the data set is divided into a training set and a test set according to 9:1. In order to demonstrate the effectiveness of the LLE method, we use confusion matrix to evaluate the effect of text classification.

The effect of logistic regression classification using LLE algorithm is shown in Table 1. For class 1, the precision is 0.88 and the recall is 0.80, f1-score is 0.84. The confusion matrix is shown in Figure 1. Limited to the length of this article, we record the precision, recall and f1-score instead of confusion matrix to evaluate the effect of text classification.

Table 1. Experimental results of classification based on logistic regression.

| class       | Precision | Recall | F1-score |
|-------------|-----------|--------|----------|
| PCA         | AVONA     | LLE    | PCA      | AVONA     | LLE    | PCA  | AVONA    | LLE    |
| 0(negative) | 0.69      | 0.79   | 0.80     | 0.72      | 0.89   | 0.88 | 0.71     | 0.84   | 0.84  |
| 1(positive) | 0.73      | 0.89   | 0.88     | 0.71      | 0.78   | 0.80 | 0.72     | 0.83   | 0.84  |
Figure 1. Confusion matrix of logistic regression model on the test set using LLE algorithm.

The feature extraction algorithm uses principal component analysis, analysis of variance (ANOVA), locally linear embedding algorithm, and the classification algorithm uses logistic regression. Table 1 shows the precision, recall, f1-score of the classification using PCA + LR, ANOVA + LR, LLE + LR.

It can be seen from Table 1 that for class 1, the recall of the LR model using LLE algorithm is 2% higher than that using ANOVA, and the f1-score is increased by 1%. At the same time, the classification effect using ANOVA is much better than using PCA. From the results of the three methods in Table 1, it can be seen that the LLE method can obtain a higher f1-score, which can improve the accuracy and recall of text classification and ensure the efficiency of classification.

4.3. Experiments based on the extreme gradient boosting
In this experiment, we choose XGBoost to classify the text.

Table 2 shows the precision, recall, f1-score of the classification using PCA + XGBoost, ANOVA + XGBoost, LLE + XGBoost.

It can be seen from Table 2 that for class 0, the XGBoost model using the LLE algorithm improves the precision by 3% and the f1-score by 1% compared with the ANOVA. Obviously, the classification effect using ANOVA is much better than using PCA. Therefore, LLE method can obtain a higher accuracy.

| class   | Precision  | Recall  | F1-score |
|---------|------------|---------|----------|
|         | PCA        | AVONA   | LLE      | PCA        | AVONA   | LLE      | PCA        | AVONA   | LLE      |
| 0(negative) | 0.64       | 0.76    | 0.79     | 0.85       | 0.85    | 0.83     | 0.73       | 0.80    | 0.81     |
| 1(positive)| 0.80       | 0.84    | 0.84     | 0.55       | 0.74    | 0.79     | 0.65       | 0.79    | 0.81     |

4.4. Performance comparison of LR and XGBoost classifier
Comparing Table 1 and Table 2, for class 0, the logistic regression classifier using ANOVA improves the precision by 3% and the recall by 4% compared with XGBoost classifier using ANOVA. In this section, based on ANOVA, we compare the classification effect of logistic regression and XGBoost on movie review data firstly. The ROC curve is shown in Figure 2. Then based on LLE algorithm, the classification effect of logistic regression and XGBoost is compared. The ROC curve is shown in Figure 3.
5. Conclusion

Manifold learning is a non-linear visualization method for dimension reduction. It has been successfully used in face recognition and pattern recognition. However, the application research in Chinese text classification has just started. In this paper, the ctrip hotel review data is classified based on three feature extraction methods and two classification models. The experimental results prove the effectiveness and feasibility of locally linear embedding algorithm in the study of Chinese text classification.

References

[1] Chen, T., Xie, Y.Q. (2006) Literature review of feature dimension reduction in text Categorization. Journal of the China Society for Scientific and Technical Information, 25: 690-695.

[2] Yang, Z., Fan, K.F., Lei, J.J. (2009) Text manifold based on Semantic Analysis. Acta Electronica Sinica, 37: 557-561.
[3] Ren, J.F., Liang, X., Li, S.H. (2012) Text categorization algorithm based on manifold learning and support vector machines. Computer Science, 39: 261-263.

[4] Wang, H.Y., Feng, L., Feng, Y., Cheng, Q.C. (2012) The manifold learning algorithm’s application in the Chinese text clustering. Journal of Shandong University (Engineering Science), 42: 8-12.

[5] Zhang, G.D., Zhang, H.Y. (2014) Text categorization algorithm based on non-linear manifold learning and k-NN. Journal of Shandong University (Engineering Science), 43: 1-6.

[6] Gao, X.F. (2009) Problems and analysis in manifold learning. Computer Science, 36: 25-28.

[7] Tenenbaum, J.B., De-Silva, V., Langford, J.C. (2000) A global geometric framework for nonlinear dimensionality reduction. Science, 290: 2319-2323.

[8] Roweis, S.T., Saul, L.K. (2000) Nonlinear dimensionality reduction by locally linear embedding. Science, 290: 2323-2326.

[9] Shi, L.K., Yang, Q.X. (2010) Manifold learning algorithm Based on the small world model. Journal of Computer Applications, 30: 2917-2920.

[10] Qun, N., Jia, H.Y. (2018) Research and implementation of based on the logistic regression model for tibetan text classification. China Computer&Communication, 5: 70-73.

[11] Chen, T., Guestrin, C. (2016) XGboost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco. pp. 785-794.

[12] Ramanathan, V., Wechsler, H. (2012) Phishing Website Detection Using Latent Dirichlet Allocation and AdaBoost. In: 2012 IEEE International Conference on Intelligence and Security Informatics, Arlington. pp. 102-107.