Journal of Visual Languages and Computing ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Journal of Visual Languages and Computing

journal homepage: www.elsevier.com/locate/jvlc

Auto-encoder based bagging architecture for sentiment analysis

Wenge Rong [a,b], Yifan Nie [c], Yuanxin Ouyang [a,b,*], Baolin Peng [a], Zhang Xiong [a,b]

[a] School of Computer Science and Engineering, Beihang University, Beijing 100191, China
[b] Research Institute of Beihang University in Shenzhen, Shenzhen 518057, China
[c] Sino-French Engineering School, Beihang University, Beijing 100191, China

ARTICLE INFO

Article history:
Received 17 September 2014
Accepted 24 September 2014
This paper has been recommended for acceptance by Shi Kho Chang.

Keywords:
Sentiment analysis
Bagging
Auto-encoder

ABSTRACT

Sentiment analysis has long been a hot topic for understanding users statements online. Previously many machine learning approaches for sentiment analysis such as simple feature-oriented SVM or more complicated probabilistic models have been proposed. Though they have demonstrated capability in polarity detection, there exist one challenge called the curse of dimensionality due to the high dimensional nature of text-based documents. In this research, inspired by the dimensionality reduction and feature extraction capability of auto-encoders, an auto-encoder-based bagging prediction architecture (AEBPA) is proposed. The experimental study on commonly used datasets has shown its potential. It is believed that this method can offer the researchers in the community further insight into bagging oriented solution for sentimental analysis.

1. Introduction

Sentiment analysis, which aims at extracting polarity orientation from a statement and identifying if the statement is positive or not, has become a hot topic in both academic and industrial sphere. As a powerful information gathering mechanism [1], sentiment analysis accumulates on-line documents ranging from twitters, blogs to customer reviews and tries to understand customers attitudes, opinion and emotion. It has been proven useful in different domains such as E-commerce [2], social media analysis [3] and political elections [4].

There are a lot of sentimental analysis approaches proposed in the literature and these methods can be roughly divided into two categories, i.e., computational linguistic approach and machine learning approach [5]. Computational linguistic approach is based on linguistic information and employs pre-defined sentiment lexicon where individual words are assigned a polarity score. The overall polarity of a statement is calculated by voting the scores of each words in the statement. Machine learning approach utilises machine learning models to perform sentiment analysis and regards sentiment analysis as a formal binary classification problem [6].

Computational linguistic approach employs pre-defined scores of words in a general context to detect the polarity information [7], while machine learning oriented solutions employs labelled data in the specific domain of the task to conduct sentimental analysis [5]. As a result machine learning approaches normally show better prediction capability as sentiment analysis is indeed a domain-dependent task [8].

Although machine learning approach presents better capability than computational linguistic approach, it still suffers from some challenges. One of them is the curse of dimensionality [9]. This is a typical phenomenon when the dimension of the feature space becomes larger, the volume

* Corresponding author at: School of Computer Science and Engineering, Beihang University, Beijing 100191, China.
E-mail address: oyyx@buaa.edu.cn (Y. Ouyang).

http://dx.doi.org/10.1016/j.jvlc.2014.09.005
1045-926X/

of the space will increases quickly as such available training data will become sparse, thereby degrading the learning algorithms for sentiment analysis due to the high dimensionality of text dataset. In order to reduce the dimension of the raw data and extract higher order features from them to perform classification, many solutions have been developed and one popular implementation is stacked auto-encoder-based prediction model.

To build a stacked auto-encoder-based prediction model, a common approach is to stack several code layers of auto-encoders and further add one classification layer on the top of the last layer, and then take the learned features as input for the classification layer [10]. Though the auto-encoder-based prediction model has shown its promising potential, it still faces several difficulties among which a notable one is how to reduce generalisation error [11], which measures how well the model generalizes to data not participated in training. To meet this challenge, a lot of approaches have been proposed among which a widely adopted one is ensemble methods [12].

Bagging (Bootstrap Aggregation) is a popular ensemble method which consists in bootstrapping several copies of the training set and then employing them to train separate models. Afterwards it combines the individual predictions together by a voting scheme for classification applications [13]. As each bootstrapped training set is slightly different from each other, each model trained on theses training sets will have different weights and focus, thereby obtaining different generalisation errors. By combining them together, the overall generalisation error is expected to decrease to some extent.

Previous works have shown that bagging works well for unstable predictors [14]. Considering stacked auto-encoder-based prediction models are also unstable predictors [15], it is intuitive to assume that applying bagging methods to auto-encoder-based models could improve classification performance. Inspired by this assumption, in this paper an auto-encoder-based bagging prediction architecture (AEBPA) is proposed to integrate stacked auto-encoder-based prediction models for sentimental analysis. The experimental study on commonly used datasets also shows its promising potential.

The main contribution our work is two-fold. (1) We integrated feature learning with bagging ensemble method on text-based datasets and empirically evaluated the performance of integration with different number of bagging sets. (2) We propose an auto-encoder-based bagging prediction architecture (AEBPA) for sentiment analysis and empirical study shows that our approach outperforms traditional methods.

The remainder of this paper is organised as follow. Section 2 will introduce the background and Section 3 will illustrate the proposed methodology. In Section 4 the experimental study will be presented and Section 5 will conclude the paper and point out possible future work.

2. Background

2.1. Sentiment analysis

Sentiment analysis is a powerful mechanism to obtain people's opinion and tell if their overall attitudes are favourable or not. It has enjoyed a huge burst of research activity in recent years [1]. For example, sentimental analysis can be used to analyse the users purchasing behaviour as more and more customers take notice of the review of other people before buying a product on-line [16]. It can also be used for predicting political election results [17]. As sentiment analysis becomes an important tool in different domains, many techniques have been proposed and can be roughly categorised into computational linguistic approach and machine learning approach.

Computational linguistic approach employs pre-explored linguistic information where a lexicon is developed [18], either from exterior [19] or through heuristics during sentiment analysis [20]. In this kind of approaches, semantic information is added into the process and every word in the lexicon is attached with a polarity score. The process will further calculate the overall polarity of the statement by voting the polarity score of each word [21]:

$$p = sign\left(\sum_{i=1}^{n} q_i + 0.5p_0\right) \tag{1}$$

where q_i stands for polarity score of each word, *sign* stands for the sign function and p_0 stands for the major polarity in the training data.

Different from computational linguistic approach, machine learning approach utilizes machine learning algorithms and treat sentiment analysis as a formal binary classification problem. Traditional methods convert a statement into a bag-of-word representation, extract other auxiliary features such as part-of-speech information, bi-grams and specific negation words, and then throw these features into SVM, MaxEnt or Naive Bayes based classifiers [22–24]. Though these methods are simple and robust, they are insufficient to reveal the underlying relationship between words and the overall sentiment polarity of a statement.

To deal with the shortcomings of simple machine learning methods, many more complex model are proposed among which Socher et al.'s recursive auto-encoder-based model [25] is a notable one and has shown its effectiveness in sentiment analysis. In this work, word indices are first mapped through a embedding matrix into semantics word vectors. Then the semantic word vectors of the words in the input sentence are lined up in their original sequence. Afterwards, the input word vectors are recursively merged into a fixed size vector representation in the following way.

Given a sequenced list of word vector nodes $X = (x_1, x_2, x_3, \ldots, x_n)$ an auto-encoder first tries to merge all neighbouring couples (x_i, x_{i+1}) from X by the following formula:

$$p = f(W_1[c_1, c_2] + b_1) \tag{2}$$

where $(c_1, c_2) = (x_i, x_j)$ and f is the activation function, p is the calculated parent node vector and W_1, b_1 are the weight matrix and bias vector of the auto-encoder, respectively. Then the auto-encoder tries to decode the parent node and calculate the reconstruction error by the following formula:

$$[c_1', c_2'] = W_2 p + b_2 \tag{3}$$

Please cite this article as: W. Rong, et al., Auto-encoder based bagging architecture for sentiment analysis, Journal of Visual Languages and Computing (2014), http://dx.doi.org/10.1016/j.jvlc.2014.09.005