Neurocomputing 207 (2016) 599–609

Contents lists available at ScienceDirect

Neurocomputing

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

ASELM: Adaptive semi-supervised ELM with application in question subjectivity identification

Hongping Fu [a], Zhendong Niu [a], Chunxia Zhang [b,*], Hanchao Yu [c], Jing Ma [a], Jie Chen [a], Yiqiang Chen [c], Junfa Liu [c]

[a] School of Computer Science and Technology, Beijing Institute of Technology, 100081, China
[b] School of Software, Beijing Institute of Technology, 100081, China
[c] Institute of Computing Technology, Chinese Academy of Sciences, 100190, China

ARTICLE INFO

Article history:
Received 27 April 2016
Received in revised form
10 May 2016
Accepted 13 May 2016
Communicated by G.-B. Huang
Available online 22 May 2016

Keywords:
ELM
Adaptive semi-supervised ELM
Question subjectivity identification
Community question answering

ABSTRACT

Question subjectivity identification in Community Question Answering (CQA) has attracted a lot of attentions in recent years. With the rapid development of CQA, subjective questions posted by users are growing exponentially, which presents two challenges for question subjectivity identification. The first one is the data imbalance between subjective and objective questions. The second one is that the amount of manually labelled training data is hard to catch up with the fast developing speed of CQA. In this paper, we propose an adaptive semi-supervised Extreme Learning Machine (ASELM) to solve those two challenges. To resolve the data imbalance problem, ASELM employs the different impacts on identification performance caused by the imbalanced data. Second, the proposed method introduces the unlabelled data, and builds a model about the ratio between the number of labelled and unlabelled data based on Gaussian Model, which is applied to automatically generate the constraint on the unlabelled data. Experimental results showed ASELM improved identification performance for the imbalanced data, and outperformed the performance of basic ELM, SELM, Weighted ELM and SS-ELM on both F1 measure and accuracy.

1. Introduction

The rapid development of Community Question Answering systems, such as Quora,[1] Yahoo! Answers,[2] Stack Exchange,[3] provides a platform for users to post questions, answer questions and interact with others, where users can share their knowledge and experience. This generates huge useful information which can offer assistance to askers and users who need help. For users using CQA as a forum to share their opinions on controversial topics, subjective questions get a high ratio in most question categories [1]. That is to say, most questions in CQA are subjective questions which expect subjective information from answers that contain answerers' attitude, opinion or emotion. Meanwhile, for Community Question Answering systems, understanding user expected information more precisely is crucial to provide better answers. Therefore, the task of question subjectivity identification is proposed. Given a question in a CQA system, question subjectivity identification is to identify whether this question is a subjective question from whose answers users want to get answerers' opinions and attitudes.

In CQA, most users want to get subjective information obtained from others' experience to guide their decision or activation. For the government, they can get users' views about some events from the subjective information in CQA systems and formulate the corresponding countermeasure. Therefore, question subjectivity identification is an important task that needs to be resolved. However, the huge useful information in CQA leads to the following two problems. First, questions in Community Question Answering systems are usually imbalanced. That fact may contain a predilection of different question types and cause negative effect on classification performance to train an identification model. It makes the question subjectivity identification challenging. Second, the huge user generated contents bring more valuable data, whereas there is not enough labelled data for question analysis. Meanwhile, manually labelling the data is time consuming and labour consuming. That also makes the question subjectivity identification challenging.

Based on the above challenges, we focus on the following two research problems: (1) Can we improve the identification performance by leveraging the different influence on it caused by

* Corresponding author.
E-mail address: cxzhang@bit.edu.cn (C. Zhang).
[1] https://www.quora.com/
[2] https://answers.yahoo.com/
[3] http://stackexchange.com/

http://dx.doi.org/10.1016/j.neucom.2016.05.041
0925-2312/

imbalanced subjective and objective questions to adapt the imbalanced data? (2) Can we utilize the ratio between labelled and unlabelled data to automatically generate the suitable constraint on unlabelled data in Community Question Answering systems?

The problems mentioned above have been noted and studied by many researchers recently. For the first problem, many previous researches [2–4] focused on how to rebalance the data by over-sampling or under-sampling, and some works handled that problem through hybrid algorithms. Also some works use the weighted method [5,6] to avoid imbalance problem. Those methods proposed at the data level may destroy the distribution of original data. To solve the second problem, many semi-supervised methods have been proposed [7–9] by introducing a large number of unlabelled data. For those methods, the importance of the unlabelled data influencing the performance needs to be considered. Therefore, a special weight is introduced in most previous studies to adjust the effect of unlabelled data, which is usually set according to experience or experiments. However, in practice, it is difficult to provide the appropriate weight manually when the amount of labelled/unlabelled data changes.

In this paper, to resolve the imbalanced data problem and lack of labelled data problem, we propose an approach Adaptive Semi-supervised Extreme Learning Machine (ASELM), which takes advantage of different impact caused by the imbalanced data and the ratio between the number of labelled and unlabelled data. To solve the first problem, we focus on how to use the imbalanced data efficiently by taking advantage of the different influence on identification performance caused by imbalanced subjective and objective questions. Other than previous works assigning different class weights based on the number of samples belonging to different classes, ASELM takes advantage of the weighting ratio between subjective and objective questions. The ratio can switched according to the required questions that users want to identify. For the second problem, ASELM brings in the unlabelled data. More importantly, it builds a model about the ratio between the number of labelled and unlabelled data based on Gaussian Model. Then, the proposed method can automatically generate the suitable parameter to control the influence of unlabelled data, rather than setting the parameter by traversing it according to the experiment results. And the adaptivity of ASELM is expressed from two aspects: (1) under the circumstances that subjective questions and objective questions are imbalanced, ASELM can improve the identification performance of the required questions (subjective or objective) according to users' requirements, which can adapt to the user requirement; (2) in the absence of labelled data, ASELM took advantage of unlabelled data, and automatically generated the constraint parameter according to the number of labelled data and unlabelled data, which can adapt to the changes of labelled and unlabelled data. The proposed method inherits the learning ability and efficiency of basic ELM learning algorithm.

The rest of this paper is organized as follows. We first overview the related works in Section 2. Then, the subjectivity question identification task is stated in detail, which was used as an application of our approach in Section 3. The proposed adaptive semi-supervised approach ASELM is introduced in Section 4 and Section 5 discusses the results of our experiments over thousands of questions in CQA systems. Finally Section 6 summarizes our conclusions and future work.

2. Related works

Question classification has been long studied in question answering and developed from factual question classification [10–12] to subjectivity question classification [13,1,14] with the quickly development of Community Question Answering systems.

There are a lot of works on the task of subjectivity question identification. Li et al. [1] used a supervised learning algorithm SVM and took advantage of both question information and answer information to identify the subjective questions. Different from works which used both question and answer information, Aikawa et al. [14] assumed that the answer information were unavailable when the questions were classified. They gave the annotation criteria to label the questions into subjective questions and objective questions, then used supervised method to detect subjective questions. Liu and Jansen [15] defined the subjective information seeking questions and objective information seeking questions in Social Q&A, and used supervised approach from Weka to classify these two types of questions. Then a comprehensive analysis was given to provide insights which can guide the information seeking process.

The data used to train a classifier are manually labelled for all the works mentioned above. But labelling enough data is time and labour consuming. To address the lack of labelled training data problem for subjective question identification, Li et al. [16] proposed a semi-supervised approach CoCQA, a learning method with the idea of co-training. CoCQA generated two classifiers exploiting the information of questions and answers respectively and teaching each other by giving their partner a newly labelled data. With the same purpose, Zhou et al. [17] utilized social information of CQA to automatically collect training data from Yahoo! Answers avoiding manual labelling, and they used question length, request word, subjectivity clue, punctuation density, grammatical modifier and entity as features to fulfill the classification. The above methods took advantage of the information of unlabelled data to solve the problem of lack of labelled data, but they left the data imbalance problem out of consideration.

3. Question subjectivity identification

In this section, we state question subjectivity identification task in detail. Here, we divide questions in CQA into two types: subjective questions and objective questions. For subjective questions, askers want to get subjective answers which contain answerers' opinions and attitudes. And for objective questions, askers want to get answers with factoid information based on common knowledge [14]. Table 1 given the examples of those two types of questions. The task is stated as follows:

Question subjectivity identification task: Given a question Q in a Community Question Answering system, our goal is to identify whether Q is a subjective question or objective question.

In this paper, we treat question subjectivity identification task as a binary classification problem, where subjective questions are considered as positive instances and objective questions are considered as negative instances. Since we aim to identify subjective questions more precisely in this task, we treat subjective questions as required questions and objective questions as auxiliary questions. In addition, the required questions can be set as any question types according to the user's requirement. For example, if we

Table 1
Examples of subjective questions and objective questions.

Question type	Examples
Subjective	• What's your favourite type of match? • To men: do you think crying is unmanly?
Objective	• What channel is the fifa world cup in Maryland? • What is the structure and classification for 1-methylcyclopentanol?