A Cross-domain Deceptive Opinion Detection by Genetic Algorithm

TANG Qiao-jing, LI Wei-hua ZHAO Jing

School of Information,Yunnan University,Kunming57000, China.

1159060893@qq.com

Keywords: Deceptive reviews, Cross-domain, Genetic algorithm, Spectral clustering

Abstract. With the maturing of the electronic commerce,many people will choose to shop online.At the same time the goods's reviews will influence the decision of people to buy goods,which has led businesses to deliberately write deceptive opinion to improve the sales of goods or undercut rivals .Aiming at the problem of the deceptive reviews,this paper proposes a deceptive reviews detection based on cross-domain and genetic algorithm.To define the data of feature in the source domain ,using genetic algorithm for feature to get the optimal feature set and constructing the feature incidence matrix M y the similarity of documents.it reducing the dimensionality of M , by using spectral clustering algorithm,to obtain a corresponding space mapping.then, with the help of sentiment classifiers,training and being classified in the target domain.Experimental results points that the feasibility and advantages of the method in recognizing the problem of deceptive reviews.

Introduction

So far, there have been many scholars to discussed the issue of sentiment analysis,and the work about cross-domain sentiment classification has been very extensive.But deceptive opinion detection and related work is still relatively less.Pan et al[3],in 2010,realized the process of reducing dimension by constructing the co-occurrence words matrix and transforming this matrix. "gold standard data set",which is about hotel reviews by Ott et al.[2],will server as the domain data sets of this paper.Ren(2014)[4],who proposed that deceptive reviews detection based on language structure and sentiment polarity,is using genetic algorithrm to extract the optimal features of language structure and sentiment polaritya and is combining non-supervision K-means clustering algorithm to identify deceptive reviews.Li F T[1]proposed Co-Training algorithm to identify deceptive reviews,which is considering the comments and reviewers.But it is requires a lot of labeled training set.So the paper will be less labeled training set as a sample set to identify the deceptive reviews for unknown domain.According to the characteristics of the data,Song[5] constructed the correlation matrix by calculating the similarity between documents.Then ,the improved K-means clustering algorithm is used to cluster the feature and there is discriminated deceptive reviews by calculating the outlier degree of category.

In this paper,we proposes a deceptive reviews detection based on cross-domain and genetic algorithm.we using genetic algorithm for feature to get the optimal feature set and constructing the feature incidence matrix by computing vector cosine for the similarity of documents.then, with the help of spectral clustering and sentiment classifiers,identifing deceptive reviews and proveing the feasibility of the method.

Deceptive Opinion Feature Definition

Deceptive opinion has some characteristics on the language structure and customer behavior.Thus, in this paper, the deceptive opinion to make the following definition:

Review enthusiasm: it is possible to use the proportion of vocabulary with the longest sentence in reviews to represent the degree of fakement.

Positive sentiment: it is considered that the positive deceptive opinion have more positive emotional vocabulary than the truthful opinion,which used the proportion of the positive words with total vocabulary in sentence.

Negative sentiment: it is considered that the negative deceptive opinion have more negative emotional vocabulary than the truthful opinion,which used the proportion of the negative words in total vocabulary in sentence.

Product attribute repeatability: The proportion of the number of goods's brand occurrenced in vocabulary of a review to express the degree of fakement.

Personal relationships: With the proportion of the two types of personal pronouns,it would be used to represent the personal relationship in reviews.

The similarity of language: With the maximum value by computing the similarity between reviews as the value of similarity of this review, the greater the value, the greater the likelihood of deceptive.

Feature Selection Based on Genetic Algorithm

In this paper,deceptive opinion of the above characteristics adopting genetic algorithm to feature selection, which is as follows in figure 1.

Encoding and recoding for individual. Each deceptive opinion is considered as an individual, which is composed of a number of false evaluation of a population. The individual called chromosomes is made up of genes,in which the vector parameter value and the feature selection status of the deceptive opinion feature is stored(The value of each feature is represented by a vector).So construction of the chromosome is based on the binary encoding ,which shown in Table 1

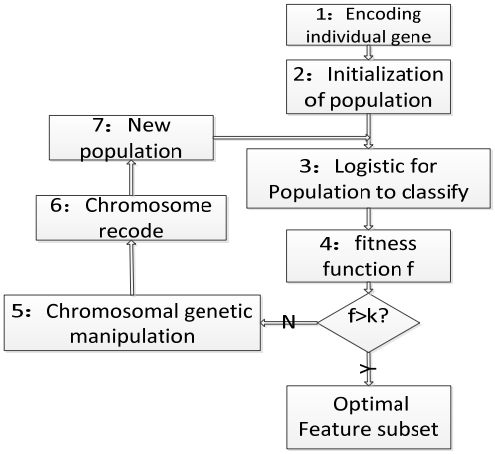

Figure 1 :genetic algorithm feature selection

Table 1: chromosome structure

N	O	S
$N_1,N_2.....N_n,ST$	$O_1,O_2.....O_n$	$S_1,S_2.....S_n$

The chromosome N represents the corresponding gene value of selected features and the deceptive status.And the chromosome O represents the corresponding gene value of all the features. The value of $\{N_1,N_2.....N_n\}$ was composed of multiple vectorvalued.For example,（N_1）O_1 represented the vector parameter value is corresponding to the first (S=1).*ST* record whether the individual review for a deceptive opinion,in which value of 1 indicates that the individual is a deceptive opinion and value of 0 indicates that the individual is a truthful opinion. Chromosome S represents the features selected state,which is represented by 0,1.The value of 1 indicates the selected feature, and the value of 0 is the opposite.

Producing the reorganization of the population,it can produce entirely new species by recoding the chromosome N before it have the population of chromosomes genetic operation.The function of chromosome N to recoding is:

$$r(N_i) = O_i * S_i \quad (1)$$

Logistic regression classification. Truthful data ,as the classification of the training data set,which is the chromosome N data in the population obtained by the anti-coding. The data set is

divided into the source domain data set D_s and the target domain data set D_t . $D_s=\{(x_i, y_i)\}|^{ns}_{i=1}$, x_i is a review of a feature vector representation, $x_i=\{n_1, n_2... n_n\}$,Where n1 is the N_1 converted from the binary to decimal truthful value. y_i indicates whether the *i* review is a deceptive opinion $y_i=ST$. *ns* is the size of the source domain data set. $D_t=\{(x_i)\}|^{nt}_{i=1}$, in which *nt* is the size of the target domain data set. What is trained by using logistic regression on the D_s to get $g(x)$ is used to classify D_t. The given D_s is defined by logistic model is as follows:

$$P_i = P\left(y_i / x_i\right) = \frac{e^{g(x)}}{1 + e^{g(x)}} \tag{2}$$

g(x)=yi*wT+b,w is the weight parameters. The joint distribution of the observed values of the training set is taking logarithm to get formula (3) by which it obtains value of w with the maximum value of the derivative.

$$L(B) = \ln\left[\prod_{i=1}^{ns} p_i^{y_i}\left(1 - p_i^{1-y_i}\right)\right] \tag{3}$$

Fitness function. Fitness function is used to judge the merits of the population,which can be determined by population classification accuracy and the number of individual characteristics. This definition is as follows:

$$f = w_a * r + w_b\left(\sum_{i=1}^{n} y_i\right)^{-1} \tag{4}$$

Among them, R is the accuracy of the classification results. A and B are the weight of the accuracy and the number of selected features. Obviously, the higher the accuracy, the fewer the number of feature representation and the higher the fitness value.

Genetic operation. Roulette algorithm is adopted to improve the individual choice,and each individual into the next generation is equal probability. For individual chromosomes O , S , were respectively randomized to single-point crossover.Then random mutation of chromosomes follow the way of "0" - "1", "0" - "1".It recoded the first segment of transformed chromosomes according to the formula (1) to get a new population.

A Cross-domain Deceptive Opinion Detection

Optimal feature with genetic algorithm would be used to represent a review. The "gold standard data set" as the research of the source domain data set $D_{src}\{(x_{srci},y_{srci})\}|^{nsrc}_{i=1}$ suitable.And the target domain data set $D_{tra}\{(x_{trai})\}|^{ntra}_{i=1}$ adopting reviews from the Amazon about digital goods. x_{srci} and x_{trai} respectively represented by the feature vector (a_1,a a_2.....a_n), (b_1,b_2.....b_n).

Document similarty. The similarity between the document T_1 and T_1 of two domain that calculate the similarity between the feature vectors. This is because each record is represented by a feature vector. $T_1=(a_1,a_2.....a_n)$,$T_2=(b_1,b_2.....b_n)$. The relationship between the size of the angle of two vectors is represented by using the cosine value of between vectors,in which the value range is [0 1] and the larger the value is, the closer the two vectors are. The expression of document similarity I is as follows:

$$I(T_1, T_2) = \cos(T_1, T_2) = \frac{\sum_{i=1}^{n} a_i * b_i}{\sqrt{\sum_{i=1}^{n} a_i^2 * \sum_{i=1}^{n} b_i^2}} \tag{5}$$

Space mapping function and deceptive opinion detection. In this paper, first we calculate the value of the similarity of documents I.Then I used to express the relationship between the value of the different domain of the document to define the two domain document. Finally, according to the above values, we have established feature correlation matrix M, which is an n*n matrices. We want to make the matrix based on the map of the dimension reduction processing, according to the algorithm 1,which can be mapped to the vector of the document that reduce the dimension to a potential space. Each document has a new way of expression, and it can be said that each document is mapped to a

point in a potential space by a mapping function, such as step 6. According to this new document expression, we recognize deceptive opinion by using the emotion classifier $Q(x)=sgn(g(x))=sgn(y_{srci}*w^T+b)$ to train the data.

Alorithm 1 a cross-domain deceptive opinion detection:

1: **Input**: source domain data sets $D_{src}\{(x_{srci},y_{srci})\}|^{nsrc}_{i=1}$,the target domain data sets $D_{tra}\{(x_{trai})\}|^{ntra}_{i=1}$, space latitude k

2: **For** i=1，2.....nsrc

For j=1,2.....ntra

According to the formula (5) computing the vaule of the similarity of x_{srci} and x_{trai}.

3: By connecting the data points of two domain with the similarity,according to the step 2 to establish a correlation matrix $M \in R(nsrc+ntra)^2$ between xsrc and xtra. M_{ij} represents the similarity of the point i and the point j. When i and j respectively D_{src} data points with D_{tra}, $M_{ij}=m_{ij}$ otherwise $M_{ij}=0$.

4: Calculate the D of the diagonal matrix M,$D_{ij}=\sum j M_{ij}$,and construct the Laplasse matrix L=D-M.

5: Calculate the k largest eigenvalues of L and corresponding eigenvectors, c_1,c_2.....c_k, and composition（$nsrc+ntra$）*k matrix $C=[c_1,c_2.....c_k]$.

6: Find each point to the potential space for the new representation:h（x_i）=C[x_i:k],

where x_i data points corresponding to the vector represented as a column of the x_i row of matrix C.

7: The new representation D_{src} training for the emotion classifiers formula (2),and classifying D_{tra}.

8: **Output** :classification results.

The Experiment and Analysis

The experimental data. "gold standard data sets" are selected as the experimental data to select deceptive feature and detect deceptive opinion. Crawling 500 review information of electronic goods on the Amazon site,which we take as a target domain data to detect cross-domain deceptive revies.

Experimental process and results analysis. In the experiments of deceptive opinion of feature selection based on genetic algorithm, the size of initial population is 50, the probability of crossover operation is 0.5. Finally, the corresponding value of the optimal feature set of chromosome S is "111010", and the number of iterations is 20 generations to converge algorithm.

Table 2:comparison results of deceptive review

	Supervised	Co-Training	SVM	In paper
Accuracy	0.549	0.630	0.693	0.737
Recall	0.521	0.589	0.427	0.506
F	0.535	0.609	0.528	0.600

The experiment was conducted to identify the deceptive opinion based on the obtained deceptive review features. In this paper, the cross-domain deceptive opinion detection algorithm and the supervised method, semi-supervised method and no transfer learning method are compared and analyzed. Supervised method adpoting the maximum entropy meth- od,which is using the data of target domain to train maximum entropy model and then the data on the target domaon were classified. No transfer method adopting SVM classifier,which using the labeled data of source domain to train the SVM classifier,and then to classify the target domain data.The accuracy rate , the recall rate and vaule of F are used to evaluate the analysis results of the deceptive opinion detection.And the experimental data of the Semi-supervised method (Co-Training) come from the literature [8].

From Table 2 the cross-domain deceptive opinion detection by genetic algorithm can effectively improve the recognition accuracy of deceptive review of the target areas.

The method in this paper for the other three methods in accuracy of deceptive recognition have improved.But in the recall, the proposed method with respect to poor maximum entropy method and Co-Training method,which shows that although transfer learning method can reduce the difference between the two domain,but not completely eliminated it. The other three methods to solve the problem is that the target domain data have a lot of data for training.When the amount of data in the target domain is sparse, the accuracy rate of deceptive recognition would be greatly affected. The proposed method can be used to solve the problem of sparse data in the target domain.

Summary

In this paper, we propose a cross-domain deceptive opinion detection by genetic algorithm, which is used to judge the deceptive review of the goods in the network. Through the genetic algorithm to extract the feature of the deceptive opinion,we establish the similarity between the two domain for the existing source domain data and the target domain data. According to this relationship we established correlation matrix,and mapped to a potential space with spectral clustering to train to get a semotion classifier. In the experiment, this algorithm can be seen that it is feasible, and there is advantage in some aspects. Future research will be looking for more the feature of a deceptive opinion.We further optimize feature representation of deceptive review and try to use multisource cross-cross manner to detect deceptive opinion.

References

[1]Li F T, Huang M, Yang Y,et al. Learning to Identify Review Spam. In: Proceedings of the 22nd International Joint Conference on Artificial Intelligence. AAAI ess,2011:2488-2493.

[2]M.Ott,Y.Choi, C.Cardie, et al. Finding Deceptive Opinion Spam by Any Stretch of the Imagination. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Portland,USA, Jun19-24,2011.Stroudsburg,PA,USA:ACL, 2011:309-319.

[3]Pan S J,Ni X C,Sun J T,et al.Cross-domain Sentiment Classification via Spectral Feature Alignment.In:Proceedings of the 19th International Conference on World Wide Web.New York,NY,USA:ACM,2010:751-760.

[4] Ren Y F,Yin L,Ji D H.Deceptive reviews detection based on language structure and sentiment polarity.Journal of Frontiers of Computer Science and technology,2014,8(3):313-320.

[5]Song Hai Xia, Yan Xin, Yu Zheng Tao, et al. Detection of fake reviews based on adaptive clustering. Journal of Nanjing University, 2013, 49(4):433-438.

[6]M. Ott, C. Cardie, J. Hancock.Negative Deceptive Opinion Spam.Proceedings of NAACL-HLT of the Association for Computational Linguistics:Georgia, Atlanta,2013:497-501.

[7] Li S, Lee S Y M, Chen Y, et al. Sentiment classification and polarity shifting. Proceedings of the 23rd International Conference on Computational Linguistics, Stroudsburg, PA, USA:Association for omputational Linguistics, 2010: 635–643.

[8] Wan X. Co-training for cross-lingual sentiment classification . ACL-IJCNLP, Stroudsburg, PA,USA: Association for Computational Linguistics, 2009:235–243.

[9]Suqin J I, Shi H. Optimized K-means clustering algorithm for massive data[J]. Computer Engineering & Applications, 2014.

[10]Pan S J, Yang Q. A Survey on Transfer Learning. Knowledge & Data Engineering IEEE Transactions on, 2010, 22(10):1345-1359.

[11] N. Jindal and B. Liu. Opinion spam and analysis.In Proceedings of the international conference on Web search and web data mining, ACM, 2008:219–230.