Decision Support Systems 50 (2011) 732–742

Contents lists available at ScienceDirect

Decision Support Systems

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

Predicting consumer sentiments from online text

Xue Bai

Department of Operations and Information Management, School of Business, University of Connecticut, Storrs, CT 06269, USA

ARTICLE INFO

Available online 19 August 2010

Keywords:
Sentiment analysis
Online reviews
Online news
Markov blanket
Heuristic search

ABSTRACT

Sentiment analysis from unstructured text has witnessed a boom in interest in recent years, due to the sheer volume of online reviews and news corpora available in digital form. An accurate method for predicting sentiments could enable us, for instance, to extract opinions from the Internet and gauge online customers' preferences, which could prove valuable for economic or marketing research, for leveraging a strategic advantage for an enterprise, or for detecting cyber risk and security threats. In this paper, we propose a heuristic search-enhanced Markov blanket model that is able to capture the dependencies among words and provide a vocabulary that is adequate for the purpose of extracting sentiments. Computational results on two collections of online movie reviews and three collections of online news show that our method is able to identify a parsimonious set of predictive features, yet simultaneously yield comparable or better prediction results about sentiment orientations, than several state-of-the-art feature selection algorithms as well as sentiment prediction methods. Our results suggest that sentiments are captured by conditional dependencies among words as well as by keywords or high-frequency words.

1. Introduction

Prior to the Internet, researchers used surveys to collect limited amounts of data in a structured form to analyze consumers' opinions on a product or service. In recent years, the advent of the Internet, and the widespread use of advanced information technologies in general (e.g. Web 2.0), have resulted in a surge of information that is freely available online in a text format. For example, many online forums and review sites exist for people to post their opinions about a product. While such consumer-generated online content offers tremendous business opportunities, potential risk and security concerns arise due to the possible malicious use of online social media [8]. Automatic and accurate understanding of *sentiments* expressed within the online text could lead to effective information retrieval, enable automated cyber risk management strategies, and improve business profits.

Researchers in the data mining field have studied the problem of text classification for more than three decades. Effective solutions have been found in the area of topic categorization and of authorship attribution. Topics are captured by sets of keywords, while authors are identified by their choices in the use of non-contextual, high-frequency words [2,3]. Pang et al. [34] showed that such solutions, or extensions of them, underperform when applied to sentiment extraction, yielding cross-validated accuracies and areas under the curve (AUC) in only between the high 70%s to low 80%s. Even more worrisome is the fact that these performances are obtained using large vocabularies, for which the discriminatory power of words is likely due to chance for many of the words. We conjecture that one reason for the failure of such approaches may be attributed to the fact that the words selected in the classification are considered to be independent of one another. We argue that their very interactions lead to the emergence of sentiments in the text. The goal of this paper is to present a machine learning method for learning predominant sentiments of online texts available in unstructured format, that is capable of selecting words that are related to one another and to the sentiment embedded in the texts significantly, i.e., beyond pure chance, and finding a minimal vocabulary that leads to good performance in categorization and prediction tasks.

In this paper, we present a two-stage prediction algorithm, the Markov Blanket Classifier, that is able to capture the dependencies among words, find a vocabulary that can efficiently extract sentiments, and ultimately, provide better predictions about sentiment expressed in a text document, compared to several state-of-the-art machine learning methods. Our two-stage Markov Blanket Classifier learns conditional dependencies among the words and encodes them into a *Markov Blanket Directed Acyclic Graph* (MB-DAG) for the sentiment variable (first stage), and then uses a *Tabu search* (TS) meta-heuristic strategy to fine-tune the MB-DAG (second stage) in order to yield a higher cross-validated accuracy. Learning dependencies allows us to capture semantic relations and dependent patterns among the words, which help us to approximate the meaning of sentences with respect to the sentiment they encode. Furthermore, performing the classification task using a Markov Blanket (MB) for the sentiment variable has two important properties: it specifies a statistically efficient prediction of the probability distribution of the sentiment variable from the smallest subset of predictors, and it

E-mail address: xue.bai@business.uconn.edu.

0167-9236/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.dss.2010.08.024

provides accuracy while avoiding over-fitting due to redundant predictors. We test our algorithm on two versions of the publicly available online movie reviews data [32,34] and on three collections of proprietary online news with different degrees of topicality [30]. The computational results show that our method is able to achieve a cross-validated accuracy and AUC comparable to the best performance of competing state-of-the-art sentiment classification methods with a parsimonious vocabulary.

The remainder of this paper is organized as follows. Section 2 surveys relevant literature. Section 3 provides brief background knowledge about Bayesian networks, Markov Blankets, and Tabu search heuristic. Section 4 introduces the methodology. Section 5 presents the data sets used in the study. Section 6.2.2 presents the experimental results with comparisons to several state-of-the-art methods. Section 7 discusses the findings and concludes with a summary of this work and future directions.

2. Literature review

2.1. Bayesian network and Markov blanket classification

Research in the Bayesian network field has sought to identify a part of the Bayesian Network that can be exploited as classifiers for the target class variable.

Friedman et al. introduced the General Bayesian Network (GBN) algorithm [14]. The basic idea of GBN is to learn a Bayesian Network structure that contains the target class node using Information Gain scores, and to perform the classification based on the Bayesian Network learned. The score driving the search of the structure for GBN measures the overall fit. The score is calculated on a weighted average schema over all the nodes in the network, not optimized for the target class node. In addition, the GBN algorithm learns the whole Bayesian Network first in order to use a portion of it for classification; hence it is feasible only for small data sets.

Madden proposed a methodology for induction of a Bayesian network structure for classification [24]. This structure is called Partial Bayesian Network (PBN). PBN is implemented using the K2 framework introduced in [10]. Learning the Partial Bayesian Network essentially reduces to a Bayesian Network learning problem using the K2 algorithm. The complexity of K2 algorithm is exponential to the number of variables; hence, PBN is also feasible only for small data sets.

Later, Madden presented an extended description of the algorithm in [24], introducing the concept of “Markov Blanket” [25]. The algorithm is called MBBC. However, the result of MBBC is a set of “Markov Blanket” variables, not a MB-DAG; each node represents a variable in the Markov blanket. This is because the results of K2 can only correctly identify the union of the set of variables adjacent to the target and the set of variables adjacent to those variables. It does not correctly identify the edge orientations of the variables in the union. By contrast, a MB-DAG contains both the variables in the Markov blanket and the relevant edges among the variables. MBBC, like PBN, is feasible only for small data sets. This is because 1) the scoring criterion used to construct the MB is a K2 metric [25]; and 2) after the MB is constructed, the conditional independence tests used to do the classification is the joint probabilities estimation proposed in the K2 algorithm, which is essentially also an exponent-based K2 metric estimation.

To the best of our knowledge, little of the previous work has dealt with real world data with a large number of variables and limited number of samples. Furthermore, for those studies that have used the notion of Markov Blanket as a minimal set of dependent variables, none of them have actually generated and retained the graphical structure of the Markov Blanket corresponding to a specific data set, nor have they used the structure for Bayesian inference in order to perform the classification. Our algorithm addresses all of these limitations.

2.2. Sentiment analysis

Sentiment analysis is also known as opinion extraction or semantic classification in the text mining literature [1,33]. A related problem is that of studying the semantic orientation or polarity of words [31]. Huettner and Subasic [19] developed a cognitive linguistic model for affection sentiments based on fuzzy logic. Liu et al. [23] proposed a method to categorize emotions using a large dictionary of common sense knowledge and on linguistic models. Das and Chen [11] constructed lexicon and grammar rules using domain knowledge to capture the “pulse” of financial markets, as expressed by online news about traded stocks, and thus achieved a classification accuracy of 62% (against the baseline accuracy of 33%). Hatzivassiloglou and McKeown [17] developed a log-linear model to predict the semantic orientation of conjoined adjectives. Turney and Littman [41] proposed a semi-supervised method for learning the polarity of adjectives starting from a small set of adjectives with known polarity. Turney [40] later applied this method to predicting the consumers' opinions about various objects (e.g., movies, cars, banks), and achieved accuracies between 66% and 84%. Pang et al. [34] applied the off-the-shelf classification methods to frequent, non-contextual words in combination with various heuristics and annotators, and achieved a maximum cross-validated accuracy of 82.9% on the IMDb data set. Dave et al. [12] classified movie reviews into positive versus negative using support vector machines on different types of semantic features based on substitutions and proximity, and achieved an accuracy of 88.9% on data sets from Amazon and CNN.Net.

Recent research tackles the problem by pairing data mining algorithms with feature selection methods [1,15,26,27,32,36,42]. Gamon [14] developed a linear support vector machine algorithm that combines two feature reduction pre-processing procedures. The features are first filtered by linguistic analysis, then by the ranking of “predictiveness.” The algorithm was evaluated on two sets of satisfaction survey data and achieved high classification accuracy. Ng et al. [27] used unigram and n-gram features combined with a log-likelihood based feature selection method and showed that the simple “bag-of-words” method was able to attain satisfactory accuracy on the movie review data set, using SVM classifiers. Reliff et al. [36] implemented a subsumption hierarchy mechanism in Information Gain criterion and showed good improvement in opinion classification using three opinion-related data sets. Abbasi et al. [1] developed a hybrid genetic algorithm that incorporated the Information Gain heuristic with the entropy metric for feature selection. Their method, namely Entropy Weighted Genetic Algorithm (EWGA), when paired with SVM classifiers, has achieved high classification performance on both the movie reviews data and the web forum postings in multiple languages. Other studies have also achieved good classification results by pairing “bag-of-words” n-gram features with various advanced feature selection schema, such as attitude and orientation features [42], topic proximity and syntactic-relations features [26], and sentence-level sentiment extraction [32].

For an excellent survey of recent work on sentiment analysis, please see [33] and [1].

3. Problem definition and background knowledge

In this section, we introduce the problem definition and briefly discuss its scope, and provide brief overviews of relevant concepts to our model.

3.1. Problem definition

The problem can be formally defined as a sentiment prediction problem as described in [4]. The data consists of a collection of N documents $\{x_{d1},\ldots,x_{dV}\}_{d=1}^{N}$, that is, of N examples of the random variables corresponding to the words, $\{X_1,\ldots,X_V\}$. The overall sentiment