Sequential Models for Sentiment Prediction

Yi Mao YMAO@ECN.PURDUE.EDU
School of Electrical and Computer Engineering, Purdue University - West Lafayette

Guy Lebanon LEBANON@PURDUE.EDU
Department of Statistics and School of Electrical and Computer Engineering, Purdue University - West Lafayette

Abstract

We examine the problem of predicting local sentiment flow in documents, and its application to several areas of text analysis. Formally, the problem is stated as predicting an ordinal sequence based on a sequence of word sets. In the spirit of isotonic regression, we develop a variant of conditional random fields that is better suited to handle this problem. Experiments are reported for both sentiment prediction and text summarization, showing the possibility of incorporating sentiment concept into a range of new applications.

1. Introduction

The World Wide Web and other textual databases provide a convenient platform for exchanging opinions. Many documents, such as reviews and blogs, are written with the purpose of conveying a particular opinion or sentiment. Other documents may not be written with the purpose of conveying an opinion, but nevertheless they contain one. Opinions may be considered in several ways, the simplest of which is varying from positive opinion, through neutral, to negative opinion.

Most of the research in information retrieval has focused on predicting the topic of a document, or its relevance with respect to a query. Predicting the document's sentiment would allow matching the sentiment, as well as the topic, with the user's interests. It would also assist in document summarization and visualization. Sentiment prediction was first formulated as a binary classification problem to answer questions such as: "What is the review's polarity, positive or negative?" Pang et al. [5] demonstrated the difficulties in sentiment prediction using solely the empirical rules (specifically, a subset of adjectives), which motivate the use of statistical learning techniques. The task was then refined to allow multiple sentiment levels, facilitating the use of standard techniques for multiclass text categorization [3].

However, sentiment prediction is different from traditional text categorization: (1) in contrast to the categorical nature of topics, sentiments are ordinal variables; (2) several contradicting opinions might coexist, which interact with each other to produce the global document sentiment; (3) context plays a vital role in determining the sentiment. Indeed, sentiment prediction is a much harder task than topic classification tasks such as Reuters or WebKB and current models achieve much lower accuracy.

Rather than using a bag of words multiclass classifier, we model the sequential flow of sentiment throughout the document using a conditional model. Furthermore, we treat the sentiment labels as ordinal variables by enforcing monotonicity constraints on the model's parameters.

2. Local and Global Sentiments

Previous research on sentiment prediction has generally focused on predicting the sentiment of the entire document. A commonly used application is the task of predicting the number of stars assigned to a movie, based on a review text. Typically, the problem is considered as standard multiclass classification or regression using the bag of words representation.

In addition to the sentiment of the entire document, which we call global sentiment, we define the concept of local sentiment as the sentiment associated with a particular part of the text. It is reasonable to assume that the global sentiment of a document is a function of the local sentiment and that estimating the local sentiment is a key step in predicting the global sen-

Appearing in *Proceedings of the ICML Workshop on Learning in Structured Output Spaces*, Pittsburgh, PA, 2006.
Copyright 2006 by the author(s)/owner(s).

timent. Moreover, the concept of local sentiment is useful in a wide range of text analysis applications including document summarization and visualization.

Formally, we view local sentiment as a function on the words in a document taking values in a set O possessing an ordinal relation $\leq$. To determine the local sentiment at a particular word, it is necessary to take the context into account. For example, due to the context the local sentiment at each of the following words `this is a horrible product` is low (in the sense of $(O, \leq)$). Since sentences are natural components for segmenting the semantics of a document, we view local sentiment as a piecewise constant function on sentences. Occasionally we encounter a sentence that violates this rule and conveys opposing sentiments in two different parts. In this situation we break the sentence into two parts and consider them as two sentences. We therefore formalize the problem as predicting a sequence of sentiments $\mathbf{y}, y_i \in O$ based on a sequence of sentences $\mathbf{x}$.

Modeling the local sentiment is challenging from several aspects. The sentence sequence $\mathbf{x}$ is discrete-time and high-dimensional categorical valued, and the sentiment sequence $\mathbf{y}$ is discrete-time and ordinal valued. Regression models can be applied locally but they ignore the statistical dependencies across the time domain. Popular sequence models such as HMM or CRF, on the other hand, typically assume that $\mathbf{y}$ is categorical valued. In this paper we demonstrate the prediction of local sentiment flow using an ordinal version of conditional random fields, and explore the relation between the local and global sentiment notions.

3. Isotonic Conditional Random Fields

Conditional random fields (CRF) [2] are parametric families of conditional distributions $p_\theta(\mathbf{y}|\mathbf{x})$ that correspond to undirected graphical models,

$$\begin{aligned} p_\theta(\mathbf{y}|\mathbf{x}) &= \frac{p_\theta(\mathbf{y},\mathbf{x})}{p_\theta(\mathbf{x})} = \frac{\prod_{c\in C}\phi_c(\mathbf{x}|_c, \mathbf{y}|_c)}{Z(\theta,\mathbf{x})} \qquad (1) \\ &= \frac{\exp\left(\sum_{c\in C}\sum_k \theta_{c,k} f_{c,k}(\mathbf{x}|_c,\mathbf{y}|_c)\right)}{Z(\theta,\mathbf{x})} \qquad \theta_{c,k}\in\mathbb{R} \end{aligned}$$

where C is the set of cliques in the graph and $\mathbf{x}|_c$ and $\mathbf{y}|_c$ are the restriction of $\mathbf{x}$ and $\mathbf{y}$ to variables representing nodes in C. It is assumed above that the potentials ϕ_c are exponential functions of features modulated by decay parameters $\phi_c(\mathbf{x}|_c, \mathbf{y}|_c) = \exp(\sum_k \theta_{c,k} f_{c,k}(\mathbf{x}|_c, \mathbf{y}|_c))$.

CRF have been mostly applied to sequence annotation, where $\mathbf{x}$ is a sequence of words and $\mathbf{y}$ is a sequence of labels annotating the words, for example part-of-speech tags. The standard graphical structure in this case is a chain structure on $\mathbf{y}$ with noisy observations $\mathbf{x}$. In other words, the cliques are $C = \{\{y_{i-1}, y_i\}, \{y_i, x_i\} : i = 1, \ldots, n\}$ leading to the model

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} e^{\sum\limits_i \sum\limits_k \lambda_k f_k(y_{i-1}, y_i) + \sum\limits_i \sum\limits_k \mu_k g_k(y_i, x_i)}. \quad (2)$$

In sequence annotation a standard choice for the feature functions is $f_{\langle\sigma,\tau\rangle}(y_{i-1}, y_i) = \delta_{y_{i-1},\sigma}\delta_{y_i,\tau}$ and $g_{\langle\sigma,w\rangle}(y_i, x_i) = \delta_{y_i,\sigma}\delta_{x_i,w}$ (note that we index the feature functions using pairs rather than k as in (2)). Given a set of iid training samples the parameters are typically estimated by maximum likelihood or MAP using standard numerical techniques such as conjugate gradient or quasi-Newton methods.

Despite the great popularity of CRF in sequence labeling, they are not appropriate for ordinal data such as sentiments. The ordinal relation is ignored in the model (2), and in the case of limited training data the parameter estimates will possess high variance and lead to a poor prediction performance. To reduce the variance of the parameter estimates, we enforce a set of monotonicity constraints on the parameters that are consistent with prior knowledge and with the ordinal structure. The resulting model is a restricted subset of the CRF in (2) and, in accordance with isotonic regression [1], is named isotonic CRF.

Since ordinal variables express a progression of some sort, it is natural to expect some of the binary features in (2) to correlate strongly (either positively or negatively) with the ordinal variables y_i. In such cases, we should expect the presence of the binary feature to increase (or decrease) the conditional probability in a manner consistent with the ordinal relation. We discuss this in greater depth below, in the context of sentiment prediction.

As mentioned before, we are interested in estimating an ordinal-valued function that is piecewise constant on sentences. We therefore denote $\mathbf{x}$ as a sequence of sentences with x_i representing the i-th sentence. The sequence $\mathbf{y}$ represents the sentiments of the sentences. We use the same graphical structure and features f as (2) and slightly modify feature functions

$$g_{\langle\sigma,w\rangle}(y_i, x_i) = \begin{cases} 1 & \text{if } y_i = \sigma \text{ and } w \in x_i \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

since in our case x_i is a sentence rather than a word.

In a preliminary step, we identify a set of words $\mathcal{M}$ that strongly indicate positive or negative sentiment. We expect words that are strongly associated with positive sentiment to have the effect of increasing the probabilities of higher sentiments more than the probabilities of lower sentiments. Since the parameters $\mu_{\langle\sigma,w\rangle}$ represent the effectiveness of the appearance of w with respect to increasing the probability of σ, they are natural candidates for monotonicity constraints. More specifically, for words w that are identified as strongly associated with positive sentiment, we enforce

$$\mu_{\langle\sigma,w\rangle} \leq \mu_{\langle\sigma',w\rangle} \quad \text{if and only if} \quad \sigma \leq \sigma'. \tag{4}$$

Similarly, for words w that are identified as strongly associated with negative sentiment, we enforce

$$\mu_{\langle\sigma,w\rangle} \geq \mu_{\langle\sigma',w\rangle} \quad \text{if and only if} \quad \sigma \leq \sigma'. \tag{5}$$

The motivation behind the above constraints is that the change in probability as a result of the addition of a word w respects the ordering of the parameters $\mu_{\langle\sigma,w\rangle}$. While this result is immediate in the non-conditional case of Markov random fields, things get more complicated in the conditional case because of the fact that the normalization term is a function of $\mathbf{x}$. In the special case of the linear CRF with the feature functions given above, the following proposition shows that the same interpretation holds (proof omitted due to lack of space).

Proposition 3.1. *Let $\boldsymbol{x}$ be a sentence sequence and w' a word such that $w' \notin x_j$ for some j. We denote by $\boldsymbol{x}'$ the sentence sequence that is obtained from $\boldsymbol{x}$ by adding w' to x_j. Then the ratio of the new probability to the original probability $\frac{p(\tilde{\boldsymbol{y}}|\boldsymbol{x}')}{p(\tilde{\boldsymbol{y}}|\boldsymbol{x})}$ for some sentiment sequence $\tilde{\boldsymbol{y}}$ has the same ordering as $\mu_{\langle\tilde{y}_j,w'\rangle}$.*

Conceptually, the parameter estimates for isotonic CRF may be found by maximizing the likelihood or posterior subject to the monotonicity constraints (4)-(5). Since such a maximization is relatively difficult for large dimensionality, we propose a re-parameterization that leads to a much simpler optimization problem. We define new features to replace the features $g_{\langle\sigma,w\rangle}$ that are subjected to monotonicity constraints

$$g^*_{\langle\sigma,w\rangle}(y_i, x_i) = \sum_{\tau:\tau\geq\sigma} g_{\langle\tau,w\rangle}(y_i, x_i). \tag{6}$$

Similarly we re-parameterize the corresponding parameters $\{\mu_{\langle\sigma,w\rangle} : \sigma, w\} \mapsto \{\mu^*_{\langle\sigma,w\rangle} : \sigma, w\}$ such that

$$\mu_{\langle\sigma,w\rangle} = \sum_{\tau:\tau\leq\sigma} \mu^*_{\langle\tau,w\rangle}. \tag{7}$$

Using the new parameterization, we can express isotonic CRF as

$$\begin{aligned} p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp \Bigg(& \sum_i \sum_{\sigma,\tau} \lambda_{\langle\sigma,\tau\rangle} f_{\langle\sigma,\tau\rangle}(y_{i-1}, y_i) \\ & + \sum_i \sum_{\sigma,w\notin\mathcal{M}} \mu_{\langle\sigma,w\rangle} g_{\langle\sigma,w\rangle}(y_i, x_i) \\ & + \sum_i \sum_{\sigma,w\in\mathcal{M}} \mu^*_{\langle\sigma,w\rangle} g^*_{\langle\sigma,w\rangle}(y_i, x_i) \Bigg) \end{aligned}$$

subject to non-negativity (or non-positivity) constraints on $\mu^*_{\langle\sigma,w\rangle}$ for $\sigma > \min(O), w \in \mathcal{M}$. The reparameterized model has the benefit of simpler constraints and its maximum likelihood estimates can be obtained by a very simple adaptation of conjugate gradient or quasi-Newton methods.

3.1. Sentiment Flow as Smooth Curves

The sentence-based definition of sentiment flow is problematic when we want to fit a model (for example to predict global sentiment) that uses sentiment flows from multiple documents. Different documents have different number of sentences and it is not clear how to compare them or how to build a model from a collection of discrete flows of different lengths. Because of this we convert the sentence-based flow to a smooth length-normalized flow that can relate to other flows in a meaningful and robust way.

We assume at this point that the ordinal set O is realized as a subset of $\mathbb{R}$ and that its ordering coincide with the standard ordering on $\mathbb{R}$. In order to account for different lengths, we consider the sentiment flow as a function $h : [0, 1) \to O \subset \mathbb{R}$ that is piecewise constant on the intervals $[0, l), [l, 2l), \ldots, [(k-1)l, 1)$ where k is the number of sentences in the document and $l = 1/k$. Each of the intervals represents a sentence and the value of the function on the interval is the sentiment of that sentence.

To create a more robust representation we smooth out the discontinuous function by convolving it with a smoothing kernel. The resulting sentiment flow is a smooth curve $f : \mathbb{R} \to \mathbb{R}$ that can be easily related or compared to similar sentiment flows of other documents (see Figure 2 for an example). We can then define natural distances between two flows, for example the L_p distance

$$d_p(f_1, f_2) = \left(\int_0^1 |f_1(r) - f_2(r)|^p \, dr \right)^{1/p} \tag{8}$$

for use in a k-nearest neighbor model for relating local sentiment flow and global sentiment.

4. Experiments

To examine the ideas proposed in this paper we implemented isotonic CRF, and the normalization and smoothing procedure, and experimented with a small dataset of 249 movie reviews, randomly selected from the Cornell sentence polarity dataset v1.0[1], all written by the same author. The code for isotonic CRF is a modified version of the quasi-Newton implementation in the Mallet toolkit. In order to check the accuracy and benefit of the local sentiment predictor, we hand-labeled the local sentiments of each of these reviews. We assigned for each sentence one of the following values in $O \subset \mathbb{R}$: 2 (highly praised), 1 (something good), 0 (objective description), -1 (something that needs improvement) and -2 (strong aversion). In the few cases where two different opinions were present in the same sentence, possibly connected by the word "but" or "however", we divided the sentence into two sentences each exhibiting a consistent sentiment.

4.1. Sentence Level Prediction

To evaluate the prediction quality of the local sentiment we compared the performance of naive Bayes, SVM (using the default parameters of SVM^{light}), CRF and isotonic CRF. Figure 1 displays the testing accuracy and distance of predicting the sentiment of sentences as a function of the training data size averaged over 20 cross-validation train-test split.

The dataset presents one particular difficulty where more than 75% of the sentences are labeled objective (or 0). As a result, the prediction accuracy for objective sentences is over-emphasized. To correct for this fact, we report our test-set performance over a balanced (equal number of sentences for different labels) sample of the labeled sentences. Note that since there are 5 labels, random guessing yields a baseline of 0.2 accuracy and guessing 0 always (middle point) yields a baseline of 1.2 distance.

As described in Section 3, for isotonic CRF, we obtained 300 words to enforce monotonicity constraints. The 150 words that achieved the highest correlation with the sentiment were chosen for non-negativity constraints. Similarly, the 150 words that achieved the lowest correlation were chosen for non-positivity constraints. Figure 1 (right) displays the top 15 words of the two lists.

The results in Figure 1 indicate that by incorporating the sequential information, the two versions of CRF perform consistently better than SVM and naive Bayes. The advantage of setting the monotonicity constraints in CRF is elucidated by the distance performance criterion. This criterion is based on the observation that in sentiment prediction, the cost of misprediction is influenced by the ordinal relation on the labels, rather than the 0-1 error rate.

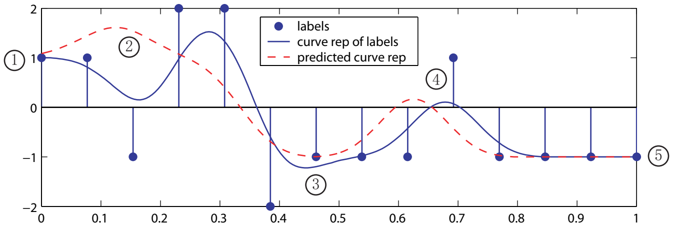


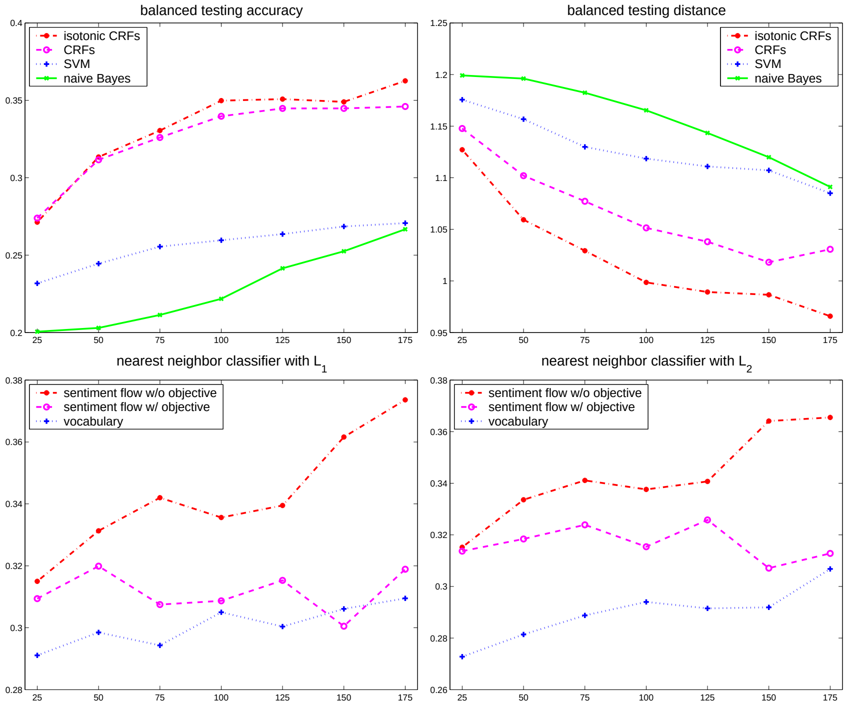
Figure 2. Sentiment flow and its smoothed curve representation. The blue circles indicate the labeled sentiment of each sentence. The blue solid curve and red dashed curve are smoothed representations of the labeled and predicted sentiment flows. Only non-objective labels are kept in generating the two curves.

4.2. Global Sentiment Prediction

We also evaluated the contribution of the local sentiment analysis in helping to predict the global sentiment of documents. We compared a nearest neighbor classifier for the global sentiment, where the representation varies from bag of words to normalized and smoothed local sentiment representation (with and without objective sentences). The smoothing kernel was a bounded Gaussian density (truncated and renormalized) with $\sigma^2 = 0.2$. Figure 2 displays an example of the discrete and smoothed versions of the local sentiments, as well as the smoothed version of the sentiment flow predicted by isotonic CRF.

Figure 1 displays test-set accuracy of global sentiments as a function of the training data size. The distance in the nearest neighbor classifier was either L_1 or L_2 for the bag of words representation or their continuous version (8) for the local sentiment curve representation. The results indicate that the classification performance of the local sentiment representation is better than the bag of words representation. In accordance with the conclusion of [4], removing objective sentences (that correspond to sentiment 0) significantly increased the performance of sentiment flow by 20.7%. This supports the conclusion that the local sentiment flow of objective sentences is largely irrelevant and removing objective sentences improves performance as the model estimates achieve lower variance with only a slight increase in bias.

[1] Available at `http://www.cs.cornell.edu/People/pabo/movie-review-data`



great	superb	memorable
enjoyable	mood	perfection
outstanding	performance	enjoyed
certain	considerable	wonderfully
worth	beautifully	delightfully
too	didnt	just
failed	unnecessary	couldnt
i	no	satire
contrived	wasnt	uninspired
lacked	boring	tended

	L_1		L_2	
vocabulary	0.3095		0.3068	
sentiment flow with objective sentences	0.3189	3.0%	0.3128	1.95%
sentiment flow without objective sentences	0.3736	20.7%	0.3655	19.1%

Figure 1. Top: Local sentiment prediction. Balanced testing results for naive Bayes, SVM, CRF and isotonic CRF (left, center) and lists of 15 words with the largest positive (right, top) and negative (right, bottom) correlation coefficients. Bottom: Global sentiment prediction (4-class labeling). Comparison of nearest neighbor classifier using vocabulary and sentiment flow (left, center) and accuracy results and relative improvement when training size equals 175 (right).

4.3. Text Summarization

We demonstrate the potential usage of sentiment flow for text summarization with a very simple example. The text below shows the result of summarizing the movie review in Figure 2 by keeping only sentences associated with the start, the end, the top, or the bottom of the predicted sentiment curve. The number before each sentence relates to the circled number in Figure 2.

1 What makes this film mesmerizing, is not the plot, but the virtuoso performance of Lucy Berliner (Ally Sheedy), as a wily photographer, retired from her professional duties for the last ten years and living with a has-been German actress, Greta (Clarkson). 2 The less interesting story line involves the ambitions of an attractive, baby-faced assistant editor at the magazine, Syd (Radha Mitchell), who lives with a boyfriend (Mann) in an emotionally chilling relationship. 3 We just lost interest in the characters, the film began to look like a commercial for a magazine that wouldn't stop and get to the main article. 4 Which left the film only somewhat satisfying; it did create a proper atmosphere for us to view these lost characters, and it did have something to say about how their lives are being emotionally torn apart. 5 It would have been wiser to develop more depth for the main characters and show them to be more than the superficial beings they seemed to be on screen.

Alternative schemes for extracting specific sentences may be used to achieve different effects, depending on the needs of the user. We plan to experiment further in this area by combining local sentiment flow and standard summarization techniques.

5. Summary

In this paper, we address the prediction and application of the local sentiment flow concept. As existing models are inadequate for a variety of reasons, we introduce the isotonic CRF model that is suited to predict the local sentiment flow. This model achieves better performance than the standard CRF as well as non-sequential models such as SVM. We also demonstrate the usefulness of the local sentiment representation for global sentiment prediction and text summarization.

References

[1] R. E. Barlow, D.J. Bartholomew, J. M. Bremner, and H. D. Brunk. *Statistical inference under order restrictions; the theory and application of isotonic regression.* Wiley, 1972.

[2] J. Lafferty, F. Pereira, and A. McCallum. Conditional random fields: Probabilistic models for segmenting and

labeling sequence data. In *International Conference on Machine Learning*, 2001.

[3] B. Pang and L. Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of ACL-05*.

[4] B. Pang and L. Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of ACL-04*.

[5] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of EMNLP-02*.