Iterative Threshold-based Naïve Bayes Classifier: an efficient Tb-NB improvement

Il classificatore Iterative Threshold-based Naïve Bayes: un efficace miglioramento di Tb-NB

Maurizio Romano, Gianpaolo Zammarchi, and Giulia Contu

Abstract While analyzing online reviews on Booking.com, we proposed an ad-hoc classification model (Threshold-based Naïve Bayes Classifier, Tb-NB) to evaluate Customer Satisfaction, starting from the reviews' content, and predicting them as positive/negative. The log-likelihood ratios attributed to each word included in a review are then used to estimate a numeric sentiment score. In this paper we propose an improved version of Tb-NB called "iterative" Tb-NB. It results in a second step of Tb-NB: starting from the output of Tb-NB and reclassifying reviews with a probabilistic approach, it refines iteratively the threshold value used to classify a given subset of reviews.

Abstract *Analizzando le recensioni online presenti su Booking.com, è stato proposto un classificatore ad hoc (Threshold-based Naïve Bayes Classifier, Tb-NB) per valutare la Customer Satisfaction che, partendo dal contenuto delle recensioni stesse, le classifica in positive/negative. La log-verosimiglianza attribuita ad ogni parola contenuta in una recensione è successivamente utilizzata per stimare un sentiment score numerico. In questo paper si propone una versione migliorata chiamata "iterative". La nuova versione si comporta come un secondo step: a partire dall'output di Tb-NB si riclassificano le recensioni con un approccio probabilistico, raffinando la regola-soglia di classificazione per un sottoinsieme di recensioni iterazione dopo iterazione.*

Key words: Threshold-based Naïve Bayes Classifier, Iterative Threshold-based Naïve Bayes Classifier, Customer Satisfaction, Sentiment Analysis, General Sentiment Decomposition

---

Maurizio Romano
University of Cagliari, Cagliari, Italy, e-mail: romano.maurizio@unica.it

Gianpaolo Zammarchi
University of Cagliari, Cagliari, Italy, e-mail: gp.zammarchi@unica.it

Giulia Contu
University of Cagliari, Cagliari, Italy, e-mail: giulia.contu@unica.it

1 Introduction

Analyzing textual data has proved to be extremely useful in many field [7, 1, 4] but, likewise the language is complex so is the data [3]. In [2, 6] we have proposed Threshold-based Naïve Bayes Classifier (Tb-NB), a framework able to produce a good sentiment classifier with a particular focus on the model interpretability. In this paper we introduce an "iterative" step able to improve the performance of Tb-NB while reclassifying subsets of observations. Furthermore, we compare the performance of the new proposal with the previous one using the same Booking.com reviews data.

2 The data

For this study, with an ad hoc web scraping Python program, we have collected two datasets from Booking.com with a total of 127 features:

- Hotels dataset (86 features): Hotel (3), Review (8), Reviewer (2), Booking's score (11), Accommodation (32), Guest (8), Length of stay (6), Other info (4), Score components (12);

- Comments dataset (41 features): Hotel (2), Comment (6), Reviewer (2), Accommodation (16), Guest (4), Length of stay (3), Other info (2), Score components (6).

More in detail, the web-scraped data, is related to:

- 619 hotels located in Sardinia;
- 66,237 reviews, divided in 106,800 comments (in Italian or English): 44,509 negative + 62,291 positive;
- Period: Jan 3, 2015 – May 27, 2018.

3 Iterative Threshold-based Naïve Bayes Classifier

Considering a Natural Language text corpora as a set of reviews $\boldsymbol{r}$ s.t.:

$$r_i = comment_{pos_i} \cup comment_{neg_i}$$

where $comment_{pos}$ ($comment_{neg}$) are set of words (a.k.a. comments) composed by only positive (negative) sentences, and one of them can be equal to $\emptyset$, the basic features of the Threshold-based Naïve Bayes classifier applied to reviews' content are as follows. For a specific review r and for each word w ($w \in$ *Bag-of-Words*), we consider the log-odds ratio of w,

$$
\begin{aligned}
LOR(w) &= \log\left[\frac{P(c_{neg}|w)}{P(c_{pos}|w)}\right] \approx \\
&\approx \log\left[\frac{P(w|c_{neg})}{P(\bar{w}|c_{neg})}\cdot\frac{P(w|c_{pos})}{P(\bar{w}|c_{pos})}\cdot\frac{P(c_{neg})}{P(c_{pos})}\right] = \ldots = \\
&\approx pres_w + abs_w
\end{aligned}
$$

where $c_{pos}(c_{neg})$ are the proportions of observed positive (negative) comments whilst $pres_w$ and abs_w are the log-likelihood ratios of the events $(w \in r)$ and $(w \notin r)$, respectively.

While computing those values for all the w ($w \in$ *Bag-of-Words*) words, it is possible to obtain an output such that reported in Table 1, where we report c_{pos}, c_{neg}, $pres_w$ and abs_w for each word in the considered *Bag-of-Words*.

Table 1 Threshold-based Naïve Bayes output

	w_1	w_2	w_3	w_4	w_5	...
$P(w_i\|c_{neg})$	0.011	0.026	0.002	0.003	0.003	...
$P(w_i\|c_{pos})$	0.007	0.075	0.005	0.012	0.001	...
$pres_{w_i}$	0.411	-1.077	-1.006	-1.272	1.423	...
abs_{w_i}	-0.004	0.052	0.003	0.008	-0.002	...

We have then used cross-validation to estimate a parameter τ such that: c is classified as "negative" if $LOR(c) > \tau$ or as "positive" if $LOR(c) \leq \tau$.

We knew now that the Threshold-based Naïve Bayes Classifier has engaging performances with respect to other algorithms [6, 2]. However, still, there are some problems during the classification in Pos/Neg of text data that have a LOR-value close to the selected value of τ. In particular, we can observe this in Fig. 1: if the LOR-value is far from the estimated threshold, there are few chances to obtain an incorrect classification.

Considering that the LOR-value is a numeric sentiment value, it is obvious than if this value is close to the threshold even a word with an irrelevant sentiment value could be classified incorrectly.

To improve the performance of the Threshold-based Naïve Bayes Classifier for LOR-values located close to the threshold, we add a new step that refines the threshold while classifying only those comments that are close to the selected value of τ. Then, we classify again those comments with a new estimated value of τ.

The "augmented" Iterative Threshold-based Naïve Bayes works as follows:

1. a proportion of observations close to τ, and located in the uncertainness area, are marked for being reclassified.
2. A new decision rule is created for those observations only, estimating a new value of τ.
3. Observations in the uncertainness area are re-classified and Steps 1-3 are repeated until the proportion of cases in the uncertainness area reduces consistently.

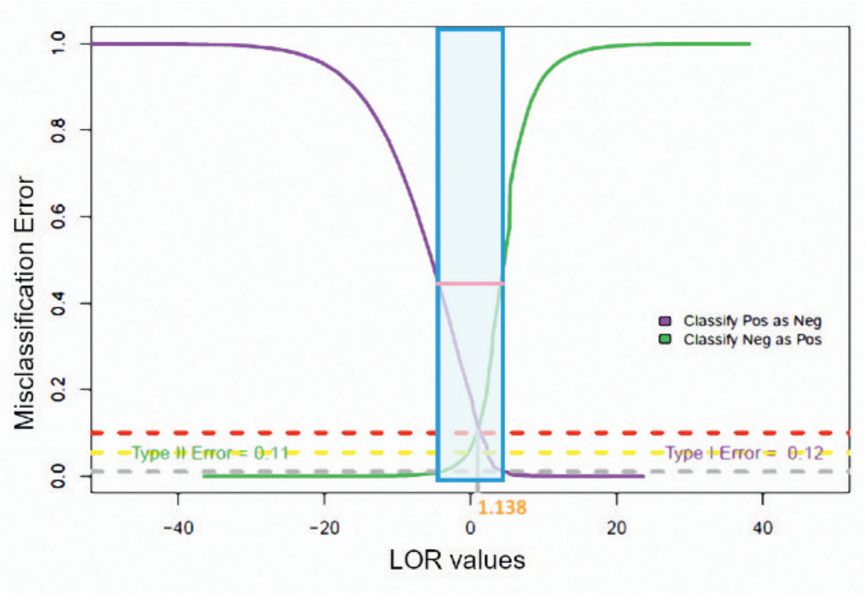

Fig. 1 Threshold-based Naïve Bayes uncertainness area

We can then transform that logic in an algorithm that better describes such a process:

---

1. Set $i = 0$
2. Map the distribution of LOR-values into two functions:

$$x = LOR(\pi_c)$$

$$f_{pos} : x \mapsto [0;1], f_{neg} : x \mapsto [0;1] \qquad \forall x \in (-\infty;+\infty)$$

   s.t. the input (x) is the LOR-value of a comment c and the output of f_{pos} (f_{neg}) is the probability that c is positive (negative).
3. Define max_{pos}, max_{neg} s.t.:

$$max_{pos} = \underset{x \in (-\infty;+\infty)}{\operatorname{argmax}} f_{pos}(x)$$

$$max_{neg} = \underset{x \in (-\infty;+\infty)}{\operatorname{argmax}} f_{neg}(x)$$

4. Compute the intersection point τ_i s.t.

$$\tau_i = \begin{cases} \underset{x \in (max_{pos}; max_{neg})}{\operatorname{argmin}} & (f_{neg}(x) - f_{pos}(x)) \quad \text{if } max_{pos} < max_{neg} \\ \underset{x \in (max_{neg}; max_{pos})}{\operatorname{argmin}} & (f_{neg}(x) - f_{pos}(x)) \quad \text{if } max_{pos} > max_{neg} \end{cases}$$

5. if $max_{pos} < max_{neg}$ update the decision rule:
   $LOR(\pi_c) > \tau_i \rightarrow$ c is classified as "Negative"
   $LOR(\pi_c) \leq \tau_i \rightarrow$ c is classified as "Positive"
6. otherwise, if $max_{pos} > max_{neg}$ update the decision rule:
   $LOR(\pi_c) > \tau_i \rightarrow$ c is classified as "Positive"
   $LOR(\pi_c) \leq \tau_i \rightarrow$ c is classified as "Negative"
7. Set $i = i + 1$
8. Repeat steps 1-7 with a proportion ω of comments located in the interval $\tau_i \pm \partial\omega$ until a stopping criterion is satisfied.

The proportion of observations marked for being re-classified is estimated while considering the uncertainness area around τ. Using a large value of ω leads to results that are similar to those obtained from Threshold-based Naïve Bayes classifier, whilst using a low value for ω might lead to consider few observations only, thus producing a decision rule which overfits the data. Our experiments suggest that, empirically, a good trade-off is reached by setting $\omega = 0.20$.

Whereas, the **stopping criterion** of the iterative Tb-NB depends on at least one of the following conditions:

1. the number of cases (positive and negative comments) used to produce a reliable distribution of LOR-values, and the associated function that map them, is too small;
2. there is no intersection point (other than zero) between $f_{pos}(x)$, *and* $f_{neg}(x)$ and $max_{pos} \neq max_{neg}$. The two distributions are well separated, thus the τ estimated by the standard Threshold-based Naïve Bayes Classifier at the previous step is the final value of τ;
3. $max_{pos} = max_{neg}$. We cannot distinguish the positive distribution from the negative one. Thus, there is not a geometric solution to this problem and the classification of those comments has to be done with the τ estimated by the Threshold-based Naïve Bayes Classifier.

Moreover, it is interesting to highlight how empirical results suggest that the stopping criterion usually is reached after no more than two iterations ($0 \leq i \leq 2$).

Furthermore, in Table 2 we have compared the performance of the proposed classifier (iTb-NB) with the previous version (Tb-NB) and that of other well known competitors, in particular: Logistic Regression (LOG), Random Forest (RF), standard Naïve Bayes (NB E1071 [5]), Naïve Bayes using kernel estimated densities (NB KlaR [8]), Decision Trees (CART), Linear Discriminant Analysis (LDA) and Support Vector Machine (SVM).

It is possible to notice that iTb-NB is the most accurate classifier with respect to four out of the five measures reported in the table.

Table 2 Performance metrics on raw data using 5-fold cross validation

Classifier	ACC	Sensitivity	Fall-out	F1	MCC
Tb-NB	0.911	**0.929**	0.117	0.926	0.813
iTb-NB	**0.925**	0.923	**0.072**	**0.939**	**0.829**
LOG	0.850	0.884	0.532	0.877	0.361
RF	0.811	0.873	0.591	0.849	0.303
NB(E1071)	0.806	0.804	0.389	0.834	0.390
NB(KLAR)	0.806	0.804	0.389	0.834	0.390
CART	0.768	0.842	0.587	0.815	0.272
LDA	0.764	0.860	0.641	0.816	0.246
SVM	0.793	**0.929**	0.290	0.771	0.621
average	0.826	0.872	0.401	0.851	0.469

Notes: ACC = Accuracy; F1 = F1-score; MCC = Matthews Correlation Coefficient

4 Conclusions

We have proposed an improved, and "iterative", Threshold-based Naïve Bayes Classifier. Furthermore, we have compared the performance of the new proposal with the previous one using the same Booking.com reviews data. Reclassifying subsets of observations has then been proved as a way to reduce the misclassification rate and hence boosting the performance of Tb-NB whilst keeping the original interpretability of the results.

References

1. Boyd, D., Crawford, K.: Critical questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon. Information, Communication & Society **15**(5), 662–679 (2012)
2. Conversano, C., Romano, M., Mola, F.: Hotel search engine architecture based on online reviews' content. In: Arbia, G., Peluso, S., Pini, A., Rivellini, G. (eds.) Smart Statistics for Smart Applications. Book of Short Papers SIS2019, pp. 213-218. Pearson, Milan (2019)
3. Goldberg, Y.: Neural Network Methods in Natural Language Processing. Synthesis Lectures on Human Language Technologies **10**(1), 1–309 (2017)
4. Halevy, A., Norvig, P., Pereira, F.: The Unreasonable Effectiveness of Data. IEEE Intelligent Systems **24**(2), 8–12 (2009)
5. Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F.: Misc. Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien (2019).
6. Romano, M., Frigau, L., Contu, G., Mola, F., Conversano, C.: Customer Satisfaction from Booking. In: Mieli, M., Volpe, C. (eds.) Selected papers Conferenza GARR_18 Data (R)evolution, pp. 111-118. Associazione Consortium GARR, Cagliari (2018)
7. Schmunk, S., Höpken, W., Fuchs, M., Lexhagen, M.: Mobile Social Travel Recommender System. In: Xiang, Z., Tussyadiah, I. (eds.) Information and Communication Technologies in Tourism 2014, pp. 3-16. Springer International Publishing, Cham (2013)
8. Weihs, C., Ligges, U., Luebke, K., Raabe, N.: klaR Analyzing German Business Cycles. In: Baier, D., Decker, R., Schmidt-Thieme, L. (eds.) Data Analysis and Decision Support, pp. 335-343. Springer-Verlag, Berlin/Heidelberg (2005)