A Game Theory Approach to Feature Selection for Text Classification

Giorgio Maria Di Nunzio and Nicola Orio

[1] Dept. of Information Engineering – University of Padua
giorgiomaria.dinunzio@unipd.it
[2] Dept. of Cultural Heritage: Archaeology and History of Art, Cinema and Music – University of Padua
nicola.orio@unipd.it

Abstract. This paper presents an initial study on how game theory can be applied to select positive and negative features for a text classification task. The proposed approach builds upon previous work where the players are the positive and negative categories, while the strategies are the choices of selecting a given feature as positive or negative. We explore how the payoff matrix can be described in a more general way and analyze how the choice of a given payoff matrix influences the effectiveness of a text classification engine.

1 Introduction

The automatic classification of textual documents [4] is a process that requires the choice of a set of representative features in order to tell whether a document belong to a particular topic or not. For text classification, a straightforward and easy solution is to use the entire set of words present in all the documents of the dataset. However, there are situations where the choice of a subset of words, rather than the entire vocabulary, leads to significant performance improvement [2, 3]. This choice is based on the output of functions usually called feature selection metrics; these functions can can be roughly divided in two sets: one-sided and two-sided metrics. One-sided metrics select the features most indicative of membership for one class only (positive features), while two-sided metrics consider the features most indicative of either membership or non-membership (positive or negative features) for a class. 'Positive' and 'negative' have two meanings in the context of text classification: i) a feature can be positive or negative depending on whether it is indicative or not for a class; ii) a class can be positive or negative. The second meaning is usually related to a binary classification problem where, given a set of classes, the positive class is the one we are interested in and the negative class is the union of all the remaining classes. Positive features are intuitively the first choice to describe the documents of a class, but also negative features are important to improve the quality of classification [2]. In fact, "a judicious combination [of positive and negative features] shows great potential and practical merits" for (probabilistic) text classifiers [6].

Following this idea, Azam and Yao [1] propose a game theory framework for selecting features that are representative for both positive and negative classes. According to their approach, the choice of including or discarding a positive/negative feature can be modeled as a two-player game. Players are the positive and negative categories, which indipendently choose whether to keep or discard a given feature. As in any game, the payoff of each player depends on the combined choices of the two categories, which thus try to maximize their payoff for each feature eventually reaching the Nash equilibrium. In this paper, we discuss a generalization of the game theory framework proposed by [1] and show an interpretation, inspired by language models, that has an additional parameter that can be optimized to achieve a better classification performance.

2 Mathematical Background

In this section, we present the mathematical notation and the basic definition of game theory to describe the problem of feature selection. Given a set of classes $C = \{c_1, \ldots, c_i, \ldots, c_n\}$ and a set of documents $D = \{d_1, \ldots, d_j, \ldots, d_m\}$, we indicate the positive class i with $c^+ = c_i$ and the corresponding negative class with $c^- = C \setminus c_i$. If documents are described by a vocabulary of words $V = \{w_1, \ldots, w_k, \ldots, w_v\}$, then the probability of word w_k given the positive (or negative) category is $P(w_k|c^+)$ (or $P(w_k|c^-)$).

In game theory, a game G is a triple $G = \{P, S, F\}$, where P is the set of players, S a set of strategies, and F a set of payoff functions. In the context of feature selection, we have two 'players' $P = \{c^+, c^-\}$, two 'strategies' $S = \{s_1 : keep, s_2 : discard\}$, where s_1 means the choice of keeping the feature and s_2 to discard it, and two sets of payoff functions $F = \{u_{c^+}, u_{c^-}\}$ one for each player. The payoff set for the two players can be defined as:

$$u_{c^+} = \left\{u(s_1^+, s_1^-), u(s_1^+, s_2^-), u(s_2^+, s_1^-), u(s_2^+, s_2^-)\right\}$$
$$u_{c^-} = \left\{u(s_1^-, s_1^+), u(s_1^-, s_2^+), u(s_2^-, s_1^+), u(s_2^-, s_2^+)\right\}$$

where, for example, $u(s_1^-, s_2^+)$ is the payoff of the 'negative' player that plays the action 'keep the feature' while the 'positive' player played the action 'discard the feature'. The payoff table used to play the game is shown in Table 1.

The typical approach in non-cooperative game theory is to analyze the payoff function in order to find strategies that form a Nash equlibrium, that is a combination of choices where no player can improve its payoff by changing its strategy unilaterally. For instance, given Table 1, the combined strategy (s_1^+, s_1^-) (both players keeping the feature) is a Nash equilibrium only if player c^+ will have a smaller payoff with the combined strategy (s_2, s_1) and player c^- will have a smaller payoff with the combined strategy (s_1, s_2). Even if it is not guaranteed that each payoff matrix has a Nash equilibrium with only deterministic choices (called pure strategies), in this initial contribution we will not discuss the case of a Nash equilibrium with probabilistic choices (called mixed strategies).

In [1], the utility of players in the payoff matrix is computed in two ways according to the particular situation: i) when the action is 'keep the feature'

Table 1. Payoff matrix for a two-player and two-action game.

		c^-	
		s_1	s_2
c^+	s_1	$u(s_1^+, s_1^-), u(s_1^-, s_1^+)$	$u(s_1^+, s_2^-), u(s_1^-, s_2^+)$
	s_2	$u(s_2^+, s_1^-), u(s_2^-, s_1^+)$	$u(s_2^+, s_2^-), u(s_2^-, s_2^+)$

Table 2. Payoff matrix for feature selection.

		c^-	
		s_1	s_2
c^+	s_1	$\frac{1}{2}\left(P(w_j\|c^+) + P(w_j\|c^-)\right)$, $\frac{1}{2}\left(P(w_j\|c^-) + P(w_j\|c^+)\right)$	$P(w_j\|c^+), P(\overline{w}_j\|c^-)$
	s_2	$P(\overline{w}_j\|c^+), P(w_j\|c^-)$	$\frac{1}{2}\left(P(\overline{w}_j\|c^+) + P(\overline{w}_j\|c^-)\right)$, $\frac{1}{2}\left(P(\overline{w}_j\|c^-) + P(\overline{w}_j\|c^+)\right)$

or 'discard the feature' for both players, the utility is the average between the probability of finding (or not finding) the word w_j in the positive and negative class and it is equal for both players; ii) when only one player decides to 'keep the feature' (or 'discard the feature') the utility for that player is just the probability of the word given the class. In Table 2, we show the utilities for each pair of strategies in for the game of feature selection.

3 Our proposal

Starting from the utilities defined in Table 2, we propose a more general interpretation for computing the payoff matrix based on the idea of language models to interpolate the probability of the word in the class and in the collection with a parameter λ [5] that, in our case, can be optimized for each pair of actions of the two players.

For example, if we use λ to weight the action of one player and $(1-\lambda)$ the action of the other player we can define the utility of 'keeping the feature' for the positive category and 'discard the feature' for the negative category as:

$$u(s_1^+, s_2^-) = \lambda_{1^+2^-} P(w_j|c^+) + (1 - \lambda_{1^+2^-}) P(\overline{w}_j|c^-) \quad (1)$$

$$u(s_2^-, s_1^+) = \lambda_{2^-1^+} P(\overline{w}_j|c^-) + (1 - \lambda_{2^-1^+}) P(w_j|c^+) \quad (2)$$

where $\lambda_{1^+2^-}$ and $\lambda_{2^-1^+}$ are two different parameters for the two actions that, in [1], are equal to one $\lambda_{1^+2^-} = \lambda_{2^-1^+} = 1$. Similarly, when the actions are concordant, for example both players keep the feature, we have:

$$u(s_1^+, s_1^-) = \lambda_{1^+1^-} P(w_j|c^+) + (1 - \lambda_{1^+1^-}) P(w_j|c^-) \quad (3)$$

$$u(s_1^-, s_1^+) = \lambda_{1^-1^+} P(w_j|c^+) + (1 - \lambda_{1^-1^+}) P(w_j|c^-) \quad (4)$$

In order to reproduce the approach presented in [1], the parameters should be equal to $\frac{1}{2}$, $\lambda_{1^+1^-} = \lambda_{1^-1^+} = \frac{1}{2}$. In Table 3, we show the payoff matrix with the interpolated probabilities for all the possible actions.

Table 3. Payoff matrix with interpolated probabilities.

		c^-	
		s_1	s_2
c^+	s_1	$\lambda_{1^+1^-}P(w_j\|c^+) + (1-\lambda_{1^+1^-})P(w_j\|c^-)$, $\lambda_{1^-1^+}P(w_j\|c^-) + (1-\lambda_{1^-1^+})P(w_j\|c^+)$	$\lambda_{1^+2^-}P(w_j\|c^+) + (1-\lambda_{1^+2^-})P(\overline{w}_j\|c^-)$, $\lambda_{2^-1^+}P(\overline{w}_j\|c^-) + (1-\lambda_{2^-1^+})P(w_j\|c^+)$
	s_2	$\lambda_{2^+1^-}P(\overline{w}_j\|c^+) + (1-\lambda_{2^+1^-})P(w_j\|c^-)$, $\lambda_{1^-2^+}P(w_j\|c^-) + (1-\lambda_{1^-2^+})P(\overline{w}_j\|c^+)$	$\lambda_{2^+2^-}P(\overline{w}_j\|c^+) + (1-\lambda_{2^+2^-})P(\overline{w}_j\|c^-)$, $\lambda_{2^-2^+}P(\overline{w}_j\|c^-) + (1-\lambda_{2^-2^+})P(\overline{w}_j\|c^+)$

4 Preliminary Analysis and Future Work

In our initial analysis, we used the toy example shown in [1] to understand how different values of the parameters can shift the Nash equilibrium from one strategy to another. This can be easily demonstrated for some of the words in that experiment, but it is less clear whether this is possible in situations where the same feature has almost the same probability of appearing in both classes. However, the analysis of this simple example may lead to superficial conclusions since the numbers used in that experiment do not reflect values of the probabilities computed for a real text collections (where the magnitude of the values of probabilities is very rarely above 10^-1). For this reason, our study will focus on the next steps:

- use standard text collections (i.e. Reuters 21578, 20 Newsgroups, Reuters RCV1) to study whether it is possible to shift Nash equilibrium by changing the values of the paramters lambda;
- find optimal lambdas by means of k-fold cross-validation approaches and test whether this values of the parameters can improve text classification performances significantly;
- study whether a Nash equilibrium with mixed strategies is possible (and effective in terms of classification performance) on real datasets.

References

1. Nouman Azam and JingTao Yao. *Proc. of Rough Sets, Fuzzy Sets, Data Mining and Granular Computing: RSFDGrC 2011*, chapter Incorporating Game Theory in Feature Selection for Text Categorization, pages 215–222. Springer, 2011.
2. George Forman. An extensive empirical study of feature selection metrics for text classification. *J. of Mach. Learn. Res.*, 3:1289–1305, 2003.
3. Vipin Kumar and Sonajharia Minz. Feature Selection: A Literature Review. *Smart CR*, 4(3):211–229, 2014.
4. Fabrizio Sebastiani. Machine Learning in Automated Text Categorization. *ACM Comput. Surv.*, 34(1):1–47, March 2002.
5. ChengXiang Zhai. Statistical Language Models for Information Retrieval A Critical Review. *Found. Trends Inf. Retr.*, 2(3):137–213, March 2008.
6. Zhaohui Zheng, Xiaoyun Wu, and Rohini Srihari. Feature Selection for Text Categorization on Imbalanced Data. *SIGKDD Explor. Newsl.*, 6(1):80–89, June 2004.