
Text classification with a Primal SVM endowed with domain knowledge

Emilio Parrado-Hernandez*
Dep. of Signal Processing and Communications
Universidad Carlos III de Madrid
Avenida de la Universidad 31
28911 Leganes (Madrid) Spain
emipar@tsc.uc3m.es

David R. Hardoon†
Computational Statistics & Machine Learning Centre
Department of Computer Science
University College London
London, WC1E 6BT, U.K.
D.Hardoon@cs.ucl.ac.uk

Abstract

In this paper we solve a document classification task by incorporating prior/domain knowledge onto the SVM. The algorithm consists in to learn a prior classifier in the primal space (words) from an ‘external’ source of information to the text classification itself: patterns of reader’s eyes movements when reading relevant words for discriminating texts. This prior weight vector is then plugged into the SVM optimisation in the primal space. Experimental results include a comparison of the proposed algorithm with plain SVM classifiers and with an alternative way of mixing textual and eye information based on the SVM-2K.

1 Introduction

A prominent disadvantage of current information retrieval systems is that these systems require the user to formulate an informative query which defines their search interest. This requirement of formulating an informative query can be extremely challenging when the true interest of the user is ambiguous to the user themselves. An illustrative example is when wanting to search “something” on a search engine and not quite sure *how* to formulate this “something” in text. Recent studies by [6, 5, 7] have explored the feasibility of inferring implicit relevancy for information retrieval tasks from eye movements to infer users intentions. This was with the ultimate goal of incorporating implicit with explicit information to improve on search query. Recent work by [4] had shown that it was possible to combine eye movements with textual information by using the eye movements to formulate an information retrieval query. The combined formulation was used to successfully rank unseen documents with respect to their relevance to the current interests of the users.

In our work we seek to extend on the feasibility study conducted by [4] in an attempt to observe whether it is possible to use eye movements, measured while reading documents in various categories, as an informative prior for a document classification task in the SVM. The novelty introduced in this paper with respect to [4] is that we use the eye movements information to infer a prior relevance for each word and for each topic. We use eye information not related to the textual category

*<http://www.tsc.uc3m.es/~emipar>

†<http://www.davidroihardoon.com>

of the analysed document since our intuition is that there might be a common eye movement pattern when readers spot words that amount the same information to decide the category of the document, regardless of the particular category.

To endow the final classifier, a SVM with linear kernel, with the domain knowledge contained in the prior relevances we need to introduce some changes in the SVM formulation. We formulate the proposed Prior-SVM in the primal space since our prior is in the input variables (words) rather than on the input patterns (documents). Therefore we adopt the Primal SVM formulation as detailed in [2] for our algorithm, that we term Prior Primal SVM.

The paper is laid out as follows; Section 2 reviews the building blocks of the text classification system of study; In Section 3 we derive and present our proposed Prior SVM technique. This Prior SVM is the central block of the text classification system described in Section 4 We demonstrate the Prior-SVM’s ability to incorporate prior/domain knowledge in Section 5 where we extend on the feasibility study presented by [4] for inferring document relevancy from eye movement. In Section 6 we bring forward our concluding discussion.

2 Background Components

In this section we review the building blocks of the text classification system. The introduction of the Prior Primal SVM, is more detailed in the next section.

2.1 SVM and SVR

The support vector machine (SVM) is a powerful classification tool that has been successfully applied in many applications [3]. The SVM aims to discriminate between two classes by finding the optimal separating hyperplane. Given an input set X let $y \in \{\pm 1\}$ and $f(\mathbf{x}) = \mathbf{w}'\mathbf{x} + b$. The standard SVM solves the following regularisation problem

$$\min_f \frac{1}{2} \|\mathbf{w}\|_2^2 + C \sum_{i=1}^n \ell(y_i f(\mathbf{x}_i)),$$

with the hinge loss $\ell(u) = H_1(u) = (1 - u)_+$ where $(u)_+ = u$ if $u \geq 0$ and 0 otherwise.

SVM’s are naturally constructed for classification but are also easily applied to regression type problem while retaining their maximal margin features. Known as Support Vector Regression (SVR) these can be achieved by introducing the ϵ -insensitive loss function in the optimisation problem. This way, the SVR aims at fitting a tube of radius ϵ to the patterns. The regularisation parameter C controls the amount of patterns that are allowed to lie outside the tube, what controls overfitting.

While our classifiers are designed in the input space (words), since we are in a high dimension (much more words than documents) scenario, for the regressors we make use of the kernel trick [9]. This trick consists in to use a kernel function $\kappa(\cdot, \cdot)$ that induces a mapping of the input space onto a feature space of higher dimension, where a linear SVR is built. We use RBF Gaussian kernels $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \exp\{-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / (2\sigma^2)\}$, where \mathbf{x}_i and \mathbf{x}_j are vectors containing eye features.

We detail the use of SVM and SVR in Section 5.

2.2 Text & Eye Movements

We use a data set consisting of documents obtained from different Wikipedia categories and eye movements that were measured whilst reading text during an information retrieval task. In the information retrieval task users were required to read through a set of truncated documents (the documents were truncated to fit the screen) and select those that related to the task’s given category. Despite the instruction to browse through the truncated documents sufficiently to infer their relevance to the given category, this in turn allows learning the association of relevance from eye movements to words. It was found that in several of the categories ‘reading’, rather than browsing, had taken place. We therefore acknowledge this uncontrolled limitation and its effect on our own analysis. In our discussion we elucidate on a potential ‘realistic’ information retrieval system which would incorporate the two modes of reading. The selected eye movement features are described

in [7] and the detailed procedure of recording and measuring the eye movements on the truncated Wikipedia documents is detailed in [4]. We work with a bag of words representation for the documents and term specific eye movement features, i.e. eye movements per a given term (word) had been extracted from recordings.

A dictionary had been constructed consisting of all the words within the truncated document corpus domain knowledge. Initially each document is represented as a bag of words, i.e. a vector with the word frequency occurrence. Term Frequency Inverse Document Frequency (TFIDF) [8] was applied on the documents in order to increase the weighting of terms that occur frequently within a document but infrequently across the document corpus. Given a corpus D and dictionary of terms T , the TFIDF can be computed for term $t \in T$ within document $d \in D$ as

$$T_{d,t} = n_{dt} \log \frac{|D|}{\{|\delta \in D | n_{\delta t} > 0\}},$$

where n_{dt} denotes the number of times term t occurs in documents d .

3 The Prior Primal SVM

We start from a training set of l labelled texts in bag-of-words representation $\{\mathbf{x}_i, y_i\}_{i=1}^l$. In addition, we build a prior weight vector from an external (but relevant) source of information to the training set: \mathbf{v} . This vector lives in the subspace of the bag-of-words that we consider.

A recent result [1] indicates that to build the SVM close to an informative prior classifier can help improve the classification accuracy. Therefore, we envisage an optimisation problem where the posterior SVM is somehow forced to resemble the informative prior:

$$\min_{\mathbf{w}, \eta} \lambda \|\mathbf{w}\|^2 + \sum_{i=1}^l \max\{0, (y_i - (\mathbf{w} + \eta\mathbf{v})^T \mathbf{x}_i)^2\} \quad (1)$$

where $\lambda = 1/C$ marks the trade-off between margin maximisation and training error minimisation and η tunes the weight of the prior classifier. We are following the primal SVM formulation of [2], that assumes a quadratic loss. Once problem (1) is solved, the posterior classifier is given by: $f(\mathbf{x}) = \text{sign}(\mathbf{w} + \eta\mathbf{v})^T \mathbf{x}$

The gradient of (1) with respect to \mathbf{w} is:

$$\nabla_{\mathbf{w}} = \lambda \mathbf{w} - XI^0 \mathbf{y} + XI^0 X^T \mathbf{w} + \eta XI^0 X^T \mathbf{v} \quad (2)$$

where $X = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_l]$, $\mathbf{y} = [y_1, y_2, \dots, y_l]^T$ and I^0 is an $l \times l$ diagonal matrix with ones in the positions corresponding to the support vectors and zeroes in the positions of the correctly classified points. The partial derivative with respect to the other variable, η , yields:

$$\frac{\delta}{\delta \eta} = -\mathbf{v}^T XI^0 \mathbf{y} + \mathbf{v}^T XI^0 X^T \mathbf{w} + \eta \mathbf{v}^T XI^0 X^T \mathbf{v} \quad (3)$$

Now (2) and (3) form a joint gradient:

$$\nabla = \begin{bmatrix} (\lambda I_l + XI^0 X^T) \mathbf{w} + \eta XI^0 X^T \mathbf{v} - XI^0 \mathbf{y} \\ -\mathbf{v}^T XI^0 \mathbf{y} + \mathbf{v}^T XI^0 X^T \mathbf{w} + \eta \mathbf{v}^T XI^0 X^T \mathbf{v} \end{bmatrix} \quad (4)$$

The resulting Hessian is

$$H = \begin{bmatrix} \lambda I_l + XI^0 X^T & XI^0 X^T \mathbf{v} \\ \mathbf{v}^T XI^0 X^T & \mathbf{v}^T XI^0 X^T \mathbf{v} \end{bmatrix} \quad (5)$$

With the gradient and Hessian, we can iteratively solve the optimisation using the following Newton step:

$$\begin{bmatrix} \mathbf{w} \\ \eta \end{bmatrix} \leftarrow \begin{bmatrix} \mathbf{w} \\ \eta \end{bmatrix} - \mu H^{-1} \nabla \quad (6)$$

where μ is a step size determined by linear search.

4 Text classification with Prior Primal SVM system

In this Section we explain the whole system for text classification using a prior obtained from eye movements.

We desire to create a prior weight vector from the eye movements to represent some prior implicit information with respect to the textual content. Also, In order to infer any relevancy from eye movements to documents, a link must be established between the eye features and the words. We adhere to the procedure detailed in [4] and compute, what is coined to be as, an ideal weight vector for each category (category vs. all others). This ideal weight vector represents the words relevance to that given category, using a SVM with a linear kernel and $C = 1$. Let these ideal weight vectors be represented by $\tilde{\mathbf{w}}_i$, for $i = 1, \dots, 25$.

Similarly, we assume that there is a link between eye movements and the relevance of words within a document. Therefore, we use a Support Vector Regression (SVR)¹ to learn the mapping between eye movements to ideal weights.

For this purpose, we access a data set of eye features where each register \mathbf{e}_k^j are the eye movement features corresponding to the k -th word in the dictionary appearing in a document relevant to a category j . When a particular word appears several times in the same document, we average the features, so that there are just a feature vector for every word in the dictionary, document and text category.

Then we construct the regressor by associating each eye feature vector with the ideal weight value of the word, within the category, the eye was fixated on. Let $f_i : \mathbf{e}_k^j \rightarrow w_{jk}$, $i \neq j$ be the learnt regression functions, where j is a text category distinct from i and w_{jk} the relevance of word k to category j . Note that we construct f_i by withholding category i , documents and eye movements, and using all the eye movements and ideal weights values from the remaining 24 categories. The SVR was used with a Radial Basis Function (RBF) kernel and 5-fold cross validation to separately tune the SVR hyperparameters C , σ and ϵ .

Finally, we construct the prior weight vector for category i , \mathbf{v}_i , passing the eye information corresponding to this category through its corresponding regressor f_i :

$$(\mathbf{v}_i)_k = \begin{cases} f_i(\mathbf{e}_k^i) & \text{if word } k \text{ appears in category } i \\ 0 & \text{otherwise} \end{cases}$$

This prior weight is fed into the Prior Primal SVM to obtain the final classifier.

Figure 1 includes a sketch of the whole system.

5 Experiments

Our data corpus consisted of 25 Wikipedia categories which amounted to 528 training documents represented by a 5306 long Term Frequency Inverse Document Frequency (TFIDF) vector and paired eye movements represented by a 26 feature vector. The number of training documents in each category is detailed in Table 1. For testing we have 244 documents, and no corresponding eye movements measurements. The ‘Natural disasters’ category consisted of only 4 test documents while the remaining 24 categories have had 10 test documents each.

The obtained hyperparameters for $f(\tilde{\mathbf{e}})$ were found to be $C = 10$, $\epsilon = 0.1$ and $\sigma = 0.5\sqrt{d}$ where d is the input data dimension. We continue to evaluate the quality of the learnt regressor functions by trying to predict the relevance of the words of the corresponding category i i.e. the one withheld from training. Or in other words we wish to assess $\|f_i(\tilde{\mathbf{w}}) - \tilde{\mathbf{w}}_i\|_2^2$. Table 2 shows the Mean Square Error (MSE) obtained when predicting each word’s relevance, lower values are better.

Now that we have substantiated the link between eye movements and inferring the relevancy of a word we are able to use our learnt regressors to infer the ideal weight vector for each category which in turn will be used as the prior in the Prior-SVM.

To evaluate the if the prior is informative or not, we have computed in Table 3 the average precision yielded by each prior classifier \mathbf{v}_i on the test set corresponding to its corresponding category i .

¹We use a Matlab wrapper 1.2 for LIBSVM ‘<http://www.csie.ntu.edu.tw/~cjlin/libsvm/>’ by Michael Vogt.

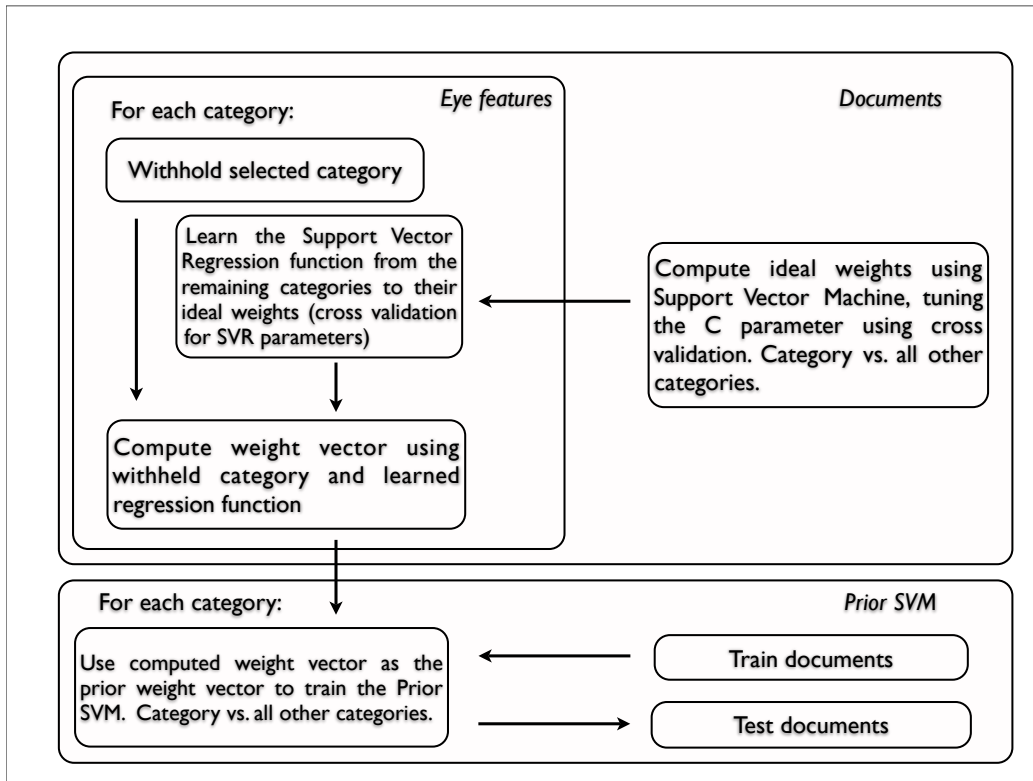


Figure 1: Sketch of the text classification using the Prior Primal SVM and the eye information

Table 1: Summary of the Training Corpus. We detail the number of documents for each of the Wikipedia categories.

Category	Search Topic	# of Documents	Category	Search Topic	# of Documents
1	Astronomy	23	14	Music	16
2	Ball games	23	15	Natural disasters	21
3	Cities	13	16	Olympics	22
4	Court systems	23	17	Optical devices	23
5	Dinosaurs	17	18	Postal system	23
6	Education	22	19	Printing	23
7	Elections	22	20	Sculpture	20
8	Family	18	21	Space exploration	23
9	Film	21	22	Speeches	23
10	Government	21	23	Television	23
11	Internet	23	24	Transportation	23
12	Languages	22	25	Writing systems	17
13	Literature	23			

Table 2: We list the mean square error obtained when predicting, per category, each word’s relevance for the two learnt regressor functions $f_i(\tilde{\mathbf{w}})$. Lower error values are better and are bold faced for emphasis.

Category	$f_i(\tilde{\mathbf{w}})$	Category	$f_i(\tilde{\mathbf{w}})$
1	0.0037	14	0.0043
2	0.0040	15	0.0044
3	0.0043	16	0.0039
4	0.0042	17	0.0040
5	0.0041	18	0.0050
6	0.0039	19	0.0041
7	0.0045	20	0.0038
8	0.0039	21	0.0042
9	0.0043	22	0.0034
10	0.0042	23	0.0037
11	0.0047	24	0.0041
12	0.0045	25	0.0040
13	0.0045		

Table 3: We list the classification error in the test set achieved by the corresponding prior inferred from the eye features. Note that the weights in this classifier are a mere hint about the relevance of the word extracted by comparing the way this word was read with the way the words used to learn the regressor where read.

Category	\mathbf{v}_i	Category	\mathbf{v}_i
1	6.15	14	7.44
2	5.06	15	28.25
3	5.81	16	5.12
4	5.12	17	7.61
5	2.65	18	4.68
6	3.78	19	4.39
7	10.23	20	3.94
8	6.69	21	8.02
9	4.51	22	4.90
10	4.33	23	5.98
11	2.99	24	15.10
12	5.14	25	4.19
13	3.30		

Note that the performance of this classifier is really poor since the features point to relevant words, without taking into account the relevance of that particular word to the textual category.

Priors \mathbf{v}_i are plugged into the Prior Primal SVM algorithm to yield the final text classifiers. We have used a regularisation parameter of $C = 1$ to keep a fair comparison with the vanilla SVM used to learn the word relevances.

Finally, Table 4 shows the text classification rate of the system and compare it with the classification achieved by a vanilla SVM using only textual information, with a SVM-2K classifier that also combines textual information and eye features as described in [4] and with a baseline classifier constructed by using the inferred ideal weights vector from the learnt regression function $f_i(\mathbf{e}^i)$ for each category.

The results show that for some categories, the introduction of the eye features helps improve the accuracy of the classifier, although in some other cases, the prior contributes to confuse the classifier. We are still working on a hypothesis that explains this behavior.

Table 4: We list the text classification rate of the system and of compared methods.

Category	Baseline	Prior SVM	Vanilla hinge loss	SVM-2K[4]
1	24.18	37.30	39.57	40.11
2	85.91	71.69	75.33	86.01
3	25.14	77.38	69.83	80.53
4	47.38	72.38	62.72	59.83
5	53.42	99.09	95.73	94.30
6	30.60	71.78	50.33	56.25
7	42.09	78.44	72.87	67.52
8	22.22	52.39	70.22	71.81
9	23.28	47.65	54.51	54.08
10	32.88	60.82	32.75	26.92
11	8.84	59.30	35.58	39.35
12	55.27	92.16	89.74	93.51
13	14.32	20.06	18.24	26.84
14	16.32	19.07	60.21	74.03
15	83.04	50.84	100.00	100.00
16	39.80	74.52	92.63	97.69
17	18.21	56.23	56.08	64.44
18	20.44	60.33	76.30	81.66
19	55.12	92.84	64.24	68.01
20	19.52	34.09	60.17	62.44
21	72.16	53.12	65.08	67.41
22	37.30	57.46	75.31	70.29
23	45.20	72.99	36.68	34.61
24	13.07	37.55	44.02	41.59
25	28.63	81.96	46.76	50.28

6 Discussion

In this paper we have presented a text classification system that is based on an SVM that combines texts represented as bag-of-words vectors with features extracted from the eye movements captured when humans read through the documents. The combination of the two sources of information has been made by building a prior classifier with the eye movement features and combining this prior classifier with the posterior SVM in the optimisation problem.

The experimental work shows that the Prior Primal SVM system sometimes achieves higher accuracy than the plain classifier. It also compares favorably against the SVM-2K, that combines text and eye features. The SVM-2K combines the two stage learning of learning a new semantic representation that maximally correlates between the text and eye features followed by SVM into a single optimisation. Ongoing research is focused on working out an explanation about when the Prior helps the posterior and when it is just a source of confusion, in order to implement hybrid systems that may switch between Prior SVM and flat SVM.

Despite the inconclusive results in using eye movements as prior information for document classification, we believe the Prior SVM, and in the presented paper its primal formulation, to be an interesting approach for incorporating prior knowledge to the SVM learning procedure. We hope to extend on our results.

References

- [1] Amiran Ambroladze, Emilio Paraado-Hernandez, and John Shawe-Taylor. Tighter pac-bayes bounds. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 9–16. MIT Press, Cambridge MA, 2007.
- [2] Olivier Chapelle. Training a support vector machine in the primal. *Neural Computation*, 19(5):1155–1178, 2007.
- [3] Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, 2000.

- [4] David R. Hardoon, Antti Ajanki, Kai Puolamaki, John Shawe-Taylor, and Samuel Kaski. Information retrieval by inferring implicit queries from eye movements. In *The 11th International Conference on Artificial Intelligence and Statistics*, 2007.
- [5] Kai Puolamäki, Jarkko Salojärvi, Eerika Savia, Jaana Simola, and Samuel Kaski. Combining eye movements and collaborative filtering for proactive information retrieval. In Gary Marchionini, Alistair Moffat, John Tait, Ricardo Baeza-Yates, and Novio Ziviani, editors, *Proceedings of SIGIR 2005, Twenty-Eighth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 146–153. ACM, New York, NY, 2005.
- [6] Jarkko Salojärvi, Ilpo Kojo, Jaana Simola, and Samuel Kaski. Can relevance be inferred from eye movements in information retrieval? In *Proceedings of WSOM'03, Workshop on Self-Organizing Maps*, pages 261–266. Hibikino, Kitakyushu, Japan, September 2003. Kyushu Institute of Technology.
- [7] Jarkko Salojärvi, Kai Puolamäki, Jaana Simola, Lauri Kovanen, Ilpo Kojo, and Samuel Kaski. Inferring relevance from eye movements: Feature extraction. In Kai Puolamäki and Samuel Kaski, editors, *Proceedings of the NIPS 2005 Workshop on Machine Learning for Implicit Feedback and User Modeling*, pages 45–67. Helsinki University of Technology, Espoo, Finland, 2006.
- [8] G Salton and M J McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Berlin (Germany), 1983.
- [9] B Schölkopf and A J Smola. *Learning with kernels*. MIT Press, Cambridge (MA), 2002.