

ASSESSMENT OF FEATURE SELECTION TECHNIQUES FOR SUPPORT VECTOR MACHINE CLASSIFICATION OF SATELLITE IMAGERY

Tarek HABIB^{†,‡}, Jordi INGLADA[‡], Grégoire MERCIER^², Jocelyn CHANUSSOT[†]

[†] GIPSA-Lab (Signal and Images Department), CNRS, INP Grenoble,

[‡] Centre national d'études spatiales (Cnes - DCT/SI/AP)

^² Institut Telecom ; Telecom Bretagne / CNRS FRE 3167 LabSTICC/CID.

ABSTRACT

The problem of focusing on the most relevant information in a potentially overwhelming quantity of data has become increasingly important. Using irrelevant or noisy features not only can affect the accuracy of the classification results obtained but also the convergence time. In this paper several feature selection algorithms used with the Support Vector Machine (SVM) algorithm are presented. The feature selection algorithms are classified as filter and wrapper approaches. Two different wrapper techniques are presented: the first one uses the generalization error estimate of the leave-one-example-out error, while the second one uses the error estimate of the leave-one-feature-out error. Filter approaches with 4 different parameters are presented, namely: the mutual information, the FScore, and two advanced entropy measures are studied. Results in the context of change detection using satellite imagery are then discussed.

1. INTRODUCTION

In recent years, change detection has received great attention from the Geoscience and Remote Sensing (GRS) community especially in the case of hazards where the spatial facilities are used for damage evaluation. Change detection between remotely sensed data is a methodological bolt in the accuracy of the overall damage evaluation procedure.

However several difficulties (*i.e.* the difference in the acquisition geometry, the difference between the instruments used for the data acquisitions, the change in the lighting and atmospheric conditions that are not related to the actual change, and *s.o.*) makes the process of change detection very challenging.

In order to build a generic, context independent change detection system, it would be interesting to build a system that uses the maximum amount of data at its input and that it identifies the most relevant data sets using some learning samples. The relevance being identified in the sense of optimizing the classification performance from an accuracy and runtime points of view.

The Support Vector Machines (SVM) algorithm is one of

the most known machine learning algorithms that can meet the above mentioned constraints. Due to its high adaptability to the input data as well as its capacity to handle large dimensional data without increasing the system's complexity, SVM binary classification has been extensively used in a wide variety of applications [1] and is suitable for the application of change detection.

In this paper several feature selection techniques that are coupled with the SVM algorithm in the context of change detection are presented. Following the categories presented in the feature selection literature as in [2], feature selection algorithms can be divided into 3 categories: 1. Filters, 2. wrappers, and 3. embedded. In this paper only filter and wrapper approaches are discussed. Since the focus is made on feature selection algorithm coupled with SVMs, some basic notions concerning SVMs are presented in section 2. In the following sections 4 and 3 the different feature selection algorithms are presented, test data and results are then introduced in section 5. Finally the conclusions are drawn in section 6.

2. SUPPORT VECTOR CLASSIFICATION

SVM is a binary classifier that has the objective of assigning a new *test* data to a class by minimizing the probability of error. The classical classification problem can be formulated as follows:

Given a set of learning data $(x_i, y_i)_{i=1}^m$ where $x_i \in \mathbb{R}^n$ are the input feature vectors and $y_i \in \{-1, 1\}$ are the set of corresponding labels, the problem is to find $y_t = f(x_t)$ for a new test vector x_t so that the probability of error is minimal. All of this under the hypothesis that the new x_t is issued from the same unknown probability density function that produced the learning set x_i .

The decision function for the SVM algorithm is given by:

$$f(x_t) = \text{sgn}\left(\sum_{i=1}^m \alpha_i y_i k(x_t, x_i) + b\right) \quad (1)$$

Where α_i and b are the solutions of the SVM optimization algorithm, and $k(\cdot, \cdot)$ is the kernel function used during the learning phase. The use of a kernel function allows much

flexibility in the SVM algorithm without increasing the system's complexity.

3. WRAPPERS

Wrappers are classifier dependent feature selection techniques. The main idea is to use the same procedure for learning and/or classification in the process of feature selection [3]. In the case of SVM, the solution of the optimization problem provides many useful parameters that can be used for feature selection. In this section the usefulness of the margin, more precisely \vec{w} is first studied. Then a second method that uses the SVM generalization error estimation presented in [4] is studied. And finally a new wrapper inspired from the SVMs optimization problem and using the kernel properties is presented.

3.1. The normal vector

A simple idea for feature selection can be developed when using a linear kernel. The idea is to obtain the vector normal to the separating hyperplane \vec{w} and then, in a step wise fashion, remove the least component in this vector and then restart from the learning step. Removing the least component comes to removing the component that is almost *parallel* to the hyperplane; and hence the one that has the least effect on the classification process (i.e. the less discriminative component).

This procedure can only be applied when using linear kernels (i.e. linear SVM), since this is the only possible case where one can get the vector \vec{w} . In the case of using other types of kernels only the margin (i.e. through $\|w\|$) can be obtained.

The feature selection algorithm works as follows: 1. Perform the SVM learning step using all the available features. 2. Calculate \vec{w} . 3. Remove the feature with the smallest contribution in \vec{w} . 4. Restart until no features are left. 5. Order features based on survival.

Ordering features based on survival means that the later a feature is removed from the set the more important it is, in other terms, the feature that is first removed is the less relevant and the one that stays until the end of this process is the most important.

While this feature selection method has the advantage of being very simple and easy to implement it suffers from two drawbacks, 1. it can only be used with linear SVM, 2. the learning process has to be restarted after the removal of a given feature. the first drawback implicitly states that the performance of this approach for feature selection is not reliable when the input data is not linearly separable in the input space. While the second drawback supposes a high computation time when a multitude of features are introduced at the input of the SVM.

In the following section 3.2 the second wrapper approach is presented. This approach solves the generality problem mentioned above since it can be applied when using any type of kernels, while the problem of multiple learning steps is addressed by the third wrapper in 3.3.

3.2. $\xi\alpha$ -estimator

The author in [4] proposes an efficient and effective approach for estimating the generalization performance of a SVM for text classification. However the proposed approach can be operational for any application since the theoretical development is based on the optimization equations of SVM. The author propose several performance estimators namely the generalization error estimation, precision, recall and $F1$. These last three are based on the error estimate. This performance estimator is called the $\xi\alpha$ -estimator since it uses the ξ and α values that are obtained during the SVM optimization.

The objective of the $\xi\alpha$ -estimator is to obtain an estimate of the leave-one-example-out error using the learning examples. Thus in [4] it is proven that for stable soft margin SVMs, the $\xi\alpha$ -estimator of the error rate is:

$$Err_{\xi\alpha}^n(h) = \frac{d}{n} \quad (2)$$

with $d = |i : (\rho\alpha_i R_{\Delta}^2 + \xi_i) \geq 1|$

where $\rho = 2$, $\vec{\alpha}$ and $\vec{\xi}$ are the solutions of the optimization problem on a set of examples of size n . R_{Δ}^2 is an upper bound on $K(\vec{x}, \vec{x}') - K(\vec{x}, \vec{x})$.

This generalization error estimation can be used for feature selection as follows: 1. Perform the SVM learning step using all the available features. 2. Calculate the generalization error using the $\xi\alpha$ -estimator. 3. Store the estimate. 4. Remove only one of the features and restart from the learning step using all the other features (i.e. repeat until each feature was *left out*). 5. Order the features based on the generalization error estimate.

This feature selection algorithm has the advantage of being applicable to any SVM classification scenario, however as was mentioned earlier, the need to perform a new learning step after the removal of one feature is not practical in the case of using a large number of inputs for the algorithm. The following feature selection scheme proposes a method to solve this problem of re-learning using the kernel properties and more precisely by introducing the notion of additive kernels.

3.3. Kernel properties

Instead of using a generalization error based on the leave-one-example-out error as in the section 3.2, in this section a generalization error in the sense of leave-one-feature-out is proposed. The function of leave-one-feature-out is performed within the kernel evaluation itself, thus it does not need the

repetition of the learning process as it will be shown later. For this leave-one-feature-out function, a specific kernel construction has to be used.

Several kernel properties can be derived from existing kernels using Mercer's theorem [5]. From these properties, it can be demonstrated that the following function is a kernel:

$$K(x, y) = \sum_{l=1}^M a_l K_l(\psi_l(x), \psi_l(y)) \quad (3)$$

where $\psi(\cdot) \in \mathbb{R}^p$ and K_l is a kernel function.

Using this additive kernel, and starting from the SVM constrained optimization problem, it can be proven that a leave-one-feature-out error occurs if the following condition is satisfied:

$$b \leq \sum_{i=1}^m y_i (\alpha_i + \text{constant}(C - \alpha_i)) k_l(x_{test}, x_i) + \sum_{i=1}^m y_i (\alpha_i + \text{constant}(C - \alpha_i)) k_l(x_i^l, x_{test}^l) \quad (4)$$

The feature selection procedure is then performed as follows: 1. Perform the SVM learning step using all the available features. 2. Evaluate the inequality in equation 4 for each set. 3. Order features based on the inequality score.

The presented wrappers propose different strategies according to the change detection scenario, the advantage of the wrapper using the $\xi\alpha$ - estimator is that estimations can be obtained using any type of kernels. However the approach using the kernel properties could be more suitable for classification as well as for feature selection purposes since it avoids the repetition of the learning step.

4. FILTERS

Filter based approaches for feature selection consists on building a separate unit for this purpose that operates before the actual learning or classification algorithm. They are called *filter* methods since they are meant to filter out irrelevant features. The main advantage of being independent of the classification algorithm is that filters can be combined with any classifier. Hence if an optimization step is required in order to properly adjust the classification algorithm's parameters, the feature selection part need not be changed.

The simplest filter based algorithm is the one that directly calculates the correlation of each feature with the target function, then orders the features using their correlation score. In the following a similar feature selection technique is presented but using the mutual information technique. Other algorithms developing this idea of correlation were presented in the literature, for example algorithms that search the feature space for minimal feature combinations that would perfectly discriminate the classes. Another technique for filter implementation is the use of principal component analysis (PCA)

in order to find the most discriminant features [6]. In the following the focus is made on filters using specific parameters, the tested parameters are:

- **Mutual information [7]:**

$$I(I; J) = \sum_{j \in J} \sum_{i \in I} p(i, j) \log \left(\frac{p(i, j)}{p(i)p(j)} \right) \quad (5)$$

- **Feature-Score(FScore) [8]:**

$$F(i) = \frac{(\bar{x}_i^{(+)} - \bar{x}_i)^2 + (\bar{x}_i^{(-)} - \bar{x}_i)^2}{f^+ + f^-} \quad (6)$$

where: $f^+ = \frac{1}{n_+ - 1} \sum_{k=1}^{n_+} (x_{k,i}^{(+)} - \bar{x}_i^{(+)})^2$ and f^- is calculated in the same manner as f^+ by replacing $+$ by $-$ in the above equation, i is the feature under study n_+ and n_- are the number of positive and negative instances $\bar{x}, \bar{x}_i^{(+)}, \bar{x}_i^{(-)}$ are the average of the i th feature of the whole, the positive and the negative data sets respectively, $x_{k,i}^{(+)}$ and $x_{k,i}^{(-)}$ are the i th feature value of the k th pixel.

- **Advanced entropy and mutual information based measures:**

$$v(I|J) = \frac{\zeta(I) - \zeta(I|J)}{\zeta(I)} = \frac{I(I, J)}{\zeta(I)} \quad (7)$$

$$k(I, J) = \frac{I(I, J)}{\zeta(I) + \zeta(J)} \quad (8)$$

where: $I(I, J)$ is the mutual information measure $I(I, J) = \zeta(I) + \zeta(J) - \zeta(I, J)$

The feature selection algorithm consists on ordering the features based on the evaluation score obtained using the different measures.

The advantage of the use of a supervised classification (*i.e.* SVM) is the availability of a labeled learning features. Thus using these labeled features, the measure's error prediction capacity can be measured, in the sense that each measure is computed using the learning features and the results obtained are generalized for the feature selection of the whole data set.

5. TEST DATA AND RESULTS

The test data set is the Goma data set. These are images acquired before and after the Nyiragongo volcanic eruption in eastern Congo on 17 January 2002. When the volcano erupted, the lava flowed to Goma in a single stream and finally covered about a fifth of the town. The change detection problem is then to detect the areas that have undergone an

Table 1. Average ranking results using the different feature selection techniques

Technique	Group Number					mean
	1	2	3	4	5	
$\xi\alpha$ -estimator	2.5	2	1.5	2.5	2.5	2.2
Kernel properties	2	2.5	1.5	2.5	2.5	2.2
Mutual information	2	2	2	2	3	2.2
FScore	3	3	2.5	2.5	3	2.8
U Metric	2.5	3	3	2.5	3	2.8
K Metric	2.5	3	3	2.5	3	2.8

abrupt change due to this volcanic eruption. Hence the classification problem is between 2 classes change, no change. Using the available images (*i.e.* the before and after SAR images of size 400*800 pixels), several features were computed, namely: difference, ratio, ratio of means, ratio of medians, correlation, mean squares, Kullback-Leibler distance, mutual information, cardinality match, gradient difference, entropy and energy. Along with the original images these images were used to form the input vectors for the SVM algorithm, with each feature representing one dimension of the vector, this gives vectors of size 14.

The results presented in this section are measured on 10 different iterations on the different data sets. Five different feature groups were specified as follows: 1. "Before" Image, "After" Image, Difference, ratio. 2. Ratio of means, ratio of medians, correlation, mutual information. 3. Entropy, energy, mean squares and Kullback-Leibler distance. 4. Cardinality match, gradient difference, "Before" image and "After" Image. 5. Difference, ratio, ratio of means and ratio of medians.

For testing purposes random masks, with the size of 1% of the whole scene, were selected for each iteration. Due to this randomness, the absolute values of the different parameters is hence not significant, and thus for the evaluation of the feature selection approaches, the interest is only in the behavior with respect to the real classification error. In this sense, the interesting information resides in the confrontation between the ranking obtained from a given feature selection technique to the ranking obtained from the real classification error.

For this testing phase, we have 5 different feature groups, and hence for each group we have 4 different *relative* rankings. *Relative* ranking means the ranking of each feature group with respect to the other 4 feature groups. In the following a feature group is awarded one point when it is properly ranked with respect to another group and zero points otherwise (*i.e.* hence the highest score for a given feature is 4). The mean score for each feature group over the 10 iterations is shown in table 1.

Based on the average rankings it can be noted that the performance of the filter approaches using the FScore, the U and the K metric outperforms the mutual information based filters

as well as the wrapper approaches. However these approaches require extra processing with respect to the use of an approach as in the wrapper using the kernel properties. Also changing the kernel function used for the SVM optimization can improve the results obtained from the wrapper based approaches without affecting those of the filter based approaches.

6. CONCLUSION

In this paper several feature selection techniques for SVMs were proposed. Based on the obtained results it can be noted that according to the data in hand, some of these techniques are more adapted than others, for the filter based results the mutual information based technique could be preferred to other techniques due to its simplicity, while for the wrapper based approaches the ones using the kernel properties provides better or comparable results than the one based on the $\xi\alpha$ - estimator, however it can be used only when the additive kernel is used.

7. REFERENCES

- [1] V. N. Vapnik, "An overview of statistical learning theory," *IEEE transactions on neural networks*, vol. 10, no. 5, pp. 988–999, September 1999.
- [2] A. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artificial Intelligence*, vol. 97, no. 1-2, pp. 245–271, 1997.
- [3] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, no. 1-2, pp. 273–324, 1997.
- [4] T. Joachims, "Estimating the generalization performance of a SVM efficiently," in *Proceedings of ICML-00, 17th International Conference on Machine Learning*, P. Langley, Ed. Stanford, US: Morgan Kaufmann Publishers, San Francisco, US, 2000, pp. 431–438.
- [5] M. Genton, "Classes of kernels for machine learning: A statistics perspective," 2000.
- [6] R. Malhi, A. and Gao, "PCA-based feature selection scheme for machine defect classification," *IEEE Transactions on Instrumentation and Measurement*, vol. 53, no. 6, pp. 1517– 1525, December 2004.
- [7] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Transactions on Neural Networks*, vol. 5, no. 4, 1994.
- [8] Y.-W. Chen and C.-J. Lin, *Feature Extraction, Foundations and Applications*, M. N. Isabelle Guyon, Steve Gunn and L. Zadeh, Eds. Springer, 2006.