Support Vector Reduction in SVM Algorithm for Abrupt Change Detection in Remote Sensing

Tarek Habib, Member, IEEE, Jordi Inglada, Associate Member, IEEE, Grégory Mercier, Senior Member, IEEE, and Jocelyn Chanussot, Senior Member, IEEE

Abstract—Satellite imagery classification using the support vector machine (SVM) algorithm may be a time-consuming task. This may lead to unacceptable performances for risk management applications that are very time constrained. Hence, methods for accelerating the SVM classification are mandatory. From the SVM decision function, it can be noted that the classification time is proportional to the number of support vectors (SVs) in the nonlinear case. In this letter, four different algorithms for reducing the number of SVs are proposed. The algorithms have been tested in the frame of a change detection application, which corresponds to a change-versus-no-change classification problem, based on a set of generic change criteria extracted from different combinations of remote sensing imagery.

Index Terms—Image classification, image matching, image processing, remote sensing.

I. INTRODUCTION

In the context of risk management and hazard assessment using satellite imagery, high-performing change detection algorithms are of great importance [1]. For this type of application, change detection problems can be viewed as a classification using two classes (change/no change). Due to its multiple advantages, the support vector machine (SVM) binary classification algorithm has been widely used for the classification of satellite imagery [2]. In this letter, the SVM algorithm is used in the context of building a generic change detection algorithm. In order to achieve the goal of generality, a large amount of change indicators have been extracted from the images and used as input in the SVMs, yielding high-dimensionality input vectors. Despite the fact that, theoretically, the SVM algorithm can handle such large data sets with high dimensionality, in practice, the following two main problems arise: 1) The computation time required for the convergence of the optimization problem during the “learning” step of the SVM algorithm is very high, and 2) the classification time increases in a polynomial fashion with the increase in the data dimensionality [3].

This letter focuses on the decrease of the classification time. In the SVM algorithm, the classification time is directly proportional to the number of support vectors (SVs) in the nonlinear case. Several methods for the SV reduction are proposed in Section III. The reduction of the SVs may logically result in a decrease in the classification accuracy; however, in the context of risk management and hazard assessment, a slight decrease in the classification accuracy can be tolerated for a gain in classification time.

This letter is divided into the following sections. In Section II, a brief introduction on the SVM binary classification algorithm is presented. In Section III, the proposed techniques for obtaining the reduced set of SVs are presented. The conducted experiments and the results obtained using the different methods are discussed in Section IV. Finally, the conclusions are drawn in Section V.

II. SVMs

The SVMs are state-of-the-art large margin classifiers that have gained much popularity within the image-processing community. The SVM is a learning-based binary classification algorithm that is based on the concept of structural risk minimization [4]. This construction has shown to generally outperform traditional learning machines like multilayer neural networks that are based on the concept of empirical risk minimization [5]. In this section, a brief review of the theory of this algorithm is presented; for further details, the reader is invited to review [6].

Consider 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 (i.e., classes). The SVM solution finds \( y_{\text{test}} = f(x_{\text{test}}) \) for a new test vector \( x_{\text{test}} \) so that the probability of the error is minimal. The SVM decision for any new vector \( x_{\text{test}} \) is obtained under the hypothesis that \( x_{\text{test}} \) is issued from the same unknown probability density function that produced the learning set \( x_i \).

If it is assumed that the two classes (i.e., the SVM is a binary classifier) can be separated by a hyperplane and that no prior information concerning the data distribution is available, then the optimal hyperplane is the one which maximizes the margin of separation between the two classes [6]. The optimal values of \( w \) and \( b \) can be obtained by solving a constrained minimization problem through the use of the Lagrange multipliers \( \alpha_i \). The decision function provided by the SVM can thus be put in the following form:

\[
  f(x_{\text{test}}) = \text{sgn} \left( \sum_{i=1}^m \alpha_i y_i K(x_{\text{test}}, x_i) + b \right). \tag{1}
\]
The input feature vectors $x_i$ having a Lagrange multiplier $\alpha_i$ that is not equal to zero are considered as SVs, hence the name SVMs. The kernel function $K(\cdot, \cdot)$ substitutes the scalar product $\langle \cdot, \cdot \rangle$ in order to allow the SVM to learn the nonlinear classifiers as, for example, polynomial classifiers or radial basis function (RBF) networks.

III. REduced Set of SV

With the goal of providing a tradeoff between the classification accuracy and the computation time, four different approaches are proposed in this section. The objective of these approaches is the selection of a representative subset of the SVs. These approaches are intended to select the SVs according to their importance in order to obtain an acceptable approximation of the SVM decision boundary.

A. Set Reduction by Lagrange Multipliers

The value of the Lagrange multipliers is a valid indicator of the importance of a given vector. Using the information provided by the Lagrange multipliers, the first approach would be to keep the SVs based on the value of the Lagrange multipliers. Thus, the SVs with the highest unbounded Lagrange multipliers (i.e., $0 < \alpha_i < C$) will be used as the reduced set, and the rest of the SVs will be eliminated. The size of the reduced set is either specified a priori by a percentage of the entire set of SVs or by specifying a threshold on the value of the $\alpha_i$ to be retained. If the percentage is higher than the available number of unbounded SVs, the set is completed using the randomly selected bounded SVs. Since this SV reduction procedure may not respect the constraints of the SVM algorithm, namely, $\sum_{i=1}^{m} \alpha_i y_i = 0$, then a new learning step has to be applied.

A modified classification algorithm that takes into consideration the SVM constraints is proposed as follows: 1) Perform a first learning step; 2) choose a subset of the SVs according to a maximum value of $f_d(x_i)$ as in the aforementioned equation; 3) reperform the learning step; and 4) classify. As described earlier, the relearned step is used in order to recompute the $\alpha_i$. As implemented in this algorithm, the SVs situated far away from the separating hyperplane are considered as better suited for set reduction since they provide a better generalization of the separating hyperplane. The situation far from the hyperplane limits the eventual overlap between the two classes and hence is expected to provide a better separability between the two classes.

C. Set Reduction by Mechanical Analogy

In [7], Schlkopf and Smola present the SVM classification problem as a mechanical problem. The structure of the SVM optimization problem is closely similar to the ones that typically arise in Lagrange’s formulation of mechanics. From this point of view, it is possible to give a mechanical interpretation to the optimal margin hyperplanes. Assume that each SV $x_i$ exerts a perpendicular force of magnitude $\alpha_i$ and direction $y_i(w/||w||)$ on a solid plane sheet lying along the hyperplane. The idea is to use this mechanical analogy to reduce the number of SVs by merging the vectors using geometrical and mechanical properties.

According to this point of view, two SVs $x_1$ and $x_2$ may be replaced by a new vector $x_{new}$ if its mean is set as the mean of the two vectors. The difficult point would be to set the value of $\alpha_{new}$, which is associated to $x_{new}$. Considering the mechanical analogy, the problem would be to identify the direction of this force (i.e., the vector that is perpendicular to the separating boundary $w$ cannot be computed directly in the nonlinear case). This problem could be avoided by merging the two vectors directly in the feature space (i.e., the space induced by the kernel function). Assuming that the system is stable around the origin of the feature space, the $\alpha_{new}$ could be obtained as follows:

$$\alpha_{new} y_{new} K(x_{new}, x_{new}) = y_1 \alpha_1 K(x_1, x_1) + y_2 \alpha_2 K(x_2, x_2).$$

Once again, in order to respect $\sum_{i=1}^{m} \alpha_i y_i = 0$, the difference (i.e., $\delta$) between the computed $\alpha_{new}$ and the required one for stability is computed by

$$\alpha_{new} y_{new} + \delta = \alpha_1 y_1 + \alpha_2 y_2. \tag{2}$$

Once the value of $\delta$ is obtained, it is distributed uniformly on the SVs of the same class of $x_{new}$ [according to the sign of $f(x_{new})$ in (1)], which guarantees the SVM stability constraint.

An algorithm based on this mechanical analogy was developed. In order to identify the SVs that should be merged together, a specific signature was computed for each vector, and then, the vectors with similar signatures are merged.

The proposed algorithm is decomposed in the following steps, which will be detailed further on: 1) Compute the signatures of the different SVs ($s_i$); 2) compute the distances between the different vectors according to their respective signature $[D(s_i, s_j)]$ by evaluating the symmetrized...
Kullback–Leibler (KL) distance; and 3) define a threshold and then merge the SVs having an interdistance that is less than the defined threshold.

The signature \( s_n \) of an SV \( x_n \) is a vector of the \( m \) components \((m \) is the number of SVs) defined, component by component, by the value of \( K(x_i, x_n) \), \( 1 \leq i \leq m \)

\[
s_n = (K(x_1, x_n), K(x_2, x_n), \ldots, K(x_m, x_n)). \tag{3}
\]

The computation of the signature by using this procedure has two reasons: 1) The interaction between the different vectors using the kernel function is indicative of the distribution in the feature space, and 2) the kernel values are already computed during the learning phase, and hence, with proper storage, no further computation will be required. Once the signature for all the \( m \) vectors is computed, the symmetrized KL distance is used to identify the vectors to be merged. The expression of this distance was developed in [8] and is defined as follows:

\[
\frac{1}{R(s_1, s_2)} = \frac{1}{D(s_1 || s_2)} + \frac{1}{D(s_2 || s_1)} \tag{4}
\]

where \( D(s_1 || s_2) = \int s_1(x) \log(s_1(x)/s_2(x)) dx \). The reason for using this type of distance instead of, for example, a simple quadratic distance is its interesting properties. This type of distance measurement is based on the KL divergence [9] that contains the discrimination information of the first hypothesis (i.e., represented by \( s_1 \)) on the second hypothesis (i.e., represented by \( s_2 \)). The addition of such a rich measurement that already contains information concerning the discrimination between the two classes can enhance the classification performance in the sense that it allows a first stage of clustering of the vectors.

The threshold definition procedure can be done manually or automatically according to the required number of SVs at the output of this reduction process.

This procedure has the advantage of not requiring a new learning step like the other algorithms; however, the computation of the signatures and the distances is a time-consuming task. Knowing that all these kernel evaluations have been already computed during the initial learning phase, storing these values greatly increases the speed of this algorithm. The total number of operations required for this algorithm is then \((4m + 1)\) for the KL distance computation in addition to the six operations required to evaluate the values of \( \alpha_{\text{new}} \) and \( x_{\text{new}} \). On the other hand, merging two vectors saves the \( m \) operations for the classification of a new test vector. Thus, if \( N > (4m + 7) \), this type of reduced set SVM can be interesting, where \( N \) is the total number of test data vectors to be classified.

D. Set Reduction by Optimization Problem

Following the same procedure as that proposed for the mechanical analogy approach, another strategy may be used to evaluate \( \alpha_{\text{new}} \).

It takes its inspiration from the SVM optimization problem in the sense that the solution of the SVM optimization problem using the new vector is equal to the solution obtained using the original vectors \( x_1 \) and \( x_2 \). Once the values of \( y_{\text{new}} \) and \( x_{\text{new}} \) are specified, the \( \alpha_{\text{new}} \) of the new SV \( x_{\text{new}} \) can be computed in order to replace the two SVs \( \alpha_{i_1} \) and \( \alpha_{i_2} \). \( \alpha_{\text{new}} \) can be obtained as follows (note that this computation can be extended to merge any number of vectors):

\[
\alpha_{\text{new}} = \alpha_{i_1} + \alpha_{i_2} - \frac{1}{2} (T_{i_1} + T_{i_2} + T_{i_1,i_2})
\]

\[
1 - \frac{1}{2} \sum_{i=1}^{m} \alpha_i y_i |y_{\text{new}}| K(x_i, x_{\text{new}})
\]

where

\[
T_{i_1,i_2} = 2 \alpha_{i_1} \alpha_{i_2} y_{i_1} y_{i_2} K(x_{i_1}, x_{i_2}) \quad \text{for} \ i = i_1 \text{ or } i_2
\]

\[
T_i = \sum_{j \neq i}^{m} \alpha_i \alpha_j y_i y_j K(x_i, x_j).
\]

The number of operations required to compute \( \alpha_{\text{new}} \) from the merging of the two vectors is \((7m - 6)\); this includes the computation of the KL distance to identify the vectors to be merged. Thus, this type of reduction is interesting when \( N > (7m - 6) \), where \( N \) is the number of data vectors to be classified. When three vectors are merged using this technique, the reduction is interesting when \( N > (6m - 2) \); thus, the procedure becomes more interesting when the number of vectors to be merged increases.

IV. EXPERIMENTS

The proposed SV reduction algorithms were tested on three data sets. The first is the Goma data set that is composed of a pair of synthetic aperture radar images that were obtained before/after the Nyiragongo volcanic eruption in eastern Congo in 2002. The second set is composed of a pair of optical Satellite Pour l’Observation de la Terre images that were obtained before/after the earthquake that hit the Algerian city of Boumerdes.

Several features were computed, namely, difference, ratio, ratio of means, ratio of medians, correlation, mean squares, KL 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 component of the vector; this yields vectors of size 14.

For testing purposes, a soft margin SVM was used, with \( C = 1 \). An RBF kernel was used, with \( \gamma = 0.5 \) (i.e., it should be noted that the exact values of the parameters are of no interest in our case since we are only interested in obtaining information concerning the overall behavior of algorithms that are independent of these parameters). Ten different full iterations were performed on the test data. A single iteration is composed of the following steps: 1) Choose a random set of learning vectors (consisting 1% of the total number of available testing vectors); 2) perform the learning phase of the SVM algorithm in order to obtain the original SV set; 3) reduce the number of SVs by following the different algorithms (for the algorithm in Section III-A, this means 10% reduction with respect to the original set’s size, whereas for the remaining three algorithms, this means increase of the threshold by 10% of the maximum distance according to the specified criteria); 4) perform the SVM classification using the new sets (four new sets following the four proposed algorithms) of SVs; and 5) restart from step 3 until the new SV set’s size is only 10% of the original SV set.
Fig. 1. Classification accuracy versus percentage (divided by ten) of the set’s size reduction with respect to the original set’s size using the reduced set algorithm of Section III-A applied on the different data sets. (a) Goma data set. (b) Boumerdes data set.

Note that, in the following figures, the green line is a simple connection between the mean values. This green line is intended to show the overall behavior of the different functions.

A. Set Reduction by Lagrange Multipliers

Fig. 1 shows the classification error using the reduced set as a function of the total number of SVs applied on the different testing data sets. Hence, the x-axis represents the percentage of SVs that was removed ($\times 10$). In the case shown in Fig. 1(a), the classification accuracy is almost constant with the removal of the SVs. However, due to the large variance of the classification error, following this method may not be robust to reduce the set from the Goma data.

In the case shown in Fig. 1(b), the evolution is convenient in the sense that the classification accuracy decreases with the removal of the SVs. Thus, it offers a tradeoff between the classification accuracy and the classification time (through the reduction of the number of SVs). However, when removing a large number of SVs (i.e., higher than 70%), the variance of the classification accuracy is very high, showing that the results are more and more noisy.

B. Set Reduction by Distance to Separating Surface

Fig. 2 shows the results obtained for the algorithm using the distance-to-hyperplane criteria. From Fig. 2(a), the results show that the classification accuracy has a large variance with respect to the removal of the SVs. On the other hand, the mean values show that this technique can be used to provide a tradeoff between the classification accuracy and the classification time since the overall accuracy decreases with the decrease of the number of SVs.

When tested on the Boumerdes data set as shown in Fig. 2(b), this approach provides irregular results, where the classification accuracy seems to be independent of the number of SVs.

C. Set Reduction by Mechanical Analogy

Fig. 3 shows the results obtained using the algorithm defined in Section III-C. It can be noticed that the error’s mean and variance increase with the number of merged vectors. As the number of SVs decreases, the separating surface becomes an approximation of the original surface and thus provides a lower classification accuracy. This shows that the reduced set using this technique can provide a tradeoff between the classification accuracy and the classification time through the reduction of the SVs.

Despite a large variance of the classification accuracy, the reduced set using the mechanical analogy provides adequate results when applied to the Goma data set. As can be seen in Fig. 3(a), the classification accuracy slightly decreases with the decrease of the number of SVs. This is maintained until
a very large number of SVs are removed (80%), where the classification result becomes unreliable due to the lack of representativeness of the remaining SVs.

Fig. 3(b) shows the result obtained when the algorithm is applied on the Boumerdes data set. It can be noted that the algorithm provides reasonable results in terms of the variance of the classification accuracy as well as the accuracy with respect to the number of SVs that remains constant until 70% of the SVs are removed.

D. Set Reduction by Optimization Problem

Fig. 4 shows the results obtained using the algorithm defined in Section III-D. From these results, it can be shown that the mean error increases with the decrease of the number of SVs. However, the variance value remains high independently from the number of SVs.

Fig. 4(a) shows the results of the application of the algorithm on the Goma data set. Despite a large variance of the classification accuracy, the algorithm provides good results, where the mean classification accuracy is being maintained at a relatively constant level with the removal of the SVs.

Similar to the tests on the Goma data set are the tests on the Boumerdes data set. Fig. 4(b) shows that the algorithm is capable of preserving the classification accuracy despite the removal of the SVs. Hence, the obtained approximation of the separating hyperplane is satisfactory.

After the inspection of the different obtained results, it can be observed that, in terms of classification accuracy, the mechanical-analogy-based algorithm provides the best results. The algorithm using the optimization problem is also efficient but suffers from the following two drawbacks: 1) high computational cost during the merging phase and 2) high error variance independently from the number of available SVs. The algorithms using the distance to the hyperplane and the direct filtering of the Lagrange multipliers provide irregular results, probably due to the relearn step that could change the shape of the separating hyperplane.

E. Classification Time

In this section, a comparison between the different methods is provided in terms of computational time. Table I shows the amount of reduction opposed to the average classification time.

As can be seen from these results, despite the computational cost added for the merging schemes represented by the optimization problem or the mechanical analogy set reduction techniques, the computation time reduction as proposed in these methods remains interesting. Note, however, that the computation time required for the filtering techniques (i.e., represented by the Lagrange multiplier set reduction and the distance-to-hyperplane set reduction techniques) has a significantly less computation time, but, as shown earlier, this large decrease of the computation time comes at the expense of the classification accuracy.

V. Conclusion

In this letter, four different SV set reduction algorithms have been proposed. The objective of these different algorithms is to accelerate the classification process for a generic SVM-based change detection algorithm in the context of risk management in order to respect crucial operational time constraints. Since the classification time is directly proportional to the number of SVs in the nonlinear case, the adopted strategy was to obtain a reduced set of the initial SVs. It was noted that the algorithms using the mechanical analogy (Section III-C) and the optimization problem (Section III-D) show good performances. The algorithm using the mechanical analogy provides the best results since it overperforms the algorithm using the optimization problem both in required computation and in the classification accuracy.

REFERENCES