ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

Reject Option Paradigm for the Reduction of Support Vectors

Ricardo Sousa[1] Ajalmar R. da Rocha Neto[2] Guilherme A. Barreto[3]
Jaime S. Cardoso[4] Miguel T. Coimbra[1] *

1- Instituto de Telecomunicações, Faculdade de Ciências da Universidade do Porto
2- Federal Institute of Ceará
3- Departamento Engenharia de Teleinformática, Universidade Federal do Ceará
4- INESC TEC and Faculdade de Engenharia da Universidade do Porto

Abstract. In this paper we introduce a new conceptualization for the reduction of the number of support vectors (SVs) for an efficient design of support vector machines. The techniques here presented provide a good balance between SVs reduction and generalization capability. Our proposal explores concepts from classification with reject option. These methods output a third class (the rejected instances) for a binary problem when a prediction cannot be given with sufficient confidence. Rejected instances along with misclassified ones are discarded from the original data to give rise to a classification problem that can be linearly solved. Our experimental study on two benchmark datasets show significant gains in terms of SVs reduction with competitive performances.

1 Introduction

It is well known that Support Vector Machines for pattern classification guarantee good generalization capabilities [1]. However, they carry a serious drawback: the run time can be considerably high due to the complexity of the decision function (related to the number of support vectors–SVs) [2]. This is particularly important on low-computational, e.g. embedded, systems. To solve this, Reduced Set (RS) algorithms have been seen as a reliable learning approach for the reduction of the number of SVs [2], often with minimal impact on performance [3, 4, 5, 6]. In [7] an online approach was devised to limit the number of SVs that a SVM should have based on different methods for aggregating SVs. Other approaches consider improving the efficiency of the SVMs prediction by reformulating the problem or by using derivatives of the SVM [8] (and references therein). Separable Case Approximation (SCA) and its Smoothed SCA (SSCA) derivation was proposed recently in [9]. This method flips the labels or removes misclassified instances that are farther than a given threshold from the hyperplane to generate a linearly separable hyperplane (flip and remove strategy). However, this has an explicit effect over the decision hyperplane which can cause generalization losses. Reject Option (RO) algorithms have recently

*This work was financially supported by FCT (Portuguese Science Foundation) grant PTDC/EIA-CCO/109982/2009, PTDC/SAU-ENB/114951/2009 and by Project I-City for Future Health NORTE-07-0124-FEDER-000068.
Corresponding authors: {rsousa, mcoimbra}@dcc.fc.up.pt, ajalmar@ifce.edu.br, gbarreto@ufc.br, jaime.cardoso@inescporto.pt

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

regained interest [10, 11] in the machine learning field based on a much older work [12]. RO algorithms automate only those decisions which can be reliably predicted, labeling the critical ones for a human expert to analyze – the rejected instances. In this work we present the RO paradigm [11] as a generalization of the SCA methods for the RS problem.

The contributions are manifold: Firstly, a new paradigm is introduced to tackle the RS problem by exploring RO mechanisms. As a consequence, the proposed approach avoids explicitly the tuning of the decision hyperplane to define a smaller training set towards a minimal number of SVs on SVMs. Moreover, and to the best of our knowledge, this is the first work that tries to obtain a minimum number of SVs and, at the same time, tries to maximize the accuracy; and, finally, we conduct several experiments where significant gains on the SVs reduction are achieved with comparable performances attesting the benefit of the proposed approach.

2 Reduced Set Methods

In this Section we summarize two recent RS methods: Separable Case Approximation (SCA) and Smoothed Separable Case Approximation presented in [9].

Separable Case Approximation (SCA): In a nutshell, SCA [9] was proposed to achieve a reduced set solution by means of a linearly separable problem. It is a method to handle classification tasks that achieves an approximation of the SVM solution from a separable problem in the feature space.

SCA, and following methods, introduces two *criteria* to attain the reduced set: *label flipping and instances removal*, respectively. SCA is reported to provide a high impact of the SVs reduction with a considerable drawback since the norm of the weight vector ($\|w\|^2$) can be arbitrary large [9]. To overcome this issue, the authors in the same work proposed a smoothed approach of the SCA which is described in the following paragraphs.

Smoothed Separable Case Approximation (SSCA): SSCA follows the same rationale of SCA with the difference of attaining a compromise between accuracy in the training set and the norm of the weight vector. Such is achieved through the incorporation of a new criterion that removes instances that are closer than a threshold D. This method has, however, an adverse effect on the decision hyperplane by removing instances blindly. The following section we will present a new concept that solves this issue.

3 Reject Option for Reduced Set Algorithms

RO aims to improve the classification's reliability in decision support systems. Basically, it consists in withholding the automatic classification of an item, if the decision is considered not sufficiently reliable. Patterns in the reject region can then be handled by a different classifier, or manually by a human. The design of classifiers with RO can be systematized in three different methods:

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

Method 1: It involves the design of a single, standard binary classifier. If the classifier provides some approximation to the a posteriori class probabilities, $\mathcal{P}(\mathcal{C}_k|\mathbf{x})$, $k = 1, 2, ..., K$, then a pattern is rejected if the largest value among the K posterior probabilities is lower than a given threshold, say β $(0 \leq \beta \leq 1)$ [11].

Method 2: The design of two, *independent*, classifiers. A first classifier is trained to output $\mathcal{C}_{-1}$ only when the probability of $\mathcal{C}_{-1}$ is high and a second classifier is trained to output $\mathcal{C}_{+1}$ only when the probability of $\mathcal{C}_{+1}$ is high. When both classifiers agree on the decision, the corresponding class is outputted. Otherwise, in case of disagreement, the reject class is the chosen one [11]. This method can be extended to handle K classes through conventional multi-class decomposition rules (e.g., one-against-one).

Method 3: The design of a single classifier with embedded RO; that is, the classifier is trained following optimality criteria that automatically takes into account the costs of misclassification and rejection in their loss functions, leading to the design of algorithms specifically built for this kind of problem [11].

To properly consider the two different types of outcomes (the predicted and rejected instances), Chow [12] devised a specific formulation through the minimization of the empirical risk $\omega_r R + E$, where R accounts for the rejection rate and E for the misclassification rate [12, 11]. This paradigm allows us to extend it to RS algorithms and present it as a generalization of the SCA.

3.1 Reject Option for Separable Approximation Algorithms

Based on the insight gained, it is possible to tailor classifiers with RO to solve the reduced set problem. Model overfitting is controlled through a cross-validation scheme by minimizing the empirical risk where the rejection cost, ω_r, induces the rejection region size. See [11] for further details.

Regardless of the chosen strategy, in a reduced set setting, data points falling in the rejection region will also be considered for achieving a reduced set. Since a classifier will be more prone to error on these instances, thus biasing the decision hyperplane during the training phase, we can choose to flip (or remove) their labels from the training set. However, removing rejected instances will generate smaller training sets but not necessarily a separable one. The final step is attained by flipping or removing misclassified instances of the training set as suggested in [9]. In the following paragraphs we will outline the incorporation of RO algorithms for reduced set problems.

Method 1 Reduced Set (M1RS): For this method, we first train a SVM model to find the best parameterization. The reject region is determined only *after* the training phase by ascertaining the threshold which best minimizes the empirical risk.
▷ **STEP 1:** Find the non-pruned solution using SVM for the non-separable case;
▷ **STEP 2:** Modify the training set by applying rules a) or b) for the instances that fall in the reject region $(\max_k \mathcal{P}(\mathcal{C}_k|\mathbf{x}) < \beta)$ and that were misclassified $y_k \left(w^T \phi(x_k) + b\right) \leq 0$: (a) Flip their labels; (b) Remove the instances.

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

▷ **STEP 3:** Given the modified training set, find the standard (without rejection) SVM solution for the separable case.

One limitation with M1RS approach is concerned with the estimation of the threshold β that best minimizes the empirical risk [11]. In the limit, M1RS is reduced to the SSCA approach when $\omega_r \to 0$ and when no cross-validation over β is conducted. The following method (M2RS) relaxes this limitation.

Method 2 Reduced Set (M2RS): For this method we first train two SVM models according to Method 2 as described in the beginning of Section 3. After obtaining the best parameterization with respect to each ω_r, the reduced set approach is conducted as follows:

▷ **STEP 1:** Find the non-pruned solutions using 2 SVMs for the non-separable case;
▷ **STEP 2:** Modify the training set by applying rules a) or b) for the instances that fall in the reject region (instances with different predicted labels by the two learners) and that were misclassified $y_k \left(w^T \phi(x_k) + b\right) \leq 0$: (a) Flip their labels; (b) Remove the instances.
▷ **STEP 3:** Given the modified training set, find *one, standard* (without rejection) SVM solution for the separable case.

Similarly to M1RS, the removal of rejected and misclassified instances suffices to produce a linearly solvable problem.

4 Experimental Study

Datasets: The performances of the proposed methods were assessed over two well-known benchmark datasets: `pendigits` and `ijcnn`.[1] The multiclass dataset `pendigits` was converted to two-class by representing the digits 1, 2, 4, 5 and 7 as -1 and 3, 6, 8, 9 and 0 as $+1$ as in [7] and the training data was rescaled to the $[-1, 1]$ range. Then, testing data were adjusted accordingly. `pendigits` is comprised by **7494** training instances and 16 dimensions whereas the `ijcnn` by **35000** training instances and 22 dimensions.

Performance Discussion: For simplicity reasons of our experimental study the misclassification trade-off and SVM margin parameter C was fixed to 100 ($C = 100$) [1]. Due to the size of the `ijcnn` dataset, we limited the training size to 27000 instances as in [7]. To obtain an estimative of the performances of the proposed methods over these datasets, we repeated the simulation 5 times and averaged the results. As mentioned, the parameterization of our proposals is chosen according the minimization of the rejection loss (w.r.t. a given ω_r). Afterwards, SVMs are trained with the RS and accuracy is assessed in the testing data (see Section 3.1). These values do not control directly the rejection rate but define how 'much does it cost to reject' (please see [11] for more details).

[1] Datasets publicly available on the UCI and LIBSVM websites.

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

Finally, in our experiments we analyzed only the remove criterion since it yielded significantly better results than the flip approach (as it was also stated in [9]). We will use the symbol "—" to represent that we obtained the same values that the non-pruned version (baseline).

Overall, our approaches attained competitive results or outperformed SCA and SSCA state-of-the-art methods without undermining accuracy rates. In some particular cases we observed, however, that our proposals performed a conservative SVs reduction. This has to do with the fact that our methods favored the accuracy over the reduction of SVs. Inspecting the results in the

Table 1: Results for `pendigits` dataset by polynomial kernel with degree 3. SCA: Separable Case Approximation; SSCA: Smoothed SCA (both from [9]); M1RS: method 1 reduced set; M2RS: method 2 reduced set. (†SVs Reduction and ‡Accuracy Variation)

Method		Acc.	SVs	SV.R.† (%)	Acc.V.‡ (%)
baseline	—	98.1	214	—	—
SCA [9]	flip	98.1	214	—	—
	remove	98.1	214	—	—
SSCA [9]	remove ($D = 0$)	97.0	160	25.2	−1.1
	remove ($D = 0.3$)	97.0	160	25.2	−1.1
	remove ($D = 0.9$)	97.0	160	25.2	−1.1
	remove ($D = 1.3$)	98.3	143	33.1	+0.2
M1RS	remove ($\omega_r = 0.44$)	98.8	210	1.8	+0.7
	remove ($\omega_r = 0.24$)	98.8	210	1.8	+0.7
	remove ($\omega_r = 0.04$)	98.8	210	1.8	+0.7
M2RS	remove ($\omega_r = 0.44$)	98.8	174	18.6	+0.7
	remove ($\omega_r = 0.24$)	98.8	159	25.7	+0.7
	remove ($\omega_r = 0.04$)	99.0	147	31.3	+0.9

dataset `pendigits` (see Table 1) we see that M1RS could not reduce more than a little less than 2% and M2RS attained similar results than SSCA methods. As we mentioned early in Section 3, RO discard instances when it is not possible to assess with sufficient confidence the prediction. When rejected instance neglect the accuracy rates our methods will take a very conservative posture. Fewer instances are removed generating more SVs candidates. Despite this demeanor we can see nevertheless an improvement of the accuracy (a gain of 0.9% of performace for the M2RS in relation to the baseline). For the `ijcnn` (see Table 2), M2RS attains a notable drop of the number of SVs in contrast to SSCA. Furthermore, M2RS attains a far more aggressive performance. Almost 80% of SVs from the non-pruned version are removed and still an accuracy improvement is registered.

5 Conclusions

We present the reject option (RO) paradigm as a generalization for the RS problem. Our analysis on the SCA state-of-the-art methods and experimental study clearly shows the benefits of using the RO paradigm for RS problems where a good compromise between accuracy and number of SVs was achieved.

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

Table 2: Results for `ijcnn` dataset by polynomial kernel with degree 3. SCA: Separable Case Approximation; SSCA: Smoothed SCA (both from [9]); M1RS: method 1 reduced set; M2RS: method 2 reduced set. (†SVs Reduction and ‡Accuracy Variation)

Method		Acc.	SVs	SV.R.† (%)	Acc.V.‡ (%)
baseline	–	95.2 ± 0.0	673.2 ± 34.6	—	—
SCA [9]	flip	95.2 ± 0.0	683.4 ± 32.1	-1.4	—
	remove	94.8 ± 0.0	670.6 ± 35.6	0.4	-0.4
SSCA [9]	remove ($D = 0$)	95.1 ± 0.0	296.4 ± 11.3	56.0	-0.1
	remove ($D = 0.3$)	95.1 ± 0.0	296.4 ± 11.3	56.0	-0.1
	remove ($D = 0.9$)	94.9 ± 0.0	296.2 ± 20.4	56.0	-0.3
	remove ($D = 1.3$)	94.9 ± 0.0	255.2 ± 10.8	62.1	-0.3
M1RS	remove ($\omega_r = 0.44$)	94.3 ± 0.0	381.2 ± 08.9	43.3	-0.9
	remove ($\omega_r = 0.24$)	94.3 ± 0.0	377.8 ± 10.2	43.9	-0.9
	remove ($\omega_r = 0.04$)	94.3 ± 0.0	362.0 ± 15.7	46.2	-0.9
M2RS	remove ($\omega_r = 0.44$)	96.2 ± 0.0	236.0 ± 3.8	64.2	$+1.0$
	remove ($\omega_r = 0.24$)	96.9 ± 0.0	215.6 ± 9.3	65.9	$+1.7$
	remove ($\omega_r = 0.04$)	95.6 ± 0.0	136.2 ± 5.9	79.3	$+0.4$

References

[1] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 1995.

[2] C. J. C. Burges. Simplified support vector decision rules. In *Proceedings of the 13th International Conference on Machine Learning (ICML'96)*, pages 71–77. Morgan Kaufmann, 1996.

[3] B. Tang and D. Mazzoni. Multiclass reduced-set support vector machines. In *Proc. of the 23rd International Conf. on Machine Learning*, pages 921–928, 2006.

[4] Wolf Kienzle, Matthias O Franz, Bernhard Schölkopf, and Gökhan H Bakir. Face detection—efficient and rank deficient. In *Advances in Neural Information Processing Systems*, pages 673–680, 2004.

[5] A. Hussain, S. Shahbudin, H. Husain, S. A. Samad, and N. M. Tahir. Reduced set support vector machines: Application for 2-dimensional datasets. In *Proc. of the Second International Conf. on Signal Processing and Communication Systems*, 2008.

[6] Ajalmar RR Neto and Guilherme A Barreto. Opposite maps: Vector quantization algorithms for building reduced-set svm and lssvm classifiers. *Neural processing letters*, 37(1):3–19, 2013.

[7] Zhuang Wang and Slobodan Vucetic. Online training on a budget of support vector machines using twin prototypes. *Statistical Analysis and Data Mining*, 3(3):149–169, 2010.

[8] Cijo Jose, Prasoon Goyal, Parv Aggrwal, and Manik Varma. Local deep kernel learning for efficient non-linear svm prediction. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 486–494, 2013.

[9] D. Geebelen, J.A.K. Suykens, and J. Vandewalle. Reducing the number of support vectors of svm classifiers using the smoothed separable case approximation. *IEEE Transactions on Neural Networks and Learning Systems*, 23(4):682 –688, 2012.

[10] Ignazio Pillai, Giorgio Fumera, and Fabio Roli. Multi-label classification with a reject option. *Pattern Recognition*, 2013.

[11] Ricardo Sousa and Jaime S. Cardoso. The data replication method for the classification with reject option. *AI Communications*, 26:281–302, 2013.

[12] C. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970.