Combining Negative Selection with Immune K-Means Algorithm for Improving the Support Vector Machines Method

Michał Bereta[1] and Tadeusz Burczyński[1,2]

[1] Cracow University of Technology, Institute of Computer Modeling, Cracow, Poland,
email: beretam@torus.uck.pk.edu.pl
[1,2] Silesian University of Technology, Department for Strength of Materials and Computational Mechanics, Gliwice, Poland, email: Tadeusz.Burczynski@polsl.pl

Abstract This paper presents a novel method of using the ideas from Artificial Immune Systems for improving the performance of the Support Vector Machines. By means of Immune K-Means algorithm a set of artificial data is generated based on the original training data. The artificial data describes the most important information from the classifiers learning point of view - the information about the boundaries among the classes remain in the artificial data. Combining the Immune K-Means algorithm with Negative Selection methods allows for further improvements of the artificial data set. The proposed approach allows to speed up the learning process of SVM when the training data set is large by extracting the most important information first. The proposed method can also be used as a data compression, especially suited when the information about boundaries among classes is an important issue. The artificial data can be created once and then used for parameters tuning of different classification methods, speeding up the learning process.

1 Introduction

Among many well know classifiers learning techniques (like Neural Networks, Decision Trees, etc.) SVM (Support Vector Machines) [10] is the one that very often achieves the best results. It is a general purpose machine learning method based on structural risk minimization instead of empirical risk minimization, which is often the base of other approaches. SVM exhibits great generalization even when there are few training samples. Many researchers worked to improve the learning techniques of SVM to make it possible to apply SVM to very large data sets. Although much effort has been made and there exist efficient algorithms to train SVM, like SMO [8], the problem is still open.

On the other hand, Artificial Immune Systems (AIS) are relatively new techniques inspired by vertebrate immune system. The biological immune system is a complex system formed by many distributed individuals. It exhibits many features that are interesting from the computational point of view: learning, memory, self-regulation, pattern recognition, diversity maintenance. Immune system has been an inspiration for computational models creation applied for engineering tasks ranging from computer security and anomaly detection to data mining and optimization [4, 3, 7, 12, 11, 1].

Immune K-Means falls to the category of artificial immune algorithms. It was first proposed in [2] showing its unique properties among others immune-inspired methods. It is less computationally demanding and is an attempt at combining the best characteristics of clonal selection algorithms and the well known K-Means clustering method. In [2] it was presented, that two different suppression mechanisms incorporated in Immune K-Means allows to evolve two different types of populations of antibodies. The first one is dedicated to discover the spatial distribution of data, the second makes it possible to evolve lymphocytes that are important for classification, thus situated close to classes boundaries.

It is well know fact that during training process SVM finds the so called *support vectors* that are these training samples that are responsible for the current shape of decision surfaces found. All other samples from the training set are irrelevant from the point of view of SVM - their presence or absence during learning does not influence the final result and decisions made by SVM. The whole process of SVM training can be described as finding these relevant training samples. However, if the training set is significantly large (hundreds or thousands of samples), the learning process becomes unmanageable.

In this paper a new methods to overcome the aforementioned problems are proposed. Immune K-Means algorithm is applied to the original training data to evolve the population of individuals as an artificial data set maintaining the most important information concerning the classification. It is shown later that the size of this artificial data set is significantly smaller and can reduce the training time of SVM providing that some conditions are satisfied. In addition, another concept from AIS, Negative Selection (NS), is applied to improve the quality of the artificial data set, allowing to model the classes boundaries more accurate.

There are almost no research works on possible ways of combining AIS with SVM. In [9] an optimization algorithm based on AIS was applied to evolve the best set of parameters of SVM. The approach presented in [9] demanded training of SVM every time when the value of the objective function was needed. It was possible, as the training data size was not big (several hundreds samples). The problem addressed in this paper is of a different nature, as the proposed method can be a useful tool when the training set is big, causing problems even when trying to train SVM only once.

The rest of the paper is organized as follows. In the next chapter basic information about SVM, Immune K-Means and Negative Selection is given. In chapter three the proposed approach of combining these three aforementioned methods is presented. In chapter four some preliminary results on artificial data sets are presented together with pointing out the possible problems and ways of solving them. Other possible applications of the proposed scheme are suggested, too. In the last chapter some conclusions are drawn and future work is discussed.

2 Support Vectors and Artificial Immune Systems

In this chapter SVM and AIS are only briefly described. For details please refer to the cited works.

2.1 Support Vector Machines

The original SVM was formulated for the linearly separable two-class case. Given the training sample set $\{(x_i, y_i)\}_{i=1}^{N}$ where N is the number of training samples, SVM finds the best separating plane. Equation $y_i(w \cdot x_i + b \geq 1),\ i = 1..N,\ y_i \in \{-1, 1\}$ where w is the normal direction and b is threshold (bias), describes all planes that correctly separates the training samples. The best plane is the one with the biggest margin, i.e., the distances of the closest training samples from both classes to the plane are the biggest among all possible. The problem of finding the best plane can be formulated as a optimization of a convex function and solved as a QP (Quadratic Programming) problem. The optimization problem is:

$$Maximize_{\alpha}\ \ W(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j), subject\ to \sum_{i=1}^{N} \alpha_i y_i = 0,\ \ \alpha_i \geq 0$$

where α_i are Lagrange multipliers. If the sample set is not perfectly linearly separable, the relaxation factors (slack variables) are introduced:$y_i(w \cdot x_i + b) \geq 1 - \xi_i,\ i = 1..N,\ y_i \in \{-1, 1\}$ In the non-linear case the sample data is projected into a high-dimensional space by means of the so called *kernel functions.* After this non-linear transform the optimal separating plane is searched in a new feature space. The problem is reformulated by replacing the term $(x_i \cdot x_j)$ in the objective function by $K(x_i, x_j)$, where $K(x_i, x_j)$ is a kernel function. Popular kernels are gaussian and polynomial functions.

2.2 Immune K-Means and Negative Selection Algorithms

Biological immune systems inspired many computational models and algorithms. Most of them are based on two main principles of AIS, clonal selection and negative selection.

Immune K-Means Algorithm In clonal selection type of evolution the individuals (lymphocytes) that are most stimulated by the presence of antigens (training data) are cloned and mutated according to their stimulation level, then suppression is performed and too similar or the least stimulated lymphocytes are removed (die). Different ways of performing all the steps of clonal selection result if different algorithms which all fall in the category of AIS but are well suited to different tasks (optimization, clustering etc.) Immune K-Means [2] has the unique ability to evolve a population of lymphocytes that concentrate themselves close to the class boundaries or close to the training samples that are important for correct classification. This is because of the new type of suppression proposed in [2]. The new suppression focuses itself more on usefulness of the lymphocytes for the correct classification then on the other types of individuals stimulation.

Negative Selection Algorithms Negative Selection (NS) is the second main principle of immune systems. Its uniqueness among others learning methods results from the fact that it uses learning samples from only one class, self-samples, and based on them builds a set of detectors (lymphocytes) that try to cover the complementary space of self samples. Thus, the correctly evolved detectors react (recognize) all except the self-samples, being able to discover any anomalies (antigens) not present in training data. In NS algorithms a problem with so called *holes* exists; after detectors generation, due to given data and receptors representation schema, there exist non-self subspaces not covered by detectors which results in possible false negative alarms.

Many NS algorithms have been developed with different representations of training samples and detectors [6, 1, 3, 7, 5]. In this work V-detector algorithm with spherical detectors was used as the real-valued NS algorithm [6].

3 Combining Immune K-Means, Negative Selection and SVM

Each of the aforementioned algorithms has its unique features. On the other hand, there are some limitations for each of them. In this chapter they will be pointing out and it will be proposed how limitations of one approach can be solved by advantages of the others.

3.1 Immune K-Means and SVM

Immune K-Means and SVM both concentrates on training samples important for classification. It is very likely that both make similar choices. However, both of them have some limitations. Immune K-Means works as a Nearest-Neighbor classifier during learning and exploratory phase which results in generalization definitely lower than generalization possible for SVM. SVM however needs much more time when the training set is big and the boundaries are complex. There is also an issue of SVM parameters that have to be carefully chosen which results in the need of applying the SVM learning several times. It can be intractable in the case of huge training set.

The solution proposed in this work is to apply Immune K-Means algorithm to the original training data first and to train SVM next not on original data but using the population of lymphocytes evolved by Immune K-Means instead. The population of lymphocytes serves as an artificial data set which is smaller than the original, however the information necessary for correct classification remains.

Figure 1 illustrates the idea. Three classes are present. It can bee seen that the boundaries found by SVM based on all samples available are similar to those found when only artificial data from Immune K-Means were used.

3.2 Immune K-Means and Negative Selection

Immune K-Means uses Nearest Neighbor method which can result in a fact that the population of evolved lymphocytes correctly classifies training samples but does not describe the class boundaries correctly. It will happen when the classes are relatively far away from each other and only a few lymphocytes are necessary for the correct classification with *NN* method. Immune K-Means finds lymphocytes better describing the boundaries when the classes are close,which is not always the case.

There are two ways to overcome this problem. The first one is to apply Immune K-Means several times and to use all populations evolved in conjunction to train SVM. It goes from the fact that Immune K-Means can result in slightly different population evolved every time due to its stochastic nature.

The second idea is to use Negative Selection on samples from each class separately. Having the NS detectors it is possible to generate artificial data for each class. Artificial samples from each class covers the complementary subspace of this class. These artificial

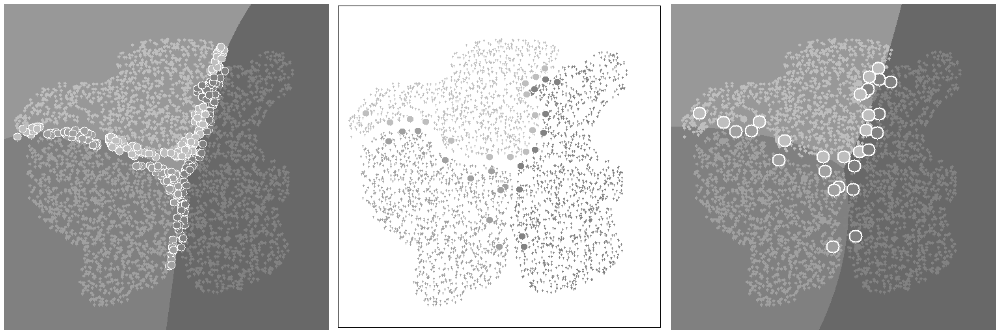

Figure 1. SVM with Immune K-Means. Left: SVM on original data. Support vectors marked as bigger circles. Middle: Population of lymphocytes found by Immune K-Means (marked as bigger circles). Right: SVM applied on artificial data from Immune K-Means. All lymphocytes marked as support vectors by circles.

samples for each class are then used in Immune K-Means which is applied for each class using training samples from this class and artificial samples from NS for this class.

The algorithm goes as follows:

ALGORITHM 1

cnum: number of classes
T_i : training data from $i-th$ class
D_i : detectors found by NS algorithm using samples only from $i-th$ class
T_i^{compl} : complementary artificial samples (from complementary subspace) for $i-th$ class
T_i^{art} : artificial samples for $i-th$ class found by Immune K-Means
T^{art} : artificial samples from all classes
For each class i :

1. Apply Negative Selection on T_i
2. Using detectors from D_i generate artificial samples for $i-th$ class:
   - $T_i^{compl} = \varnothing$
   - Generate random point. If it is recognized by any detector from D_i, add it to T_i^{compl}, else discard it. Repeat until sufficient number of artificial samples exists in T_i^{compl}.
3. Apply Immune K-Means using $T_i \bigcup T_i^{compl}$ as training data. Save in T_i^{art} only those lymphocytes from the resulting population which represents T_i class.

Finally, construct $T^{art} = \bigcup_{i=1}^{cnum} T_i^{art}$

Figure 2 shows the result of applying the algorithm to 2D data. NS algorithm used was V-detector from [6].

3.3 Negative Selection, Immune K-Means and SVM

Based on the previous sections it is easy to propose a general method that combines NS, Immune K-Means and SVM. The algorithm follows:

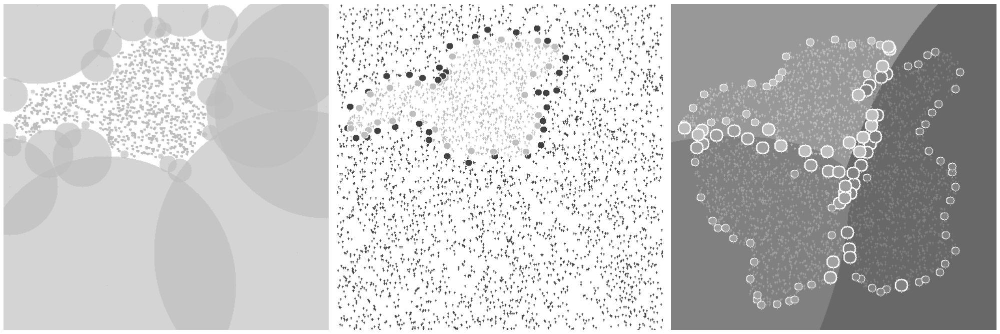

Figure 2. Presentation of the proposed method. Left: Samples from one of the classes and spherical detectors generated by means of Negative Selection (V-detector algorithm). Middle: Samples from one of the classes and artificially generated new samples from complementary space together with lymphocytes near the boundaries evolved by Immune K-Means. Right: All classes with lymphocytes generated by NS and Immune K-Means for each class separately together with the result of applying SVM on artificial samples. The circles represents artificial data (lymphocytes), the biggest circles are those that were considered as support vectors.

ALGORITHM 2

```
1. Using original training samples $T$ construct $T^{art}$ by means of ALGORITHM 1.
2. Train SVM using $T^{art}$.
```

Next section presents some preliminary results of the proposed methodology.

4 Preliminary Results

The proposed method was tested on artificially generated 2D data with different class boundaries complexity. All examples were three-class classification problems. Figure 3 shows the example results of applying SVM on artificial data generated by **ALGORITHM 1**. The learning algorithm of SVM was SMO [8].

Figure 4 shows the times needed by each approach when the size of the training set grows as well as the errors made by each approach. All data was generated from $[0,1]^2$. The parameters of the V-detector algorithm were: $\alpha = 0.995$, $selfradius = 0.005$ and $Tmax = 5000$. Starting size of the population in Immune K-Means was set to 30 and the number of iterations was set to 15. In each experiment there were three passes of Immune K-Means if it was used alone (without Negative Selection). In SVM gaussian kernel was used with $\sigma = 1$. All results presented were averaged over 10 runs.

In this set of experiments the proposed methodology does not outperforms the pure SVM when the training data set is small, however when the size of the training set grows, the superiority of the proposed method is evident. It can be observed that the smaller size of artificial data set used in SVM training results often in higher error rates, however the profit in computational time is the key positive effect especially as the growth in error rates is not big. It can be also observed that the proposed method in the first data set made unnatural big errors when the data set size was small. This is the result of poor choices of parameters values of Negative Selection, especially $selfradius$, which was

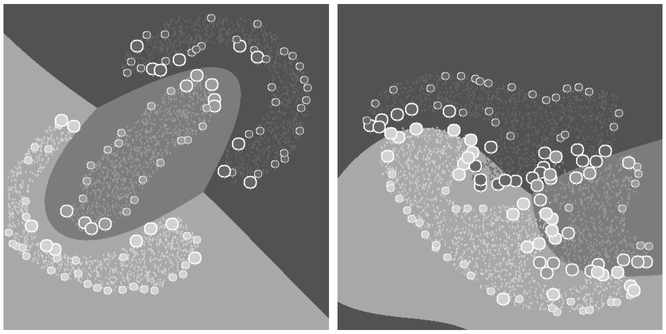

Figure 3. Two sets of three-class classification problems. Results obtained by means of the proposed approach.

set to small value and detectors inside the classes were generated resulting in difficulties in Immune K-Means while evolving lymphocytes and further while generating artificial data sets.

5 Conclusion

A novel method of using ideas from AIS to improve the performance of SVM was proposed. By first applying NS and Immune K-Means algorithms, a set of artificially generated samples are used for training SVM resulting in significantly smaller computational times. The proposed method is not able to outperform standard SVM when the size of the training set is small, and also when the class boundaries are simple. However, there are many situations when SVM training is intractable due to the complexity of the problem and huge training sets. The proposed method can be used to compress the original data first, remaining the information important for classification.

Another important way of using the proposed methodology is for parameters tuning of SVM which can be done on artificial data set, when applying SVM on original data several times to try different parameters is not possible due to the computational demands. The proposed method is a possible solution in such cases.

Negative Selection algorithm is a key feature of the proposed approach. Generating a correct set of complementary samples for each class allows Immune K-Means to find proper lymphocytes close to real class boundaries. Developing better Negative Selection algorithms will result in better performance of the presented method, as it is independent on actual NS algorithm used.

Bibliography

[1] M. Bereta and T. Burczyński. Hybrid immune algorithm for feature selection and classification of ECG signals. In T. Burczyński, W. Cholewa, and W. Moczulski, editors, *Recent Developments in Artificial Intelligence Methods*, AI-METH Series, pages 25–28, Gliwice, 2005.

[2] M. Bereta and T. Burczyński. Immune k-means: A novel immune algorithm for

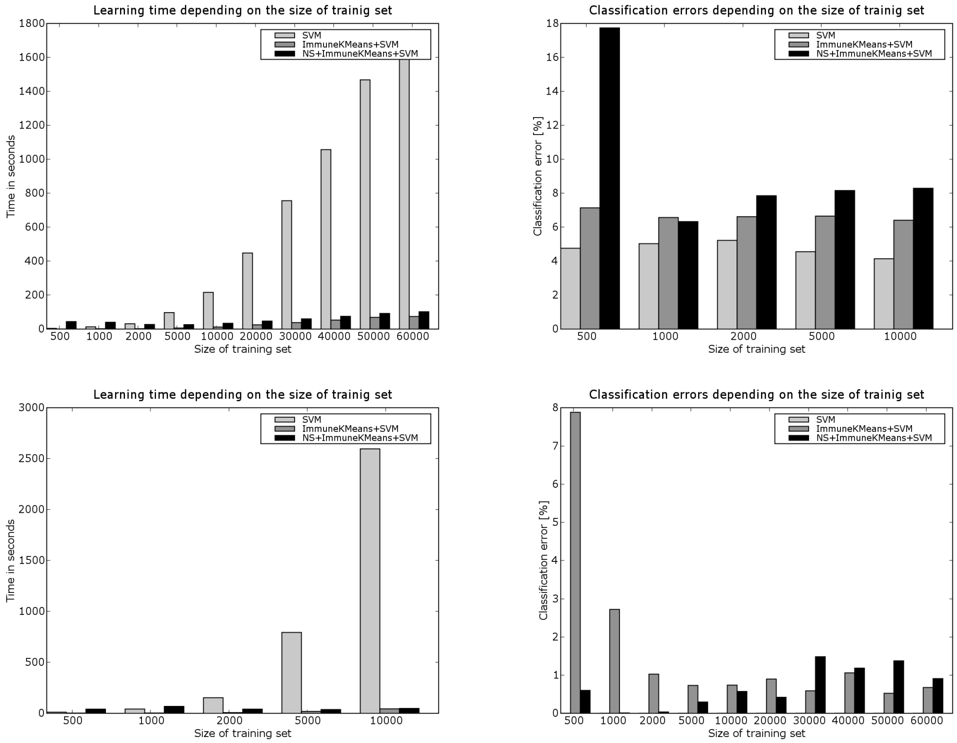

Figure 4. Results for classification problems presented in Figure 3. First column: Dependence of times needed by each approach on the size of training set. Second column: Errors made by each method.

data clustering and multiple-class discrimination. In *Proceedings of the IX Conference Evolutionary Computation and Global Optimization, KAEIOG 2006, Prace Naukowe PW s. Elektronika z.156, Oficyna Wydawnicza PW, Warszawa 2006*, pages 49–60, 2006.

[3] D. Dasgupta and S. Forrest. Novelty detection in time series data using ideas from immunology. In *ISCA 5th International Conference on Intelligent Systems*, pages 19– 21, Reno, Nevada, June 1996.

[4] L.N. de Castro and J. Timmis. *Artificial Immune Systems: A New Computational Approach.* Springer-Verlag, London. UK., September 2002.

[5] Fabio A. González, Dipankar Dasgupta, and Luis Fernando Niño. A randomized real-valued negative selection algorithm. In *ICARIS, Artificial Immune Systems, Second International Conference, ICARIS 2003, Edinburgh, UK, September 1-3, 2003*, pages 261–272, 2003.

[6] Z. Ji and D. Dasgupta. Real-valued negative selection algorithm with variable-sized detectors. In Deb K. et al., editor, *International Conference on Genetic and Evo-*

lutionary Computation (GECCO-2004), pages 287–298, Seattle, Washington USA, June 26-30 2004. Springer-Verlag.

[7] P.D'haeseleer, S. Forrest, and P. Helman. An immunological approach to change detection: algorithms, analysis, and implications. *In Proceedings of the 1996 IEEE Symposium on Computer Security and Privacy*, 1996.

[8] J. Plat. Sequential minimal optimization: A fast algorithm for training support vector machines. technical report 98-14. Technical report, Microsoft Research, Redmond, Washington, http://www.research.microsoft.com/ jplatt/smo.html, 1998.

[9] Y. Shengfa and C. Fulei. Fault diagnosis based on support vector machines with parameter optimisation by artificial immunisation algorithm. *Mechanical Systems and Signal Processing*, 21:1318–1330, 2007.

[10] Vladimir N. Vapnik. *The nature of statistical learning theory.* Springer-Verlag New York, Inc., New York, NY, USA, 1995.

[11] Andrew Watkins, Jon Timmis, and Lois Boggess. Artificial immune recognition system (AIRS): An immune-inspired supervised machine learning algorithm. *Genetic Programming and Evolvable Machines*, 5(3):291–317, September 2004.

[12] S. T. Wierzchoń. *Artificial Immune Systems. Theory and Applications.* Akademicka Oficyna Wydawnicza EXIT, Warszawa, 2001.