Silesian
University
of Technology

MASTER THESIS

Data classification using artificial immune systems

Artur OLEKSIŃSKI
Student identification number: 293694

Programme: Control, Electronic, and Information Engineering
Specialisation: Data Science

SUPERVISOR
dr hab. inż. Robert Czabański, prof. PŚ
DEPARTMENT Department of Cybernetics, Nanotechnology and Data Processing
Faculty of Automatic Control, Electronics and Computer Science

Gliwice 2024

Thesis title

Data classification using artificial immune systems

Abstract

This study focuses on improving a popular artificial immune system algorithm for classification called CLONCLAS and introduces a new supervised machine learning classification algorithm based on biological immune systems called InteliAIS. The research evaluates the performance of the modified CLONCLAS methods and the new InteliAIS algorithm on different parameters using the original and diagnostic Wisconsin Breast Cancer datasets. In addition, the study compares the proposed algorithm with selected machine learning procedures such as decision trees, support vector machines, and multilayer perception. InteliAIS shows promising results, achieving the highest F1-score among all algorithms analyzed.

Key words

Artificial Immune Systems, Machine Learning, Classification, Metaheuristics, Supervised Learning

Tytuł pracy

Klasyfikacja danych z użyciem sztucznych systemów immunologicznych

Streszczenie

W pracy rozważano możliwości ulepszeniu popularnej metody klasyfikacji CLONCLAS, należącej do grupy sztucznych systemów immunologicznych. Poza modyfikacjami oryginalnego klasyfikatora CLONCLAS, wprowadzono nowy algorytm uczenia nadzorowanego wzorowanego na biologicznych systemach immunologicznych o nazwie InteliAIS. Na podstawie badań oceniono jakość klasyfikacji uzyskiwanej za pomocą zaproponowanych modyfikacji CLONCLAS oraz nowego klasyfikatora InteliAIS dla różnych parametrów kontrolnych tych metod i przy zastosowaniu dwóch baz danych testowych: oryginalnego i diagnostycznego zbioru Wisconsin Breast Cancer. Ponadto w badaniu porównano InteliAIS z wybranymi (i uznanymi) algorytmami uczenia maszynowego, takimi jak klasyfikator Bayesowski, drzewo decyzyjne, maszyna wektorów nośnych i wielowarstwowy perceptron. Stosując InteliAIS uzyskano obiecujące wyniki, osiągając najwyższą wartość wskaźnika klasyfikacji (F1-score) spośród wszystkich analizowanych metod.

Słowa kluczowe

Sztuczne Systemy Immunologiczne, Uczenie maszynowe, Klasyfikacja, Metaheurystyka, Uczenie Nadzorowane

5.2 Comparison of CLONCLAS-based algorithms

A comparison of the performance of all AIS-based (CLONCLAS-based) algorithms is performed with the help of the diagnostic Wisconsin breast cancer dataset and each of the algorithms analyzed obtained a mean balanced accuracy greater than 80% (see Table 5.11). This shows that all of the methods are able to correctly classify the WDBC dataset successfully. The original CLONCLAS algorithm is behind all the proposed CLONCLAS modifications and the InteliAIS. It obtained the lowest average scores of all metrics analyzed and the highest standard deviation values. With the use of the Top-3 classification procedure, proposed in this work, the classification accuracies as well as the F1-scores for all the methods considered increased. It confirms that the Top-3 approach improves the overall quality of the classification.

The first proposed modification, called CLONCLAS with center estimation, increased the classification quality of the original CLONCLAS. It is indicated by the increase in all quality metrics except precision, which decreased when comparing results using the Top-3 approach. Similarly, as in the case of original CLONCLAS, the Top-3 showed improvement across the majority of quality metrics, however, there is a decrease in the value of recall when compared with the single classification method. The CE modification also reduced the standard deviation, indicating better reproducibility of the classification results. This feature can be seen in the violin plot presented in Figure 5.16. The starting points of "violins" show the lowest obtained result, the shape represents the approximated distribution of results, and the upper part of "violins" marks the highest obtained values. The methods that obtained "violins" with short range, located higher in the graph, and having a distribution closer to the top are considered superior in terms of the classification quality (simply, they perform better). Furthermore, the mean values and standard deviations of the algorithm obtained during the experimentation are presented in Figure 5.17.

The experimental approach of CLONCLAS without affinity did not yield satisfactory results. Most of the quality scores are lower than the scores obtained by CLONCLAS with a center estimation. Only the value of precision is marginally improved. Considering the high increase in training time and overall lack of quality, the method is not recommended for usage in real-life scenario problems. With all drawbacks in mind, the stability (repeatability of classification results) of CLONCLAS without affinity proved to be the highest among all AIS algorithms analyzed.

The InteliAIS obtained the highest classification quality among all AIS algorithms analyzed with the majority of metrics. Only the precision metric shows a significant decrease and the lowest value among the AIS algorithms. However, this is balanced by an increase in Recall (of primary importance in the case of medical data) and ultimately F1-score. The repeatability of classification results is high compared to the CLONCLAS-based algorithms. As InteliAIS obtained the best overall score, it will be the algorithm compared with the reference algorithms. To increase the validity of the comparison, two sets of Wisconsin breast cancer original and diagnostic will be used for comparison.

All of the AIS-based algorithms considered proved their ability to successfully classify the diagnostic Wisconsin breast cancer (WDBC) data. Most of the modifications introduced to the original CLONCLAS algorithm increased the quality of classification. However, even with additional improvements, CLONCLAS-based methods were not able to achieve the levels of quality offered by the reference machine learning algorithms.

Table 5.11: The mean values of the quality metrics of the AIS algorithms (CLONECLAS and its modifications) executed on the diagnostic WBC dataset.

Model	ACC	BACC	F1S	PRR	REC
CLONCLAS (Single)	0.8256	0.8056	0.7545	0.9115	0.6765
CLONCLAS (Top-3)	0.8695	0.8479	0.8019	0.9627	0.7116
CLONCLAS CE (Single)	0.8774	0.8698	0.8291	0.9124	0.7884
CLONCLAS CE (Top-3)	0.8909	0.8748	0.8435	0.9585	0.7646
CLONCLAS WA (Single)	0.8830	0.8624	0.8301	**0.9755**	0.7327
CLONCLAS WA (Top-3)	0.8844	0.8667	0.8331	0.9604	0.7457
InteliAIS	**0.9055**	**0.9075**	**0.9195**	0.8916	**0.9636**

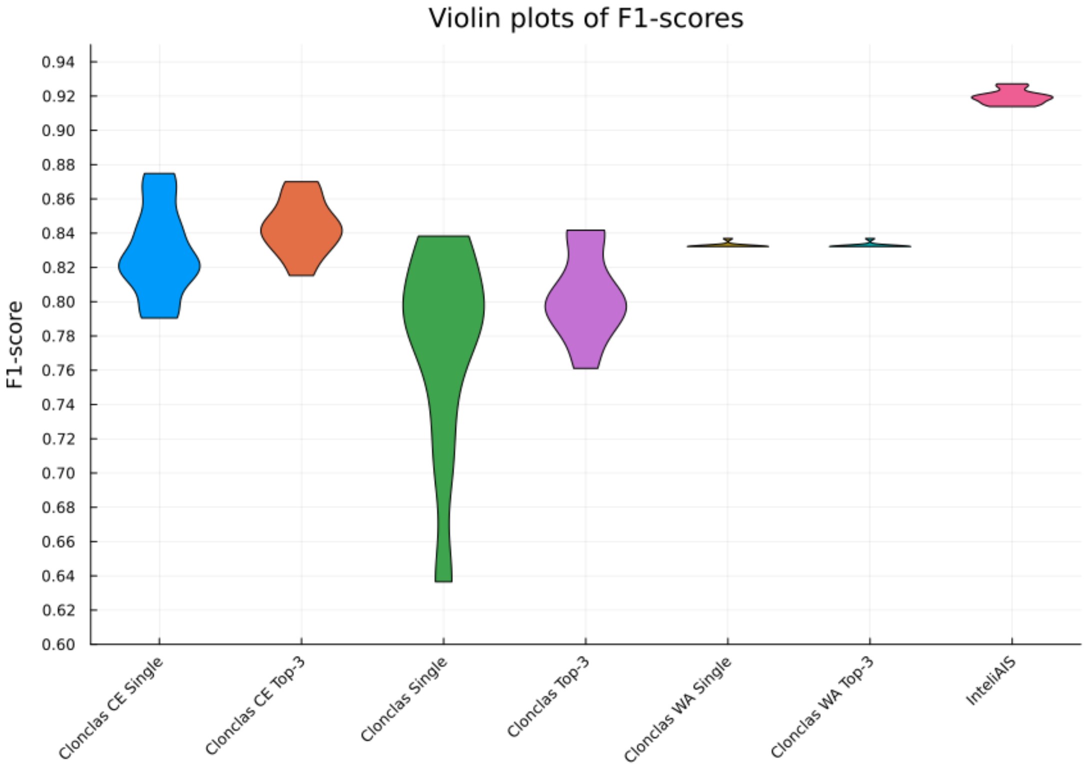

Figure 5.16: The violin plot representing the values of F1-scores of 50 executions of 5-fold cross-validation. The shape of the violins represents the distribution of values of the F1-score.

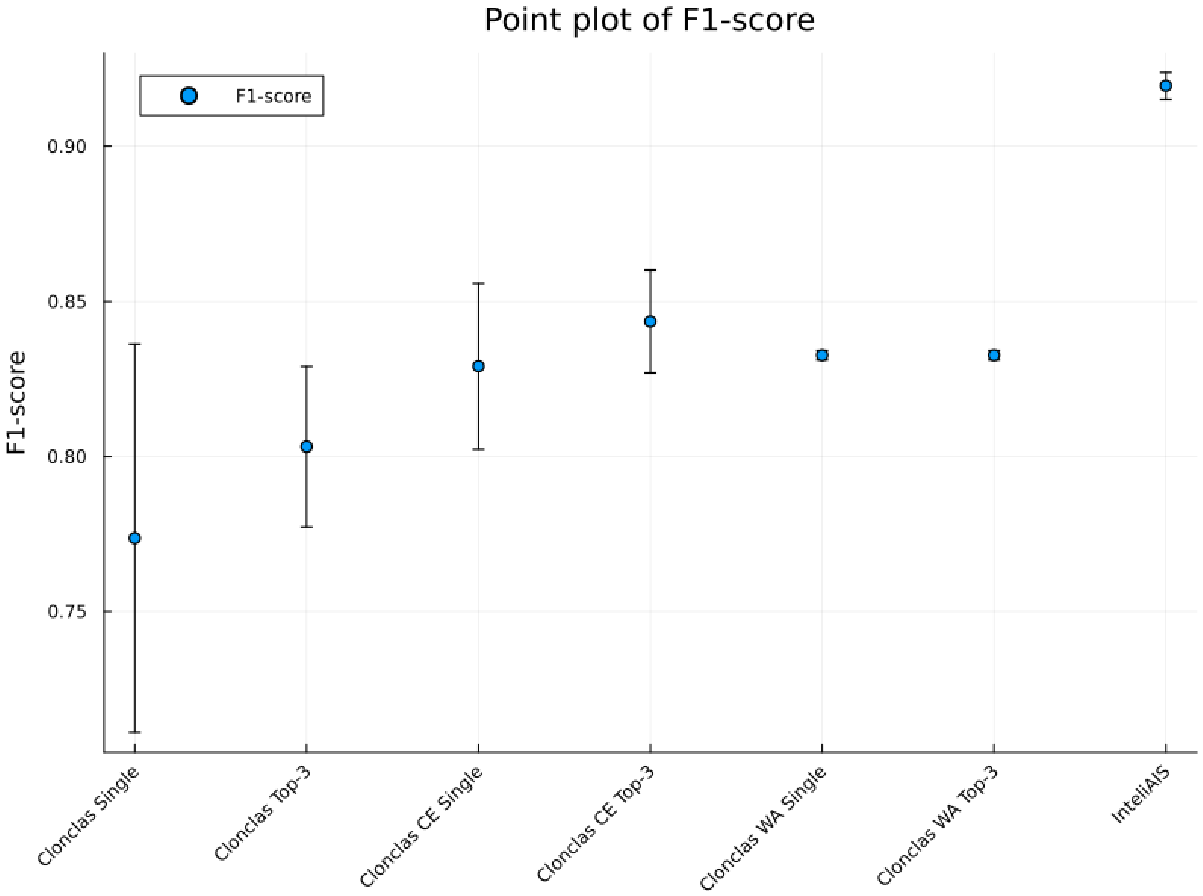

Figure 5.17: The point plot of mean F1-scores and their standard deviations for all AIS algorithms.

5.3 Comparison of InteliAIS with reference methods

Comparing InteliAIS with different families of classifiers is crucial to verify the real effectiveness of the proposed solution. InteliAIS is run with the default parameters presented in Table 5.7 and the reference algorithm uses the parameters shown in Table 5.12:

Table 5.12: Parameters of the reference machine learning methods.

Model	**Parameters**
Naïve Bayes	Gaussian Naïve Bayes (GaussianNB) var_smoothing = 1e-09

Decision Tree	DecisionTreeClassifier
	criterion = "Gini"
	splitter = "best"
	max_depth = None
	min_samples_split = 2

Support Vector Machine	SVC
	C = 1.0
	kernel = "rbf"
	degree = 3
	gamma = "scale"
	coef0 = 0.0
	shrinking = True
	tol = 1e-3

Multi-layer Perceptron	MLPClassifier
	hidden_layer_sizes = (100,)
	activation = 'relu'
	solver = 'adam'
	alpha = 0.0001
	batch_size = 'auto'
	learning_rate='constant'

5.3.1 WDBC data classification

The results of the InteliAIS evaluation compared to the reference algorithms are presented in the form of Table 5.13. As in the previous examples, the violin plot of the F1-score is shown in Figure 5.18.

The InteliAIS method did not obtain the highest classification quality with the Wisconsin diagnostic data set. With an accuracy of 90.55%, it falls behind the reference algorithms. Interestingly, InteliAIS is the only algorithm with a higher balanced accuracy than the overall classification accuracy. It also obtains the highest score of the F1-score equal to 0.9195. The second highest F1 score is obtained with Naive Bayes, equal to 0.9169 which is a small difference. However, InteliAIS achieved the highest recall score. This is an especially crucial metric when applying the algorithm to medical data (such as the WDBC and OWBC datasets). Recall measures how sensitive the algorithm is to positive cases, which means the InteliAIS has a lower chance of incorrect classifications of patients who actually have the disease. Analysis of the violin plot also indicates that the variability of the classification results obtained with the proposed algorithm also worsened compared to Naïve Bayes and SVM. The low value of precision appears to be the main problem of the current implementation of the InteliAIS algorithm and should be further analyzed.

The Naïve Bayes algorithm is the best in terms of accuracy and balanced accuracy. However, when evaluating algorithms on medical data for diagnosis, the priority is to identify the highest number of positive cases possible. Although InteliAIS did not perform as well in terms of accuracy, balanced accuracy, and precision compared to some reference algorithms, it did achieve the highest values of F1-score and recall. These metrics are important when dealing with medical data. Ultimately, InteliAIS obtained the best classification results for WDBC data when considering support for medical diagnosis.

Table 5.13: The mean values of the quality metrics of the AIS algorithms executed on the diagnostic WBC dataset.

Model	ACC	BACC	F1S	PRR	REC
InteliAIS	0.9055	0.9075	**0.9195**	0.8916	**0.9636**
SVM	0.9103	0.8873	0.8655	**0.9550**	0.7970
Decision Tree	0.9188	0.9162	0.8934	0.8778	0.9109
MLP	0.9270	0.9181	0.9004	0.9276	0.8791
Naïve Bayes	**0.9402**	**0.9300**	0.9169	0.9467	0.8909

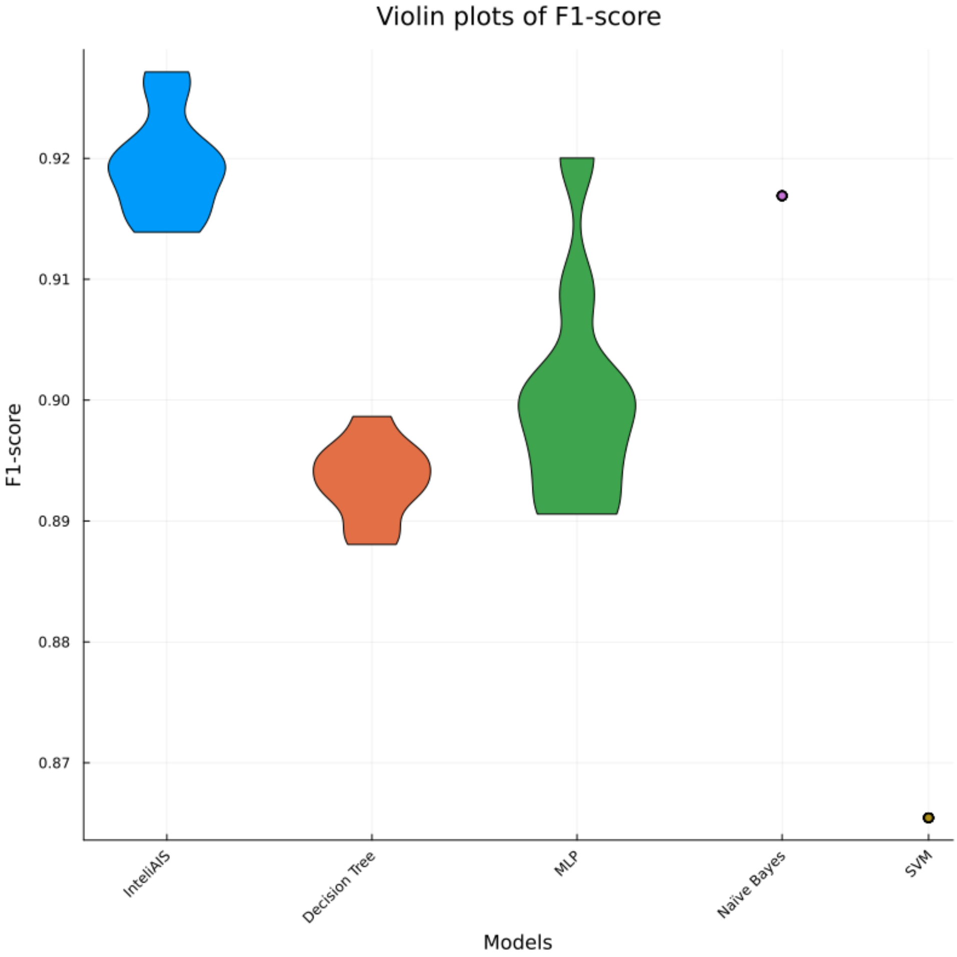

Figure 5.18: The violin plot representing the values of F1-scores of 10 mean executions of 5-fold cross-validation of each algorithm on the diagnostic WBC dataset. The shape of the violins represents the distribution of values of the F1-score.

5.3.2 OWBC data classification

InteliAIS and the reference algorithms were also tested on the original Wisconsin Breast Cancer dataset (OWBC) and the results are presented in Table 5.14. In addition, the violin plots that illustrate the performance of the algorithms are presented in Figure 5.19. For the OWBC dataset, the InteliAIS obtained the best overall classification results among all algorithms analyzed: superior results for accuracy, F1-score, Precision, and Recall. Only the naïve Bayes classifier obtains slightly higher Balanced Accuracy.

The analysis of the violin plot shows that the InteliAIS provides a stable range of results showing better stability than the decision tree and multi-layer perception, however, it is still inferior to the naïve Bayes and support vector machine which obtained a uniform value from all *k*-fold cross-validations and had to be denoted in the plot as points instead of "violins". Taking into consideration all obtained results it is safe to state the fact that the InteliAIS method is a competitive algorithm for classification tasks and can be compared with popular, well-grounded machine learning algorithms.

Table 5.14: The mean values of the quality metrics of all analyzed algorithms executed on the original WBC dataset.

Model	ACC	BACC	F1S	PRR	REC
InteliAIS	**0.9631**	0.9608	**0.9690**	**0.9659**	**0.9731**
SVM	0.9546	0.9488	0.9328	0.9413	0.9291
Decision Tree	0.9328	0.9263	0.9042	0.9198	0.8868
MLP	0.9552	0.9465	0.9343	0.9502	0.9198
Naïve Bayes	0.9590	**0.9617**	0.9436	0.9189	0.9707

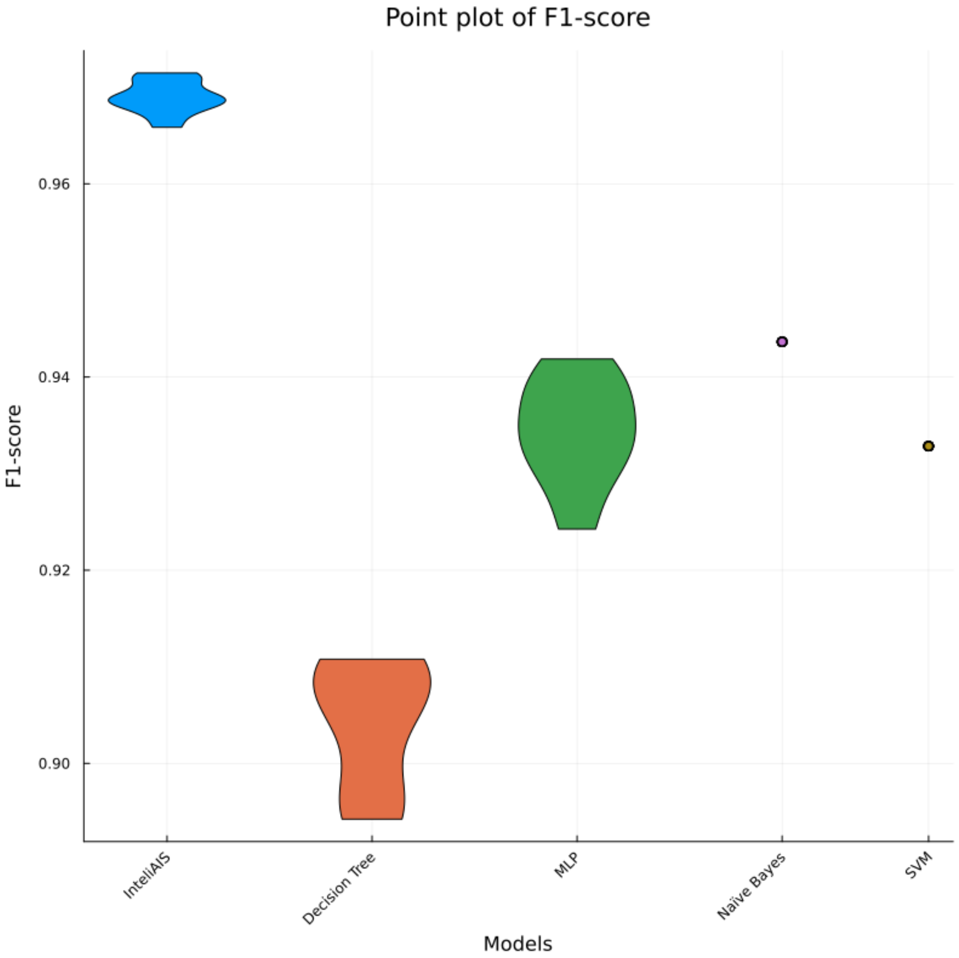

Figure 5.19: The violin plot representing the values of F1-scores of 10 mean executions of 5-fold cross-validation of each algorithm on the original WBC dataset. The shape of the violins represents the distribution of values of the F1-score.

In addition, Table 5.15 presents a comparison of the performance of the analyzed algorithms with other studies conducted using the OWBC dataset. The accuracy and F1-scores provided are evaluations of models trained on the original dataset "as is", i.e., without filtering and feature selection. The studies also compare several machine learning algorithms that were not analyzed in this thesis: Random Forest (RF), Linear Regression (LR), Neural Network (NN), J48, Sequential Minimal Optimization (SMO), and Instance-

Based for K-Nearest neighbor (IBK).

The highest classification accuracy of 97.0% is achieved with the SMO algorithm in the studies by Gouda I. Salama [39] and Siham A. Mohammed [33]. However, InteliAIS is the second-best method with an accuracy of 96.3%. The overall best F1-score of 96.9% is attained with the InteliAIS method, with RF evaluated by Yixuan Li [27] as the second best method, providing 95.5%.

Further analysis and the introduction of common filtering and feature selection techniques will be necessary to explore the potential performance of InteliAIS.

Table 5.15: Comparison of results with other studies performed on the OWBC database.

Study	**Validation**	**Algorithms**	**ACC**	**F1S**
Yixuan Li [27]	70/30 train-test-split	DT	0.961	0.941
		SVM	0.951	0.934
		RF	0.961	0.955
		LR	0.937	0.938
		NN	0.956	0.945
Gouda I. Salama [39]	10-fold cross-validation	Naïve Bayes	0.960	NA
		MLP	0.953	NA
		J48	0.951	NA
		SMO	**0.970**	NA
		IBK	0.946	NA
Siham A. Mohammed [33]	10-fold cross-validation	Naïve Bayes	0.960	NA
		J48	0.946	NA
		SMO	**0.970**	NA
This Thesis	10-times 5-fold cross-validation	InteliAIS	0.963	**0.969**
		SVM	0.955	0.933
		Decision Tree	0.933	0.904
		MLP	0.955	0.934
		Naïve Bayes	0.959	0.944

The proposed method of classification, the InteliAIS, as a proof-of-concept achieved higher overall quality than all CSAs. Furthermore, it obtained results at a similar or even higher level of classification quality when compared to the reference algorithm, indicating that the InteliAIS is competitive against commonly used classification algorithms.

Chapter 6

Summary

The main goal of this thesis was to analyze possible ideas to improve the performance of the CLONCLAS algorithm in classification scenarios. The additional challenge posed during the implementation of clonal selection algorithms was to create a new supervised machine learning algorithm that incorporates the ideas of humane immune systems into its learning and classification procedures.

Multiple algorithms of artificial immune systems were implemented from scratch and evaluated for the purpose of this study. The paper focuses on improving the learning procedures designed for CLONCLAS, a supervised classification algorithm inspired by clonal selection. The modifications that improve the learning procedures of the CLONCLAS algorithm are: CLONCLAS with center estimation and CLONCLAS without affinity. The new procedures and the original CLONCLAS algorithm were tested with different control parameters settings and showed significant improvements in classification quality compared to the original method for Wisconsin breast cancer benchmark data.

This work also proposes a brand new supervised machine learning algorithm for classification inspired by the biological immune system called InteliAIS. The InteliAIS introduces an alternative way of classification to the one used in the mentioned clonal selection algorithms. The intuitive learning procedure and highly explainable classification are introduced "by design" making it trivial to understand the main idea behind the algorithm. The method also has the capability of fitting any shape of data as it is not bounded by the center or mean, but by the groups of objects in the class. The CLONCLAS with center estimation significantly increased the value of all metrics of the classification quality. At the same time, the repeatability of the results obtained increased considerably compared to the original version of CLONCLAS. However, the CLONCLAS without affinity is a failed attempt to increase the performance of the CLONCLAS algorithm. As a variation of CLONCLAS with center estimation, its goal is to improve classification quality at the cost of longer training time. The high increase in training time, which is not reflected in the improvement of classification quality, indicates the low efficiency of the method.

In addition, the thesis introduced an alternative method of AIS-based classification called Top-3 antibodies. This improved the quality of the classification of all the clonal selection-based algorithms. While increasing the quality of classification, it also significantly decreased the classification time. With no impact on the training procedure, it did not influence the complexity of training. Because it is an alternative to the single antibody classification method, it can be used in a plug-and-play form without any disadvantages found during the analysis carried out. This means that changing the single antibody classification method with Top-3 is a direct update that does not show a negative impact on any classification performance. The InteliAIS also seems to have the ability to match the complex shape of the class boundary, as the method does not focus on the center or average of the class objects, but analyzes the groups of objects located near the class border. However, more experiments are needed to confirm this claim.

This work also proposes a new supervised machine learning algorithm for classification inspired by the biological immune system called InteliAIS. The InteliAIS introduces an alternative approach to classification compared to that used in the considered clonal selection-based algorithms. The intuitive learning procedure and highly comprehensible classification approach were explained, making it simple to understand the main idea of the algorithm. The InteliAIS proof-of-concept classification method proved to be a success. The proposed algorithm obtained a higher quality of classification than the reference methods for both benchmark datasets. The only metric of quality that obtained lower values is precision, which will be an important topic in future development. InteliAIS also has longer classification times compared to other AIS-based methods. This problem will be addressed in the future. With these cons in mind, InteliAIS at this early stage of development obtained competitive classification results when compared to popular machine learning algorithms, always obtaining the highest value of the F1-score.

The objective of this thesis was reached. This work obtained significant improvements with modifications of the CLONCLAS classification method. The additional goal of this thesis is also a success. The proposed algorithm improved the classification quality of all other CLONCLAS-based algorithms. Furthermore, it proved to be competitive with popular machine-learning algorithms, which shows its development potential. Furthermore, it proved to be competitive with the most popular and effective classifiers, indicating its possible application in real-life problems.

Future work

Most of the future efforts will be dedicated to the improvements and further development of InteliAIS. The analysis carried out in this thesis revealed some drawbacks of the current version of the method that should be addressed during the development process.

To increase the precision of the classification, it is planned to introduce an iterative modification of the Gaussian distributions that generate the search cells. At each iteration, the distributions will be adjusted with the use of mature antibodies that have already found the objects of the 'hostile' class. In this way, the method should become better at inserting search cells in the most important regions. This will also create the possibility of introducing a stopping condition based on the changes in the distributions. If the change of shape and position of the distribution is smaller than a certain threshold, then the procedure can be stopped.

The problem with increasing classification time can be resolved with an additional rule to the search cell promotion requirement, which will require the cell to be at some distance from a few closest antibodies. This will successively reduce the formation of new antibodies in regions that have already been well 'protected' (with a well-determined boundary between classes).

An important element of future analysis will be increasing the number and diversity of datasets used for algorithm evaluation. This work focused on two WBC data sets, but the algorithm may perform differently when used with other data. This requires in-depth analysis, as the algorithm has properties that may allow it to solve complex, nonlinear classification tasks, as it searches for the actual shape of the interclass boundary in the original feature space.

In addition, various code optimization improvements are planned, including parallelization of calculations, GPU support, and distributed computing capabilities. A beneficial change would be to replace the KD-tree with a hierarchical navigable small worlds data structure that could increase performance and allow the algorithm to work with the vector databases. Finally, public software distribution should be prepared once the first stable algorithm release is developed.