A Strategy for Training Set Selection in Text Classification Problems

Maria Luiza C. Passini, Katiusca B. Estébanez, Grazziela P. Figueredo, Nelson F. F. Ebecken
COPPE/UFRJ
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil

***Abstract*—An issue in text classification problems involves the choice of good samples on which to train the classifier. Training sets that properly represent the characteristics of each class have a better chance of establishing a successful predictor. Moreover, sometimes data are redundant or take large amounts of computing time for the learning process. To overcome this issue, data selection techniques have been proposed, including instance selection. Some data mining techniques are based on nearest neighbors, ordered removals, random sampling, particle swarms or evolutionary methods. The weaknesses of these methods usually involve a lack of accuracy, lack of robustness when the amount of data increases, overfitting and a high complexity. This work proposes a new immune-inspired suppressive mechanism that involves selection. As a result, data that are not relevant for a classifier's final model are eliminated from the training process. Experiments show the effectiveness of this method, and the results are compared to other techniques; these results show that the proposed method has the advantage of being accurate and robust for large data sets, with less complexity in the algorithm.**

Keywords—text mining; data reduction; classification problems; feature selection

I. INTRODUCTION

Nowadays most of the information is stored electronically, in the form of text databases. Text databases are rapidly growing due to the increasing amount of information available in electronic form, such as electronic publications, various kinds of electronic documents, e-mails, and the World Wide Web.

Text mining, also known as knowledge discovery from textual databases, is a semi-automated process of extracting knowledge from a large amount of unstructured data. Traditional information retrieval techniques become inadequate for the increasingly vast amounts of text data. Typically, only a small fraction of the many available documents will be relevant to a given individual user. Without knowing what could be in the documents, it is difficult to formulate effective queries for analyzing and extracting useful information from the data. Users need tools to compare different documents, rank the importance and relevance of these documents, or find patterns and trends across multiple documents. Thus, text mining has become an increasingly popular and essential theme in data mining (Feldman 1995).

There are many types of statistical and artificially intelligent classifiers, as it can be seen in [1],[2]. One of the main issues in classification problems involves the choice of good samples to train a classifier. A training set capable to represent well the characteristics of a class has better chances to establish a successful predictor.

II. OBJECTIVES

This paper proposes a new approach for addressing the training data reduction in text mining classifications problems. This new algorithm was inspired by suppression mechanisms found in biological immune systems [3]. The suppression concept is applied to the training process to eliminate very similar data instances and to keep only representative data. The propose consists in a non-statistical method to select samples for training. The main objectives of this work are to find a subset of samples for training without spending excessive processing time and to simultaneously maintain good accuracy.

In order to do this, this paper is set out as follows. The Section 2 presents a literature review of what has been done to solve the reduction problem as well as the features and problems associated to each of them. Section 3 introduces a detailed description of the algorithm proposed and the suppression mechanism. Section 4 explains the methodology used in the experiments. Finally, Section 5 points out the conclusions and gives some direction of future work.

III. PREVIOUS WORK

An important contribution in the area of data reduction for structured data (data mining) can be found in (Cano et al. 2003). In this work, the authors present a review of the main instance selection algorithms. In addition, they perform an empirical performance study that compares the classical instance selection methods with four major evolutionary-based strategies. The authors divide the instance selection methods into four sets. The first set involves techniques based on nearest neighbor (NN) rules. These techniques are Cnn [4], Enn, Renn [5], Rnn [6], Vsm [7], Multedit [8], Mcs [9], Shrink, Icf [10], Ib2 [11], and Ib3 [12]. The second set involves methods based on ordered removal. These methods are Drop1, Drop2 and Drop3 [13]. There are two methods based on random sampling that were considered, i.e., Rmhc [14] and Ennrs [15]. The evolutionary-based methods are the generational genetic algorithm (GGA) [17] and [17], the steady-state genetic algorithm (SGA) [18], and the CHC adaptive search algorithm [19]. The authors in [19] claim that the execution time associated with evolutionary algorithms (EAs) represents a greater cost compared to the execution time

of the classical algorithms. However, when compared to non-EAs that have a short execution time, EA-based algorithms offer more reduction without overfitting. The authors concluded that the best algorithm corresponds to the CHC, whose time is lower compared to the rest of the EAs, the probabilistic algorithms and some of the classical instance selection algorithms. The classical and evolutionary algorithms are affected when the size of the data set increases, whereas CHC is more robust. In CHC, the chromosomes select a small number of instances from the beginning of the evolution, so that the fitness function based on 1-NN has to perform a smaller number of operations. There are many other strategies in the literature [20], [21]. [22], [23], [24], [25], [26] and [27].

IV. The Suppression Mechanism

The suppression concept for proposed algorithm SeleSup (selection by suppressor) is employed in the training set to eliminate very similar data instances and to keep those instances that are truly representative of a certain class [28]. To perform such tasks, the mechanism divides the training database into two subsets. The first subset represents the white blood cells (WBCs) or antibodies in the organism, representing the training set. The second subset represents a set of pathogens or antigens that will select the higher affinity with WBCs; hence, this method performs suppression. The algorithm starts with the idea that the system's model must identify the best subset of WBCs to recognize pathogens, i.e., the training set, and to be able to identify new pathogens that are presented.

Both antibodies and antigens were represented as vectors containing the most relevant terms of the documents. Each vector was normalized to belong to the same scale of values which is mapped to the interval [0,1]. The affinity between antibodies and antigens was determined by the cosine distance. This measure is commonly used to measure the level of similarity between two documents.

Given two vectors representing documents, *WBC* and *Pathogen*, their cosine will describe the similarity.

As the angle between the vectors shortens, the cosine angle approaches 1, meaning that the two vectors are getting closer, or more similar.

According to [28] the algorithm aims to identify the best subset of antibodies to recognize the antigens, i.e., the new training set must be able to identify new antigens. Finally, the antibody survivors are represented by an evaluation measure (fitness value) and are selected to be a part of the new reduced training set.

In other words, those WBCs able to recognize pathogens from the suppression set remain while the others are eliminated from the population. The signals for a WBC's survival are represented by a fitness variable. Each time the nearest WBC recognizes a same class-label pathogen, the survival signal is sent and the fitness is incremented. Every WBC with a fitness greater than zero is selected to be part of the new suppressed repertoire. The pseudo-code for this technique can be seen in Algorithm 1.

Algorithm 1: The Suppressive Algorithm

---

input:The normalised (in[0, 1]) full training data set T
and the fraction f of WBCs (default f =0.9)
output: A reduced training data set T

// Initialisation phase
Shuffle T and assign [f ·|T|] samples as WBCs (training set); the remaining samples are assigned as pathogens (suppression set);

for all the WBCs **do** fitness = 0;
// Suppression phase
for each pathogen **p do**
NearestWBC ← Find the nearest WBC with regard to p;
if NearestWBC's class = p's class then
// NearestWBC was able to recognize the pathogen
Increment the NearestWBC's **fitness** by one;
endif;
end;

// Output phase
Eliminate those WBCs whose **fitness** value is 0;
Output the set of surviving WBCs as the reduced training set T

V. Experimental Study

In this section, the experiments presented aims to evaluate the reduced training instances selected by the SeleSup algorithm in four data sets (shown in **Error! Reference source not found.**) frequently used in information retrieval research.

TABLE I. Data Sets Features

Data set	Instances Total	Number Instance Train	Number Instance Test	Number Attributes	Number Classes
Reuters-4	1337	888	449	2833	4
Reuters-10	6689	4416	2273	2833	10
Original Reuters	8250	5169	2680	2833	62
NewsGroup	18300	16470	1830	1154	20

The Reuters-21578 Text Collection contains documents collected from the Reuters newswire in 1987. It is a standard text categorization benchmark that contains 135 classes. The collection was divided it in two subsets: one consisting of the four more balanced classes, which was identified as Reuters-4, and the other consisting of the ten most frequent classes, which was identified as Reuters-10. The third datasets consists of the sixty two classes, which was identified as Reuters-Original.

The last data set, the NewsGroup (20NG) dataset contains approximately 20000 articles evenly divided among 20 Usenet

newsgroups. Over a period of time 1000 articles were taken from each of the newsgroups, which makes an overall number of 20000 documents in this collection. Except for a small fraction of the articles, each document belongs to exactly one newsgroup (Joachims 1997).

The performance of the two classification algorithms Naive Bayes and Support Vector Machine (SVM) over the resulting reduced training and test subsets of SeleSup is compared to the performance over the subsets selected by the CHC algorithm, which is based on genetic algorithms [19] and random sampling (RS) based on the reduction percentages of experiments of each algorithm.

For each one of these subsets, the algorithms SeleSup and RS of each method were run out ten times and the reduced sets of training data were submitted to the classification algorithms (Naive Bayes and Support Vector Machine). The CHC percentage reduction, obtained in just one execution, due to computational cost was adopted. The RS was run 10 times. The average was obtained as final result for each experiment.

VI. The Data Sets

A. REUTERS

The first experiment performed in this paper makes use of the Reuters collection (Zeidat et al. 2006; Yang et al.1996; Schapire 1990; Schapire et al. 2000; Sebastiani 2002). The Reuters-21578 collection is a collection of documents from the Reuters news agency that was released in 1987. By 1990, the collection was given to the scientific community to perform research related to text categorisation. The rights of authorship belong to Reuters Ltd. and the Carnegie Group, which promoted its free distribution for research activities. The document basis consists of 21578 Reuters articles that consist of files in the SGML language.

These documents are grouped into 22 separate files. Each document possesses several attributes that indicate different characteristics. The attributes used in this work are: Lewissplit (related to the information of the experiments done by Lewis who defines the values Test, Training and Not-Used); Oldid, which represents the identification number of the collection (before the Reuters- 21578); D, which represents the categories or classes; and Body, which presents the text content of major news. The number of documents per class varies from class "earnings" (3964 documents) to class "castor-oil" (which contains a single document). Furthermore, some documents are not associated with any of the classes, and others are associated with up to 12 of the classes.

The SGML files were transformed into XML format and were pre-processed in Microsoft Excel, joining all documents in one single file. The resulting file was considered as the format for the input file for the mining process containing a collection of 8250 records sorted into 62 categories.

Then, the usual text mining data preparation techniques were performed. From this subset it was partitioned other two subsets: Reuters-4 and Reuters-10 as explained in next section. The four more balanced and the ten most frequent classes are indicated in Table 2 and 3.

TABLE II. FOUR MORE BALANCED CLASSES OF REUTERS DATA SET.

Class name	Samples
1 - Grain	375
2- Crude	362
3- Money-fx	313
4-Trade	287

TABLE III. TEN MOST FREQUENT CLASSES OF REUTERS DATA SET.

Class name	Samples
1 - Earn	3126
2 - Acq	1744
3 - Grain	375
4 - Crude	362
5 - Money-fx	313
6 – Trade	287
7 – Interest	154
8 – Ship	150
9 - Sugar	90
10 – Coffee	88

B. Newsgroup Data

The 20 Newsgroups data set is a collection of approximately 20000 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups. This collection has become a popular data set for experiments in text applications of machine learning techniques, such as text classification and text clustering.

Some of the newsgroups are very closely related to each other (e.g. comp.sys.ibm.pc.hardware / comp.sys.mac.hardware), while others are highly unrelated (e.g misc.forsale / soc.religion.christian). The **Error! Reference source not found.** presents a list of the 20 newsgroups, partitioned (more or less) according to subject matter (Table 4).

TABLE IV. NEWSGROUPS CLASSES

comp.graphics comp.os.ms-windows.misc comp.sys.ibm.pc.hardware comp.sys.mac.hardware comp.windows.x	rec.autos rec.motorcycles rec.sport.baseball rec.sport.hockey	sci.crypt sci.electronics sci.med sci.space
misc.forsale	talk.politics.misc talk.politics.guns talk.politics.mideast	talk.religion.misc alt.atheism soc.religion.christian

C. Parameters

The parameter setting is given in Table 5 and remained constant throughout the experiments. It was used stopwords and stemming in the document preparation stage. In additional, it was performed a filter on keywords with more than 50% significance and keyword´s relevance was used to generate the vector space model.

TABLE V. PARAMETER SETTING

Algorithm	Parameter	Value
SeleSup	fraction of training samples (WBCs)	0.9
Random Supression[1]	fraction of training samples	from reduction rate of SeleSup and CHC
CHC[2]	population's size number of evaluations alfa equilibrate factor percentage of change in restart 0 to 1 probability in restart 0 to 1 probability in diverge	50 100/1000 0.5 0.35 0.25 0.05

[1] Implementation from POLYANALYST v 6 - http://www.megaputer.com

[2] Implementation from KEEL v2.0 rev. 2010-05-13 - http://www.keel.es

D. Significance Test

Statistical evaluation of experimental results has been considered an essential part of validation of the new machine learning methods [29],[30]. The statistical test has the objective of reject a false null hypothesis [31].

This paper shows a comparison between nonparametric tests, Wilcoxon signed rank test [32] and Mann-Whitney test [33] for comparing of two classifiers, Naïve Bayes and SVM. [29] mentions Wilcoxon signed rank test as safe and robust non-parametric tests for statistical comparisons of classifiers.

It was used data sets with high dimension space, which demand a high processing time. So, it was chosen the training data set of the each one of the four data sets (see Table 1), which have been run on 10-fold cross-validation method to obtain a random sample of 10 results. The test is two-tailed with significance level of 0.05. The results have been obtained through the KEEL software [34], [30] and [29].

Generally when the p value is greater than 0.05, the null hypothesis is accepted resulting as no evidence that the samples are significantly different. However, if the null hypothesis is rejected ($p < 0.05$) denotes that the samples are statistically significant.

VII. RESULTS AND ANALYSIS

The first experiment was carried out in the Reuters-4 data set. This data set is characterized by balanced classes (see Table 6 and 7). The accuracy of SeleSup is just as good as results of CHC-100 and with the same data set without reduction, the results presented are very similar. The CHC-100 produces the best performance. Therefore, CHC-100 hasn't nearly as high reduction rate as SeleSup.

The CHC-1000 has a bigger reduction, but comparing with SeleSup the accuracy don't nearly produce as good results as its. In the tests, there was only one case (CHC1000) where the performance hasn't shown significantly different.

TABLE VI. RESULTS FOR REUTERS-4 DATA SET

Reuters - 4	Reduction (%)	Naïve Bayes Accuracy Test (%)	SVM Accuracy Test (%)	Execution Time (s)
None	0.00	92.89	*93.56*	*00:00:00*
SeleSup	90.43	88.38	88.96	00:00:06
Random Sampling		88.64	89.67	00:00:00
CHC_100	77.11	*93.11*	92.22	00:00:04
Random Sampling		91.16	92.27	00:00:01
CHC1000	*97.18*	72.89	79.11	00:01:45
Random Sampling		74.53	74.71	00:00:01

TABLE VII. MANN-WHITNEY U AND WILCOXON TESTS COMPARING BAYES VS SVM FOR REUTERS-4 DATA SET

Reuters - 4	Mann_Whitney p-value	Wilcoxon p-value
None	1.57E-4	0.0055
SeleSup	4.39E-4	0.0055
CHC_100	2.12E-4	0.0055
CHC_1000	1.2662	0.7037

The second experiment was carried out with the Reuters-10 data set. This data set is characterized by an imbalance on its classes (see **Error! Reference source not found.**).

Therefore, as can be seen in Table 8, all the classifiers produced satisfactory results when their learning process used all the training and test data set. In addition, as expected, the same behavior occurs when suppression mechanism is applied.

The accuracy of SeleSup is just as good as results with the same data set without reduction, Random Sampling and CHC-100. The results are very similar between the classifiers. Therefore, CHC-100 has not nearly as high reduction rate as SeleSup.

It can be noticed that if the number of evaluation increases, the accuracy test of CHC-1000 decreases and consumes a high time execution (more than 50 higher). So, the CHC-1000 doesn't produce nearly as good results as SeleSup.

The results (Table 9) indicate that the Wilcoxon test is more powerful than the Mann-Whitney test according to [29].

TABLE VIII. RESULTS FOR REUTERS-10 DATA SET

Reuters - 10	**Execution Time (s)**	**Naïve Bayes** Accuracy **Test (%)**	**SVM** Accuracy **Test (%)**	**% Reduction**
None	*00:00:00*	*92.92*	*93.53*	*0.*
SeleSup	00:01:46	90.13	90.21	91.
Rand. Samp.	00:00:00	89.35	89.16	91.
CHC_100	00:58:29	91.95	91.29	77.
Rand. Samp.	00:00:01	92.00	92.06	77.
CHC1000	01:58:12	84.43	83.77	97,
Rand. Samp.	00:00:01	84.70	82.29	97.

TABLE IX. MANN-WHITNEY U AND WILCOXON TESTS COMPARING BAYES VS SVM FOR REUTERS-10 DATA SET

Reuters - 10	**Mann_Whitney p-value**	**Wilcoxon p-value**
None	1.57E-4	0.0055
SeleSup	1.57E-4	0.0055
CHC_100	1.57E-4	0.0055
CHC_1000	2.12E-4	0.0055

TABLE X. RESULTS FOR ORIGINAL REUTERS DATA SET

Original Reuters	**Red. (%)**	**Naïve Bayes** Accuracy **Test (%)**	**SVM** Accuracy **Test (%)**	**Exec. Time (s)**
None	0.00	*83.62*	*87.01*	*00:00:00*
SeleSup	91.82	78.02	78.66	00:02:30
Random Sampling		77.48	78.00	00:00:00
CHC_100	76.42	81.98	83.99	01:00:33
Random Sampling		81.54	83.57	00:00:00
CHC1000	*97.12*	72.61	71.83	02:43:27
Random Sampling		72.65	71.61	00:00:00

TABLE XI. MANN-WHITNEY U AND WILCOXON TESTS COMPARING BAYES VS SVM FOR ORIGINAL REUTERS DATA SET

Original Reuters	**Mann_Whitney p-value**	**Wilcoxon p-value**
SeleSup	1.57E-4	0.0055
CHC_100	1.57E-4	0.0055
CHC_1000	1.57E-4	0.0055

The third experiment was carried out with the Reuters Original data set. This data set is characterized by a great imbalance on its classes and high dimensionality (Table 10 and 11). SeleSup produced results almost as good as CHC-1000 in the training set, but the Reuters Original without suppression produces the best results in the test set.

It can be noticed once more that the CHC-1000 produces the best data reduction percentages, but it isn't nearly as fast as SeleSup. According to (Cano et al. 2003) the main limitation of CHC is its long processing time, which makes it difficult to apply this algorithm to very large data sets.

This experiment shows the limitations of the SVM with the larger dataset (Original Reuters) which were omitted.

Finally, the last experiment was carried out using the Newsgroup data set. This data set is an example of a very large data set with 18300 instances (see Table 12). This is the largest data set in our experiments.

The SeleSup and CHC obtained results are very similar in accuracy. In addition, the algorithm SeleSup was easily applied in this data set and its results were just as good as CHC-1000. Its processing time has been very meaningful when compared with the CHC that produces a very similar percentage of reduction (92,09% and 93,29%).

It can be observed that the RS had in general results very similar to the algorithms SeleSup and CHC, but it has a clear disadvantage of not reducing data by itself. Therefore, another algorithm has to be used to define the reduction percentage.

TABLE XII. IT IS ALSO POSSIBLE TO NOTICE THAT THERE IS NO STATISTICAL DIFFERENCE BETWEEN THE METHODS APPLIED IN THIS DATASET (TABLE 13). RESULTS FOR NEWSGROUP DATA SET

News group	**Reduction (%)**	**Naïve Bayes** Accuracy **Test (%)**	**SVM** Accuracy **Test (%)**	**Exec. Time (s)**
None	0.00	*88.8*	*93.01*	*00:00:00*
SeleSup	92.09	79,2	91,84	00:13:00
Random Sampling		79.5	91,18	00:00:00
CHC_100	77.01	85.1	93.55	17:12:00
Random Sampling		85.2	93.45	00:00:00

CHC_100	*93.29*	80.1	90.55	13:48:05
Random Sampling		78.1	90.30	00:00:00

TABLE XIII. MANN-WHITNEY U AND WILCOXON TESTS COMPARING BAYES VS SVM FOR NEWSGROUP DATA SET

Newsgroup	**Mann_Whitney p-value**	**Wilcoxon p-value**
None	1.57E-04	0.0055
SeleSup	1.57E-04	0.0055
CHC_100	1.57E-04	0.0055
CHC_1000	1.57E-04	0.0055

VIII. CONCLUSION

To carry out efficiently the training of classifiers of large collections of text the selection of the training set must be done carefully. If it is used an excessive number of documents the computational effort can make the task impossible. Using a very small sample leads to the inaccuracy of the classifier.

This paper presented a new method for instance selection (IS) by suppressing data in the original training set. IS can be very useful to reduce costs, improve computational performance and eliminate non-informative data. The proposed technique was designed to work together with different types of classifiers. The goal was to improve the performance related to the time spent on training without losing accuracy. This approach was inspired by the suppression mechanisms found in biological immune systems.

The experiments were conducted by testing the SeleSup algorithm in four data sets. The performance of three classification algorithms over the resulting training subsets of SeleSup was compared with the performance over the subsets selected by the CHC algorithm and random sampling (RS).

In order to test whether the algorithms' performances were significantly different or not, it was adopted a comparison between non-parametric tests Mann–Whitney U and Wilcoxon signed rank. In the tests, there were only one case where the performances haven't shown significantly different. Therefore, the statistical tests have provided strong evidence concerning the results obtained when comparing the evaluated algorithms.

The SeleSup algorithm significantly reduces the data set size. This algorithm is just as good as CHC algorithm and it offers the advantage of being faster. Then, it consumes less processing time. Although CHC has a higher reduction rate, it does not produce the best results with high dimensionality data sets and it showed high time execution. Moreover, on the contrary of CHC, the presented approach was applied to all the data sets on a less power computer, and overall, its results were better than RS.

IX. FUTURE WORK

An alternative method for performing a faster test would be inserting into the WBCs' population the pathogen-specific WBC whose distance is the minimum distance. This technique should provide the system with the capability of keeping rare cases or rare classes in the training set.

An additional improvement to the original algorithm could be to insert some probabilistic information on the choice of the WBCs to be eliminated. The way that the mechanism works currently is deterministic with regard to data selection.

ACKNOWLEDGMENT

The authors acknowledge the support provided by CNPq, the Brazilian Research Agency, FAPERJ, the Rio de Janeiro Research Foundation and CAPES, Coordination for the Improvement of Higher Level Education.

REFERENCES

[1] J. Han, and M. Kamber, "Data Mining: Concepts and Techniques" Morgan Kaufmann Publishers, San Francisco , CA, 2001.

[2] T. M. Mitchell, "Machine Learning" Mc Graw-Hill Series in Computer Science, USA, 1997.

[3] J. Timmis, "Artificial Immune Systems: A Novel Data Analysis Technique Inspired by the Immune NetWork Theory." PhD Thesis, Universityos Whales, Department os Computer Science, AlberystWyth, Ceredigion, Wales, 2000.

[4] P.E. Hart, "The condensed nearest neighbor rule" IEEE Transactions on Information Theory, 14, pp .515-516, 1968.

[5] D.L. Wilson, "Asymptotic properties of nearest neighbor rules using edited data." IEEE Transaction on Systems Man Cybernetics, 2, pp.408-421, 1972.

[6] G.W. Gate, "The reduced nearest neighbor rule." IEEE Transactions on Information Theory, 14, pp. 431-433, 1972.

[7] D.G. Lowe, "Similarity metric learning for a variable-kernel classifier", Neural Computation, 7, pp. 72-85 1995.

[8] P.A. Devijver and J. Kittler, "Pattern recognition: A statistical approach" , Prentice-Hall International, 1982.

[9] C.E. Broadley, "Automatic algorithm/model class selection", Proceedings of the Tenth International Machine Learning Conference, pp. 17-24.

[10] H. Brighton and C. Mellish, "Advances in instance selection for instance-based learning algorithms". Data Mining and Knowledge Discover, 6 pp. 153-172, 2002.

[11] D. Kibber, D.W. Aha, "Learning representative exemplars os concepts: An initial case of study." Proceedings of 4th International Machine Learning Workshop, pp. 24-30, 1987.

[12] D.W. Aha and M.K. Albert D, "Instance based learning algorithms" Machine Learning, 6 pp. 37-66, 1991.

[13] D.R. Wilson and T.R. Martinez, "Instance pruning techniques". In Proceedings of 14 th International Conf. Machine Learning, pp. 404-417, 1997.

[14] D.B. Skalak, "Prototype and feature selection by sampling and random mutation hill climbing algorithms". In Proceedings os 11 th International Conference on Machine Learning, New Brunswick, NJ Morgan Kaufmann, 1994.

[15] D.R. Wilson and T.R. Martinez, "Reduction techniques for instance-based learning algorithms". Machine Learning, 38 pp. 257-268.

[16] D.E. Goldberg, "Genetic Algorithms in Search Optimization, and Machine Learning." Addison-Wesley longman Publishing Co., Boston, Mass, 1989.

[17] J.H. Holland, "Adaptation in Natural and Artificial Systems". University of Michigan Press, Ann Arbor, MI, 1975.

[18] D. Whitley, "The genitor algorithm and selective preasure: Why rank based allocation of reproductive trials ins best." In Proceedings os 3 rd Int.Conf. Gas, pp. 116-121, 1989.

[19] J.E. Cano and M. Lozano F., "Using evolutionary algorithms as instance selecton for data reduction in KDD: An experimental study". IEEE Transaction on Evolutionary Computation, 7 pp. 561-575, 2003.

[20] A. Franco, D. Maltoni and L. Nanni, "Data pre-processing through reward-punishment editing." Pattern Analysis and Applications, 13, pp. 367-381, 2010.

[21] J. Kittler, M. Hatef and J. Duin R, " On combining classifiers." IEEE Transactions on Pattern Analysis and Machine Intelligence, 20 pp. 226-239, 1998.

[22] L. Nanni and A. Lumini, "Particle swarm optimization for prototype reduction." Neurocomputing, 72, pp. 1092-1097, 2009.

[23] L. Nanni, "Experimental comparison of one-class classifiers for online signature verification", Neurocomputing, 69, pp. 869-875, 2006.

[24] R. Parades and E. Vidal, "Learning Prototypes and distances: a prototype reduction technique based on nearest neighbor error minimization." Pattern Recognition, 39, pp. 180-188, 2006.

[25] C. Pedreira, "Learning Vector quantization with training data selection". IEEE TPAMI, 18 pp. 157-162, 2006.

[26] J. R. Cano, F. Herrera and M. Lozano F. "On the combination of evolutionary algorithms and stratified strategies for training set selection in data mining". Applied Soft Computing, 6 pp. 323-332, 2006.

[27] T. Joachims, "Text categorization with Support Vector Machines: Learning with many relevant features. In Machine Learning: ECML-98", Tenth European Conference on Machine Learning, pp. 137-142, 1998.

[28] G.P. Figueiredo, N.F.F. Ebecken and H.J.C. Augusto D.A. " An Immune-inspired Data Selection Mechanism for Supervised Classification, Memetic Computing, v. 4, pp. 135-147, 2012.

[29] J. Demsar, "Statistical comparison of classifiers over multiple data sets." Journal of Machine Learning Research, 7, pp.1-30, 2006.

[30] S. Garcia, A. Fernández, J. Luengo and F. Herrera, "Advanced nonparametrictests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power". Information Sciences. DOI: 10.1016/j.ins. 2009.12.010.

[31] S. Garcia and F. Herrera, "An extension on statistical comparisons of classifiers over multiple data sets for all pairwise comparisons". Journal of Machine Learning Research, 9, pp. 2579-2596, 2008.

[32] F. Wilcoxon, "Individual Comparisons by Ranking Methods". Biometrics 1, pp. 80-83, 1945.

[33] H.B. Mann and D.R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other". Annals of Mathematical Statistics, 18, pp. 50-60, 1947.

[34] J. Alcalá-Fdez, L. Sánchez, S. García, Del, M.J. Jesus, S. Ventura, J.M. Garrell, Romero, J. Otero, C. Romero, Rivas J. Bacardit, J.C. Fernández and F. Herrera, " Keel: a software tool to assess evolutionary algorithms to data mining problems." Soft Computing, 13 (3), pp. 307-318, 2009.