Investigation of data clustering preprocessing algorithm on independent attributes to improve the performance of CLONALG

Dr. S.Chitra [1] , B.MadhuSudhanan[2] ,DR.M.Rajaram [3], DR.S.N.Sivanandham[4]

ABSTRACT

It is a popularly held belief that preprocessing of data generally improves the classification efficiency of data mining algorithms. We study the effects of preprocess by utilizing an algorithm to cluster points in a data set based upon each attribute independently, resulting in additional information about the data points with respect to each of its dimensions. Noise, data boundaries are identified and the cleaned data subset is used to study the performance of CLONALG data mining algorithm against unprocessed dataset.

I. ARTIFICIAL IMMUNE SYSTEMS

The Artificial Immune Systems (AIS) are a relatively new area of research with considerable potential in helping solve a myriad of difficulties. Its growth has allowed the proposal of new techniques and approaches for solving known problems.

The aim of this technology is to model defence mechanism characteristics and functionalities of living beings. The defence mechanism allows an organism to defend against invasion from foreign substances. The recognition of these substances is based on the key and lock analogy, in which the objective is to find antibodies that have the best immune response to the invading antigens [1].

The natural immune system stores the best antibodies in its genetic memory. These are later used to identify antigens that have previously invaded the organism, thereby obtaining a quicker, more efficient response.New functionalities observed in the biological environment were studied for the modelling of this new immunological approach, principally the organization and clustering of similar antibodies (Ab) throughout the process. It is believed that these functionalities may improve the recognition capacity of artificial immune algorithms.

II. CLONAL SELECTION ALGORITHMS

Clonal selection algorithms have taken inspiration from the antigen driven affinity maturation process of B cells and the associated hypermutation mechanism. These AIS also often use the idea of memory cells to retain good solutions to the problem being solved. [4] highlight two important features of affinity maturation in B cells that can be exploited from the computational viewpoint. The first feature is that the proliferation of B cells is proportional to the affinity of the antigen that binds it, thus the higher the affinity, the more clones that are produced. Secondly, the mutations suffered by the antibody of a B cell are inversely proportional to the affinity of the antigen it binds. Applying these two features, [2] developed an AIS called CLONALG, which has been used to performed the tasks of pattern matching and multi-modal function optimisation [3]. For the example of pattern matching, a set of patterns, S, to be matched are considered to be antigens. The task of CLONALG is to then produce a set of memory antibodies, M, that match the members in S.

Input: S = set of patterns to be recognised, n the number of worst elements to select for removal

Output: M = set of memory detectors capable of classifying unseen patterns

begin

Create an initial random set of antibodies, A

For all patterns in S do

Determine the affinity with each antibody in A

Generate clones of a subset of the antibodies in A with the highest affinity. The number of clones for an antibody is proportional to its affinity

Mutate attributes of these clones inversely proportional to its affinity. Add these clones to the set A, and place a copy of the highest affinity antibodies in A into the memory set, M

Replace the n lowest affinity antibodies in A with new randomly generated antibodies

End

End

Algorithm : CLONALG for pattern .

Clonal selection algorithms share many similarities with evolutionary algorithms [8], although importantly the selection and mutation mechanisms are influenced by the affinities of antibody-antigen match- ing [3]. Due to this similarity, many of the theoretical approaches applied to evolutionary algorithms are applicable to clonal selection algorithms also. [11] summarises much of the theoretical work done on AIS to date. This includes the work of [Clark et al. 2005] who develop an exact Markov chain model of the clonal selection algorithm called the B-cell algorithm (BCA) [13], proving its convergence. They go on to show how the model can be applied to give insight into optimal parameter set- tings for the BCA in a function optimisation landscape. Other AIS that have been

inspired by the adaptive immune mechanisms of B cells are AIRS [14], a supervised learning algorithm, and IA that has been used in numerous applications and well studied [9].

IMMUNE NETWORK ALGORITHMS

Immune network algorithms have their basis in the continuous ordinary differential equation models used by theoretical immunologists to explore the perceived behaviour of real immune networks. Examples include the models by [7] and [8]. One of the main differences between the discretised immune network algorithms is that they interact with their environ- ment (i.e. antigens), whereas the continuous models typically do not [4].

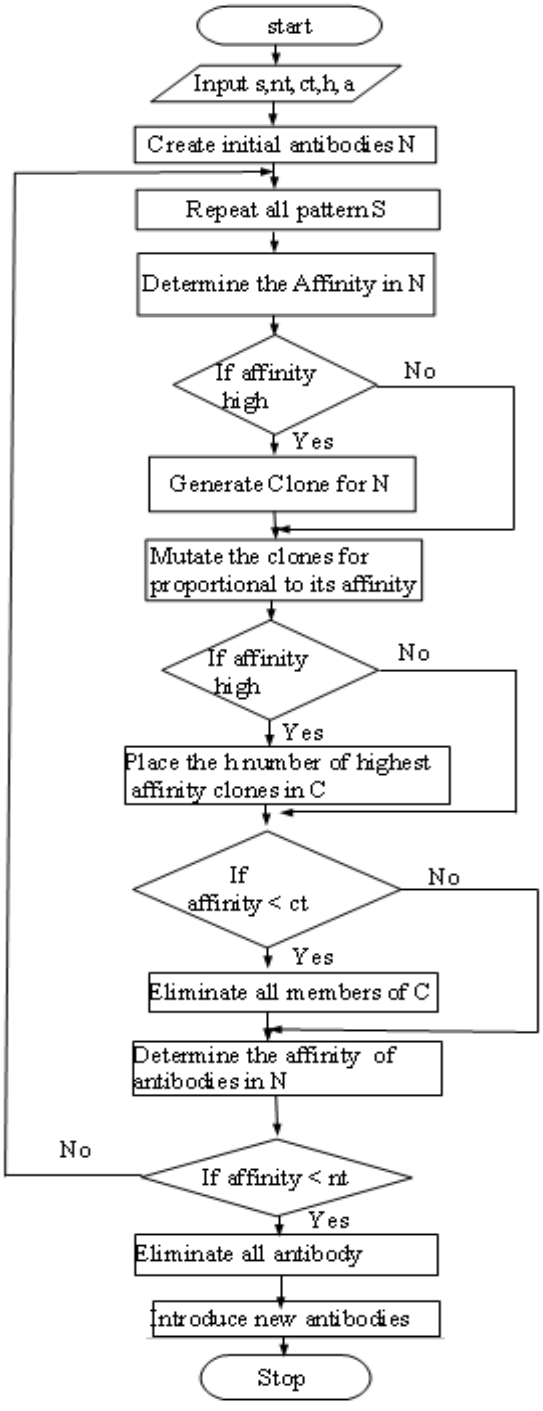

Figure 1. Filter algorithm process flow

The main difference between immune network algorithms and other immune algorithms is that the components of the system not only interact with antigenic components, but with the other components in the AIS. Two examples of im- mune network algorithms are RAIN [10] and aiNet [2], which attempt to use the basic concepts of immune network theory to solve problems such as pattern recognition and data clustering. aiNet consists of a network of antibody components that adapt to match a population of input components (antigens) to be clustered. aiNet is essentially a modified version of CLONALG (described above) with an added mechanism of suppressive interactions between the antibody components. The resulting set of network antibodies that is generated represents an internal image of the antigens to which they have been exposed. aiNet has found wide use in the area of optimisation, and many adaptations have been made to the algorithm such as [14]. [3] provides a good review of the different immune networks that appear in the literature.

Input: S = set of patterns to be recognised, nt network affinity threshold,

ct clonal pool threshold, h number of highest affinity clones, a

number of new antibodies to introduce

Output: N = set of memory detectors capable of classifying unseen patterns

begin

Create an initial random set of network antibodies, N

Repeat all patterns in S do

Determine the affinity with each antibody in N

Generate clones of a subset of the antibodies in N with the highest affinity. The number of clones for an antibody is proportional to its affinity

Mutate attributes of these clones inversely proportional to its affinity, and place the h number of highest affinity clones into a clonal memory set, C

Eliminate all members of C whose affinity with the antigen is less than a pre-defined threshold (ct)

Determine the affinity amongst all the antibodies in C and eliminate those antibodies whose affinity with each other is less than a pre-specified threshold (ct)

Incorporate the remaining clones in C into N

end

Determine the affinity between each pair of antibodies in N and eliminate all antibodies whose affinity is less than a pre-specified threshold nt

Introduce a number (a) of new randomly generated antibodies into N

Until a stopping condition has been met end

Relatively little theoretical work exists for immune network algorithms, al- though [12] do highlight various issues. Due to their network structure, immune networks would be open to theoretical techniques used in the study of other networks such as small-world and scale-free networks. Additionally many of the techniques used to study the continuous theoretical immune models are also relevant.

RESULT

Without Clustering filter

Instances: 438

Attributes: 11

LOC_BLANK
BRANCH_COUNT
LOC_CODE_AND_COMMENT
LOC_COMMENTS
CYCLOMATIC_COMPLEXITY
DESIGN_COMPLEXITY
ESSENTIAL_COMPLEXITY
HALSTEAD_DIFFICULTY
HALSTEAD_ERROR_EST
HALSTEAD_LEVEL
PROBLEM

Test mode: split 40% train, remainder test

Correctly Classified Instances 217 82.5095 %

Incorrectly Classified Instances 46 17.4905 %

Kappa statistic 0

Mean absolute error 0.1749

Root mean squared error 0.4182

Relative absolute error 62.1479 %

Root relative squared error 110.0434 %

Total Number of Instances 263

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0	0	0	0	0	YES
1	1	0.825	1	0.904	NO

=== Confusion Matrix ===

a	b	<-- classified as
0	46	a = YES
0	217	b = NO

With Filter

Instances: 415

Attributes: 11

LOC_BLANK
BRANCH_COUNT
LOC_CODE_AND_COMMENT
LOC_COMMENTS
CYCLOMATIC_COMPLEXITY
DESIGN_COMPLEXITY
ESSENTIAL_COMPLEXITY
HALSTEAD_DIFFICULTY
HALSTEAD_ERROR_EST
HALSTEAD_LEVEL
PROBLEM

Test mode: split 40% train, remainder test

Correctly Classified Instances 179 71.8876 %

Incorrectly Classified Instances 70 28.1124 %

Kappa statistic -0.002

Mean absolute error 0.2811

Root mean squared error 0.5302

Relative absolute error 104.2276 %

Root relative squared error 144.3928 %

Total Number of Instances 249

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0.175	0.177	0.159	0.175	0.167	YES
0.823	0.825	0.839	0.823	0.831	NO

=== Confusion Matrix ===

a	b	<-- classified as
7	33	a = YES
37	172	b = NO

III. CONCLUSIONS

Though data reduction has been achieved by removing data boundaries using clustering filter we see that the classification result has not improved but deteriorated.

IV. FUTURE WORK

More analysis is required to understand the effects of pre processing for specific data set.

REFERENCES

[1] De Castro, L. N. & Timmis, J. I. (2002). Artificial Immune Systems: A Novel Paradigm for Pattern Recognition, In : Artificial Neural Networks in Pattern Recognition, L. Alonso, J. Corchado, C. Fyfe, 67-84, University of Paisley.

[2] De Castro, L. N. & Von Zuben, F. J. (2000). The Clonal Selection Algorithm with Engineering Applications, Proceedings of Genetic and Evolutionary Computation Conference, Las Vegas, Nevada, USA, July, 2000, pp. 36-37.

[3] Leandro N. de Castro and Jon Timmis(2002). An artificial immune network for multimodal function optimization. In IEEE Congress on Evolutionary Computation (CEC), pages 699–704,.

[4] Leandro N. de Castro and Jon Timmis(2002). Artificial Immune Systems: A New Computational Intelligence Approach. Springer,.

[5] Burnet, F. (1959), The clonal selection theory of acquired immunity. Cambridge University Press. The Origins of the Clonal Selection Theory of Immunity,A Case Study for Evaluation in Science By D. R. FORSDYKE (From FASEB. Journal 1995, vol 9, 164-166 wit.

[6] Carter, J. H. (2000), 'The immune systems as a model for pattern recognition and classification'. Journal of the American Medical Informatics Association 7(1).

[7] Chotirat "Ann" Ratanamahatana Dimitrios Gunopulos ,'Scaling up the Naive Bayesian Classifier', Computer Science Department University of California Riverside, CA 92521 1-909-787-5190.

[8] Mingxi Wu, Christopher Jermaine, 'A Bayesian Method for Guessing the Extreme Values in a Data Set', Department of Computer and Information Sciences and Engineering, University of Florida, Gainesville, FL, USA.

[9] Mingxi Wu, Christopher Jermaine, 'A Bayesian Method for Guessing the Extreme Values in a Data Set', Department of Computer and Information Sciences and Engineering, University of Florida, Gainesville, FL, USA.

[10] Sommerville, I. Software Engineering, 6th edition. Addison-Wesley, 2001.

[11] Timmis, J. and M. Neal (2001), 'A Resource Limited Artificial Immune System'. Knowledge Based Systems 14(3/4), 121–130.

[12] Timmis, J., M. Neal, and J. Hunt (2000), 'An Artificial Immune System for Data Analysis'. Biosystems 55(1/3), 143–150.

[13] Tom M. Mitchell (2005), 'Generative and discriminative classifiers: naive bayes and logistic regression', Lecture notes on Machine learning.

[14] Watkins, A. (2001), 'A Resource Limited Artificial Immune Classifier'. Master's thesis, Mississippi Sate University.

[15] Xin Jin1, Rongfang Bie and X. Z.Gao (2006), 'An Artificial Immune Recognition System-based Approach to Software Engineering Management: with Software Metrics Selection', Proceedings of the Sixth International Conference on Intelligent Systems Design.

[16] Y.K.Malaiya and P.K.Srimani,Eds (1990), 'Software Reliability Models:Theoretical Developments,Evaluation and Applications', IEEE computer society press.

ABOUT THE AUTHORS

[1]Dr .S.Chitra, she is the head of the department of computer science and engineering in M. Kumarasamy College Of Engineering, Karur. She has 16 years experience in teaching. She completed her BE and ME in Computer Science And Engineering. She is doing her research in the area Software Reliability in Software Engineering. She presented more than 17 papers including national, international conferences and journals.

[2]B.Madhusudhanan , He is Student in Master of Engineering (Computer Science and Engineering), M. Kumarasamy College Of Engineering, Karur.

[3]Dr.M.Rajaram, Professor/Head/EEE Government College of Engineering, Thirunelveli. He is guiding 12 research scholars and 4 have been awarded doctorate. He presented more than 105 papers including national, international and journals and he has 21 years experience in teaching.

[4]Dr.S.N.Sivanandham, Professor/Head/CSE. PSG College of Technology, Coimbatore. He is guiding 7 research scholars and 27 have been awarded doctorate. He authored/ co-authored more than 750 papers including national, international and journals and he has 42 years experience in teaching.