We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS BOOK CITATION INDEX INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

16

Artificial Immune Network: Classification on Heterogeneous Data

Mazidah Puteh[1], Abdul Razak Hamdan[2], Khairuddin Omar[2] and Mohd Tajul Hasnan Mohd Tajuddin[1]
[1]*Universiti Teknologi MARA*
[2]*Universiti Kebangsaan Malaysia*
Malaysia

1. Introduction

Classification is one of the important tasks in data mining that can extract knowledge from real world data sets. It helps in forecasting the future knowledge from the available knowledge or information. It also helps people in making better decision in the future based on the history and existing knowledge. With the classification algorithm, people can repeatedly make a forecast on the accumulated knowledge in new situations.

2. Immune System

2.1 Natural Immune System

A biological immune system has two broad response systems. One is innate immunity, which is general and exists in our body since we are born. The other one is an adaptive immunity that is based on two kinds of antibody cells in the body: T-cells, so named because they originate in the thymus gland and B-cells originate in bone marrow (de Castro & Timmis, 2002). When a pathogen invades the body, special cells called antigen are available. An individual T-cell or B-cell responds to the antigens by cloning and mutating to match the antigen. This is the concept of clonal selection theory (Burnet, 1959) where the binding of antibody with the antigen will activate the antibody and the clonal expansion of the antibody occurs. The closer the match, the affinity of that T-cell or B-cell from the antigen (Hunt & Cook, 1996) becomes stronger. B-cells that do not match any antigens will be eliminated. From immune network theory, (Jerne, 1974) antibody also interacts with the neighbour antibodies to form a network. If the antibody do not stimulate with the neighbour antibodies, it eventually die. After the process of generating antibodies and combating the antigens and a body has successfully defended against a pathogen, a comparatively small number of memory cells remain in the body for very long time. These memory cells recognize antigens similar to those that originally cause the immune response, so that the body's response to a future and very similar invader is much faster and powerful than to a never-before-seen invader.

2.2 Artificial Immune System

An artificial immune system is a bio-inspired computational model that uses idea and concepts from the natural immune system. Although there are about four concepts that are explored in the immune system (de Castro & Timmis, 2002), the concepts that are discussed in this paper are the interaction between antigen and B-cells (stimulation and suppression) as in the clonal selection theory (Burnet, 1959) and also the interaction between antibody and antibody as in the immune network theory (Jerne, 1974). Both theories involve cloning and mutating process (de Castro & Von Zuben, 2000). It can offer strong and robust information processing capabilities for solving complex problems. Applications of AIS include supervised and unsupervised machine learning, pattern recognition, intrusion detection and security (Dasgupta, 2006). Among the early models on supervised machine learning is Immunos81 (Carter, 2000) and AIRS (Watkins, 2001; Watkins et al., 2004). However, the former model uses significantly different and complex approach. The later model is the first straightforward immune-inspired supervised learning algorithm and has subsequently gone through a period of study and refinements (Watkins & Timmis, 2002; 2004; Hamaker & Boggess, 2004). However, many of these studied classification models concentrate on the population-based or clonal selection algorithm and ignore the important network feature (Timmis, 2001) of the immune system. The models also require numerical representation of data and mostly are tested only on numerical dataset. Some of the applications on classification with AIS models are summarized in Table 1.

Concept	Objective	Referrences
Immune Network	DNA Classification, Text Classification	Hunt & Cook, 1996; Secker et al, 2004
Clonal Selection	Numeric Data Classification	Carter 2000; Leandro 2000; Sahan et al 2005; Leung et al, 2006; Peng et al 2007
Clonal Selection with resource limited	Numeric Data classification, Text Classification, Heterogeneous data classification	Watkins 2001;2002;2004; Hamaker 2004; Secker 2007; Puteh et al 2008
Clonal Selection with resource limited and parallel	Numeric data classification	Watkins 2004
Negative Selection	Binary classification	Igawa et al 2005

Table 1. Classification applications with AIS models

As suggested in (Watkins, 2001; Freitas & Timmis, 2007; Hart & Timmis, 2008; Timmis, 2006), methods of using other types of data need to be explored to allow for greater applicability of this learning paradigm. (Hamaker & Boggess, 2004) has explored variety of similarity measurements in generating classifiers with clonal selection concept or population-based AIS algorithm but a more comprehensive experiment on many problems with heterogeneous types is required in order to prove a high quality classification technique for heterogeneous data types. (Puteh et al., 2008) has introduced a classification

model using clonal selection for heterogeneous data that is called Flexible Artificial Immune Recognition System (FAIRS) to experiment the heterogeneous data in its original types. FAIRS has shown some improvement in the accuracy compared to the existing AIS classification models. To further experiment on AIS algorithm and to overcome the limitation mentioned in the previous research, there is a need for developing the AIS classifier with the network feature and be able to accept heterogeneous data without the need for the data transformation. In order to accept various types of data, all processes involving these data must consider appropriate and suitable affinity measurement, mutation method and the correct data structure implementation.

2.2 Distance Metrics

There are many learning systems depend on good distance function to be successful such as the nearest neighbour techniques (Cover & Hart, 1967; Hart, 1968; Dasarathy & Belur, 1991), and memory-based reasoning methods (Stanfill & Waltz, 1986). Such algorithms have had much success on a wide variety of applications (real-world classification tasks). Many of these metrics work well for numerical attributes but do not appropriately handle nominal attributes (Wilson & Martinez, 1997). The common distance metrics that are used for numerical attributes and binary attributes are the Euclidean metric and the Hamming metrics as shown in equation 1 and 2.

$$\text{Euclidean (x,y)} = \sqrt{\sum_{a=1}^{m}(x_a - y_a)^2} \tag{1}$$

$$\text{Hamming(x,y)} = \sum_{i=1}^{L} \cdots \delta_i \; where \; \delta_i = \begin{cases} 1 \; if \; x_i \neq y_i \\ 0 \; other \; wise \end{cases} \tag{2}$$

The value difference Metric (VDM) (Stanfill & Waltz, 1986) was introduced to define an appropriate distance function for nominal attributes as shown in equation 3.

$$\mathrm{vdm_a(x,y)} = \sum_{c=1}^{C} \left| \frac{N_{a,x,c}}{N_{a,x}} - \frac{N_{a,y,c}}{N_{a,y}} \right|^q \tag{3}$$

where Na,x is the number of training records in T that has the value x for an attribute a; Na,x,c is the number of records in T that has the value x for attribute a and class c; C is the number of classes in the problem domain; q is a constant, usually value 1 or 2.

This distance metric work well in many nominal domains, but they do not handle continuous attributes directly. Instead, they rely upon process of discretization which can degrade generalization accuracy (Ventura et al., 1995). Many real-world applications have both nominal and numeric attribute as shown in the UCI MLR (Merz & Murphy, 1998). The distance function that is used in the proposed model is Heterogeneous Value Difference Metric (HVDM). It can take heterogeneous data where it uses normalized VDM for nominal data and normalized difference for linear data. HVDM has shown a good potential to be the distance metric for heterogeneous data without the need for any transformation of data into any specific type. HVDM has become the choice for the algorithm in this research. The discussion of the distance metrics can be found in (Wilson & Martinez, 1997). As mentioned in the previous section, the Euclidean distance function is inappropriate for nominal

attributes, and VDM is inappropriate for continuous attribute, so neither is sufficient on its own for use on a heterogeneous application, i.e. one with both nominal and continuous attributes. So, HVDM is used as shown in equation 4,5,6,7.

$$\text{HVDM}(x, y) = \sqrt{\sum_{a=1}^{m} d_a^{\,2}(x_a, y_a)} \tag{4}$$

where m is the number of attributes. The function d_a (x, y) returns a distance between the two values x and y for attribute a and it is defined as:

$$d_a(x, y) = \begin{cases} 1, \textit{if } x \textit{ or } y \textit{ is unknown; otherwise} \\ \textit{normalized\_vdm}_a(x, y), \textit{if } a \textit{ is} \text{ nominal} \\ \textit{normalized\_diff}_a(x, y), \textit{if } a \textit{ is linear} \end{cases} \tag{5}$$

where normalized vdm and normalized diff are defined as follows:

$$\text{normalized\_vdm}_a(x,y) = \sqrt{\sum_{c=1}^{C} \left| \frac{N_{a,x,c} - N_{q,y,c}}{N_{a,x} - N_{a,y}} \right|^2} \tag{6}$$

and

$$\text{normalized\_diff}_a(x,y) = \frac{|x - y|}{4\sigma_a} \tag{7}$$

where x and y are 2 input vectors for attribute a and σ is a standard deviation value for a.

Distances are often normalized by dividing the distance for each variable by the range of that attribute, so that the distance for each input variable is in the range 0..1 and this is employed by algorithm in (Hamaker & Boggess, 2004). However, dividing by the range allows outliers (extreme values) to have a profound effect on the contribution of an attribute. A more robust alternative in the presence of outliers is to divide the values by the standard deviation to reduce the effect of extreme values on the typical cases. The situation for HVDM is more complicated because the nominal and numeric distance values come from different types of measurements: numeric distances are computed from the differences between two linear values, normalized by standard deviation, while nominal attributes are computed from a sum of C differences of probability values (where C is the number of output classes). It is therefore necessary to find a way to scale these two different kinds of measurements into approximately the same range to give each variable a similar influence on the overall distance measurement (Wilson & Martinez, 1997).

3. Proposed Algorithm (FINERS)

In the real world situation, there are many data set comprise both numerical and nominal data types. This paper investigates the use of HVDM distance metric for heterogeneous datasets that are composed of nominal, discrete or continuous data types or the combination of them without the need for the transformation of the data into any specific type. The algorithm in the proposed model considers an appropriate data structures to suit the complexity of recognizing heterogeneous data in its original types.
The FINERS algorithm works as follows:
1st stage:

- Calculate Affinity Threshold (AT) by calculating average affinity (distance) between all pairs among antigens
- MemoryCell (MC) initialization, usually starts with null

For each antigen do

2nd stage:

- Search for mcmatch from MC, if unavailable, antigen as mcmatch
- Clone and mutate mcmatch
- Generate first generation antibodies (AB)
- Create a network among antibodies with affinity greater than network affinity threshold (NAT)

3rd stage:

- Clone and mutate antibody from AB randomly until average stimulation is greater than stimulation threshold.
- Generate the final AB
- Create a network among antibodies with affinity greater than network affinity threshold (NAT)

4th stage:

- Search for mccandidate (most stimulated) from AB
- Compare mccandidate to mcmatch, if mccandidate is more stimulated, it is added to MC. If affinity between mccandidate and mcmatch is less than affinity threshold scalar times affinity threshold then mccandidate replaces mcmatch inside MC
- Create a network among antibodies with affinity greater than network affinity threshold (NAT)

Basically, FINERS is a one shot algorithm where each antigen is processed only in one generation. At the end of the algorithm, set of rules or classifier and output of accuracy and number of rules are generated.

4. Experiments and Discussions

Experiment on FINERS is carried out on 8 datasets from UCI MLR (Merz & Murphy, 1998). The datasets are carefully selected to represent heterogeneous data types and non-heterogeneous data types. The heterogeneous data sets are Australian Credit (CRX), German Credit (GC), Hepatitis (HP), Cleveland Heart Disease (HD) and Ljubljana Breast Cancer (BC), the non-heterogeneous data sets are Iris Plant (IRIS), Zoo Animals (ZOO), Wisconsin Breast Cancer (WBC) (Zwitter & Milan Zoklic, 1998). The description of each data set is shown in Table 1.

	CRX	GC	BC	HD	HP	IRIS	ZOO	WBC
Continuous	6	7	0	2	6	4	0	0
Nominal	9	11	6	8	13	0	16	0
Discrete	0	2	3	3	0	0	0	9
Class	2	2	2	2	2	3	7	2
Training	562	900	249	267	132	135	91	629
Testing	62	100	28	30	15	15	10	70

Table 2. Heterogeneous and Non-heterogeneous Data Set

The dataset is distributed into 10 fold cross validation with 90% data for training and 10% data for testing with no overlapping. The data is tested in its original types as provided in the databases. For a consistent condition and comparison on FINERS and FAIRS (Puteh et al., 2008) and other immune algorithms from WEKA toolbox (Witten & Frank, 2005), they are tested using the same sets of 10-fold CV data. The selected immune classifiers from WEKA toolbox are AIRS1 (Watkins, 2001; Watkins et al., 2004; Brownlee, 2005), AIRS2 (Watkins & Timmis, 2002; Brownlee, 2005), AIRS2Parallel (AIRS2P) (Watkins & Timmis, 2004; Brwonlee, 2005), IMMUNOS1 (Brownlee, 2005; Carter, 2000) and CLONALG (Brownlee, 2005; de Castro & Von Zuben, 2000). The average accuracy is calculated from the 10 sets for each dataset and the significant difference is analyzed using paired T-Test using standard statistical package. Table 2 shows the comparison of the accuracy rates and Table 3 shows the comparisons of the rules reduction between FINERS and the other immune algorithms on heterogeneous data. Sig value in 2nd column shows the statistically significant value in differences. The value in bold is the highest accuracy in the table for each data set. The difference is significant if the significant value is less than 0.05 with 95% confidence (Coakes & Steed, 2003). NA is not applicable which means that these classification models do not test the value.

		ACCURACY				
	Sig	CRX	GC	BC	HD	HP
FINERS		**87**	**75**	**73**	**89**	**88**
FAIRS	0.070	**87**	74	72	88	**88**
AIRS1	0.000	80	67	68	82	83
AIRS2	0.001	83	71	68	82	84
AIRS2P	0.006	81	71	67	80	85
IMMUNOS1	0.027	85	68	71	86	80
CLONALG	0.025	63	70	68	71	75

Table 3. Accuracy (%) of heterogeneous data

		RULES REDUCTION				
	Sig	CRX	GC	BC	HD	HP
FINERS		30	**35**	**71**	**43**	30
FAIRS	0.006	11	28	50	34	12
AIRS1	0.900	**62**	20	45	42	**34**
AIRS2	0.045	29	13	33	18	22
AIRS2P	0.021	22	11	23	14	16
IMMUNOS1		NA	NA	NA	NA	NA
CLONALG		NA	NA	NA	NA	NA

Table 4. Rules Reduction (%) of heterogeneous data

The result shows that FINERS gives higher accuracy rate and higher rules reduction percentage in most of the heterogeneous datasets compared to other immune algorithms with statistically significant in differences.

Table 4 shows the comparison of the accuracy rates and Table 5 shows the comparisons of the rules reduction between FINERS and the other immune algorithms on non-heterogeneous data. Sig value in 2nd column shows the statistically significant value in differences. The value in bold is the highest accuracy in the table for each data set. The difference is significant if the significant value is less than 0.05 with 95% confidence (Coakes & Steed, 2003). NA is not applicable which means that these classification models do not test the value.

		ACCURACY		
	Sig	IRIS	ZOO	WBC
FINERS		**97**	89	**98**
FAIRS	0.681	**97**	95	97
AIRS1	0.644	96	**98**	97
AIRS2	0.057	94	89	96
AIRS2P	0.381	94	**98**	96
IMMUNOS1	0.181	**98**	96	85
CLONALG	0.240	92	94	94

Table 5. Accuracy (%) of non-heterogeneous data

The result shows that the differences are not statistically significant which means no improvement in accuracy rates by FINERS compared to the previous classification models on non-heterogeneous data. But, for rules reduction, FINERS shows an improvement compare to FAIRS.

		RULES REDUCTION		
	Sig	IRIS	ZOO	WBC
FINERS		39	61	**74**
FAIRS	0.044	35	57	53
AIRS1	0.166	**65**	47	46
AIRS2	0.337	**65**	**81**	52
AIRS2P	0.133	53	72	48
IMMUNOS1		NA	NA	NA
CLONALG		NA	NA	NA

Table 6. Rule reduction (%) of non-heterogeneous data

5. Conclusion

This paper has proposed a new AIS immune network classifier called Flexible Immune Network Recognition System (FINERS) that uses HVDM as a distance metric for heterogeneous data type without the need for the discretization or transformation of the data into specific type. The experimental results show that the immune network model produces a better accuracy in most of the heterogeneous datasets and it also generates less rules compared to previous immune classification models. Comparing FINERS to FAIRS, although there are no differences in the accuracy for the heterogeneous data, using network feature from the immune system decreases the number of rules in the classifiers. The study solves some limitation shown in (Watkins, 2001; Freitas & Timmis, 2007; Hart & Timmis, 2008; Timmis, 2006). However, FINERS does not show a significant different or improvement on the accuracy and rules reduction on non-heterogeneous data compared to the previous AIS classification models. In conclusion, the results suggest that the use of network feature and to process data in its original types can increase accuracy performance while reducing the number of rules in heterogeneous data. Furthermore, it is significant to process the data in its original types to avoid degradation of data accuracy and it decreases the time in pre processing of data. For the future investigation, other AIS algorithm can employ HVDM function for other tasks such as optimization and clustering. FINERS could also be further refined to make it dynamic and be able to process dynamic data such as time series data. With the result, we hope to derive a more stable and flexible AIS classifier.

6. References

Brownlee, J. (2005). Artificial Immune Recognition System (AIRS) A review and Analysis, CISCP, Faculty of ICT, Swinburne University of Technology, Australia, Technical Report 1-02

Brownlee, J. (2005). Clonal Selection Theory & ClonalG and The Clonal Selection Classification Algorithm (CSCA), CISCP, Faculty of ICT, Swinburne University of Technology, Australia, Technical Report 1-02

Brownlee, J. (2005). Immunos-81 The Misunderstood Artificial Immune System, CISCP, Faculty of ICT, Swinburne University of Technology, Australia, Technical Report 1-02

Burnet, F. M. (1959). The Clonal Selection Theory of Acquired Immunity, Cambridge University Press

Carter, J. H. (2000). The Immune Systems as a Model for Pattern Recognition and Classification, Journal of the American Medical Informatics Association 7 (1)

Coakes, S. J. & Steed, L.G. (2003). SPSS Analysis without Anguish Version 11.0 for Windows, John Wiley & Sons Australia, Ltd

Cover, T. M. & Hart, P. E. (1967). Nearest Neighbor Pattern Classification, IEEE Transactions on Information Theory, Vol 13, No. 1, pp. 21-27

Dasarathy & Belur, V. (1991). Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques , IEEE Computer Society Press Los Alamitos, CA

Dasgupta, D. (2006). Advances in Artificial Immune Systems, IEEE Computational Intelligence Magazine

de Castro, L. N. & Von Zuben, F. (2000). The Clonal Selection Algorithm with Engineering Applications. *Proceedings of GEC2000*, Las Vegas, pp.36-37

de Castro, L. N. & Timmis, J. (2002). Artificial Immune Systems: A New Computational Intelligence Approach, Springer-Verlag

Freitas, A. & Timmis, J. (2007). Revisiting the Foundations of Artificial Immune Systems for Data Mining, IEEE Transactions on Evolutionary Computation, 11:4

Hamaker, J. & Boggess, L. (2004). Non-Euclidean Distance Measures in AIRS, an Artificial Immune Classification System, *Proceedings of CEC2004*

Hart, E. & Timmis, J. (2008). Applications of Artificial Immune Systems: The Past, the Present and the Future, Journal of Soft Computing, vol 8(1) 191-201

Hart, P. E. (1968). The Condensed Nearest Neighbor Rule, IEEE Transactions on Information Theory, Vol 14, pp.515-516

Hunt, J. E. & Cooke, D. E. (1996). Learning Using an Artificial Immune System, Journal of Network Computer Applications, 19:189-212

Jerne, N. K. (1974). Towards a Network Theory of the Immune System, Ann. Immunol. (Inst. Pasteur) 125C, 373 – 389

Merz C. J. & Murphy, P. M. (1998). UCI Machine Learning Repository, University of California, Irvine, USA, http://www.ics.uci.edu/~mlearn/MLRepository.html

Puteh, M. ; Hamdan, A. R. ; Omar, K. & Abu Bakar, A. (2008). Classifying Heterogeneous Data with Artificial Immune System, *Proceedings of IEEE International Symposium on Information Technology (ITSIM2008)*

Stanfill, C. & Waltz, D. (1986). Towards Memory-based Reasoning, ACM, Vol 29, pp. 1213-1228

Timmis, J. (2001). Artificial Immune Systems: A Novel Data Analysis Technique Inspired by the Immune Network Theory, Ph. D. thesis, Department of Computer Science, University of Wales, Aberystwyth

Timmis, J. (2006). Challenges for Artificial Immune System, LNCS 3931, pp355-367, Springer Verlag

Ventura ; Dan & Martinez, T. R. (1995). An Empirical Comparison of Discretization Methods, *Proceedings of the 10th International Symposium on Computer and Information Sciences*, pp. 443-450

Watkins, A. (2001). A Resource Limited Artificial Immune Classifier, Master's Thesis, Mississippi State University

Watkins, A. ; Timmis, J. & L. Boggess, L. (2004). Artificial Immune Recognition System (AIRS): An Immune-Inspired Supervised Learning Algorithm, Genetic Programming and Evolvable Machines, vol 5, pp291-317

Watkins, A. & Timmis, J. (2002). Artificial Immune Recognition System (AIRS): Revisions and Refinements, *Proceeding of ICARIS2002*, pp173-181, Springer Verlag

Watkins, A. & Timmis, J. (2004). Exploiting the Parallelism Inherent in AIRS, an Artificial Immune Classifier, *Proceeding of ICARIS2004*, pp427-438, Springer Verlag

Wilson, D. R. &. Martinez, T. R. (1997). Improved Heterogeneous Distance Functions, Journal of Artificial Intelligence Research 6, 1- 34

Witten, I. H. & Frank, E. (2005). Data Mining, Morgan Kaufmann Publishers, USA

Zwitter, M. & Milan Soklic, M. (1998). Institute of Oncology, University Medical Center, Ljubljana, Yugoslavia

Machine Learning

Edited by Yagang Zhang

ISBN 978-953-307-033-9
Hard cover, 438 pages
Publisher InTech
Published online 01, February, 2010
Published in print edition February, 2010

Machine learning techniques have the potential of alleviating the complexity of knowledge acquisition. This book presents today's state and development tendencies of machine learning. It is a multi-author book. Taking into account the large amount of knowledge about machine learning and practice presented in the book, it is divided into three major parts: Introduction, Machine Learning Theory and Applications. Part I focuses on the introduction to machine learning. The author also attempts to promote a new design of thinking machines and development philosophy. Considering the growing complexity and serious difficulties of information processing in machine learning, in Part II of the book, the theoretical foundations of machine learning are considered, and they mainly include self-organizing maps (SOMs), clustering, artificial neural networks, nonlinear control, fuzzy system and knowledge-based system (KBS). Part III contains selected applications of various machine learning approaches, from flight delays, network intrusion, immune system, ship design to CT and RNA target prediction. The book will be of interest to industrial engineers and scientists as well as academics who wish to pursue machine learning. The book is intended for both graduate and postgraduate students in fields such as computer science, cybernetics, system sciences, engineering, statistics, and social sciences, and as a reference for software professionals and practitioners.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen