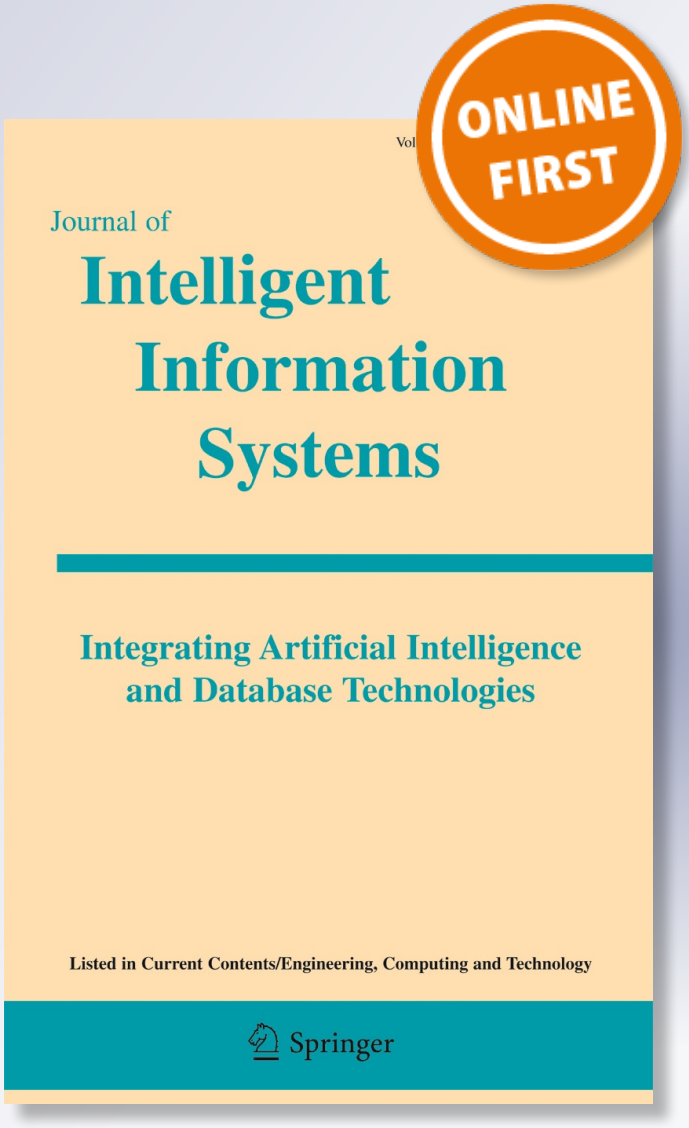
Artificial immune network with feature selection for bank term deposit recommendation

Xiao-Yong Lu, Xiao-Qiang Chu, Meng-Hui Chen, Pei-Chann Chang & Shih-Hsin Chen

Journal of Intelligent Information Systems
Integrating Artificial Intelligence and Database Technologies

ISSN 0925-9902

J Intell Inf Syst
DOI 10.1007/s10844-016-0399-2



Springer

J Intell Inf Syst
DOI 10.1007/s10844-016-0399-2

Artificial immune network with feature selection for bank term deposit recommendation

Xiao-Yong Lu[1] · Xiao-Qiang Chu[2,3] · Meng-Hui Chen[2] · Pei-Chann Chang[2] · Shih-Hsin Chen[4]

Received: 3 October 2015 / Revised: 27 January 2016 / Accepted: 15 March 2016

Abstract Artificial immune systems (AIS) have been widely utilized for pattern recognition and data analysis in various fields of science and technology, and artificial immune networks (AIN) are based on AIS. In this study, an artificial immune network is used for collaborative filtering as a classification model for bank term deposit recommendations, once feature selection has been applied to filter out key features for classification purposes. AIN is used to represent a network of customers with bank term deposits, and it can be adopted as a group decision-making model in predicting whether a new customer will have a term deposit or not. Formulae for calculating the affinity between an antigen and an antibody, and the affinity of an antigen to an immune network are also developed. A series of experiments are conducted, and the results are very encouraging. Despite the class imbalance problem in the test dataset, the proposed model outperformed other models, achieving the highest accuracy in testing.

Keywords Financial product · Term deposit · Artificial immune system · Collaborative filtering · Recommendation system

✉ Pei-Chann Chang
iepchang@saturn.yzu.edu.tw

Meng-Hui Chen
bbb422@hotmail.com

Shih-Hsin Chen
shchen@csu.edu.tw

[1] School of Software, Nanchang University, Nanchang, China

[2] Department of Information Management, Yuan Ze University, Chung-Li, Taiwan

[3] School of Economic & Management, Nanchang University, Nanchang, China

[4] Department of Information Management, Cheng Shiu University, Kaohsiung City, Taiwan

Published online: 28 March 2016

1 Introduction

Insight into how individuals make choices is very important for economists, as it can be used to develop economic models. Consumers, business managers and government policy-makers use these models every day to help make choices. People are complex organisms with a multitude of specialized cells, which perform certain categorical functions with a fair division of labor. Each cell, tissue, organ, and organ system has its own function, and the human brain is the most complex of all. As a result, people are able to celebrate and solve quandaries, decide between right and wrong, invent things and habituate themselves to their circumstances. If techniques used by such a successful biological system are used in mechanical systems, the results obtained can outperform manual-experience based techniques. The human immune system is a system of biological structures and processes which fights against diseases by identifying and killing pathogens and tumor cells. It detects a wide variety of agents, known as pathogens, which include viruses and parasitic worms. The idea of boosting a system's immunity is enticing, but the ability to do so has proved elusive for several reasons. The immune system is a system, not a single entity. To function well, it requires balance and harmony. Artificial immune networks (AIN) try to replicate this balance and harmony found in nature, imitating a human immune system for application to mechanical real world systems. This paper explores one such artificial immune network, and gives examples of how it may be applied in real world systems. The various algorithms used in business models are an important part of understanding the current business climate and learning how to apply economic concepts to a variety of real-world events. With the rapid development of social economics and information technology, customer numbers and financial products have rapidly increased. Multiple and complex marketing activities have significantly reduced public enthusiasm for economic participation. In addition, because of increasing economic pressures and the intensification of competition with each other, financial institutions have had to acknowledge the vital nature of efficiency and cost when they recommend products. When it comes to people, resources are always limited, and thus optimization is very important. Therefore, in the big data environment, identifying potential customers accurately and taking a highly individualized approach to recommend products to them can reduce costs and improve work efficiency. At the same time, customers who are not interested will not be excessively disturbed. This will not only help the institution individually, but will also add to the convenience of the general public. It will also help to retain customer loyalty, which is very important to such institutions.

Financial institutions customer information databases are very extensive, containing up to millions of entries for each company, with each customer record containing hundreds of gigabytes of information. Data mining technology in big data analysis is thus very useful to financial institutions, especially banks. At present, one of the main businesses of banks is term deposits; term deposits are in fact banks lifelines. However, with increasing investment required in human resources, material resources and time costs, many marketing activities are delivering ever decreasing profits. To solve these problems, an efficient recommendation model which can accurately predict whether a customer will subscribe a term deposit is necessary.

A recommender system is mainly based on the basic information, interests, past purchase behavior and other related historical information of users for data mining in order to find useful information for predictions and recommendations of certain objects (Hu 2005). Such systems also have significant implications for big data analysis. There is a huge amount of available information online today, but only some of it will be useful for various financial projects. This study, therefore, combines an artificial immune system with collaborative

filtering to create a recommendation model to assist bank managers to improve marketing to potential customers most likely to subscribe term deposits so that resources are not allocated unnecessarily to marketing to customers unlikely to subscribe. This is achieved using a real-life dataset collected from a Portuguese retail bank. The artificial immune system can be defined as an implementation learning technique inspired by the human immune system (Nasir et al. 2009). AIS makes use of genetic models to include an implementation of a genetic algorithm - a bio-inspired evolutionary algorithm.

This research focuses on proposing a prediction model to judge whether a client will subscribe a term deposit. Moro used a decision support system (DDS) based on a data-driven model to predict the success of bank telemarketing, and this study uses the same dataset that he used (Moro et al. 2014). In his research, four data mining models were compared: LR, DT, NN and SVM. He compared the performances of these models using two metrics: Area of the receiver operating characteristic curve (AUC) and area of the LIFT cumulative curve (ALIFT). The results showed that his method was better. In addition, many scholars now use a data mining approach to predict customer churn, credit card fraud and so on.

Hu (Hu 2005) discussed challenging issues such as highly skewed data, time series data unrolling, leaker field detection, etc., and the procedure of a data mining project for attrition analysis for retaining bank customers. The credit card fraud detection problem was solved by artificial immune system (Gadi et al. 2008). Logistic regression and decision trees were used to build a churn prediction model using credit card data collected from a real Chinese bank (Nie et al. 2011). Frameworks for customer churn prediction using longitudinal behavioral data, a hierarchical multiple kernel support vector machine model, and a three phase training algorithm have also been developed (Chen et al. 2012). A hybrid approach for extracting rules from SVM for customer relationship management (CRM) purposes was developed (Farquad et al. 2014), which deals with large-scale unbalanced data with respect to rule extraction from SVM.

There are several popular classification models, such as decision trees (DT), support vector machines (SVM), logistic, naive Bayes (NB) classifiers, artificial immune recognition system (AIRS1) and immunos99. The DT, logistic, and NBdata mining techniques can be easily understood by people. SVM, however, is more flexible. AIRS1 and immunos99 are branches of immune algorithms. Therefore, this research uses all the above methods as the comparative methods.

The paper is further organized as follows: Section 2 presents the literature review; Section 3 describes the methods used in this research; the experimental results are given in Section 4; finally, conclusions are drawn in Section 5.

2 Literature review

In this study, feature selection will be applied first to filter out key features for classification purposes. Then, an artificial immune network will be applied to collaborative filtering as a classification model for term deposit recommendation. AIN used to describe a network of customers with term deposits, and it can be adopted as a group decision-making model in predicting whether new customers will subscribe a term deposit. Formulas were developed to calculate the affinity between an antigen and an antibody, and the affinity of an antigen to an immune network. In addition, a modified similarity estimation formula based on the weighted Pearson correlation coefficient was also developed. A series of experiments were conducted, and the results are very encouraging.

2.1 Feature selection and extraction

Dimension reduction serves to reduce the number of variables under consideration in machine learning. There are two methods used in dimension reduction; one is feature selection, and the other is feature extraction. Feature selection is an approach which tries to find a subset of the original variables. Feature extraction is an approach which transforms the data in the high-dimensional space to a space of fewer dimensions. This paper adopts principal component analysis and information gain for dimension reduction to enhance the performance of the prediction model.

2.1.1 Principal component analysis

Tipping and Bishop developed a Probabilistic Principal Component Analysis (PCA) which is a ubiquitous technique for data analysis and processing (Tipping and Christopher 1999). They utilized the probabilistic approach because the earlier PCA was not predicated on a probabilistic model. They demonstrated how the principal axes of a set of observed data vectors may be tenacious through maximum likelihood estimation of parameters in a latent variable model that is approximately cognate to factor analysis. Utilizing illustrative examples, they conveyed the advantages of this probabilistic approach to PCA.

A better overview was made by Euerby and Patrik, where they used PCA to perform Chromatographic relegation as well as the comparison of inverted-phase liquid chromatographic columns available commercially (Euerby and Patrik 2003). Sato applied principal-component analysis on the near-infrared spectroscopic data of vegetable oils to relegate them (Sato 1994). The PCA pseudo code for selecting features as used in this paper is as follows:

Algorithm 1 Main procedure

$Input$: $x_1, \ldots, x_n$; d length attributes; k
$Output$: Transform matrix R

1: $X \leftarrow n \times d$ data matrix with x_i in each row of attributes
2: $\bar{x} \leftarrow \frac{1}{n} \sum_{i=1}^{n} x_i$
3: $X \leftarrow$ subtract $\bar{x}$ from each row x_i in X
4: $COV \leftarrow X$ Compute eigenvalue $e_1, \ldots, e_d$ of COV, and sort them
5: Compute matrix V which satisfy $V^{-}1 \times COV \times D$, $V = D$, D is the diagonal matrix of eigenvalue of COV
6: $R \leftarrow$ the first k column of V

2.1.2 Information gain

Lee and Geunbae (Lee and Gary 2006) proposed a feature selection method striving to reduce redundancy between features while maintaining information gain in selecting appropriate features for text categorization, since most previous works on feature selection emphasized the reduction of the high dimensionality of the feature space. They gave empirical results on some datasets, demonstrating the effectiveness of their feature selection method. Uethuz (Uğuz 2011) used two-stage feature selection and feature extraction in order to improve text categorization performance. Each term within the document was ranked in the first stage, depending upon their importance for the classification using the information

gain method. In the second phase, genetic algorithm (GA) and principal component analysis (PCA) feature selection and feature extraction methods were applied separately to those terms which were ranked in order of decreasing importance. A dimension reduction was also carried out. Thereby, during text categorization, terms of lesser importance were ignored, and extraction and feature selection methods were applied to the terms with the highest importance, thus reducing the complexity and computational time of the categorization.

The pseudo code for information gain selectionof features in this paper is shown below:

Algorithm 2 Main procedure

$Input$: R: a bank term deposit dataset of non-target attributes; C: the target attribute; S: training data.
$Output$: Transform matrix R

Initialize to the order of original data attributes
if S is empty then **then**
    Return a single attribute failure value
end if
if R is empty then **then**
    Return a single attribute with its value as the most common importance attribute value of the target attribute found in S
end if
$D \leftarrow$ the attribute that has the largest $Gain(D, S)$ among all the attributes of R
$\{d_i\, j = 1, 2, \ldots, m\} \leftarrow$ Attribute values of D
Return whose attribute is D and the impacts are labeled by $d_1, d_2, \ldots, d_m$ and going to another attrib-utes $ID3(R - D, C, S1)$, $ID3(R - DC, S2)$, .., $ID3(R - D, C, S_m)$
$R \leftarrow$ the first k column of V

2.2 AIS for recommendation systems

2.2.1 Artificial immune system

Despite the constant growth of science and technology, nature has been a great source of inspiration for mankind. Biological systems have always been proved better than mechanical systems regarding efficiency, efficacy and processing time. When it comes to human beings, resources are always limited, and there is thus a continuous quest for optimization. Similarly, this paper deals with the filtering of recommendations provided to customers, which saves time, money and effort. People have long been fascinated by the concept of artificial intelligence, but in recent times, technology has advanced sufficiently for it to become a real possibility. Artificial intelligence has helped achieve business goals involving different firms. There are several instances where people fail to outperform computer counterparts. It has been seen in recent times that Artificial Immune Systems have outperformed other algorithms. An excellent example of artificial intelligence is AIS, which mimics biological systems present in the human body. Unlike Genetic algorithms, which use a combination of natural selection, recombination and mutation of encoded genetic data to evolve the solution to the problem, AIS uses genetic models and includes an implementation of genetic algorithms. The most important types of AIS are based on the concepts of negative selection, clonal selection, and the immune network. In this research, the method used

belongs to the immune network (Dudek 2012). Immune network theory was first proposed by Jerne (Jerne 1974) in 1974. He suggested that in a real immune system when a biological organism encountered an invasion from outside pathogens, the biological organism would develop its protective agents known as antigens. The immune system would produce antibodies to combine with antigens to fight pathogens as an immune response. In recent years, many scholars have studied and improved on this understanding. This study focuses on antigens, and searches for an antibody that can be combined with it to for the solution to the problem (Dudek 2012). By using the adaptive immune response, this algorithm can be used to search for the best solution to an optimization problem. An algorithm that uses memory cells tuned using the magnitude of the standard deviation obtained with average affinity variation in each generation was proposed (Aydin et al. 2012). A self-evolving artificial immune system achieved by coordinating the T and B cells in the immune system to build a block-based artificial chromosome was also developed to reduce computation time and to improve performance with problems with different complexities (Chen et al. 2013). The imbalance of data is one problem in this research. Some studies have used artificial immune systems combined with fuzzy algorithms to create a classification model (Alatas and Akin 2005). A class of imbalanced problems (Dai 2015) was solved by using bioinformatics based fuzzy k –NN algorithm with and without cross-validation (Sengur 2009).

The use of an Artificial Immune System (AIS) has been in the research limelight over the last few decades. Simon and Hemamalinideveloped clonal selection, which was based on Artificial Immune System (AIS), and served to solve the dispatch problems that generated units and had a valve point effect. They compared the proposed technique with other methods, highlighting the advantages of using AIS (Hemamalini and Sishaj 2011). A better overview of the technique was made by Rahman and Hamid where a forecasting model was presented using Artificial Immune System (AIS) (Chang and Yeh 2012; Abdul and Rahman 2010; Pang and Coghill 2015). It was seen that AIS provided a comparable forecast to Artificial Neural Network (ANN) as a learning algorithm. Chong and Chen forecasted dynamic needs of customers, making efficient use of Artificial Immune System (AIS). Dynamic customer requirements were forecasted and satisfied by minimizing the risk of product development for shifting markets. Wu (Wu 2010) used Artificial Immune System (AIS) with back-propagation neural networks (BPNNs) to minimize an energy function based on the errors occurring between actual and desired outputs. The conventional BPNN's weight optimization was highly dimensional and unconstrained in nature. This limitation was overcome by making use of AIS algorithm-based BPNN (named AIS-BPNN). Suliman and Rahman used AIS for the prediction of voltage stability in power systems (Suliman and Rahman 2010). The voltage stability evaluation could also be used as an early warning system so that necessary actions could be taken to avoid actual voltage collapse. Cacheda et al. compared different related techniques in the literature and studied their characteristics, highlighting the pros and cons of each (Cacheda et al. 2011).

This research will use AIRS and immunos99 as the comparative methods. AIRS was developed from the AINE immune network (Knight and Timmis 2001). Immunos99 is a classification algorithm based upon the principles of AIS. Immunos99 provides one alternative. However, the algorithm and implementations have some room for performance improvement (Taylor et al. 2013).

2.2.2 Collaborative filtering

Collaborative filtering (CF), as a kind of personalized recommendation technique, has been widely used in many domains. However, collaborative filtering also suffers from a few

issues such as instance, the cold start problem, data sparseness, scalability and so on. These problems severely devalue user experience (Liu et al. 2014). To solve the data sparseness problem, an item based collaborative filtering recommendation algorithm using rough set theory prediction was developed (Su and Ye 2009). This method employs rough set theory to fill the vacant ratings of the user-item matrix wherever necessary. An approach that combines the advantages of these two kinds of approaches by joining the two methods was also presented (Gong et al. 2009). Firstly, the proposed approach employs memory-based CF to fill the vacant ratings of the user-item matrix. Then, it uses the item-based CF as model-based to form the nearest neighbors of every item. Finally, it produces the prediction of the target user to the target item in real time. Improving the accuracy is always the target of CF. Liu (Liu et al. 2014) presented a new user similarity model that not only considers the local context information of user ratings but also the global preference of user behavior to improve the recommendation performance when only a few ratings are available to calculate the similarities between each user. A novel method to find the neighbor users based on the users' interest patterns was also developed (Ramezani et al. 2014). The main idea is that users who are interested in the same set of items share similar interest patterns.

This is very similar to a group making a decision. Many scholars have studied this in detail and have come up with different results. Xia et al. studied the aggregation of fuzzy information (Xia et al. 2013). They proposed a series of several aggregation operators and discussed their connections. In order to reflect the relationship among the present aggregation arguments, they proposed two methods for determining the weight vectors. On the basis of support degrees between aggregation arguments, the weight vectors of decision makers were obtained more objectively. To deal with the correlation of criteria, they applied the Choquet integral method to obtain the weights of various criteria.

3 Methodology

The main idea in this research simulates the innate immune system to define each training data as an antigen. The immune system will generate antibodies as per the classification rules by the antigen invading the immune system. At the beginning of the training phase, there are two antigens from different classes to be treated as the initial antibodies. Next, whether a new antigen is defined as an antibody or not is decided by calculating the affinity between the antigen and each antibody. If the affinity calculated is smaller than a predefined threshold, the antigen is treated as a new antibody. When each antigen has invaded the immune system, the training procedure finishes. After the training phase, the antibodies are as per the classification rules to predict the class of each testing data. The forecasted class of the testing data is defined by the class of the antibody which has the highest affinity between the testing data and the antibody. The framework of the proposed model is shown in Fig. 1.

The training phase is based on AIN to generate the classification rules. In order to evaluate the performance of the forecasting model, the data used is different from the training data, as the testing data is to be predicted by the proposed model. More details of the training and testing phase are shown as follows.

3.1 Training phase

As mentioned above, the purpose in the training phase is to generate antibodies as the classification rules. In this research, each data is defined as an antigen to invade the immune

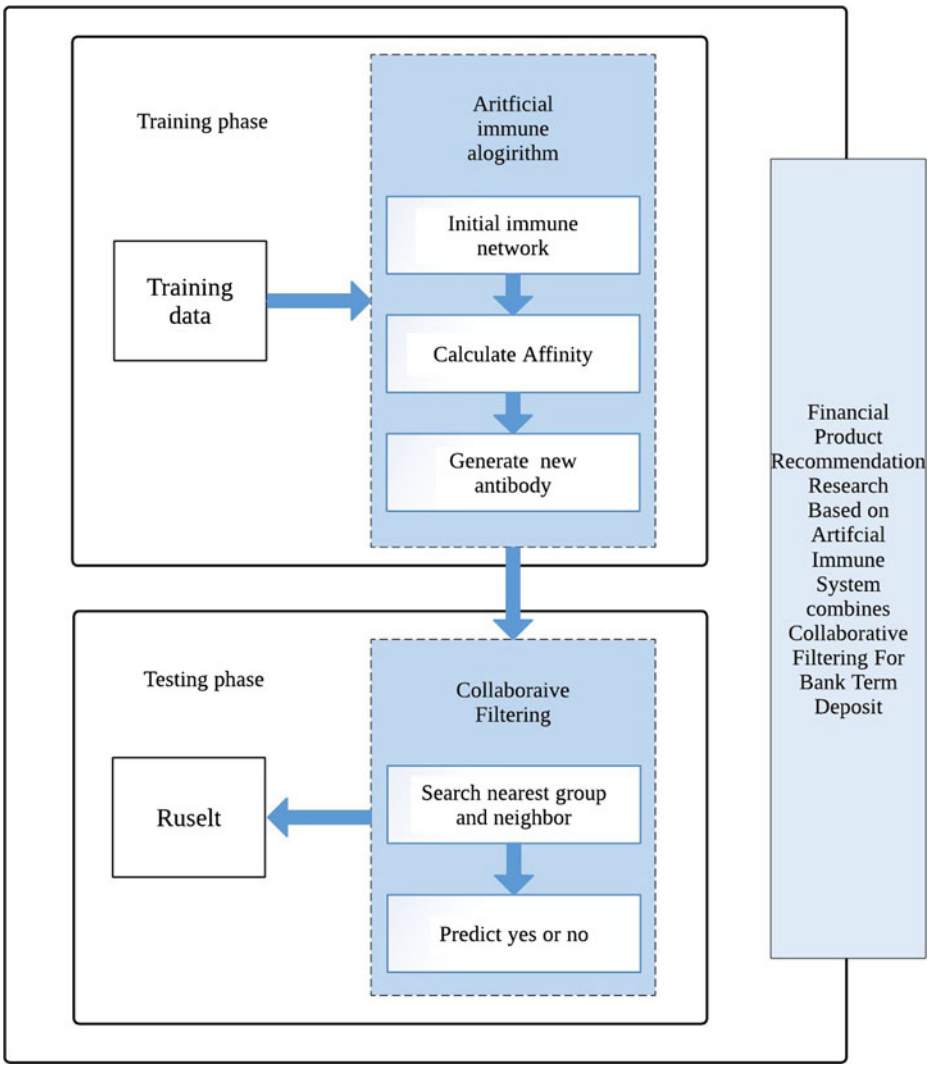


Fig. 1 The structure of proposed approach

system. Then, the process finds antibodies with high affinity with the invading antigen. If the affinity is larger than the affinity threshold, the antigen and the antibody are considered as the same antibody to construct an immune network, otherwise, if the antigen and the antibody are from different classes, the antigen is defined as a new antibody. Unless each antigen is invaded, the training procedure won't finish. The process of the training phase is shown in Fig. 2.

This section is further organized as follows: Section 3.2 presents how to evaluate the affinity between antigen and antibody; Section 3.3 describes the procedure of constructing new immune networks (Fig. 3).

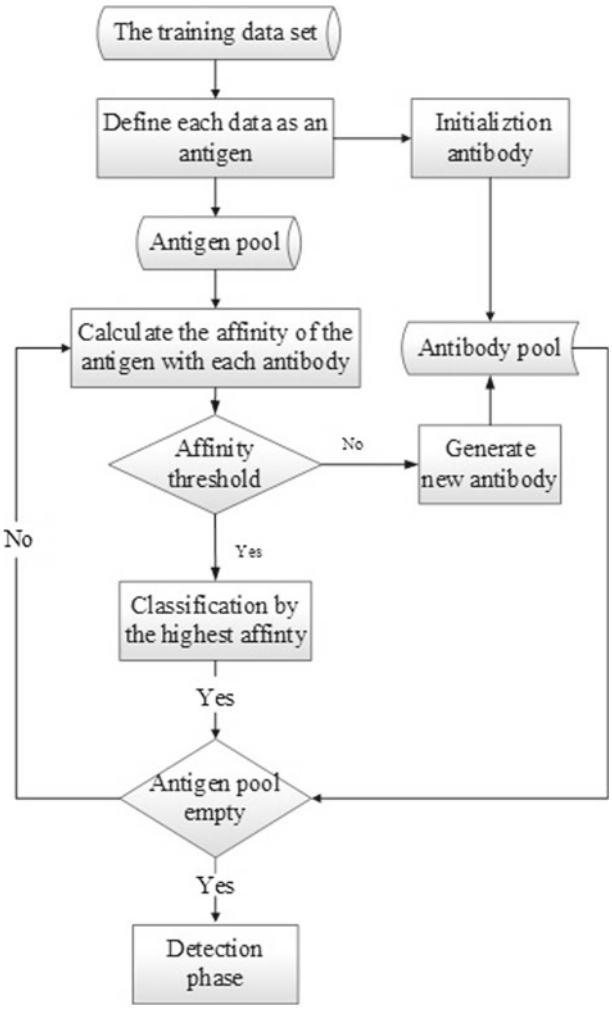


Fig. 2 The flowchart of AIN algorithm for training phase

3.2 Affinity

Affinity chromatography is a method for separating biochemical mixtures, and is based on highly specific interaction such as that between antigen and antibody, enzyme and substrate or receptor (Chen et al. 2013). The affinity in this research is defined as the degree of similarity between the antigen and antibody, and evaluated by the Hamming distance. The approach is represented as follows:

$$Affinity(Ag_k, m_n) = \sum_{v=1}^{p} H(Ag_{k,F}, Ab_{v,F})/p \tag{1}$$

Here, p is the number of Ab contained in m_n. $H(Ag_k, F, Ab_v, F)$ is the calculation function of Hamming distance. Ag_k, F and Ab_v, F is the basic information in set F of antigen

```
1.  Define T as the threshold of clustering the m_n by client
2.  Begin
3.  For each antigen Ag_k
4.      For each m_n
5.        Affinity(Ag_k, m_n)
6.      End For
7.      Find the m_n with the best affinity of |M|
8.      If (Affinity(Ag_k, m_n) > T ) Then
9.        Add the new Ab_p to m_n by reproducing Ag_k
10.     Else
11.       Generate m_n and Add the new Ab_p to m_n by reproduc-
    ing Ag_k
12.     End If
13.   End For
14. End
```

Fig. 3 The pseudo code of our AIN algorithm for generating new antibody

and antibody. But this research has made some modifications to the calculation function of Hamming Distance, and the results have been normalized. The approach is represented as follows:

$$H(Ag_{k,F}, Ab_{v,F}) = \sum_{y=1}^{q} h(Ag_{k,f_y}, Ab_{v,f_y})/q \tag{2}$$

where, q is the number of f in the F set, $h(Ag_k, F, Ab_v, F)$ is the modified calculation function of Hamming distance, if $Ag_k, F = Ab_v, F$, the result is 1, otherwise it is 0.

3.3 Generating new immune networks

In AIS, a clonal selection is used to increase and store positive information. Thus, when an antibody with high affinity to immune networks occurs, the antibody will be cloned to the immune network. After finishing the initial immune network, new immune networks are generated or reproduced by Ag_k. Therefore, T defines the threshold deciding that Ag_k is used to generate a new immune network or reproduce related immune networks. The steps and pseudo code of generating new antibodies are represented as follows:

Step1: Calculate the affinity between all Ag_k and $|M|$.
Step2: Find the m_n with the best affinity of $|M|$.
Step3: If (Affinity (Ag_k, m_n) > T) then add the new Ab to m_n by reproducing Ag_k. Else generate m_n and add the new Ab to m_n by reproducing Ag_k.
Step4: Repeat Step1 to Step3 until all the antigens are distinguished.

In this research, whether a new customer will subscribe a term deposit or not is predicted by group decision-making. Thus, the proposed AIN model can generate various immune networks by training data invaded. Some of these immune networks may be the same class, but they can be considered as different types by the different level of affinities to the target customer. In addition, the affinity between the target customer and the center of the

immune networks is calculated to yield the weight. If the weight is large, it means the class of the immune network has a high degree of influence on the target customer. The detail of predicting the class of customers is shown as follows.

3.4 Testing phase

After generating the immune networks, the testing data is used to search the related immune networks; furthermore, related neighbors are found in the recent immune network. Finally,

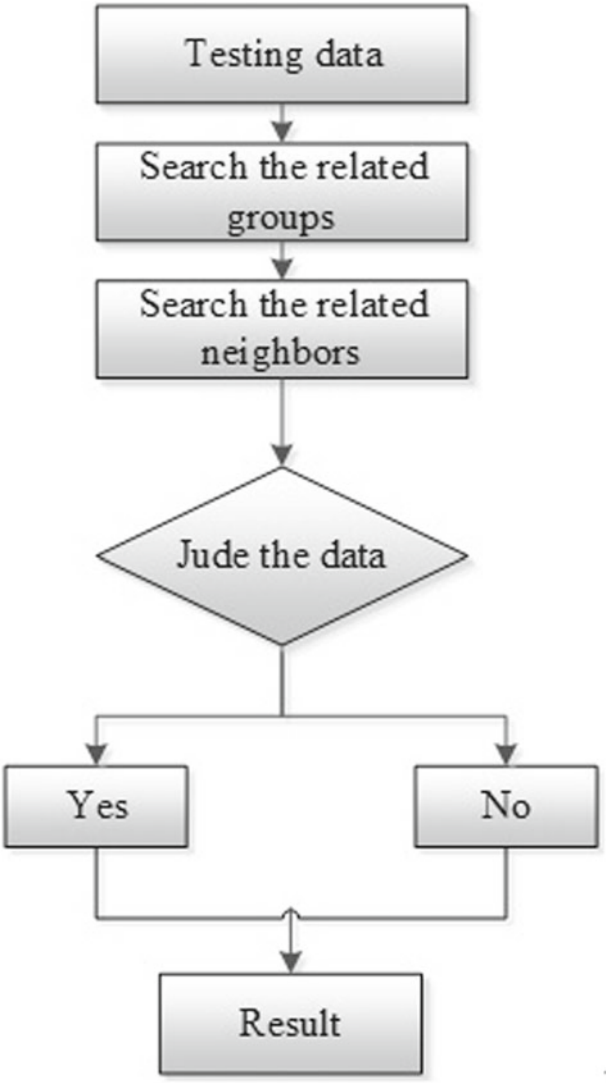


Fig. 4 The flowchart of AIN algorithm for testing phase

```
1.  Define T' as the threshold of client similarity by client
2.  Define T'' as the threshold of immune networks simi-
      larity by client
3.  Define C_temp=Ø
4.  Begin
5.    For each client c_v
6.      For each group g_k
7.        If (group similarity(c_v,g, g_k)>T'') Then
8.          For each c_z
9.            If (similarity(c_v, c_z)>T') Then
10.             C_temp←c_z
11.           End If
                        12.  End For
13.       End If
14.     End For
15.     Prediction(c_v, C_temp)
16.     C_temp=Ø
17.   End For
```

Fig. 5 The pseudo code of our AIN algorithm for searching the related groups and neighbors

whether a user subscribes a term deposit or not will be determined by comparison with related neighbors. Fig. 4 is the flowchart of the proposed AIN approach for the testing phase (Fig. 5).

The purpose here is to find similar immune networks with the target client and then based on these immune networks find the most similar client. In the end, immune networks are selected based on the decisions of this client, so as to predict the target client's movements. The steps of searching the related immune networks and neighbors are represented as follows:

Step1: Calculate the similarity of immune networks among the immune networks which c_v belongs to.
Step2: Search the related immune networks: if the similarity of immune networks $(c_{v,g}, g_k) > T$", then calculate the similarity of c_s , else repeat step1.
Step3: Search the related neighbors: if the similarity of neighbors $(c_v, c_s) > T$', then copy u_s to C_{temp}, else repeat step2.
Step4: Repeat Step1 to Step3 until all the g_k are calculated.
Step5: Use C_{temp} to predict the movement of c_v.
Step6: Clean C_{temp}.
Step7: Repeat Step1 to Step6 until all the c_v are predicted.

4 Experimental results

The experiments use a PC running Microsoft Windows 7,and the proposed AIN model is developed in C++. In addition, DT, SVM, logistic, NB, AIRS1 and immunos99 are used as comparative methods to make prediction performance comparisons with AIN. All are run in

Table 1 Detail of attributes

		Attributes	Description
Input	Bank client data	Age	The age of the client
		Job	Type of job
		Marital	Marital status
		Education	Educational status
		Default	Has credit in default?
		Housing	Has housing loan?
		Loan	Has personal loan?
	Last contact of the current campaign	Contact	Contact communication type
		Month	Last contact month of year
		Day-of-week	Last contact day of the week
		Duration	Last contact duration
	Previous data of contact-ing with the client	Campaign	Number of contacts performed during this campaign and for this client
		pdays	Number of days that passed by after the client was last contacted from a previous campaign
		previous	Number of contacts performed before this campaign and for this client
		poutcome	Outcome of the previous marketing campaign
	Social and economic context attributes	emp.var.rate	Employment variation rate - quarterly indicator
		cons.price.idx	Consumer price index - monthly indicator
		cons.conf.idx	Consumer confidence index - monthly indicator
		euribor3m	Euribor 3 month rate - daily indicator
		nr.employed	Number of employees- quarterly indicator
Output		y	Has the client subscribed a term deposit?

open source software WEKA (3.6.6). The cross-validation method will be used to evaluate the prediction efficiency (Table 1).

Table 2 Confusion matrix on subscribe a term deposit

		The actual situation	
		Subscribe	Not subscribe
Test results	Subscribe	TP	FP
	Not subscribe	FN	TN

Table 3 5-Fold cross validation method

	$Subset_1$	$Subset_2$	$Subset_3$	$Subset_4$	$Subset_5$
$Fold_1$	Testing	Training			
$Fold_2$	Training	Testing	Training		
$Fold_3$	Training		Testing	Training	
$Fold_4$	Training			Testing	Training
$Fold_5$	Training				Testing

4.1 Bank marketing data

This study considers real data collected from a Portuguese retail bank, from May 2008 to November 2010, with a total of 41,188 phone contacts. Each contact has 20 input variables and one output variable. But the dataset is unbalanced, as only 4640 (11.27

4.2 Measurement

This research presents accuracy, specificity, precision and recalls as the measurements (Stephen 1997), and a confusion Matrix is used to quantify these measurements. In a confusion matrix there are four different situations: if the client has subscribed the term deposit, it was also tested that he would subscribe it. This is called a true positive (*TP*). If the client has subscribed the term deposit, and the test result indicated non-subscription. This is called a false negative (*FN*). If the client hasn't subscribed the term deposit, and the test result indicated that he had, it is called a false positive (*FP*); else if the client hasn't subscribed the term deposit, and in fact, the test result indicated that he has not, it is called a true negative (*TN*). Table 2 shows the confusion matrix.

Accuracy: the proportion of true results in the population.

$$Accuracy = (TP + TN)/(TP + FP + FN + TN) \tag{3}$$

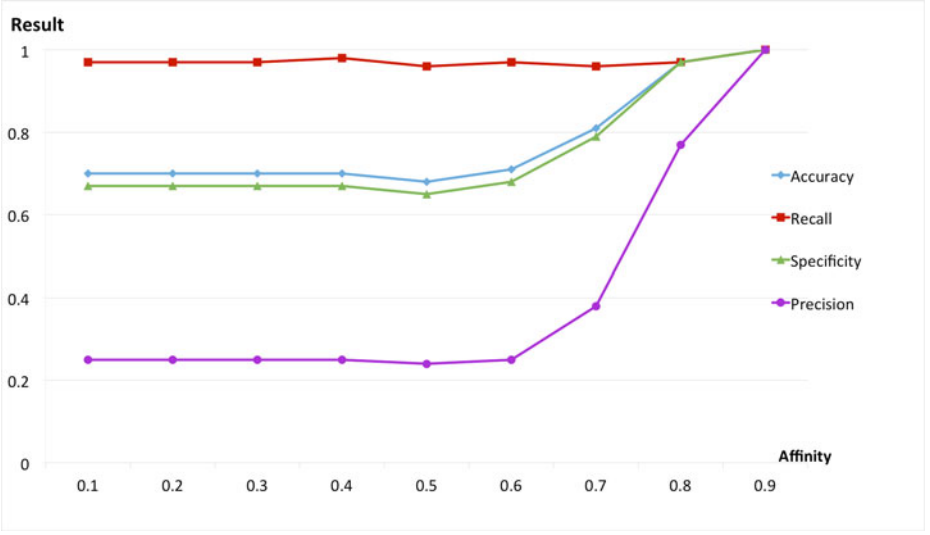


Fig. 6 Results for AIN on Different Affinity

Table 4 Results for AIN confusion matrix in 5-fold cross validation

	F1	*F2*	*F3*	*F4*	*F5*	*Average*
TP	160	193	309	655	1918	647
FN	99	254	173	257	622	281
TN	7952	7743	7708	7214	5639	7251.2
FP	26	47	47	111	61	58.4

Precision: The fraction of retrieved instances that is relevant. In this research, it means the correct rate of the selected customers who will subscribe.

Recall: The fraction of relevant instances that is retrieved. In this research, it means the proportion of the selected clients who will subscribe the term deposit in the population which will, in fact, purchase.

$$Recall = TP/(TP + FN) \tag{4}$$

Specificity: The proportion of negatives in a binary classification test which is correctly identified. In this research, it means the correct rate of the selected customers who will not subscribe.

$$Specificity = TN/(TN + FP) \tag{5}$$

This research uses a 5-fold cross validation method to evaluate the prediction efficiency. The original data set is randomly partitioned into five datasets, where the sizes of each dataset are equal. For the testing data and training data, one subset is used for testing, and the remaining four subsets are used for training, i.e. $subset_1$ for testing and $subset_{2,3,4,5}$ for training in $fold_1$. The cross-validation process is then repeated five times, with each of the five datasets. Table 3 shows the 5-fold cross validation method for training and testing data.

4.3 Experiments and results

In this section, two experiments are conducted. The first experiment is to use the original data as training data. The second experiment is to reselect the dataset using feature selection and extraction approaches.

4.4 Original data

Affinity is used as the threshold to construct antibodies in the immune network in this study. The values of the affinity are defined as (Hu 2005). If the threshold is set to a lower value, there will be more members in the immune network. Conversely, if the threshold is set to

Table 5 Results for AIN prediction performance in 5-fold cross validation

	F_1	F_2	F_3	F_4	F_5	*Average*
Accuracy	0.9848	0.9635	0.9733	0.9553	0.9171	0.9588
Recall	0.6178	0.4318	0.6411	0.7182	0.7551	0.6328
Specificity	0.9967	0.9940	0.9939	0.9848	0.9893	0.9917
Precision	0.8602	0.8042	0.8680	0.8551	0.9692	0.8713

Table 6 Comparison results with other methods

	Accuracy	Recall	Specificity	Precision
DT	0.8953	0.3478	0.9624	0.5430
LOGISTIC	0.8864	0.4268	0.9645	0.5334
NB	0.7836	0.468	0.7488	0.4644
SVM	0.7385	0.2682	0.8194	0.4176
AIRS1	0.7422	0.2172	0.8155	0.5162
IMMUNOS99	0.6489	0.4342	0.6522	0.3364
AIN	0.9588	0.6328	0.9917	0.8713

a higher value, there will be very few members of the immune network. To find the ideal threshold value, a value of affinity rages from 0.1 to 0.9. First, the experiment attempts to find the best affinity value by testing the training data with the 5-fold cross validation method. Fig. 6 shows the results for AIN with different affinity values. It shows that when the value is close to 0.9, the proposed approach has higher accuracy, recall, specificity and precision. 0.9 is thus used as the value of the affinity for testing the testing data.

As the testing data, F_1, F_2, F_3, and F_4 has 8237 samples, respectively, and F_5 has 8240 samples. It shows the result by confusion matrix in Table 4 with the affinity threshold 0.9. From Table 4, we can see *FN* and *FP* are very low. It means our prediction gives wrong results very rarely. And *TN* is much higher than *TP*, *FN* and *FP*. It means our recommendation model can highly predict the person who would not subscribe the term deposit.

However, using Eqs. (3–6), it is possible to calculate accuracy, recall, specificity and precision. This is shown in the calculations in Table 5.

Table 6 shows the comparison results of AIN with DT, SVM, logistic, NB, AIRS1 and immunos99. The cost parameter of DT (J48) is set to 0.1. The M parameter of logistic is set to 20. The gamma and cost parameters of SVM are set to 2^{-12} and 150. The affinity, clonal rate and k-nn parameter of AIRS1 are set to 0.9,10 and 3. The minimum fitness threshold parameter of immunos99 is set to -5.0. In addition, all methods have the best results in this instance. It is clear that AIN produces better accuracy, recall, specificity and precision,at 95.88.

Table 7 The attributes are selected by different methods

	IG	PCA
Attributes	Month	previous
	Duration	poutcome
	pdays	emp.var.rate
	previous	euribor3m
	poutcome	nr.employed
	emp.var.rate	
	cons.price.idx	
	cons.conf.idx	
	euribor3m	
	nr.employed	

Table 8 The comparison of different methods with feature selection

	Accuracy	Recall	Specificity	Precision
DT+IG	0.8968	0.5558	0.9146	0.3746
SVM+IG	0.8853	0.4779	0.8941	0.1074
AIN+IG	0.9620	0.6516	0.9926	0.8916
DT+PCA	0.8874	0	0.8801	0
SVM+PCA	0.7103	0.262	0.8793	0.1855
AIN+PCA	0.9857	0.8804	0.9960	0.9554

4.5 The training data with feature selection

This section uses the information gain (IG) from WEKA (3.6.6) to select ten attributes and five attributes by principal component analysis (PCA), shown in Table 7. The comparison result is shown in Table 8.

According to Table 8, the AIN proposed in this paper achieves higher performance than those of DT and SVM. If the results of Table 6 are compared with those of 8, the training data processed by feature selection can be used to enhance the performance of the proposed approach. The results of recall and precision also show that the proposed approach can avoid the class imbalance problem.

5 Conclusion

With the rapid development of social economies and information technology, more and more financial products are becoming part of our daily lives. In a big data environment, the ability to predict the preference of customers is very important to every financial service company.

This research presented an artificial immune classification model combining a collaborative filtering approach to recommend a term deposit to potential customers. This can be of great help for firms wanting to save time, effort and money.Despite the imbalanced data, the number of customers who will not subscribe term deposits is much larger than the number of customers who will, and the proposed approach can generate many classification rules by each training data set. In this way, it can avoid imbalanced class distribution. According to the experiment result, the proposed approach has better performance than those of DT, SVM, logistic, NB, AIRS1 and immunos99, especially in recall and precision index. In other words, the proposed approach could avoid the confusion of being overwhelmed by the majority class and ignoring the minority class. Thus, the prediction accuracy is very high in the proposed model.

Although the results presented here have demonstrated the effectiveness of the proposed approach, it could be further developed in two ways: One is to improve rating prediction by collaborative filtering incorporated with preference regularization. The other is to consider social media in this research. It is believed that these will improve the performance of the model described in this study. On the whole, the human immune system does a remarkable job of defending against disease-causing microorganisms. The idea of boosting real world mechanical systems is enticing, but the ability to do so has proved elusive for several reasons. The immune system is precisely a system, not a single entity. To function well, it requires balance and harmony. There is still much that researchers don't know about the

intricacies and interconnectedness of the immune response. It is hoped that this study, and its further work, will in some way contribute to the advancement of that knowledge.

References

Abdul, H., & Rahman, T.K.A. (2010). Short term load forecasting using an artificial neural network trained by artificial immune system learning algorithm Computer Modelling and Simulation (UKSim), 2010 12th International Conference on IEEE.

Alatas, B., & Akin, E. (2005). Mining fuzzy classification rules using an artificial immune system with boosting. *Lecture Notes in Computer Science*, *3631*, 283–293.

Aydin, I., Karakose, M., & Akin, E. (2012). An adaptive artificial immune system for fault classification. *Journal of Intelligent Manufacturing*, *23*(5), 1489–1499.

Cacheda, F., Carneiro, V., Fernández, D., & Formoso, V. (2011). Comparison of collaborative filtering algorithms: Limitations of current techniques and proposals for scalable, high-performance recommender systems. *ACM Transactions on the Web (TWEB) 5.1*, 2.

Chang, S.Y., & Yeh, T.Y. (2012). An artificial immune classifier for credit scoring analysis. *APPLIED SOFT COMPUTING*, 611–618.

Chen, M.H., Chang, P.C., & Lin, C.H. (2013). *A self-evolving artificial immune system II with T-cell and B- cell for permutation flow-shop problem* (Vol. 25, pp. 1257–1270). New York: Springer Science & Business Media.

Chen, Z.Y., Fan, Z.P., & Sun, M. (2012). A hierarchical multiple kernel support vector machine for customer churn prediction using longitudinal behavioral data. *European Journal of operational research*, *223*(2), 461–472.

Dai, H.L. (2015). Class imbalance learning via a fuzzy total margin based support vector machine. *APPLIED SOFT COMPUTING*, 172–184.

Dudek, G. (2012). An artificial immune system for classification with local feature selection. *IEEE Transaction on Evolutionary Computation*, *16*(6), 847–860.

Euerby, M.R., & Patrik, P. (2003). Chromatographic classification and comparison of commercially available reversed-phase liquid chromatographic columns using principal component analysis. *Journal of Chromatography A 994.1*, 13–36.

Farquad, M.A.H., Ravi, V., & Raju, S.B. (2014). Churn prediction using comprehensible support vector machine: an analytical CRM application. *Applied Soft Computing*, *19*, 31–40.

Gadi, M.F.A., Wang, X.D., & Lago, A.P.D. (2008). Credit card fraud detection with artificial immune system. *Artificial Immune Systems*, *5132*, 119–131.

Gong, S., Ye, H., & Tan, H. (2009). Combining memory-based and model-based collaborative filtering in recommender system. *Pacific-Asia Conference on Circuits, Communications and System*, 690–693.

Hemamalini, S., & Sishaj, P.S. (2011). Dynamic economic dispatch using artificial immune system for units with valve-point effect. *International Journal of Electrical Power & Energy Systems 33*, *4*, 868–874.

Hu, X.H. (2005). A data mining approach for retailing bankcustomer attrition analysis. *Applied Intelligence*, *22*(1), 47–60.

Jerne, N.K. (1974). Towards a network theory of the immune system. *Collect Ann Institut Pasteur, 125C*, *1-2*, 373–389.

Knight, T., & Timmis, J. (2001). AINE: An immunological approach to data mining. *Proceedings 2001 IEEE International Conference on Data Mining*, *69*, 297–304.

Lee, C., & Gary, G.L. (2006). Information gain and divergence-based feature selection for machine learning-based text categorization. *Information processing & management 42.1*, 155–165.

Liu, H.F., Hu, Z., Mian, A., Tian, H., & Zhu, X.Z. (2014). A new user similarity model to improve the accuracy of collaborative filtering. *Knowledge-Based Systems*, *56*, 156–166.

Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, *62*, 22–31.

Nasir, A.N.M., Selamat, A., & Selamat, H. (2009). An artificial immune system for recommending relevant information through political weblog. *Proceedings of iiWAS2009*, 420–424.

Nie, G., Rowe, W., Zhang, L., Tian, Y.G., & Shi, Y. (2011). Credit card churn forecasting by logistic regression and decision tree. *Expert System with Applications*, *38*(12), 15273–15285.

Pang, W., & Coghill, G.M. (2015). QML-ainet: An immune network approach to learning qualitative differential equation models. *APPLIED SOFT COMPUTING*, 148–157.

Ramezani, M., Moradi, P., & Akhlaghian, F. (2014). A pattern mining approach to enhance the accuracy of collaborative filtering in sparse data domains. *Statistical Mechanics and its Applications*, *408*, 72-84.

Sato, T. (1994). Application of principal-component analysis on near-infrared spectroscopic data of vegetable oils for their classification. *Journal of the American Oil Chemists' Society 71.3*, 293–298.

Sengur, A. (2009). Prediction of protein cellular localization sites using a hybrid method based on artificial immune system and fuzzy k-NN algorithm. *Digital Signal Processing*, *19*(5), 815–826.

Stephen, V.S. (1997). Selecting and interpreting measures of thematic classification accuracy. *Remote Sensing of Environment*, *62*(1), 77–89.

Su, P., & Ye, H. (2009). An item based collaborative filtering recommendation algorithm using rough set prediction. *International Joint Conference on Artificial Intelligence*, 308–311.

Suliman, S.I., & Rahman, T.K.A. (2010). Artificial immune system based machine learning for voltage stability prediction in power system. Power Engineering and Optimization Conference (PEOCO), 2010 4th International IEEE.

Taylor, P., Polack, F.A.C., & Timmis, J. (2013). Accelerating immunos 99. *Artificial Immune System – ICARIS*, 893–898.

Tipping, M.E., & Christopher, M.B. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 61*, *3*, 611–622.

Uğuz, H. (2011). A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm. *Knowledge-Based Systems 24.7*, 1024–1032.

Wu, J.Y. (2010). Forecasting Chaotic Time Series Using an Artificial Immune System Algorithm-based BPNN, Technologies and Applications of Artificial Intelligence (TAAI), 2010 International Conference on IEEE.

Xia, M., zeshui, X., & Na, C. (2013). Some hesitant fuzzy aggregation operators with their application in group decision making. *Group Decision and Negotiation 22.2*, 259–279.