Mach Learn (2015) 101:85–103
DOI 10.1007/s10994-014-5447-y

Measuring the accuracy of currency crisis prediction with combined classifiers in designing early warning system

Nor Azuana Ramli · Mohd Tahir Ismail · Hooy Chee Wooi

Received: 16 December 2013 / Accepted: 12 May 2014 / Published online: 19 June 2014

Abstract Is the prediction accuracy affected by the method used in the ensemble of the classifiers? This paper is a sequel of our experiment in order to find an answer for such question. Previously, we had conducted an experiment by using single classifiers in the machine learning against traditional statistical methods. The results showed that single classifiers in machine learning perform well compared to the traditional statistical methods. Still, we believe that there is another way to increase the prediction accuracy of these classifiers. In this paper, we conducted another experiment by combining these classifiers in predicting currency crisis of 25 countries. The combined classifiers are support vector machine with k-nearest neighbor, logistic regression with k-nearest neighbor and finally LADTree with k-nearest neighbor. These three combined classifiers are tested on 13 chosen macroeconomic indicators which the data is taken from first quarter 1980 to third quarter 2012. The results of this experiment showed that these three different combined classifiers averagely have same higher accuracy and quite comparable. Our proposed method, nearest neighbor tree has the highest area under ROC curve number among these three combined classifiers although in terms of computational time it took longer running times than the others.

Keywords Machine learning · Combined classifiers · Currency crisis · Early warning system · k-Nearest neighbor method

Editors: Vadim Strijov, Richard Weber, Gerhard-Wilhelm Weber, and Süreyya Ozogur Akyüz.

N. A. Ramli (✉) · M. T. Ismail
School of Mathematical Sciences, Universiti Sains Malaysia, Penang, Malaysia
e-mail: ajue.ramli@gmail.com

M. T. Ismail
e-mail: mtahir@cs.usm.my

H. C. Wooi
School of Management, Universiti Sains Malaysia, Penang, Malaysia
e-mail: cwhooy@usm.my

1 Introduction

Currency crisis is like a never ending episode in the economics story. It is one of the financial crisis but different from other crises like debt and banking crisis. Currency crisis also has another name which is balance-of-payments crisis. It usually occurs when there is an unexpected devaluation of currency that regularly ends with a speculative attack on the foreign exchange market. A currency crisis also happened due to never-ending balance-of-payments deficits or when the government is unable to pick up the currency value of its country from the market speculation.

The first episode of the currency crisis happened in the 1990s was back in 1992 where most of European countries were facing Exchange-Rate Mechanism (ERM) crisis. Next, the episode was about the huge crisis that happened in Mexico once peso folded at the end of December 1994. The crisis started when a massive decision to devalue their currency by 15 percent was made by the Mexican government which then caused the peso to crash down within a few days, dragging this crisis into a bigger crisis where at that time real gross national product (GNP) per capita had fallen to 9.2 %, average causing manufacturing wages to fall 21 % and the unemployment rate increase to 7.6 % compared to a year before when the rate was just 3.2 %. The effect of the crises that happened to Mexico and the European countries only can be seen in some regions, while most currency crises after that have a bigger impact where it affects the whole world economy.

The biggest crisis was the Asian Financial Crisis which happened in 1997. This crisis was the first to prove that crisis can be contagious to other countries. Back in third quarter of 1997, this crisis began with the violent devaluation of the Thai Baht. There were lots of opinions and theories regarding this crisis. Some economists categorized this crisis as a financial crisis since it was not just about a speculation on Thai currencies but also crash on stock markets that happened in South Korea. Eventually, this crisis had caused several South East Asian countries to collapse in their economic growth rates. The financial crisis which started with currency devaluations from Thai followed by Malaysia, South Korea, Indonesia, the Philippines, Singapore and Taiwan, causing interest rates to rise sharply and lots of companies to declare their bankruptcy as the increasing cost of borrowing. As a result, foreign and domestic investors had withdrawn their sponsors. These countries were not just experiencing a collapse in the level of economic activity but also the number of bankruptcies and escalated level of private sector debt.

This financial crisis that had occurred in South East Asian then had spread further to regions such as China, Russia and Brazil. The financial effect on countries outside South East Asia was the chief initial concern of governments outside the region. This crisis does not just gave bad impact on emerging economies such as those in Eastern Europe and Latin America, but it also made the real sector of the western industrialized countries to suffer since trade flows and foreign direct investment had been decreasing. The good thing out of this is we can always learn from history. The previous crises had opened our eyes to see the signs of crisis in advance so that we can avoid the crisis before it occurs or becomes worse. Therefore, we need a tool to deal with the probability of any crisis occurring in the future like an early warning system since crisis cost is very high based on its effect to the percentage of unemployment, economic reduction and requirement to restructure financial process.

An early warning is a system that indicates an alarm whenever a measurement exceeds the threshold. There are various types of early warning systems and its application is very wide in any field and crises since it had been developed. An initiative to build this model not only comes from researchers in an academic field, but also from the national government, private sectors, non-governmental organization (NGO) and other various organizations. They have

built more or less a hundred models from various methodologies including statistical methods until the latest method which is machine learning to detect worldwide crises. However, the application of early warning systems in preventing economics, financial or currency crisis are quite behind to be compared to others such as natural disasters, the spread of diseases and even in business. Therefore, we are driven to find not just a suitable method but also a method that can predict a crisis accurately, so that we can use it in modeling our early warning system later.

Similar to other paper, before we introduce all the combined classifiers in Sect. 3, a short summary on previous research involving early warning system for currency crisis will be written consecutively in Sect. 2. Then, the results of the combined classifiers are presented in Sect. 5. The selection of indicators, sample and data and also performance evaluation for each combined classifier will be clarified in Sect. 4. In the last part, we will be presenting the conclusion of the experimental study and a proposal that we have which hopefully can be done for future research.

2 Literature review

There are only two models for early warning systems in predicting currency crisis. The first one is the theoretical models which are based on the economics theory to predict crisis which comprise of three different theories. Next, we have empirical models that had been developed actively until today since researchers found that it is not sufficient to depend on just theory to predict the upcoming crisis. Since the paper focuses more on empirical than the theoretical models, the historical part on empirical models will be enlighten more.

Most of these empirical models aimed to predict crises by assessing their potential economic and financial indicators. So that, policymakers can use these models to prevent future crises by detecting the causes earlier. There have been numerous studies in the literature on the leading indicators of currency crisis. The two famous models are signaling approach developed by Kaminsky et al. (1998) and the probit or logit methods suggested by Frankel and Rose (1996) and Eichengreen et al. (1996).

Kaminsky et al. (1998) tested 15 macroeconomic indicators on the signaling approach method by optimizing the estimated threshold for each country. The aim of signaling approach is to maximize the correct signal and minimize the false alarm. This can be achieved by setting signal horizon at 24 months and defined currency crisis as a sharp depreciation of the currency or a huge drop off in international reserves. From this study, they discovered that the indicators that have shown such performance in predicting crises are output, real effective exchange rate, exports, ratio of broad money to gross international reserves, and equity prices. These indicators provide signals in advance so that preemptive policy measures can be done.

Kaminsky and Reinhart (1999) put forward the analysis by constructing leading composite indicators as a weighted sum of the signaling indicators, where each indicator is weighted by the inverse of its noise-to-signal ratio. These composite indicators provide some information on the vulnerability of an economy in an upcoming crisis. However, this kind of approach did have some weaknesses like it is unable to apply with any of standard statistical evaluation methods such as the significance tests. Soon after, study on early warning by using replication of Kaminsky et al. (1998) results had been done by Edison (2003). In her research, besides expanding the number of indicators and the country coverage, she also made an observation on regional differences and compared the existing algorithm. By using data sets consisting of 14 indicators for 20 developing and industrial countries from 1970–1998, the results of the study showed that the performance of the model was robust to various sensitivity tests

and helped in identification of vulnerabilities. The downside of the research was the model gave too many false alarms making it a failed reference of early warning systems.

Next to signaling approach is probit and logit models which had been used extensively in previous research. Probit and logit regressions are limited dependent variable models which means both of these models can be used to identify the causes of crises. Similar to the signaling approach, these methods also model the currency crisis indicators as zero-one variables. The difference between these models with signaling approach is the chosen indicators enter the model in a linear fashion. When a crisis occurs, the value is equal to one and contradicts when a crisis does not occur.

Sachs et al. (1996) used this probit model and their analysis was based on 20 emerging countries that were vulnerable to the contagion effect after the 1994 Mexican crisis. They applied the weighted summation of the percentage decreasing in reserves and the percentage depreciation of the exchange rate from November 1994 to April 1995 as their crisis index. In their results, they discovered that short-term capital inflows is not important when other variables such as reserves and fundamentals are strong at the same time as government consumption and current account deficits only matter in the countries with weak fundamentals and weak reserves.

Berg and Patillo (1999) tested previous models to observe whether these models have a good predictability on the Asian financial crisis. From their findings, the models developed by Sachs et al. (1996) and Frankel and Rose (1996) were vain in predicting the crisis. On the other side, models by Kaminsky et al. (1998) showed successful results. The probabilities for the occurrence of a crisis produced by using the signaling approach for the episode between mid 1995 and end 1996 were statistically significant over the following 24 months. Moreover, the prediction results for cross-country ranking of crisis severity provided by this model are a significant predictor of the actual ranking.

Due to poor prediction performance, several methodological issues had been raised by some researchers that then led us to the diversity of innovation models in developing an early warning system. Abiad (2003) used Markov-switching models with time-varying transition probabilities as an early warning system for currency crisis. In his study, the Markov-switching model had been estimated by using monthly data from 1974 to 1998 for the four Southeast Asian countries which are Indonesia, Malaysia, Thailand and the Philippines. The results of his findings proved that this method was quite successful in identifying crisis even with only a few chosen indicators. It also gave some warning about the occurrence of the crisis.

Peltonen (2006) was the first one that proposed multilayer perceptron artificial neural network (ANN) as a new method in developing early warning systems to forecast currency crisis. In his study, he used 13 variables for independent variables and 2 variables for crisis index and the data were taken from 1980–2001. A comparison with a probit model was made to determine which model has better predictability. Based on the results from his study, both models were able to signal in-sample correctly and it also found a better explanation on contagion effect. However, the ability of the propose model for signal currency crisis out-of-sample was found to be weak since only 1 out of 24 samples got correctly called.

There are many other researches that had been done by using empirical models as the list goes on. Results from all these studies suggested that popular methods in predicting currency crisis such as signaling approach, probit, logit and ordinary least squares were all well perform, with some variation when market fundamentals are explosive. However, when the situation is changed, all the methods performed quite badly. Therefore, in our study, we are trying to experiment with new method which involved some of the models from machine learning system. Since we already tested single classifiers on the previous research, we are

hoping by experimenting on combined classifiers, the results that we get will be improved as the combination of two or more classifiers never been applied in modeling early warning system to detect currency crisis.

3 Methodology

A combination of two or more base classifiers, also known as an ensemble of classifiers. By combining two of base classifiers, we could get higher prediction accuracy since the individual outputs are combined in some way to classify new predictions. It was proven in previous study by Dietterich (1997) and Gams et al. (1994) that the combination of two different base classifiers has higher accuracy than the individual classifier. This is because different classifiers will look at the same problem from different points of view. There are three different ways to combine two or more classifiers which are voting (bagging and boosting), cascading and stacking. In our experiment, we chose to use an approach where different types of classifiers were combined in order to get a new classifier that has higher accuracy. These two different base classifiers were combined by using a method called stacking.

3.1 Stacking of models

Stacking is just another method in the ensembles and it is used to combine two or more classifiers generated by using different base classifiers, $L_1, \ldots, L_N$ and a meta-level classifier on any data sets, D, which consists of examples $d_i = (x_i, y_i)$, where x_i is a pair of feature vector and y_i is its classification. It was introduced by Wolpert (1992). Even though voting is a more famous method than stacking when it comes to combine classifiers in machine learning, this study would like to use a different approach than other studies that had been conducted. The difference between stacking with other global technique in ensembles such as voting is that the ensemble by using stacking need not require the base classifiers to be linear since it learned through a combiner system. Furthermore, there are two other advantages of using stacking which is trained rule is more flexible and less bias plus there is no need to normalize classifier outputs.

Unlike voting, stacking has two phases in their system wherein the first phase, a set of base-level classifiers is generated. Then, a meta-level classifier is learned to combine the outputs from the base-level classifiers in the second phase. In stacking, the combiner model cannot be trained by using training data since the base-level classifiers possibly memorized the training set. That is why stacking estimates and corrects whenever there is any bias in the base-level classifier. Usually, the unused data are employed to train the ensemble of classifiers by using stacking. The complete architecture of the combiner model by using stacking is as shown in Fig. 1.

3.2 An ensemble of support vector machine classifiers

The k-nearest neighbor classifier (k-NN) which is also known as lazy learning because of its training is held up to run time is one of the most straightforward and simplest classifier in machine learning. This is due to the classification of the data set is based on its nearest neighbors class. Since in the ensembles the base classifiers are chosen for their simplicity, k-NN suits well to take a place as one of the base classifiers. Additionally, k-NN had outperforms the other three classifiers which are support vector machine, neural network and logistic

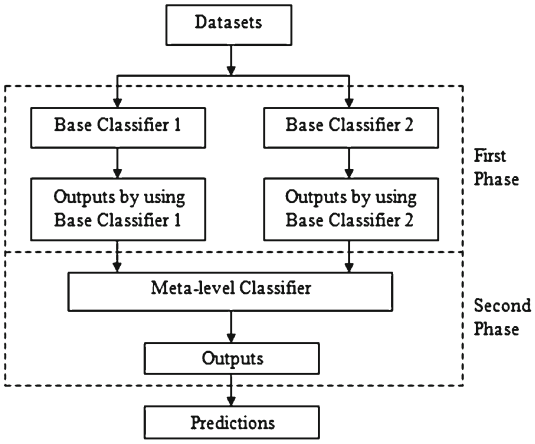


Fig. 1 The architecture of the ensemble of two different classifiers by using stacking

regression from the previous study that we have done (Ramli et al. 2013). In this study, we used an ensemble of classifiers called the kNN–SVM ensemble classifiers.

There are a few researches that have already tested support vector machine (SVM) ensembles as an experiment but most of them were in biomedical application. Liqi et al. (2011) had proposed kNN–SVM ensemble in their study to predict proteins from gene ontology. However, in their study they combined these two classifiers by using voting system whereas in our study, we combined both classifiers through stacking.

Besides the method of combining two different classifiers, the key to the formulation of a powerful ensemble model also depends on the selection of parameters. For k-NN classifier, we used the settings which are $k = 4$ and Manhattan distance based on the previous results that we got. For support vector machine, the optimization problem is extended since we have high dimensional data set and we need more than a simple linear classifier to classify the data which is not possible to be separated linearly (Gunn 1998). By adding non-negative slack variable, ξ_i in the equation below

$$y_i\,[(w.x) + b - 1] \geq 0, i = 1, 2, \ldots, n \tag{1}$$

where y_i is the class for crisis, x is the set of points, w is the normal vector to the hyperplane, b is the bias and n is the number of points; then, Eq. 1 becomes

$$y_i\,[(w.x) + b] \geq 1 - \xi_i, i = 1, 2, \ldots, n \tag{2}$$

Then, we got a new optimization problem which is

$$\min \left\{ \frac{1}{2} \parallel w \parallel^2 + C \left| \sum_{i=1}^{n} \xi_i \right| \right\} \tag{3}$$

subject to any $i = 1, 2, \ldots, n$, where $C > 0$ is the parameter that will determine the values of error penalty due to misclassification of the data and the value of C is defined by user. The value of C controls the tradeoff between margin maximization and error minimization. Large value of C gives solutions with less misclassification errors but a smaller margin while small value of C gives solutions with bigger margin and more classification errors.

Yet, it is still incomplete to deal with nonlinear complex system. That is why an inner product function which is also known as kernel function is needed in order to transform the input space into a high dimensional space by an inner product defined nonlinear trans-

form function. Many kernel mapping functions can be used in support vector machine. The generally used kernel functions are:

- linear: $K(x, x') = \langle x, x' \rangle$
- polynomial: $K(x, x') = (\gamma \langle x, x' \rangle + r)^d, \gamma > 0$
- radial basis function (RBF): $K(x, x') = e^{-\gamma \|x - x'\|^2}, \gamma > 0$
- sigmoid: $K(x, x') = \tanh(\gamma \langle x, x' \rangle + r)$.

For certain parameters, the linear kernel is a special case of RBF kernels while the sigmoid kernel behaves like the RBF kernel. When the data are linearly inseparable, a non-linear kernel that maps the data into the feature space non-linearly can handle the data better than the linear kernels. As the polynomial kernel requires more parameters to be chosen, the RBF kernel is a reasonable first choice of kernel function. When using the RBF kernel, the parameters like C and γ have to be decided. For the value of C (or γ), a possible interval can be provided with the grid space. All grid points of C and γ are tested to find the one giving the highest CV accuracy. By using library for support vector machine (LIBSVM) through WEKA, the best C and γ values that produced the highest CV accuracy are $C = 16$ and $\gamma = 3.5$. These values will later be used to train the whole training set and generate the final model. More advanced parameter selection methods are not consider to be used in this study since the number of grid points is not too large for two parameters only.

To analyze the prediction of the currency crisis and the accuracy of the ensemble of classifiers, we used the data sets and run the analyses by using the latest version of WEKA. WEKA is a well-known collection of machine learning software written in Java. As stated earlier in this methodology section, stacking system has two phases (refer to Fig. 1). In the first phase, k-NN and SVM had been chosen as our base classifiers. Therefore, we decided to use Linear Regression as a meta-level classifier for the second phase. For the cross-validation value, it is depending on the framework but the bigger the value the better the results user will get. Here, ten fold cross validation is used since it provides results close to the optimal ones based on the Zhang's (1993) findings. In ten fold cross validation, 10 % of the data from the whole data set are chosen randomly as a test set while the remaining 90 % are used as training set. The performances of kNN–SVM ensemble by stacking will be compared with single SVM classifier and kNN–SVM ensemble by a voting system.

3.3 Logistic regression ensembles (LORENS)

Lim et al. (2010) had introduced Logistic Regression Ensembles which is also known as LORENS in short. LORENS works by classifying binary responses based on high-dimensional data and in this ensemble, the logistic regression model (logit) is used as a base classifier. In their study, they combined the results from multiple logit models by taking the average of predicted values within an ensemble to achieve a higher accuracy. However, we made a little bit of modification in our study by ensemble logit model with k-nearest neighbor method instead of combined multiple logit models.

Generally, logit is a uni or multivariate technique which allows for estimating the probability that an event occurs or not, by predicting a binary dependent outcome from a set of independent variables. In our case, we involved the occurrence of the crisis where the dependent variable is whether crisis occurs or not in a relation to macroeconomic indicators. The linear probability model described by Wooldridge (2010),

$$\mathrm{P}_i = \mathrm{E}(Y = 1|X_i) = \beta_1 + \beta_2 X_1 + \cdots + \beta_n X_i \tag{4}$$

where X_i is the indicators and $Y = 1$ means that there is a crisis occurs. The probability of the occurrence of the crisis also can be written as:

$$\mathrm{P}_i = \frac{1}{1 + e^{-(\beta_1 + \beta_2 X_1 + \cdots + \beta_n X_i)}} \tag{5}$$

This equation is also known as the cumulative logistic distribution function.

Same as the ensemble of SVM classifiers, we also used WEKA to run this ensemble. Ten fold cross validation (CV) was conducted for each data set. This model works by training the data set using logit model since logit works as base-level classifier in this model and at the final stage, the outputs from both models will be combined to produce predictions of currency crisis. For this model, we experimented by only using one base classifier which means k-NN is used as a meta-level classifier.

3.4 Nearest neighbor tree (NNT)

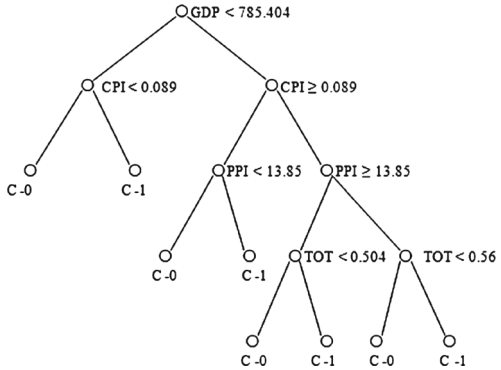
Decision tree is one of the most famous predictor used in machine learning and it is frequently used as a base classifier in constructing an ensemble of classifiers. In the decision tree algorithm, the approximated target function is represented as a tree-like structure. Generally, it works by sorting down the tree branch from the root to some leaf nodes. Each internal node represents a specific test of instance attribute and each branch represents one of the possible test results. Decision tree is an efficient method and it is also interpretable as it can cover the disadvantage of an ensemble of classifiers. However, this algorithm would probably produce the low accuracy and high variance which means that it is at its best when it performs in ensembles rather than as a single classifier.

Bagging and boosting are two popular methods and mostly used in researches involving comparison of accuracy between different combined classifiers where the base classifier is decision tree. Both of the methods are well established procedures in improving the performance of classification algorithms. That's why it is quite hard to find previous study on ensembles in decision tree that used stacking. Therefore, we have decided to use stacking instead of bagging or boosting and experiment it on our 13 macroeconomics data.

In our proposed methodology, we chose LADTree (Holmes et al. 2002) which is a method that is available in WEKA under decision tree to be one of our base-level classifiers. The LADTree learning algorithm is a class for generating a multi-class alternating decision tree using the logistic boosting algorithm. In this decision tree, a single attribute test is chosen as the splitter node for the tree at each iteration. Figure 2 shows the classification tree by using LADTree in WEKA for the macroeconomic data. This experiment is then continued with the combination of outputs from LADTree with k-NN to the second phase where Linear Regression combined these two classifiers and produced a linear equation in order to find the probability of crisis. The results of this nearest neighbor tree will be compared to boosting and bagging method and will be discussed it further in Sect. 5.

4 Design and analysis

Before we started our experiment, these three things had to be in our checklist: (1) data sets consisting of macroeconomic indicators, countries sample and years of the event, (2) how the data will be collected, and finally (3) how do we analyze it.



Note:

GDP – Gross Domestic Product
CPI – Consumer Price Index
PPI – Producer Price Index
TOT – Terms of Trade
C0 – No Crisis
C1 – Crisis

Fig. 2 The classification tree that is tested on macroeconomic data for Chile by using LADTree

Table 1 The list of selected indicators and its detail

Indicator	Details
Real effective exchange rate	Not seasonally adjusted, in USD, quarterly
Unemployment rate	Rate or quantity series, seasonally adjusted, annually
Exports of goods and services	At current prices (nominal), in USD, quarterly
Imports of goods and services	At current prices (nominal), in USD, quarterly
Foreign direct investment	At current, quarterly
M2 money multiplier	At current, in USD, quarterly
Consumer price index	Not seasonally adjusted, in USD, quarterly
Foreign exchange reserves	At current, in USD, quarterly
General government final consumption	At current, in USD, quarterly
Industrial production index	Seasonally adjusted, in USD, quarterly
Producer price index	Not seasonally adjusted, in USD, quarterly
Gross domestic product per capita	At current, in USD, quarterly
Terms of trade	Not seasonally adjusted, in USD, monthly

4.1 Selection of the indicators

In developing early warning systems, suitable indicators need to be chosen. We have selected 13 macroeconomic indicators in this study and the details about the data set for each indicator are as shown in Table 1. Those 13 indicators had been chosen based on the availability of the data, previous literatures and economics point of view. Amongst all indicators, real effective exchange rate seems to play a very significant role since it is related to definition of currency crisis itself and currency devaluation depends on the exchange rate regime in place. The two indicators which are exports and gross domestic product (GDP) were chosen based on Kaminsky et al. (1998) findings since these macroeconomic variables play quite a significant role as individual leading indicators. From the results of the study, they found that these indicator issued at least one signal 24 months prior to a crisis.

Foreign direct investment, government consumption and foreign exchange reserves had been added in the list based on their (Kaminsky et al. 2000) latest study in 2000. Other indicators such as imports, terms of trade and M2 multiplier were selected based on second

generation models from Obstfeld (1994). Indicators like consumer price index, producer price index and industrial production index are red-hot economic indicators along with the unemployment rate. Red-hot economic indicators refer to the indicators that are getting great attention in the financial markets. Consumer price index, producer price index and industrial production index are related to each other. These three indicators are important indicators because inflation affects everyone. It will start with industrial production index and then producer price index to consumer price index. Producer price index measures changes in prices that manufactures and wholesalers pay for goods during various stages of production. If business owners have to pay more for goods then consumers have to pay with much higher costs. Consumer price index determines how much consumers pay for goods and services, which definitely will affect the cost of doing business and causes chaos to personal and corporate investments and lastly will affect the retirees' quality of life.

Unemployment rate is a significant indicator in predicting currency crisis. It was used by Krugman (2000) in his study and he stated that even though unemployment is not the main factor of currency crisis, it played a role along with other more traditional fundamentals. His study concluded that heightened devaluation expectations can in turn increase unemployment by adding a devaluation premium to interest rates. It is easy to see how an external shock can have a major impact through these feedback and the effects on both unemployment and the stability of the currency. All of these variables are the same where the data are index based except for unemployment where the type of data used in this study is rate based. Because of that, we have to standardize the data by using standardization. This technique will transform the data by using the equation below:

$$x_{new} = \frac{x - \mu}{\sigma} \tag{6}$$

where μ represents the mean and σ represents the standard deviation. By using this technique, we assumed that our data have been generated with a Gaussian law. All of these data, which are definitely quantitative or numerical data will be our input to be mapped into a categorical output where we take yes (crisis occurs) or no (no crisis occurs) as results.

4.2 Sample and data

We tested our three different ensembles of classifiers on 25 countries and the period taken were from first quarter 1980 until third quarter 2012. All of the data for macroeconomic indicators were downloaded under analysis via DataStream. We chose quarterly data instead of annually and monthly because certain countries' data for some indicators are only available for the quarter and above. Even though it's not a problem to access data annually, somehow there are weaknesses such as annual data make the prediction values of leading indicators less accurate. By using data annually, it also makes the precision of crisis time unclear since we could not really know whether the crisis takes place at the beginning of the year or the end. For example, the currency crisis for Asian that started with Thailand had happened on third quarter in 1997. If we took the data as annually, the crisis could be assumed to happen earlier or later than that in 1997.

4.3 Data analysis

A test on the data needs to be run before we perform any experiment on the ensembles of classifiers. To do so, we analyzed the correlation between the selected 13 macroeconomics indicators to check if there is any multicollinearity. Multicollinearity is a statistical phe-

Table 2 Collinearity results when consumer price index taken as dependent variable

Indicator	Tolerance	VIF
Exports	0.025	**40.187**
Foreign direct investment	0.792	1.263
Foreign exchange reserves	0.105	9.486
Government consumption	0.045	**22.303**
Imports	0.092	10.92
Industrial production index	0.013	**76.271**
M2 money multiplier	0.023	**42.79**
Real effective exchange rate	0.012	**85.518**
Gross domestic product	0.023	**42.757**
Terms of trade	0.068	**14.706**
Unemployment	0.126	7.908
Producer price index	0.009	**115.197**

The bold refer to the VIF values more than 10 which indicate that the variables have higher correlation

nomenon where two or more predictor variables in a multiple regression model are highly correlated which means one can be linearly predicted from the others with a non-trivial degree of accuracy. We use a formal detection-tolerance or the variance inflation factor (VIF) in this study to check multicollinearity;

$$\text{tolerance} = 1 - R_j^2, \ \text{VIF} = \frac{1}{\text{tolerance}} \tag{7}$$

where R_j^2 is the coefficient of determination of a regression of explanatory j on all the other explanators. A tolerance of less than 0.20 or a VIF of 10 and above indicates a multicollinearity problem. By using SPSS software, we found that there is multicollinearity in our data since the VIF values are >10. Tables 2 and 3 showed the results that we got from SPSS even when we kept repeating the analysis by changing the independent variables. There are a few solutions to solve multicollinearity such as by dropping one of the variables or obtain more data as that is the preferred solution. Both of the solutions are inapplicable in our study since we will lose information because one variable had been dropped and the data used in this study are quite limited. Thus, we chose to standardize the independent variables.

4.4 Performance evaluation

For classification, especially for two-class problems like in our case, a variety of measures can be proposed. The first would be by using false alarm (type II) where in the statistical pattern recognition it is known as a false positive. This false alarm can be computed based on a confusion matrix that we got from results by using WEKA. There are four possible cases in the confusion matrix which is shown in Table 4. For a crisis that is predicted to be occurred, if the prediction is also there and a crisis occurring, that means the prediction is true (which means, if a signal given for crisis, then that signal is a correct signal) . If we predict the crisis to occur but no crisis occurs that indicates that our prediction is false positive on the contrary (which means if a signal given for a crisis, then that signal is a false alarm).

Percentage of accuracy is quite easy to calculate. It can be measured in two different ways. First is by using formula; percentage of accuracy = [(Correctly predicted data)/(Total testing data)]100. Another way is by using the results that we obtained from confusion matrix

Table 3 Collinearity results when exports taken as dependent variable

Indicator	Tolerance	VIF
Exports	0.776	1.289
Foreign direct investment	0.083	**12.107**
Foreign exchange reserves	0.044	**22.824**
Government consumption	0.123	8.153
Imports	0.017	**60.122**
Industrial production index	0.025	**40.705**
M2 money multiplier	0.012	**82.428**
Real effective exchange rate	0.032	**31.647**
Gross domestic product	0.066	**15.193**
Terms of trade	0.114	8.744
Unemployment	0.001	**772.194**
Producer price index	0.001	**756.812**

The bold refer to the VIF values more than 10 which indicate that the variables have higher correlation

Table 4 Confusion matrix for two classes

Crisis	Crisis prediction		Total
	Yes	No	
Yes	*tp*: true positive	*fn*: false negative	p
No	*fp*: false positive	*tn*: true negative	n
Total	p'	n'	N

and then using the formula; percentage of accuracy = (1 − error) 100, where error = total false positive and false negative/ total number. Besides false alarm and the percentage of accuracy, the confusion matrix gives rise to a number of graphs that can be used to assess the relative utility of a model like the ROC curve. ROC curves have long been used in machine learning and signal detection theory to describe the tradeoff between the false positive value and the sensitivity value (Fawcett 2006). It is usually used to measure the prediction accuracy of a model. The graph is based in four conditional frequencies that can be derived from a model and the choice of a cut-off point for its scores:

- the observations predicted as crisis and effectively such (sensitivity)
- the observations predicted as crisis and effectively no crisis
- the observations predicted as no crisis and effectively crisis
- the observations predicted as no crisis and effectively such (specificity)

The ROC curve is obtained representing for any fixed cut-off value, a point in the Cartesian plane having as x-value the false positive value (1-specifity) and y-value the sensitivity value. Each point in the curve corresponds therefore to a particular cut-off. In terms of model comparison, the best curve is the one that is leftmost, the ideal one coinciding with the y-axis.

However, the area under the ROC curve (abbreviated as AUC) is a more convenient way to compare classifiers since it represents numbers than figures. AUC gained importance in the classification community as a mean to compare the performance of classifiers because most classification methods do not optimize this measure directly. Ling et al. (2003) in their study had done a comparison amongst classifiers by using AUC and found that AUC should replace accuracy in measuring classification systems. Bradley (1997) has found the same results in

Table 5 Classification for the accuracy of the AUC

Area under the ROC Curve	Classification
0.90–1.00	Excellent
0.80–0.90	Good
0.70–0.80	Fair
0.60–0.70	Poor
0.50–0.60	Fail

Source: Thomas G. Tape (http://gim.unmc.edu/dxtests/ROC3.htm)

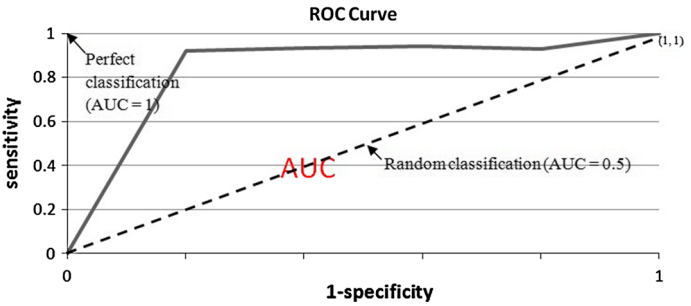


Fig. 3 An example of the area under the ROC curve (AUC) with explanation

his study and stated that AUC has increased sensitivity in Analysis of Variance (ANOVA) tests, is independent to the decision threshold and is invariant to a priori class probability distributions. That is why AUC is a useful metric for classifier performance. A guideline on classifying the accuracy for the AUC is as shown in Table 5 which is taken from Tape's notes that available online. The area under the curve which is as can be seen in Fig. 3 is used to determine the capability of the test to correctly classify the data with or without currency crisis.

Another way to gain insight into the behavior of these ensemble methods is by constructing the κ-error diagrams (Margineantu and Dietterich 1997). These diagrams help visualize the accuracy and diversity of the individual classifiers constructed by the ensemble methods. To define the κ statistic, firstly suppose there are M classes and C is an $M \times M$ square array such that C_{ij} contains the number of test examples assigned to class i by the first classifier and into class j by the second classifier. ϕ_1 is defined as

$$\phi_1 = \frac{\sum_{i=1}^{M} C_{ii}}{m} \tag{8}$$

where m is the total number of test examples. This is an estimate of the probability that the two classifiers agree. ϕ_1 could be used as a measure of agreement but it has a difficulty where in problems involving one class is much more common that the others, all reasonable classifiers will tend to agree with one another simply by chance, so all pairs of classifiers will obtain high values for ϕ_1. The κ statistic corrects for this by computing

$$\phi_2 = \left(\sum_{j=1}^{M} \frac{C_{ij}}{m} \cdot \sum_{j=1}^{M} \frac{C_{ij}}{m} \right) \tag{9}$$

Table 6 Performance measures of three different SVM classifiers (average)

Performance Measure	SVM with RBF	kNN–SVM (stacking)	kNN–SVM (voting)
Percentage of accuracy (%)	96.49	**97.00**	96.49
Sensitivity	0.977	**0.983**	0.977
Specificity	0.983	0.983	0.983
AUC	0.894	0.941	**0.969**
Root mean squared error	0.171	**0.144**	0.155

The bold refer to highest value for percentage of accuracy, sensitivity and AUC also lowest root mean squared error which to highlight that kNNSVM (stacking) has more advantage (bold) than others

which estimates the probability that the two classifiers agree by chance, given the observed counts in the table. To be specific, $\sum_{j=1}^{M} \frac{C_{ij}}{m}$ is the fraction of examples that the first classifier assigns to class i, and $\sum_{j=1}^{M} \frac{C_{ij}}{m}$ is the fraction of examples that the second classifier assigns to class i.

If each classifier chooses which examples to assign to class i completely randomly, then the probability that they will simultaneously assign a particular test example to class i is the product of these two fractions. In such cases, the two classifiers should have a lower measure of agreement than if the two classifiers agree on which examples they both assign to class i. With these definitions, the κ statistic can be computed by

$$\kappa = \frac{\phi_1 - \phi_2}{1 - \phi_2} \tag{10}$$

$\kappa = 0$ when the agreement of the two classifiers equals that expected by chance, and $\kappa = 1$ when the two classifiers agree on every example. Negative values occur when agreement is less than expected by chance which is there exist systematic disagreement between the classifiers.

5 Results and discussion

Before we discuss further about these three ensembles, the results for each ensemble of classifiers will need to be justified first. It is shown in Table 6 that kNN–SVM ensemble by stacking has slightly higher accuracy than single SVM classifier and kNN–SVM by voting. It also has the highest sensitivity among others even though kNN–SVM from voting has the highest value of AUC. The values of specificity for all these three are the same as can be seen in Table 6. In short, there is not much difference between kNN–SVM ensemble by voting or stacking. We noticed that if we analyze the results that we get by country, different country respond differently to each classifier. Due to this information, we assumed that this is a reason why single Support Vector Machine classifier and Support Vector Machine ensembles results are very close. kNN–SVM can be used if a user looking for an ensemble of classifiers that has higher accuracy and minimal error.

For our proposed method, nearest neighbor tree, the results are somehow almost the same with previous kNN–SVM ensembles case which, in other words, there is not much difference among others. We took the average values of performance measures calculated for each city in doing the comparison. The average percentage of accuracy for NNT is 96.3 % while 96.04

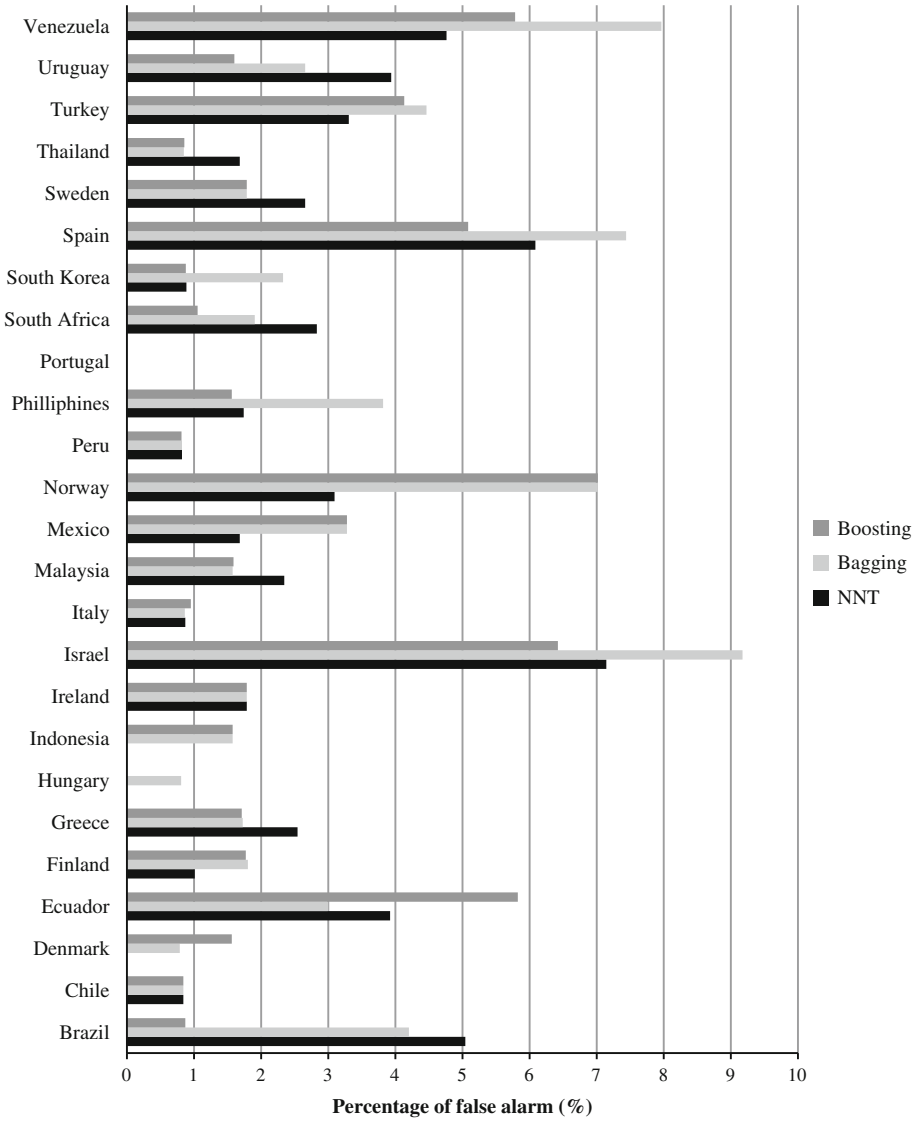


Fig. 4 Percentage of false alarms between NNT, boosting and bagging for 25 countries

and 96.64 % for bagging and boosting respectively. The average values of AUC only have slightly difference where 0.964 for NNT, 0.967 for bagging and 0.975 for boosting. The AUC values for all combined classifiers are excellent based on Table 5. A histogram on the average of percentage of false alarms between these three different types of ensemble on 25 countries is plotted as in Fig. 4 to see the comparison clearly. Based on the histogram, bagging has shown highest percentage values of false alarms in predicting crisis for 6 countries out of 25 while 5 out of 25 for NNT and boosting only has 3 countries out of 25. If we take an average value on the percentage of false alarms, NNT and boosting have less percentage of false alarms compared to bagging.

Table 7 Percentage of accuracy and AUC values for three different ensembles for 25 countries

Country	Percentage of accuracy (%)			AUC		
	kNN–SVM	LORENS	NNT	kNN–SVM	LORENS	NNT
Brazil	96.95	92.37	93.89	0.981	0.967	0.961
Chile	97.71	96.95	98.47	0.913	0.944	0.913
Denmark	100.00	98.47	99.24	1.000	0.843	0.996
Ecuador	94.66	95.42	96.18	0.946	0.965	0.988
Finland	99.24	98.47	97.46	0.965	0.949	0.990
Greece	97.71	98.47	96.95	0.991	0.984	0.993
Hungary	92.37	96.95	99.24	0.737	0.850	1.000
Indonesia	100.00	100.00	100.00	1.000	1.000	1.000
Ireland	97.71	97.71	96.95	0.993	0.992	0.996
Israel	92.37	90.84	91.60	0.978	0.918	0.959
Italy	98.47	98.47	98.47	0.997	0.955	0.999
Malaysia	96.95	96.95	96.95	0.605	0.989	0.834
Mexico	98.47	98.47	96.95	0.977	0.914	0.991
Norway	94.66	95.42	92.37	0.976	0.937	0.980
Peru	97.71	96.95	97.71	0.982	0.989	0.985
Philippines	98.47	97.71	97.46	0.997	0.992	0.996
Portugal	99.24	99.24	100.00	0.994	0.995	1.000
South Africa	95.42	96.18	95.42	0.986	0.988	0.986
South Korea	97.71	97.71	97.46	0.803	0.662	0.761
Spain	93.13	92.37	89.31	0.963	0.917	0.939
Sweden	99.24	96.18	96.95	0.992	0.981	0.991
Thailand	99.24	98.47	98.47	0.932	0.926	0.944
Turkey	92.37	88.55	93.89	0.940	0.873	0.962
Uruguay	97.46	96.18	95.42	0.896	0.992	0.987
Venezuela	97.71	95.42	90.84	0.987	0.980	0.960
Average	96.998	96.397	96.306	0.941	0.940	0.964

Finally, to answer the very last question in this paper which is how well NNT performs compared to LORENS and kNN–SVM ensemble by stacking, we have gathered the results of AUC and the percentage of accuracy for 25 countries as in Table 7. All of these three ensembles have averagely almost the same figure in terms of percentage of accuracy. NNT has highest average value of AUC although in terms of computational time it takes longer than the other two ensembles. All the runtime experiments were conducted on a personal computer with Intel ® Core ™ i5 CPU 2.30 GHz, 4 GB RAM. The average computational time for kNN–SVM is 0.059 CPU seconds, 0.085 CPU seconds for LORENS and 0.101 CPU seconds for NNT. Last but not least is by assessing the performance of these three ensembles of classifiers through the κ-error diagrams. It is illustrative of the diagrams in most of the other domains. We can see from Fig. 5 that NNT indicates high values for κ and low error rates which indicate that the classifiers are accurate but not very diverse. LORENS and kNN–SVM both gave more $\kappa = 0$ values than NNT. This can be concluded as the agreement of the two classifiers for both of these ensembles equals that expected by chance.

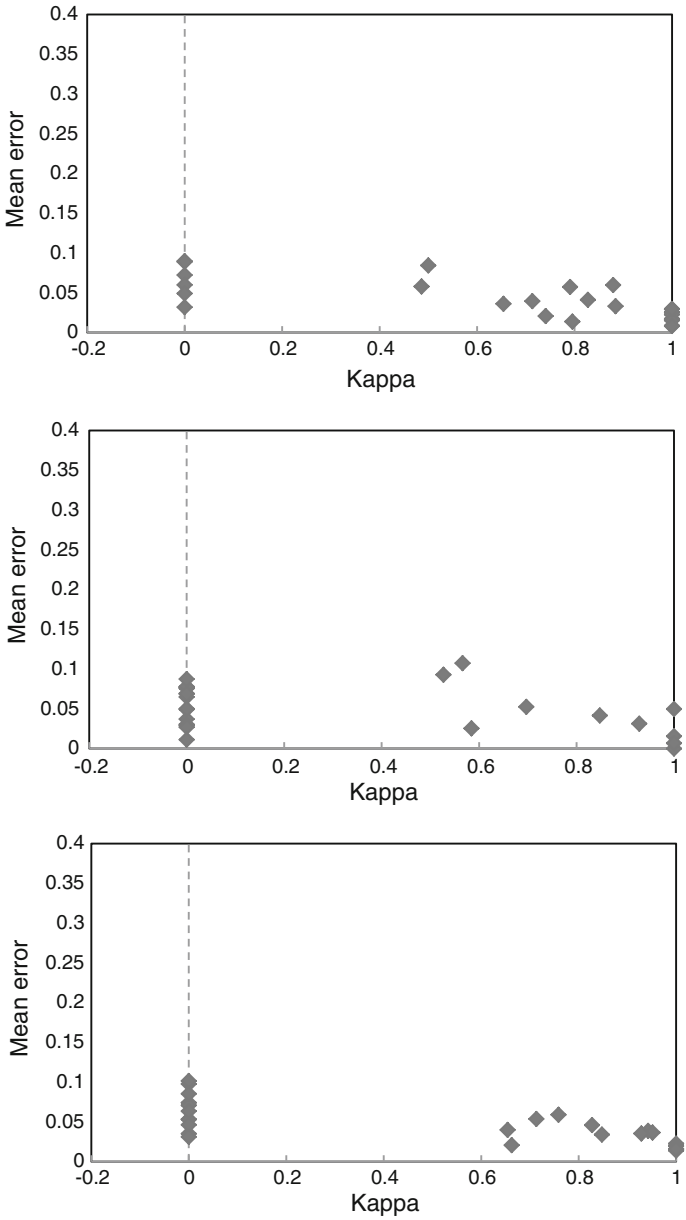


Fig. 5 κ-Error diagrams for the macroeconomic data set for 25 countries using NNT (*top*), LORENS (*middle*), and kNN–SVM (*bottom*). Accuracy and diversity increase as the points come near the origin

6 Conclusion

The results that we got from these three different experiments showed that an ensemble of classifiers by stacking produced outstanding results in terms of minimized percentage of false alarms and with the highest accuracy, AUC and sensitivity. These can be an additional

advantage of using stacking besides the trained rule is more flexible and less bias than voting. Even so, our aim in this paper is to find the best ensemble of classifiers that can predict the currency crisis on 25 countries well. Nearest neighbor tree probably is the best ensemble of classifiers that can be used to predict currency crisis in the future based on the comparison of the performance of classifiers and will be used as our methodology in modeling an early warning system. It has higher accuracy and AUC, decrease the percentage of false alarms and gives high values for κ and low error rates which indicate that the classifiers are accurate but not very diverse although in terms of the computational times, it took longer running times than others. Nearest neighbor tree also has an advantage that can cover this one disadvantage of ensemble which is its interpretable.

Further work is needed both in application of ensemble methodology and currency crisis. Since the percentage of false alarms results that we obtained are differences between countries, therefore an investigation on indicators should also be involved. Furthermore, an experiment of this method on different fields should be conducted to see if this nearest neighbor tree can performs better if applied in other fields.

References

Abiad, A. G. (2003). *Early warning systems: A survey and a regime-switching approach*. IMF Working Paper.

Berg, A., & Patillo, C. (1999). Predicting currency crises: The indicators approach and an alternative. *Journal of International Money and Finance*, *18*, 561–586.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*, 1145–1159.

Datastream. http://extranet.datastream.com.

Dietterich, T. G. (1997). Machine-learning research: Four current directions. *Al-Magazine*, *18*(4), 97–136.

Edison, H. J. (2003). Do indicators of financial crises work? An evaluation of an early warning system. *International Journal of Finance and Economics*, *8*, 11–53.

Eichengreen, et al. (1996). Contagious currency crises: First tests. *Scandinavian Journal of Economics*, *98*, 463–484.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*, 861–874.

Frankel, J., & Rose, A. (1996). Currency crashes in emerging markets: An empirical treatment. *Journal of International Economic*, *41*(3), 351–366.

Gams, M., Bohanec, M., & Cestnik, B. (1994). A schema for using multiple knowledge. In S. J. Hanson, T. Petsche, M. Kearns, & R. L. Rivest (Eds.), *Computational learning theory and natural learning systems, 2* (pp. 157–170). Massachusetts: The MIT Press.

Gunn, S. R. (1998). *Support vector machines for classification and regression*. Southampton, UK: University of Southampton.

Holmes et al. (2002). www.cs.waikato.ac.nz/~bernhard/.../ecml2002.pdfy.

Kaminsky, G. L., & Reinhart, C. M. (1999). The twin crises: The causes of banking and balance-of-payments problems. *The American Economic Review*, *89*(3), 473–500.

Kaminsky et al. (1998). Leading indicators of currency crises. *International Monetary Fund Staff Papers, 45*(1), 1–48.

Kaminsky et al. (2000). *Methodology for an early warning system: Thesignal approach*. Notes from Chapter 2 Assessing Financial Vulnerability: An Early Warning System for Emerging Markets. Washington: Institute for International Economics.

Krugman, P. (2000). *Currency crises*. London: The University of Chicago Press.

Lim, et al. (2010). Classification of high-dimensional data with ensemble of logistic regression models. *Journal of Biopharmaceutical Statistics*, *20*, 160–171.

Ling, et al. (2003). AUC: A better measure than accuracy in comparing learning algorithms. *Lecture Notes in Computer Science*, *2671*, 329–341.

Liqi, et al. (2011). Prediction of eukaryotic protein subcellular multilocalisation with a combined KNN–SVM ensemble classifier. *Journal of Computational Biology and Bioinformatics Research*, *3*(2), 15–24.

Margineantu, D., & Dietterich, T. G. (1997). Pruning adaptive boosting. In *Proceedings of the Fourteenth International Conference on Machine Learning* (pp. 211–218). San Francisco: Morgan Kaufmann.

Obstfeld, M. (1994). *The logic of currency crises*. NBER Working Paper, p. 4640.

Peltonen, T. A. (2006). *Are emerging market currency crises predictable?* A test. ECB Working Paper, p. 571.
Ramli, N. A., Ismail, M. T., & Wooi, H. C. (2013). Designing early warning system: Prediction accuracy of currency crisis by using K-nearest neighbor method. *World Academy of Science, Engineering and Technology*, *79*, 023–1028.
Sachs et al. (1996). *Financial crises in emerging markets: The lessons from 1995*. NBER Working Paper No. 2276, pp. 1–51.
Thomas, G. Tape. http://gim.unmc.edu/dxtests/ROC3.htm.
WEKA. www.cs.waikato.ac/nz/ml/weka.
Wolpert, D. H. (1992). Stacked generalization. *Neural Networks, 5*(2), 241–259.
Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. Massachusetts: The MIT Press.
Zhang, P. (1993). Model selection via multifold cross validation. *The Annals of Statistics*, *21*(1), 299–313.