2013 International Conference on Advances in ICT for Emerging Regions (ICTer): 217 - 221

Comparing Support Vector Regression and Random Forests for Predicting Malaria Incidence in Mozambique

Orlando P. Zacarias
Department of Computer and Systems Sciences
Stockholm University
Forum 100, SE-164 40 Kista, Sweden
Department of Mathematics and Informatics - Faculty of Science
Eduardo Mondlane University
Main Campus, POBox 250, Maputo, Mozambique
Emails:si-opz@dsv.su.se or ozacas@uem.mz

Henrik Boström
Department of Computer and Systems Sciences
Stockholm University
Forum 100, SE-164 40 Kista, Sweden
Email:henrik.bostrom@dsv.su.se

Abstract—**Accurate prediction of malaria incidence is essential for the management of several activities in the ministry of health in Mozambique. This study investigates the comparison of support vector machines (SVMs) and random forests (RFs) for this purpose. A dataset with records of malaria cases covering the period 1999-2008 was used to evaluate predictive models on the last year when developed from one up to nine years of historical data. Mean squared error (MSE) was used as the performance metric. The scheme for estimating variable importance commonly employed for RFs was also adopted for SVMs. SVMs developed from two years of historical data obtained the best prediction accuracy. Hence, if we are interested in predicting the actual number of malaria cases the *support vector machines* model should be chosen. In the analysis of variable importance, Indoor Residual Spray (IRS), the districts of Manhiça and Matola and month of January turned out to be the most important predictors in both the SVM and RF models.**

Keywords—predictions, support vector machines, random forests, malaria incidence cases

I. Introduction

In Mozambique, malaria affects around 90% of the population [1] and is considered together with Human Immunodeficiency Virus Infection/Acquired Immunodeficiency Syndrome (HIV/AIDS) to be the main causes of deaths. Within the hospitalized patients, malaria accounts for around 30% of all deaths [1]. Moreover, access to health care in Mozambique is very low and it is estimated that 50% of the population lives more than 20 kilometres from the nearest health facility. This situation actually implies no access to health services for a wide part of the population, especially in rural areas. Malaria affects also pregnant women in rural areas, with around 20% experiencing the parasite infection [1]. The burden of malaria in the economy has not been scientifically investigated in the country, but it is clear that malaria contributes to high economic losses, with high rates of school absenteeism and poor agricultural productivity, the main livelihood of the majority of the rural population.

Accurate numeric prediction of future malaria cases is of paramount importance, particularly for the National Malaria Programme of the Ministry of Health in Mozambique. Health policy makers, public and private health institutions, medical and other interested stakeholders need an effective tool (framework), for prediction of new cases of diseases, particularly malaria in order to assist in the management and handling of health policies. This includes the creation of appropriate logistical conditions for the realization of activities that could lead to improved prevention of the disease, e.g. through a timely scheduling of activities, provision of malaria drugs, assignment of personnel, and better delivery of health services in general. The provision of accurate estimates of the number of expected malaria cases in different regions can assist health authorities in planning interventions and control of possible disease outbreaks.

Malaria is a disease highly sensitive among other factors to weather conditions and vector (mosquito) control. However, different patterns of incidence of malaria can be seen on a daily, weekly and seasonal basis. For example, the daily patterns can distinguish activities occurring at daylight from activities of malaria vector occurring at night (most common), weekly patterns distinguishing weekdays from the weekends considering the levels of occupancy of homes due to occasional visitors, while seasonal patterns can distinguish winter from summer in terms of the distribution and incidence of malaria.

In this study, we assess the capability of SVMs as an approach for predicting cases of malaria incidence and compare them to the results of our previous study in [2] that uses random forests (RFs), in order to determine the best predictive modeling approach for future adoption within the Mozambique health sector. The study also aims to identify relevant predictors of the incidence of malaria to provide a decision tool for health authorities to improve their performance and delivery of high quality services to the public and population.

The paper is organized as follows: Section II describes the data and the employed methodological issues of SVMs and the parameter selection procedure we propose. Also includes a brief description of the random forests learning scheme. The results are presented in Section III while Section IV summarizes main findings and conclusions.

978-1-4799-1276-6/13/$31.00 ©2013 IEEE

II. METHODOLOGY

This study is focused on the performance analysis of two commonly used data mining techniques: Support Vector Machines (SVMs) [3] and Random Forests (RFs) [4] using the regression approach. The input data used for both models are the number of malaria cases, percentage of indoor-residual sprayed houses and climatic factors such as temperature, rainfall and humidity. The data is monthly aggregated and geographical distributed in eight administrative districts of the Maputo province. It covers a ten years period, ranging from 1999 to 2008, where the first nine years are taken as the training set, while the last year is the test set.

To investigate the effect of variation of malaria incidence over different periods, the training set was further divided into several sub-sets starting from the set of nine years, eighth, etc., to the first year (the year preceding the year of test set); yielding nine training samples (modelling frames). The effectiveness of both regression models is then empirically analysed.

Several applications of SVMs and RFs on health datasets exist. However, applications of SVMs in malaria studies are mostly related to protein sequence similarities, interactions and estimation of malaria parasitemia [5–7] to name a few. Up to date, SVMs have not yet been studied for prediction of malaria incidence cases particularly in Mozambique.

A. Collected Data and Pre-Processing

Historical data for the ten years period 1999-2008 with the number of malaria cases together with climatic data were obtained from the Ministry of Health[1] - Maputo Provincial Directorate of Health and the National Meteorological Institute[2] respectively. The records were collected in each administrative district and locality health center and it includes both microscopically tested and clinically confirmed malaria cases [1, 8]. The Indoor-residual spraying (IRS) data comprises the percentage of homes which had undergone the IRS campaign in each district in the study period. This data was also provided by the Ministry of Health.

The number of malaria cases, climatic factors and IRS are monthly aggregated for the study period, yielding 960 examples divided in training (864 instances, i.e. nine years; 1999-2007) and testing sets (96 examples, i.e. one year - 2008). The data was than pre-processed to meet the SVMs requirements in three stages:

- In the training set, missing values (corresponding to 16.7% of all predictor measurements) were replaced employing two strategies:
  1) by the mean of each attribute.
  2) using multiple imputation technique employed in [8].
- Transformation of each numerical variable by scaling to the interval [-1, 1].
- Nominal predictors (administrative district and month of the year) were transformed into binary variables. This was useful in order to allow for further analysis of malaria cases incidence in each particular district and month including the investigation of importance of each of these variables.

The model developed using replacement strategy one is below referred to as SVM-1 and the model derived employing strategy two is designated SVM-2.

B. Support Vector Machines

Support Vector Machines is a supervised learning technique used for classification and regression problems. For a given set of training data, the SVM algorithm constructs a multidimensional hyperplane that optimally differentiates two classes by maximizing the margin between the two data clusters [3]. In many cases, however, it is not possible to linearly separate the given dataset. One solution to this problem is to map the data into a higher dimensional space through a nonlinear transformation, such that the two classes are separable in this new space [9]. The transformations that allow for this mapping to be done efficiently are called kernel functions. Through the use of kernels, the optimization is performed in the input space rather than the potentially high dimensional feature space. Various types of kernel functions exist, where each of these may be suitable for different tasks. In problems with large amount of features such as DNA problems and text classification for instance, the use of a *linear kernel* is often considered to be most appropriate [10]. Other well-known kernel functions are *polynomial, radial basis* and *sigmoid.*

Having a k-dimensional vector, the SVM basically constructs a (k-1)-dimensional separating hyperplane to discriminate the classes in a k-dimensional space. For a two-dimensional space, a straight line will be the separating hyperplane which divides the higher dimension space in half. In cases where more dimensions are involved, the SVM performs the search for an optimal hyperplane that separates the margins at their maximum degree. The points that fall close to each side of the hyperplane are called support vectors. Actually, it is the distance from the hyperplane to these points that should be maximized. As training patterns, support vectors are samples defining the optimal hyperplane that separates the data, being the most difficult patterns to classify [9].

Support Vector Regression (SVR) is a sub-class of SVMs that uses a different loss function. The most used function is the so called $\epsilon-$insensitive loss function [11], and is employed in this study.

C. Permutation Accuracy Importance

Permutation accuracy importance was introduced by Breiman [4], to estimate the level of influence of predictor variables within the RFs learning method. The estimation of variable importance involves the use of Out-Of-Bag samples. However, unlike the random forests technique, support vector machines do not employ bagging and hence do not use any type of Out-Of-Bag estimation. Furthermore, SVMs do not explicitly estimate or represent variable importance [12].

To overcome this problem, we propose a modification of the random forests approach [4] for estimating variable importance. This is achieved by randomly permuting the values

[1] www.misau.gov.mz
[2] www.inam.gov.mz

of one feature at a time and measuring the effect on predictive performance using the test set. Basically, the idea is that a random permutation of the values of predictor variables is supposed to simulate the absence of the variable from the model [13]. The importance measure is determined as the difference in the prediction accuracy before and after permuting the corresponding predictor variable. The program accomplishing this task is written in the R language within the R-package tool [14].

D. Random Forests

In [2], we propose a random forests regression procedure for predicting the number of incidence malaria cases. Random forests consist of ensembles of classification or regression trees, which are created by employing bootstrap sampling on the training set and random feature selection during tree induction. For a given number of features (say M), random forests samples $m << M$ to split at each creation of a tree node [4]. The resulting predictions are obtained through aggregation (majority voting or averaging). Settings of random forests parameters and further description and discussion of this technique to the application of prediction of future malaria incidence cases are given in [2].

E. Setting Parameters of SVMs

To generate an optimal SVM, the kernel and its parameters cost, gamma, degree, etc., must be selected. The parameter cost controls the model over-fitting by specifying misclassification tolerance in the model. Using the R-Statistical tool [14], the libsvm is accessed through the package e1071 [15] interface.

Performance of SVMs depend on the selection of parameter Cost (C)-soft margin and the kernel, i.e., the kernel's parameters. Best results can be obtained by using a grid search procedure [16] for exhaustive parameter search. The SVM library within [16] generates for each parameter settings a cross-validation (CV) accuracy and returns parameters with highest CV accuracy. Hence, a 10-fold cross validation using two training sets: one with all missing values replaced by the mean of corresponding feature and other with same missing values replaced through multiple imputation [8] is employed to determine kernel and soft margin parameters. Four kernels, *polynomial, radial, sigmoid* and *linear* were considered. The step function of the grid is an exponentially growing sequence of $2^{-5}, 2^{-4}, \ldots . 2^{5}$, and $10^{-5}, 10^{-4}, \ldots, 10^{5}$ for bases two and ten respectively. The parameter degree of the polynomial kernel was set to take values of 2 and 3. Whereas, the ϵ-insensitive band parameter was tried for three different values of 0.1 - the default, 0.01 and 0.001 on each kernel. The insensitive band value of $\epsilon = 0.01$ revealed best performance on all the kernels. Table I shows the kernels and values of parameters with best performance determined in different tuning optimization sessions. Parameters of the model obtaining the smallest estimation error were chosen. As a result, the parameters selected to apply for further analysis of the support vector machines in both datasets are from the base ten parameterization with $cost = 10$; $gamma = 0.01$ for use with the *Radial Basis Function (RBF) kernel*. The obtained relative MSE corresponds to a normalized error considering a prior normalization of response and predictor variables by adjusting measured values to a common scale in the range $[-1.0, 1.0]$, and thus emphasizing the scatter in the entire dataset. It could be re-calculated in the original scale for practical decision making purposes.

TABLE I. TUNING KERNEL PARAMETERS

Kernels with best performance - Using the mean replacement approach					
Kernel	Cost	Gamma	Degree	Epsilon	Best Performance
Linear	1	-	-	0.01	0.02739058
Polynomial	2	0.03125	3	0.01	0.03096156
Radial	10	0.01	-	0.01	0.02205058*
Sigmoid	1e+05	1e-04	-	0.01	0.0274066
Kernels with best performance - Using multiple imputation strategy					
Linear	1	-	-	0.01	0.02732616
Polynomial	2	0.03125	3	0.01	0.03005217
Radial	4	0.03125	-	0.01	0.02206949
Sigmoid	1e+05	1e-05	-	0.01	0.02761128

* - kernel with best performance parameters

III. RESULTS

Results considering predictive performance of models on test sets using different time frames, i.e. from one to nine years of historical data is shown in Figure 1 below. The models were compared on basis of their normalized relative mean squared errors. We use the Decision Tree classifier [17] as the baseline for comparison. The two-years time frame models generated by SVMs using both strategies to handle missing values and the RFs model, accomplished the most effective results. The SVM-1 model (using mean of each attribute to replace missing values) obtained the smallest MSE of 0.0032669, while the SVM-2 (replaces missing values employing a multiple imputation technique [8]) got 0.0337215 and the RFs based model developed in [2] acquired 0.0171. Then, for the current malaria learning problem, SVMs are more effective than both RFs and the Decision Tree. Therefore, subsequent analysis and results presentation is based on SVM-1 model.

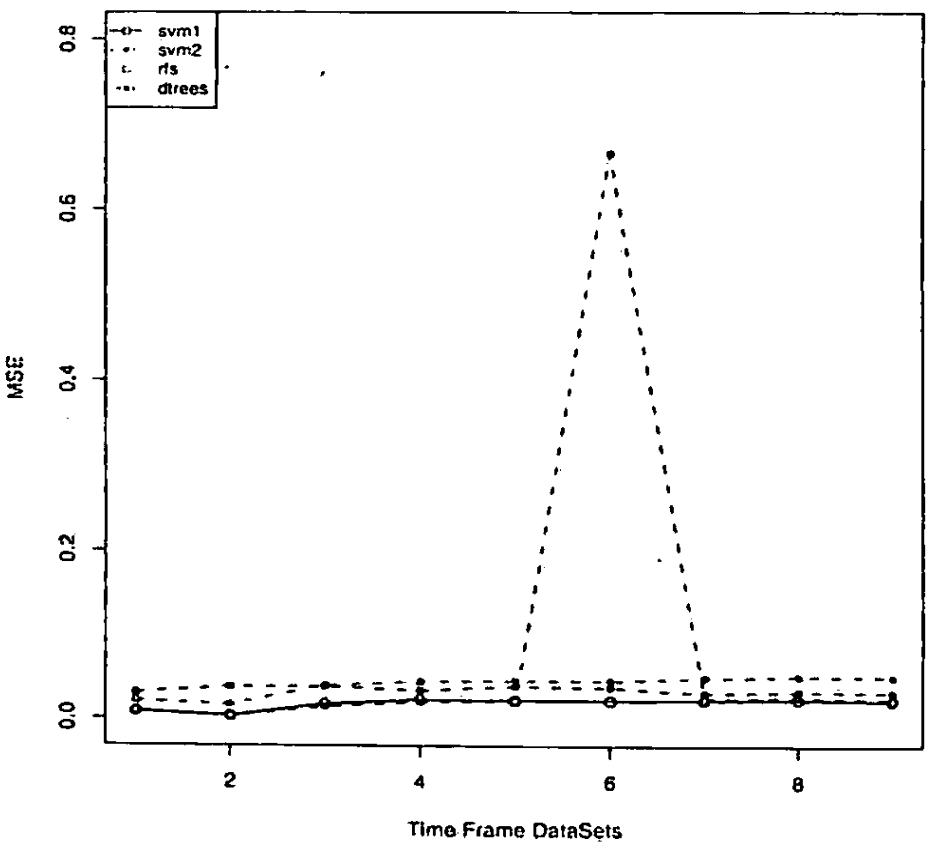

Fig. 1. Performance Comparisons for Model Selection

978-1-4799-1276-6/13/$31.00 ©2013 IEEE

On the other hand, the results suggest an improvement on predictive performance of models using SVMs over the RFs models studied in [2]. However, the obtained mean squared errors of models using RFs are higher than the baseline results for the first seven cases, i.e. from nine to three years time frame models, as it can be seen in Figure 1. Models based on SVMs show more or less the same overall results, except for the case of the data time window of six years, where the discrepancy is extremely high. This may be due to the fact that the missing pattern of climatic factors data was very high in this year in most of the stations. Furthermore, the models generated by SVMs, i.e., using the SVM-1 and the RFs show a statistically significant difference where $p-value < 2.2E-16$, as calculated by the paired t-test.

Variable importance analysis was conducted in order to compare the relative ranking of predictor variables for both random forests and support vector machines. The most effective two-years time frame SVM-1 model was used to investigate the importance of each variable by adopting the technique applied in [4], i.e., employing random permutation of each predictor values using the test data. As such, they represent the quantity by which the model accuracy is reduced due to permuting values of the corresponding variable. A negative variable importance score suggests that the variable may have an adverse impact on the model, and conversely, a variable with a positive score does contribute to the predictive performance.

The calculated relative importance of each predictor variable of the best time frame SVM-1 model is shown in Table 2. Details of the relative importance of the predictors of the RFs can be found in [2]. The most important predictor is IRS for both support vector machines and random forests. In fact, the results rate these campaigns very high as means to improve control and reduce malaria incidence in the study region. Both models rank the districts of Manhiça and Matola highly from a spatial perspective and the month of January from a temporal perspective. The results of SVM-1 model, suggest however that monthly predictors of February, March, May and June and the climatic factors humidity, minimal temperature and temperature variation (determined as the difference between maximal and minimal temperature) contribute to predictive performance, i.e. they are influential model predictors. In contrast, for the RFs case, the monthly predictor for February, March, April, November and December are not important variables for the model. In Table 2, we also note that the measures of variable importance of the model SVM-1 are relatively lower than those of RFs obtained in [2]. Furthermore, predictors for the months of August and September obtained similar importance scores.

Considering the number of influential predictors in each of the models of SVM-1 and RFs respectively, we find that the latter considered twenty-one predictors as important while eleven variables only were considered as important by the first model. The importance scores obtained for monthly period of January to March by the SVM-1 model, is crucial as this can relate to seasonality of malaria incidence predictions. This is due to the fact that it coincides with the period of the rainy season, where environmental and climatic conditions are suitable for most cases of disease to occur.

Preceding year averages of malaria cases records are usually taken into consideration for planning of future malaria related activities, and also as surrogate for new disease's episodes at the Ministry of Health. Therefore, we assume the *default predictive model* (predictive averages of the year before the year of testing set, per region and month), to be the average number of malaria incidence cases of year 2007. The default model performance was compared to the prediction model SVM-1 and all evaluated using the test set data regarding:

- Their relative mean squared errors.
- The number of effective predictions, i.e., equal or higher predictions than the actual observed cases, after conversion of obtained prediction cases to the original scale.

TABLE II. VARIABLE IMPORTANCE OF SVM

Variable Name	Importance Value	Variable Name	Importance Value
IRS	36.48374	MaxTemp	-0.7657
June	24.4501	Rain	-0.2108
Manhiça	9.791	Boane	-1.0939
Matola	5.7981	Dec	-1.0939
Humidity	5.1417	Marracuene	-1.7501
MinTemp	3.2273	Nov	-1.8049
Feb	2.8444	Oct	-1.9691
Jan	2.7349	Aug	-2.516
May	1.4223	Sep	-2.516
March	1.0393	July	-3.1724
TempVar	0.9846	Namaacha	-5.0322
Magude	-0.4922	Matutuine	-5.6338
April	-0.5469	Moamba	-5.7979

The normalized MSE values of both models show a better predictive performance for the derived SVM-1 model over the default prediction model. The SVM-1 model got 0.0032669 while the latter model obtained much higher value 0.1663. Moreover, the default model was able to effectively predict only 44 cases; whereas, the derived SVM-1 actually estimated 52 out of 96 possible from the malaria testing set cases.

IV. CONCLUSION

In this study, datasets with records of malaria cases covering the period 1999-2007 were used to develop prediction models using random forests and support vector machines techniques. These were evaluated for their predictive performance on dataset of year 2008. Their development followed a time frame strategy covering one to nine years of historical data, within the population of the Maputo province.

The SVM-1 model was compared to the results of using random forests approach developed from the same datasets, where we compared their predictive performance including the analysis of their variable importance ranking. The use of sophisticated techniques of data replacement like multiple imputation brought very little improvements on prediction accuracies, as shown when comparing SVM-1 (mean value imputation) and SVM-2 (multiple Imputation) models. However, the result of SVM-2 when using six-year time frame, could have been caused by high missing pattern of climatic factors in months February to October, covering both weather seasons in Mozambique. The models using support vector machines outperformed the models developed by random forests methods

in this investigation. Both techniques determined as the best model to be the generated from a two years time-frame, with the most effective model being the SVM-1 which obtained the lowest 0.003266941 value as its normalized relative mean prediction error.

While analyzing the relative variable importance, the most valuable predictors for both the SVM-1 and RFs models are related to the implementation of indoor-residual spraying campaigns and malaria incidence development in administrative districts of Manhiça and Matola. At the same time, less influential predictors of SVM-1 model were found to be the months ranging from July to December, as well as April. Moreover, the support vector machine regression model was able to relate its results to the effect of seasonality illustrated by the relative importance of months January to March, where the environmental conditions are adequate for most malaria cases to occur.

The investigation of usefulness of the default prediction model compared to the derived SVM-1, showed a predictive performance for the derived SVM-1 model over the default prediction model as determined by their normalized relative mean squared errors. This is also evidenced by the number of correctly estimated predictions in most of the months of the tested year. The findings show the applicability of the support vector regression modeling approach for prediction of future incidence of diseases in order to improve health prevention and control of a complex disease like malaria, when considering that the disease is a major public health problem. Therefore, the SVM prediction model developed in this study is of great potential as a tool for decision making within the health sector in Mozambique. Consequently, having major interest in predicting the actual number of malaria cases, the support vector machines approach should be adopted.

Some ways of improving this study would be:

- Investigate the generalization of developed models to the prediction of future malaria cases numbers in other regions.
- To study the variability in the importance scores between the SVM and RFs as to determine similarities of the results based on these two techniques.

ACKNOWLEDGMENT

The authors would like to thank the Mozambique Ministry of Health and the National Institute of Meteorology for their support and provision of data to conduct this study. We also thank Sida/Sarec and Eduardo Mondlane University project - *Global Research in Mathematics, Statistics and Informatics* for funding this research.

REFERENCES

[1] T. U. S. G. H. Initiative. (2012) Mozambique-national health strategy 2011-2015. [Online]. Available: http://hingx.org:8080/svn/main/eHealth Regulation/

[2] O. P. Zacarias and H. Boström, "Predicting the incidence of malaria cases in mozambique using regression trees and forests," *International Journal of Computer Science and Electronics Engineering (IJCSEE)*, vol. 1, pp. 50–54.

[3] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.

[4] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.

[5] F. Binka and P. Adongo, "Acceptability and use of insecticide impregnated bed nets in northern ghana," *Tropical Medicine and International Health*, vol. 2, p. 499507, 1997.

[6] L. Liao and W. S. Noble, "Combining pairwise sequence similarity and support vector machines for detecting protein evolutionary and structural relationships," *Journal of Computational Biology*, vol. 10, no. 6, 2003.

[7] L. Y. Han, C. Z. Cai, S. Lin Lo, M. C. Chung, and Y. Z. Chen, *Prediction of RNA-binding proteins from primary sequence by a support vector machine approach*, vol. 10, pp. 355–368, 2004.

[8] O. P. Zacarias and M. Anderson, "Spatial and temporal patterns of malaria incidence in mozambique," *Malaria Journal*, vol. 10:189, 2011.

[9] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. John Wiley, 2001.

[10] N. L. S. Mishra, "De novo svm classification of precursor micrornas from genomic pseudo hairpins using global and intrinsic folding measures," *Bioinformatics*, vol. 23, pp. 1321–1330, 2007.

[11] D. Basak, S. Pal, and C. Patranabis, "Support vector regression," *Neural Information Processing - Letters and Reviews*, vol. 11, no. 10, oct 2007.

[12] D. Barbella, S. Benzaid, J. Christensen, B. Jackson, X. V. Qin, and D. Musicant, "Understanding support vector machine classifications via a recommender system-like approach," in *Proceedings of the International Conference on Data Mining (DMIN)*, Las Vegas, USA, july, 13-16, 2009.

[13] C. Strobl, A.-L. Boulesteix, T. Kneib, and T. Augustin, "Conditional variable importance for random forests."

[14] R-statistical tool for data analysis. Accessed-September 16, 2012. [Online]. Available: http://CRAN.R-project.org/

[15] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingesse, and F. Leisch. Manual of package e1071. Accessed - October 27, 2012. [Online]. Available: http://CRAN.R-project.org/=e1071

[16] C.-W. Hsu, C.-C. Chang, and C.-J. Lin. A practical guide to support vector classification. [Online]. Available: http://www.csie.ntu.edu.tw/cjlin

[17] R. J. Quinlan, *C4.5: Programs for Machine Learning*.

978-1-4799-1276-6/13/$31.00 ©2013 IEEE