Comparing Classification Performance of Decision Trees and Support Vector Machines: A Small Data Scenario

Liew Chin Ying[1,2], Jane Labadin[1], Wang Yin Chai[1]
[1]Faculty of Computer Science and Information Technology
UNIMAS, 94300 Kota Samarahan, Malaysia.
[2]Faculty of Computer and Mathematical Sciences,
UiTM, 94300 KotaSamarahan, Malaysia.
liewchinying@hotmail.com,{ljane,ycwang}@fit.unimas.my

Andrew Alek Tuen[3], Cindy Peter[3]
[3]Institute of Biodiversity and Environmental Conservation
UNIMAS, 94300 Kota Samarahan, Malaysia.
aatuen@ibec.unimas.my, cindycharity.peter@gmail.com

Abstract—**Decision Trees and Support Vector Machines are two widely used supervised machine learning algorithms which have been applied successfully across broad research domains. They are commented as powerful and feasible classifiers model regardless of data sizes. This study aims to compare the classification performance of DT and SVM models built for sightings of Irrawaddy dolphin at Kuching Bay. The data used is a small balanced real-world dataset with 152 data points. Classification performance of these two models is further compared through assessment of classification prediction onto an independent one-class dataset consisting of 38 data points. The results show that these two models are equally competitive in their classification performance in terms of ROC-AUC values, sensitivity, specificity, G-mean, and CCI. Nevertheless, the SVM model reports much better prediction power with 15.69 percent higher in CCI recorded for the independent one-class dataset. Present study thus suggests that SVM is more preferable when ecology research scenario with small data size is of concerned.**

Keywords—Classification Performance; SVM; Decision Trees; Small Data; Modeling; Irrawaddy dolphin

I. Introduction And Background

Decision Trees (DT) is one of the most commonly applied approaches for classifications because it is practical for inductive inference, robust towards noisy data, capable in learning disjunctive expressions, and produce easy-to-interpret results [1]. Although commented as usable with scarce data, performance of DT algorithms deteriorates greatly when compared with other machine learning algorithms (MLA) in cases where data size is reduced drastically [2]. Support Vector Machines (SVM) is a relatively new MLA for classification which originates from [3]. It is based strongly in mathematical theories which is explained and illustrated extensively by [4]. SVM is capable to be applied in research problems with all sizes of data, including complex data of high dimensionality and nonlinearity [5]. In the same survey scenario [2], classification performance of SVM is reported as more stable and ranked the second place when data size is reduced drastically. The most reported advantage of SVM over other classifiers is that it requires minimal model tuning as only a few parameter settings are involved [5].

DT is a widely applied MLA across vast domains of research areas [6]. In the field of ecology, particularly in modeling population dynamics and habitat suitability for various types of species in different ecosystems under diverse environmental challenges and conditions, DT is commented as one of the most widely used MLA [7]. As for SVM, it is well known for its application in medical, image processing and text analysis related studies. However, employing SVM in ecology related studies is scarce [5] and its use in surveys particularly related to marine mammals can hardly be found [8].

Research data available to current survey is collected by Sarawak Dolphin Project (SDP) of Institute of Biodiversity and Environmental Conservation (IBEC) in UNIMAS. The current main survey focus area is at Kuching Bay of Sarawak and ID is the most frequently sighted dolphin species. ID (*Orcaella brevirostris*) is categorized as a vulnerable species [9, version 2013.2] and thus poses the urgent need in researching various aspects of ID. Being able to estimate the sightings of ID at a location point is pertinent to the researchers in identifying possible locations to observe the individuals.

Utilizing this dataset, present survey intends to formulate two ID-sighting models using DT and SVM, and compare their classification performance. Their prediction power of classification is further assessed and compared using an independent one-class dataset. This paper is organized as follows: section one introduces the research topic and its background; section two explains the survey methods and materials; section three presents the research result and its corresponding discussion; and section five concludes the paper.

II. Material And Methods

A. Dataset

Two sets of data are involved in this study: i) dataset used to model the sightings of ID at Kuching Bay, and ii) dataset used to assess the prediction power of the models formulated.

This work is part of the doctorate postgraduate study of Liew Chin Ying in UNIMAS, supported by SLAI-UiTM scholarship.

Both of these datasets are small real-world dataset that records sightings and/or non-sightings of Irrawaddy dolphin (ID) at Kuching Bay. They are collected between August 2008 and September 2010 through a boat survey observation following the pre-determined parallel transects survey methods along the coastal areas of Kuching Bay, in the rivers and channels that are interconnected during all tidal states, and down to the centre of the river. At the start, midpoint and end of each transect leg, and whenever sightings of ID are confirmed, related data are recorded and collected.

Dataset used to model the sightings of ID is a balanced dataset that consists of 152 data points which will be divided into training and testing dataset. It comprises of sighted and non-sighted data. Each data point refers to a specific point of location in Kuching Bay. At each of these points, latitude and longitude of the specific point of location, sea water depth (in meter, m), sea water surface temperature (^{0}C), and sea water salinity (in PSU) are collected. Therefore, each data point of this dataset is characterized by the above five attributes.

Dataset used to assess the prediction power of classification for the models formulated consists of 38 data points. This is the unseen data. It is an independent one-class dataset where every data point records sighting of unique individual ID which is identified following standard mark-recapture technique through photo-identification. This dataset is excluded from the modeling-dataset mentioned above where no overlapping of data points exists ensuring it to be an independent dataset. Since these are individual ID data which are sighted data points, it is a one-class dataset. Similarly, all data points in this dataset are characterized by the five attributes as the modeling-dataset.

B. *Classifications Performance Metrics*

In this paper, classification performance is referred as the characteristics and how successful the models formulated using DT and SVM learning algorithms are in accurately classifying data points from the testing dataset and/or the independent one-class dataset. The metrics used for this purpose are area under the receiver operating characteristic curve (ROC-AUC value), sensitivity or true positive rate, specificity or true negative rate, geometric mean (G-mean), and correctly classified instances (CCI). These metrics have their basis in the confusion matrix [10] and are expressed in terms of percentage.

ROC-AUC value is a score obtained from the receiver operating characteristic (ROC) analysis. ROC analysis produces plot of true positive rate as a function of false positive rate that change when threshold settings is varied [5,11]. As visual comparison is very subjective, area under the curve of these plots is usually calculated and the values obtained is reported as the ROC-AUC value ranged between 0 and 1. Models with ROC-AUC values above 0.9 indicates excellent prediction, 0.7 to 0.9 good, 0.5 to 0.7 poor, and below 0.5 is remarked as no better than a random guess [11]. It is a consistent measure even with highly skewed class distribution [12].

Sensitivity is the true positive rate and specificity is the true negative rate that are defined below respectively,

$$\text{Sensitivity} = TP/(TP+FN), \tag{1}$$

$$\text{Specificity} = TN/(TN+FP). \tag{2}$$

The metric that combine both the sensitivity and specificity is defined as the G-mean [13],

$$\text{G}-\text{mean} = \sqrt{\text{Sensitivity} \times \text{Specificity}}\,. \tag{3}$$

It computes their geometric mean and is generally considered as a more comprehensive metric in a learning problem. CCI, defined in (4) and known as the overall accuracy, is the most commonly used metric in classification learning problem.

$$\text{CCI} = (TP+TN)/(TP+FN+TN+FP) \tag{4}$$

Standard critical value of CCI signifying good prediction can hardly be found in literature. Nonetheless, models with at least 70 percent CCI are commonly considered by ecology scientific community as reliable [5]. In these equations, TP refers to True Positive (number of positive labeled data points that are predicted correctly), TN refers to True Negative (number of negative labeled data points that are predicted correctly), FP refers to False Positive (number of negative labeled data points that are predicted wrongly) and FN refers to False Negative (number of positive labeled data points that are predicted wrongly).

In order to determine models that are acceptable to be included into both DT and SVM models formulated in this survey, ROC-AUC values and CCI are used. Threshold values for these two metrics used in present survey are 70 and 60 percent, respectively. Threshold value of 60 for CCI is used here because i) the best CCI attainable by the models developed in present survey never exceed 70 percent; and ii) the data size of testing dataset with 46 data points is small where every data point significantly constitute 2.17 percent. Moreover, the 70 percent of CCI generally accepted by the ecology scientific community serves as a rule of thumb and a general guideline to refer to where reliability and prediction power of a model would not drastically downgrade to intolerant level with a decrease of 10 percent or 4.6 data points that are being classified correctly by a model in this study.

C. *Classifications Using DT*

The model constructed here is termed as DT ID-sighting Model. The modeling dataset need to be formatted as required by the specific DT learning algorithm chosen before being divided into training and testing dataset following the proportion of 70 percent and 30 percent respectively. Frequently used DT learning algorithms include Classification and Regression Trees (CART) [14], ID3 package [15], C4.5 [16], J48 package, C5.0 package, and CHAID (chi-squared automatic interaction detection) [17]. CART has advantages over other DT learning algorithms as it i) works with either categorical or continuous targets and also either categorical or continuous predictors; and ii) provide pruning option for the tree constructed via validation testing against an independent data set or through *n*-fold cross-validation; which makes CART a more accurate DT models [18]. Consequently, this present study resolves to use `classregtree` class from the

Statistics Toolbox in MATLAB® R2007b [19] that follows the CART algorithms.

Using this DT MLA, training dataset (106 data points) is used to construct DT model. In this study, the DT function features used are latitude, longitude, depth, temperature, and salinity and the response of DT function is decided as 'Y' referring to ID-sighted data points or 'N' the opposite. As too much or too little pruning may cause under- and/or over-fitting of trees constructed, pruning level one to three are implemented. The Gini's diversity index optimization criterion and stopping rule of when the node is pure are employed here. Four DT models, which are the full tree and the trees obtained from pruning level one to three, are eventually produced.

Classification performance of these DT models is then assessed with the testing dataset (46 data points) using classifications performance metrics described earlier. DT models with more than 70 percent of ROC-AUC values and more than 60 percent of CCI are accepted and included as the DT ID-sighting Model. In this study, all the four DT models fulfill these requirements and are included into the DT ID-sighting Model.

D. Classifications Using SVM

The model developed using SVM is termed as SVM ID-sighting Model. *C*-SVM formulation of LIBSVM (version 3.17) package [20], a library for SVM, is employed for the implementation of SVM learning algorithm. LIBSVM tool is user-friendly and is currently one of the most popular SVM software package with more than 250,000 downloads between 2000 and 2010 [20]. MATLAB® R2007b and Microsoft© Windows 7 Enterprise systems are used as the interface of implementing LIBSVM. Modeling dataset is first formatted where ID-sighted data points is labeled as positive one (+1) and termed as positive class whereas ID-not-sighted data points is labeled as negative one (–1) and termed as negative class. Five attributes are used to characterize each data point. Then, the same 70 and 30 percentage is used to divide the formatted modeling dataset into training and testing dataset. Implementation of SVM learning algorithm requires scaling of training dataset before it is utilized for the formulation of SVM model.

Gaussian radial basis function (RBF) is chosen as the kernel function here. Advantages of RBF include (i) ability to work in an infinite dimensional feature space; (ii) ability to handle non-linearity nature of a learning problem; (iii) having one parameter, γ(> 0), that reduces the complexity of model selection; and (iv) having few numerical difficulties as the kernel function value always lies between 0 and 1 [21]. Next, values of SVM cost parameter c and kernel parameter γ are determined through the execution of the automatic classification tool and parameter selection tool of LIBSVM library where ten folds cross-validation checking technique is used. The pairs of parameters (c,γ) generated are then used to train the scaled training dataset and to generate corresponding SVM models. Eighteen SVM models are eventually produced from these different pairs of parameters generated. Their classification performance is then assessed using testing dataset that have been scaled beforehand base on the scaling parameter generated during the scaling process of training dataset. Model with more than 70 percent of ROC-AUC values and more than 60 percent of CCI are accepted and included as the SVM ID-sighting Model. In this study, seven SVM models fulfill these requirements and are included into the SVM ID-sighting Model.

E. Prediction Power of Classification

Prediction power in this survey is referred to as how successful DT and SVM ID-sighting Models are in accurately classifying data points in the independent one-class dataset. Upon formulation of DT and SVM models, the independent dataset is sent through these models and the percentage of correctly classified instances (CCI_{ind}) is determined.

III. Result And Discussion

Table I reveals the classification performance of the four DT models in terms of the five metrics used in this study, including the prediction power of these models in classifying an independent one-class dataset in terms of CCI_{ind}. Likewise, the same is presented in Table II for the seven SVM models. Table I shows that the values of the five metrics for these four models are very similar, with a difference approximately ranged between one to five percents. CCI_{ind} values in Table I reveals that the prediction power of classification for these models increases as the pruning level increases. DT model 4, with the highest sensitivity and CCI, and lowest specificity managed to attain the highest (almost 85%) classification accuracy of the independent dataset.

As for SVM ID-sighting Model, Table II shows that the prediction power, implied by the ROC-AUC values, of these models is about the same. The prediction power of classification (CCI_{ind}) for all models in SVM ID-sighting Model turns out to be good, recording at least 85% accuracy. The first six models have the same CCI_{ind} regardless of the different corresponding classification performance of these models. CCI_{ind} of sub-model 7 turns out to be the highest although it has the lowest CCI and sensitivity.

TABLE I. Classification Performance and Prediction Power of DT ID-sighting Model

Model	*ROC-AUC value*	*Sensitivity*	*Specificity*	*G-mean*	*CCI*	*CCI_{ind}*
1	70.51	73.91	52.17	62.10	63.04	60.53
2	72.21	78.26	52.17	63.90	65.22	65.79
3	71.46	78.26	52.17	63.90	65.22	71.05
4	70.60	82.61	47.83	62.86	65.22	84.21
Average	71.20	78.26	51.09	63.19	64.68	70.40

To further compare the DT and SVM models, the respective classification performance metrics are averaged. The result is displayed in the last row of Table I and II respectively. It is observed that the classification performance of both DT and SVM ID-sighting Models are equally competitive where the average values for all the five metrics are similar to each other. SVM Model records higher values for all metrics except sensitivity. Nonetheless, the prediction power of classification for SVM Model appears to be much higher than DT Model. The former model manages to successfully classify more than 15 percent more of the data points in the independent one-class dataset.

TABLE II. CLASSIFICATION PERFORMANCE AND PREDICTION POWER OF SVM ID-SIGHTING MODEL

Model	*ROC-AUC value*	*Sensitivity*	*Specificity*	*G-mean*	*CCI*	CCI_{ind}
1	73.72	73.91	60.87	67.07	67.39	84.21
2	76.18	73.91	65.22	69.43	69.57	84.21
3	76.56	73.91	52.17	62.10	63.04	84.21
4	74.67	78.26	52.17	63.90	65.22	84.21
5	76.75	73.91	60.87	67.07	67.39	84.21
6	76.75	73.91	52.17	62.10	63.04	84.21
7	76.75	65.22	56.52	60.71	60.87	97.37
Average	75.91	73.29	57.14	64.63	65.22	86.09

IV. CONCLUSIONS

This study has shown that both MLA of DT and SVM are feasible and capable to be used in modeling the sightings of ID at Kuching Bay under the scenario of a small data size. The results demonstrate that both models are equally competitive though SVM model slightly outperformed DT model in terms of the average ROC-AUC value, specificity, G-mean and CCI. The prediction power of classification of these two models are further assessed and compared with an independent one-class dataset consisting of only 38 data points. The outcome reveals that prediction power of SVM model is much stronger than DT model in classification. Present study thus suggests that SVM learning algorithm is more preferable than DT for a better classification prediction particularly in a learning problem with small data scenario. Availability of data for latitude, longitude, sea water depth, sea water surface temperature, and sea water salinity of a location point at Kuching Bay enable sightings of ID to be determined using the SVM ID-sighting Model developed.

Acknowledgment

The authors thank Universiti Malaysia Sarawak for providing all of the resources needed in conducting this study.

References

[1] P. Boets, K. Lock and P.L.M. Goethals, “Modelling habitat preference, abundance and species richness of alien macrocustaceans in surface waters in Flanders (Belgium) using decision trees,” Ecological Informatics, vol.17, pp. 73-81, 2013.

[2] V. Rodriguez-Galiano, M. Sanchez-Castillo, M. Chica-Olmo, and M. Chica-Rivas, “Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines,” Ore Geology Reviews, in press, 15 pages, 2015.

[3] Cortes and V. Vapnik, “Support-vector networks,” Machine Learning, vol.20, issue 3, pp. 273-297, 1995.

[4] N. Cristianini and J. Shawe-Taylor, An Introduction to Support Vector Machines and other Kernel-based Learning Methods, UK: Cambridge University Press, 2001.

[5] H. H. Thu, K. Lock, A. Mouton and P. L. M. Goethals, “Application of classification trees and support vector machines to model the presence of macroinvertebrates in rivers in Vietnam,” Ecological Informatics, vol.5, issue 2, pp. 140-146, 2010.

[6] Y. Hang and S. Fong, “Countering the concept-drift problems in big data by an incrementally optimized stream mining model,” The Journal of Systems and software, vol.102, pp. 158-166, 2015.

[7] M. Debeljak and S. Džeroski, “Decision trees in ecological modeling,” in F. Jopp, H. Reuter and B. Breckling (Eds), Modelling Complex Ecological Dynamics: An Introduction into Ecological Modelling for Students, Teachers & Scientists, Springer Berlin Heidelberg, pp. 197-209, 2011.

[8] C.Y. Liew, J. Labadin, Y.C. Wang, A.A. Tuen, and C. Peter, “Modeling using support vector machines on imbalanced data: A case study on the prediction of the sightings of Irrawaddy dolphins,” in AIP Conference Proceedings 1660, 050011, doi:10.1063/1.4915644, 28-30 May, 2014.

[9] R. R. Reeves et al., “Orcaella brevirostris” in IUCN 2011. IUCN Red List of Threatened Species, Version 2011.1, http://www.iucnredlist.org/, 2008, accessed on 26 July 2011.

[10] A.H. Fielding and J.F. Bell, “A review of methods for the assessment of prediction errors in conservation presence/absence models,” Environmental Conservation, vol.24, issue 1, pp. 38-49, 1997.

[11] H. Reiss, S. Cunze, K. König, H. Neumann, and I. Kröncke, “Species distribution modeling of marine benthos: a North Sea case study,” Marine Ecology Progress Series, vol.442, pp. 71-86, 2011.

[12] R. Potolea and C. Lemnaru, “A comprehensive study of the effect of class imbalance on the performance of classifiers,” in 13th International Conference on Enterprise Information Systems, Conference Proceedings ICEIS 2011, 1 DISI, pp. 14-2, 2011.

[13] M. Kubat and S. Matwin, “Addressing the curse of imbalanced training sets: one-sided selection,” in Fourteenth Internation Conference on Machine Learning, Conference Proceedings, pp. 179-186, 1997.

[14] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone, Classification and Regression Trees, Belmont, CA: Wadsworth International Group, and Boca Raton, FL: CRC Press, 1984.

[15] J.R. Quinlan, “Induction of decision trees,” Machine Learning, vol.1, issue 1, pp. 81-106, 1986.

[16] J.R. Quinlan, C4.5: Program for Machine Learning, San Mateo, CA: Morgan Kaufmann Publishers, 1993.

[17] G.V. Kass, “An exploratory technique for investigating large quantities of categorical data,” Applied Statistical, vol.29, issue 2, pp. 119-127, 1980.

[18] M. Chambers and T.W. Dinsmore, “Predictive analytics techniques,” in Advanced Analytics Methodologies: Driving Business Value with Analytics, USA: Pearson FT Press, pp. 147-192, 2014.

[19] MathWorks Inc. MATLAB® Statistics Toolbox™: User’s Guide, R2014b. Natick, MA: The MathWorks, Inc, 2014, http://www.mathworks.com/help/pdf_doc/stats/stats.pdf , accessed on 21 January 2015.

[20] C.C. Chang and C.J. Lin, “LIBSVM: A Library for Support Vector Machines,” ACM Trans. Intell. Syst. Technol. 2, vol.3, Article 27, 27 pages, April 2011.

[21] W. Hsu, C. C. Chang and C. J. Lin, “A practical guide to support vector classification”, Technical Note, Department of Computer Science and Information Engineering, National Taiwan University, Taiwan, 2010.