Global Journal of Pure and Applied Mathematics.
ISSN 0973-1768 Volume 11, Number 6 (2015), pp. 4629-4637

http://www.ripublication.com

Classifying Infection Status of Hookworm in Cats using SMOTE Support Vector Machines and Boosting Support Vector Machines

Winalia Agwil[1], I Made Sumertajaya[2], Indahwati[3], Etih Sudarnika[4]

[1, 2, 3]Department of Statistics Bogor Agricultural University
Jalan Pajajaran, Kampus IPB Baranangsiang, Bogor 1615, Indonesia
[4]Department of Veterinary Public Health Bogor Agricultural University
Jalan Agatis, Kampus IPB Dramaga, Bogor 16680, Indonesia
1. winalia. agwil@gmail. com, 2. imsjaya@yahoo. com,
3. indahwati_43@yahoo.co.id, 4. etih23@yahoo. com

Abstract

Helminth infection is one kind of the diseases in cats that need attention. This infection not only causes a health problem for cats, but also it can be transmitted to human especially to a cat owner. One of helminth (worm) that infect cat is hookworm. Prevalence of hookworm infection in cat in Denpasar, about 36.2%, it is needs an early detection of hookworm infection status in cat. Early detection of hookworm infection status can be conducted using a classification analysis such as support vector machine (SVM). In this case, there were imbalanced datasets, it would be increasing misclassification because the classification disposed to majority class. Misclassification of the imbalanced dataset could be minimized by synthetic minority over-sampling technique (SMOTE) and boosting method. The goal of this study is to compare the classification performance of SVM, SMOTE SVM and boosting SVM based on AUC value, sensitivity and specificity. SVM method which applied on imbalanced datasets gave specificity value about 100% and sensitivity only 0%. After applying SMOTE method on preprocessing dataset, specificity value was 78% and sensitivity value was 46%, it shows that SMOTE can be improved accuracy in minority class. Whereas, boosting SVM gave the highest sensitivity value was 69%, but specificity value just 56%. Boosting SVM gave the best performance of classification that provides a stable classification and gives the highest AUC value.

Keywords: Hookworm Infection, Imbalanced Datasets, SVM, SMOTE, Boosting.

1 Introduction

Hookworm is a parasitic nematode that lives in the its intestine host, including cats, dogs and humans. Hookworm species that are commonly found in cats are *Ancylostoma tubaeform* and *Ancylostoma braziliense*. It causes the disease named ancylostomiasis. According to Oktaviana et al. (2014), prevalence of ancylostomiasis in cats in Denpasar, about 36.2%. These infection are affected by some factor such as environment condition, feeding nutrition, maintenance system and sanitation.

Hookworm is not only able to infect cats, but also humans. The infection on human called creeping eruption. Human can be infected by hookworm through penetration of the skin because of direct contact with sand, soil and cat feces that have been infected. Symptom of a hookworm infection in human is itchiness and a small rash caused by an allergic reaction as the larvae enters human skin.

Cats are popular pet at this time, so that a high level of hookworm infection in cats will be increasing the hookworm infection to human. Beveridge and Jones (2002) states roughly at least one in five people in the world infected by hookworm. Because of the large number of hookworm infection, it is necessary to do early detection status hookworm infection in cats. One of method can be used to detect a hookworm infection status in cats is a classification analysis.

Support vector machine (SVM) is one of classification method can be used to classify new objects into a class based on attribute value. SVM method basically found the most hyperplane of any possible hyperplane. So it is able to separate most of observation into the two classes and produces a high accuracy. In addition, this method can be used for large scale of features and greater robustness [12].

In classification analysis, imbalanced dataset problem are used to found in medical diagnosis [13], credit scoring [3] and others. Imbalance dataset is a serious problem, because in modeling produces a biased classifier that have higher predictive accuracy over the majority class, but less predictive accuracy over the minority class [4]. Whereas, in many cases, misclassifying the minority class object could have a bigger problem than misclassifying the majority object. SMOTE is one of methods to overcome imbalance dataset.

Imbalanced dataset can be solved by combination of standard classifier and ensemble method. One of ensemble method that be used is boosting. The combination of boosting and SVM named boosting SVM. It will give the higher predictive accuracy than single classifier. The goal in this study, compare classifying infection status of hookworm in cats using SVM, SMOTE SVM and boosting SVM.

2 Materials And Methods

2.1 Synthetic Minority Over-sampling Technique (SMOTE)

Imbalanced dataset exist when there is domination on the whole dataset named majority classes and other class becoming minority. Chawla *et al.* (2002) propose a method for handling the imbalanced dataset named Synthetic Minority Oversampling Technique (SMOTE). The basic idea of this method is adding a number of examples in the minority class with generating synthetic samples based on information from *k*-nearest neighbor. In the nearest neighbor computations for the continuous features

using Euclidean distances (Eq. 1) and the nominal features using Value Difference Metric (VDM) (Eq. 2).

$$\Delta(X,Y) = \sqrt{(x-y)'(x-y)} \quad (1)$$

$$\Delta(X,Y) = W_x W_y \sum_{i=1}^{N} \delta(x_i y_i)^r \quad (2)$$

$$\delta(x_1,y_2) = \sum_{i=1}^{S} \left| \frac{C_{1i}}{C_1} - \frac{C_{2i}}{C_2} \right|^k$$

Where $\Delta(X,Y)$ is distances between *X* and *Y* instances, W_x and W_y are weights of observation, N is the total number of features, *r* = 1 yields Manhattan distance and *r* = 2 Euclidean distance, $\delta(x_i,y_i)$ is distances between two value for a specific features, C_{1i} is the number of times x_1 was classified into category *i*, C_1 is the total number of times value 1 occurred, *S* is the total category of a features, *i* is the total category in response variable and *k* is a constant, usually set to 1.

The following procedure to generate synthetic sample:

1. Continuous features
   a. Take a difference between a features vector and one of k nearest neighbor has chosen randomly.
   b. Multiply difference from previous stage (a) by a random number between 0 and 1.
   c. Add value from previous stage (b) to the features value of the original features vector, thus creating a new features vector.
2. Nominal features
   a. Take majority vote the features vector based on k-nearest neighbor information.
   b. Assign that value to new synthetic minority class sample.

2.2 Support Vector Machines (SVM)

Support Vector Machines is one of several data mining techniques used in prediction, either classification or regression. SVM introduced in 1992 by Vapnik. The basic concept of SVM is a linear classifier but has been developed to non-linier problems by using a kernel trick. In simply concept SVM is finding the best hyperplane of all possible hyperplane [6], by maximizing the distance between hyperplane of each closest point (support vector) for each class. The total distance between two closest points from each class usually named margin.

Given a dataset $x_i \in \{x_1, \ldots x_p\}$ is features in p-dimensional space dan $y_i \in \{+1, -1\}$ is class labels for all i = 1, 2, …, n. The following is the formula to hyperplane :

$$\boldsymbol{w}\boldsymbol{x} + b \quad (3)$$

where $\boldsymbol{w}$ is an (1xp) vector that is perpendicular to hyperplane and *b* is a threshold. The decision function is defined as follows:

$$f(\boldsymbol{x}) = sign(\boldsymbol{w}\boldsymbol{x} + b) \quad (4)$$

If $\boldsymbol{wx} + b > 0$ the observation would be classified to class 1 and if $\boldsymbol{wx} + b < 0$ then observation would be classified to class -1 [8]. The maximal margin hyperplane is the solution to the primal problem [5]:

$$min\ \frac{1}{2}\|\boldsymbol{w}\|^2 \quad (5)$$

Subject to,

$$y_i(\boldsymbol{w}\boldsymbol{x}_i + b) \geq 1, i = 1, \ldots, n \quad (6)$$

The SVM problem can be solved by quadratic programing, using equation (5) and equation (6) intolagrangian (primal problem) formula.

$$L(\boldsymbol{w}, b, \alpha) = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{n} \alpha_i\{(y_i(\boldsymbol{w}\boldsymbol{x}_i + b)) - 1\} \quad (7)$$

Where Lagrange multipliers $\alpha_i \geq 0$. Minimizing Langrangian function to $\boldsymbol{w}$ and b obtains,

$$\sum_{i=1}^{n} \alpha_i y_i = 0$$

$$\boldsymbol{w} = \sum_{i=1}^{n} \alpha_i y_i x_i \quad (8)$$

Langrangian L (primal problem) can be change to dual problem L_d by replacing $\boldsymbol{w}$ equation (8) in the lagrangian equation (7).

$$L_d(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i{}' x_j \quad (9)$$

Maximum hyperplane is obtained by solving the following equation:

$$\max L_d(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i{}' x_j \quad (10)$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0, \alpha_i > 0$$

α_i is used to find the vector *w* and *b*. Each observation of the training dataset has a α_i value, the observation corresponding to non-zero α_i values are called support vectors. In SVM classification model is only influenced by support vectors. Illustration of linier SVM on separable case can be seen in Figure 3.

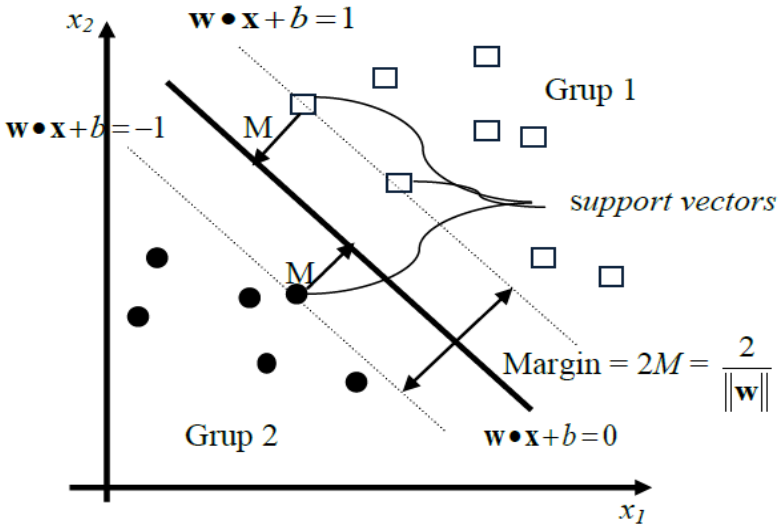

Figure 3 Illustration of Linear SVM

In some case, there is the dataset that cannot be separated linearly. Non-linear SVM can be used in this case. Figure 4 illustrates that dataset cannot be separated in a linear (left) and the transformation to a high-dimensional space (right). SVM is able to mapping the dataset to a higher-dimensional space using the kernel, so the dataset can be separated linearly in this space with the transformation ϕ [2]. The possible kernel functions in SVM [10]:

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \gamma\left((\boldsymbol{x}_i{}'\boldsymbol{x}_j) + b\right)^d \text{ (polynomial)} \quad (11)$$

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = exp\left(-\gamma|\boldsymbol{x}_i - \boldsymbol{x}_j|\right)^2 \text{ (radial basis function)} \quad (12)$$

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = tanh\left[v(\boldsymbol{x}_i\boldsymbol{x}_j) + b\right] \text{ (sigmoid)} \quad (13)$$

Where γ is gamma, d is a degree of polynomial.

In general, kernel method has two main parts. The first part is a module that transforms data from input space to high-dimensional space. The second part is an algorithm that serves to find a linear pattern in new space [6].

In a similar way as for the linear SVM, we can write hyperplane as following formula:

$$\boldsymbol{w}\boldsymbol{\phi}(\boldsymbol{x}) + b \quad (14)$$

with decision function in high-dimensional space:

$$f(\boldsymbol{x}) = sign(\boldsymbol{w}\boldsymbol{\phi}(\boldsymbol{x}) + b) \quad (15)$$

In non-linear SVM, $\boldsymbol{w}$ is linear combination of support vector in high-dimensional space.

$$\boldsymbol{w} = \sum_{i=1}^{n} \alpha_i y_i \phi(\boldsymbol{x}_i) \quad (16)$$

$\phi(x)$ must be sufficient following equation,

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{\phi}(\boldsymbol{x}_i)\boldsymbol{\phi}(\boldsymbol{x}_j) \quad (17)$$

$$\phi : \mathbf{x} \longrightarrow \phi(\mathbf{x})$$

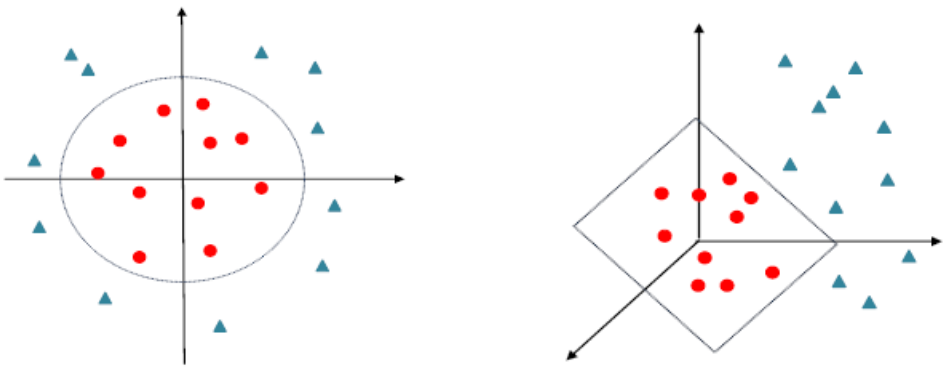

Figure 4 Illustration of Non-linear SVM

2.3 Boosting Support Vector Machines (BSVM) Algorithm

Misclassification caused imbalance dataset can be overcome by using boosting method. Generally, boosting is focuse on modeling the classification at each iteration

in this study using SVM. Each iteration is using different training dataset. At the first iteration, training dataset has same weight for each sample or object. But in the next iteration, the weight value of the incorrectly classified object are increasing and the weight value of the correct classified object are decreasing. Boosting SVM algorithm are as follow [9]:

1. Determine the initial weights of each observation, $p_1(x_i) = {}^1/_n$ for $i = 1, \ldots, n$
2. Suppose r is the number of iterations, then for $r=1,\ldots, R$ do the following:
   a) Train SVM ($f_r(x)$) with the weight value $p_r(x_i)$
   b) Calculate misclassification error ε_r

$$\varepsilon_r = \frac{\sum_{i=1}^{n} p_r(x_i)\, I(y_i \neq \hat{y}_i)}{\sum_{i=1}^{n} p_r(x_i)} \quad (18)$$

   c) Calculate α_r

$$\alpha_r = 1/2 ln\left(\frac{1-\varepsilon_r}{\varepsilon_r}\right) \quad (19)$$

   d) Update:

$$p_{r+1}(x_i) = \begin{cases} \frac{p_r(x_i)}{Z_r}\exp(-\alpha_r), f_r(x_i) = y_i \\ \frac{p_r(x_i)}{Z_r}\exp(\alpha_r), f_r(x_i) \neq y_i \end{cases} \quad (20)$$

   Where Z_r is normalization constant.
3. Final Classifier is

$$F(x) = sign\left(\sum_{r=1}^{R} \alpha_r f_r(x)\right) \quad (21)$$

2. 4 Performance Measurements

Performance of classification measured by confusion matrix as illustrated in Table 1. Confusion matrix is tabulation of the actual and prediction class. True positive (TP) is the number of correctly classified positive object, true negative (TN) is the number of correctly classified negative object, false negative (FN) is the number of misclassified positive object and false positive (FP) is the number of misclassified negative object.

Table 1 *Confusion Matrix*

Prediksi	Aktual	
	Positif	Negatif
Positif	*True Positive*(TP)	*False Positive*(FP)
Negatif	*False Negative*(FN)	*True Negative*(TN)

Sensitivity is the proportion of positive object that correctly classified and Specificity is the proportion of negative object that correctly classified.

$$Sensitivity = \frac{TP}{TP+FN}$$

$$Specificity = \frac{TN}{TN+FP} \quad (22)$$

Another performance measure that commonly used is AUC (Area Under ROC Curve) ranges 0-1. ROC curve is a plot of percentage of false positive (1- specificity) with the percentage of true positive [6].

3 Result and Discussion

This study used a dataset from survey research by Murniati, student of Graduate School Bogor Agricultural University, Department of Veterinary Public Health. Dataset are containing 243 data of cats, with 14 features (independent variables) and one response variable. The response variable in this study is the infection status of hookworm in cat with -1 as positive infected and 1 as negative infected. From 243 cats, there are 45 cats (18.5%) found positive infected hookworm and 171 cats (81.5%) were negative. Because of this imbalanced dataset between positive infected and negative infected, it would be solve by using SMOTE method and boosting method.

The first step is splitting dataset to training and testing dataset with proportion 0.7: 0.3, 171 training dataset and 72 testing dataset. Training dataset was applied to modeling and testing dataset to validation. After splitting, proportion of class dataset to be 19% for positive infected and 81% for negative infected.

The result of SVM method that applied on imbalanced dataset, about 100% for specificity and sensitivity are 0% (Table 2). Specificity value shows percentage of negative infected is correctly classified, so concluding that all negative infected objects are correctly classified. Sensitivity value indicates percentage of positive infected is correctly classified, so showing that none of positive infected objects are correctly classified. Sensitivity and specificity value are similar for all kernel function, i.e. linear, radial basis, polynomial and sigmoid. None of the kernel function can provide accuracy in positive infected class if there is an imbalanced dataset.

Table 2 Performance of classification SVM for all kernel function

Kernel	*Specificity*	*Sensitivity*	AUC
Radial (c=1, g=0. 026)	100.00%	0.00%	64.21%
Linear	100.00%	0.00%	**64.9%**
Polynomial (degree = 3)	100.00%	0.00%	64.86%
Sigmoid(default)	100.00%	0.00%	64.4%

Imbalance dataset is modified by applying SMOTE method on the training dataset, using information 5-nearest neighbor and over-sampled at 200% to produce training data as many as 224 with proportion of the categories to 43% (positive infected) and 57% (negative infected). After modified, training dataset was modeling using SVM method. SMOTE SVM method with linear kernel function was the best kernel function based on AUC and sensitivity value (Table 3). This method can be correctly classified 46.2% of total positive infected objects and 77.9% of total negative infected objects. This indicates that SMOTE methods apparently improvement in sensitivity, while decrease in specificity value at the same time.

Table 3 Performance of classification SMOTE SVM for all kernel function

Kernel	*Specificity*	*Sensitivity*	AUC
Radial (gamma = 0.026)	86.44%	38.46%	**60.50%**
Linier	77.96%	46.15%	**60.50%**
Polinomial (degree = 3)	100.00%	0.00%	50.10%
Sigmoid (default)	96.61%	0.00%	56.32%

The other method to overcome imbalanced dataset problem is boosting method. In boosting algorithm, incorrectly classified object was given higher weight than correctly classified objects. Analyzing Table 4, it is apparent that radial kernel function higher sensitivity value than the other kernel function and as well in AUC value. Sensitivity value is 69.23%, which meaning 69.23% positive infected object was correctly classified and specificity value is 55.85%, which meaning 55.85% negative infected object was correctly classified.

Table 4 Performance of classification Boosting SVM for all kernel function

Kernel	*Specificity*	*Sensitivity*	AUC
Radial (gamma=10)	55.93%	69.23%	**65.25%**
Radial (gamma=1)	52.54%	53.85%	55.74%
Linier	55.93%	53.85%	59.12%
Polinomial (degree = 3)	69.49%	23.08%	40.35%

Table 5 show the sensitivity, specificity, and AUC value achieved by varying method that used in this experiment. The best performance of classification method can be seen from the highest AUC value. Furthermore, the best classification method can be assessed from the highest sensitivity and specificity. Nevertheless, sensitivity was more important than specificity in this study. It would be harmful when the cats were predicted negative infected whereas it was positive infected. In other hand, the specificity was less important than the sensitivity. The cats which predicted positive infected but actually negative infected could make a good preventing for the infection. Based on these considerations, boosting SVM is the best method in this case.

Table 5 Comparison of AUC value and other performance measure

Method	Specificity	Sensitivity	AUC
SVM	100.00%	0.00%	64.9%
SSVM	77.97%	46.15%	60.6%
BSVM	55.93%	**69.23%**	**65.3%**

4 Conclusion and Remark

In imbalanced case, SVM method shows none of positive infected is correctly classified. By applying SMOTE and boosting method, sensitivity value is increasing, but boosting method gives higher sensitivity value. Moreover, AUC value of boosting method is the highest. Finally, it can be said that boosting SVM method is better than SMOTE SVM for classifying infection status of hookworm in cats if there are imbalanced dataset.

References

[1] Beveridge, I., and Jones, M. K., 2002, "Diversity and biogeographical relationships of the Australian cestode fauna", International Journal of Parasitology., 32, pp. 343-351.

[2] Burges, C., 1998, "A Tutorial on Support Vector Machine for Pattern Recognition, "Data Mining and Knowledge Discovery, 2, pp. 121-167.

[3] Brown, I., and Mues, C., 20012, "An Experimental Comparison of Classification Algorithms for Imbalanced Credit Scoring Data Sets, "Expert Systems with Applications, 39(3), pp. 3446-3453.

[4] Chawla, N. V., Browyer, K. W., Hall, L. O., and Kegelmeyer, W. P., 2002, "SMOTE : synthetic minority over-sampling technique", Journal of Artificial Intelligence Research., 16, pp. 321-357.

[5] Cortes, C., and Vapnik, V., 1995, "Support Vector Network, "Machine Learning, 20(3), pp. 273 - 297.

[6] Cristianini, N., and Shawe, T. J., 2004, Kernel Methods for Pattern Analysis, Cambridge University Press, Cambridge, UK.

[7] Fawcett, T., 2006, "An introduction to ROC analysis, " Pattern Recognition Letters, 27, pp. 861-874.

[8] James, G., Witten, D., Hastie, T., and Tibshirani, R., 2013, "An Introduction to Statistical Learning with Applications in R, " New York, Springer.

[9] Kim, H. C., Pang, S., Je, H. M., Kim, D., and Bang, S. Y., 2003, "Constructing support vector machine ensemble, "Pattern Recognition, 36, pp. 2757-2767.

[10] Meyer, D., 2014, "Support Vector Machines: The Interface to Libsvm in Package e1071, " FH Technikum Wien, Austria.

[11] Oktaviana, P. A., Dwinata, M., and Oka, I. B. M., 2014, "Prevalensi Infeksi Cacing *Ancylostoma spp* pada Kucing Lokal (*Felis Catus*) di Kota Denpasar, " Buletin Veteriner Udayana, 6(2), pp. 2085-2495.

[12] Steinberg, D., and Colla, P., 1995, CART: Tree-Structured Nonparametric Data Analysis, Salford Systems, San Diego.

[13] Yap, B. W., Khatijahhusna, A. R., Hezlin, A. A. R., Simon, F., Zuraida, K., and Nik, N. A., 2014, "An Application of Oversampling, Undersampling, Bagging and Boosting in Handling Imbalanced Datasets, "Proceedings of the First International Conference on Advanced Data and Information Engineering, Singapore.