An Improved Classification Algorithm Applied on Landslide Dam Disaster Events Detection

Bai Hua[1], Wang Shaoyu[2]
[1] School of Management, Harbin Institute of Technology
[2] School of Architecture, Harbin Institute of Technology
Harbin, P. R., 150001, China
baihua1727@163.com

ABSTRACT: *Landslide dam is formed when a river is blocked by some kind of mass wasting such as the debris and rocks. Landslide dams frequently fail soon and lead to upstream and downstream flooding, which could cause high casualties and economic losses. So it is important to predict landslide dam stability for reasonable subsequent disposal. This study proposes an improved model for landslide dam disaster events detection based on Support Vector Machine and Ridge Regression. The improved model introduces Support Vector Machine method into traditional Ridge Regression algorithm and get an combinational algorithm, named Combinational Ridge Regression- Support Vector Machine(CRR-SVM) algorithm. This research chooses a record dataset about landslide dam's variables to test the effectiveness and superiority of the new method; experiment result shows that the boosting approach is more effective than previous methods.*

Keywords: Landslide dam stability, Disaster events detection, Support vector machine (SVM), Ridge regression (RR), CRR-SVM model

Received: 10 April 2015, Revised 20 May 2015, Accepted 28 May 2015

1. Introduction

Landslide dams are a common phenomenon. They form when a landslide reaches the bottom of a river valley, causing a blockage (Ermini and Casagli, 2003). Landslide dams occur in numerous regions worldwide and result in considerable flooding hazards which could pose a higher risk to the upstream and downstream area .

Landslide dams occur in many countries all over the world, they have formed when some blockage reaches the bottom of a river (Ermini and Casagli, 2003) and also might pose a higher risk of casualties and property losses to the nearby area. After its formulation, the nature lake often fail in a short time as a result of the outburst flood (Schuster and Costa 1986). Regarding the enormous flooding disaster caused by landslide dam's failure, a rapid assessment method for its stability is very necessary (Schuster and Costa, 1986). However, the material characteristics influencing their stability are hardly to get in short time (Casagli et al., 2003), thus a lot of geomorphic approaches are widely presented. Ermini and Casagli (2003) provided an geological index called Dimensionless Blockage Index(DBI) to predict the stability of landslide dam. This index combined three landslide dam geomorphic variables including dam's height, volume and landslide lake's area. But the subjective selection process of the variables resulted in a limited separated accuracy of 64.9%. Korup (2005) suggested that the risk level of landslide dam failing could be predicted by the function of some spatial variables. Dong et al.(2009) used discriminate analysis method to identify the

key variables influenced landslide dam stability based on the Japanese dataset (Tabata inventory). He constructed an discriminated regression model to rapidly evaluate the stability of landslide dams. The accuracy of his model was about 70.1 %. Furthermore, Dong et al.(2011) proposed the logistic regression model for quick assessment of landslide dam stability. In comparison with discriminate analysis method, the logistic approach got a little better ability to classify the landslide dams into stable and unstable groups(76.9%). Nevertheless, logistic regression generally works on equal size groups. Besides, there is obviously collinearity between the variables constructed logistic regression model, this collinearity could lead to the regression coefficients unreliable.

In order to obtain a higher detection accuracy for the landslide dam stability to mitigate the catastrophic consequence effectively, this paper considered the characteristic of data then introduced Support Vector Machine(SVM) non-linear transform method into traditional Ridge Regression(RR) model, boosting a novel Combinational Ridge Regression- Support Vector Machine(CRR-SVM) algorithm for landslide dam stability detection.

Support vector machine is a research focus of machine learning, which has recently been extensively applied in the field of classification and detection. Although the basic theory of SVM has been well established, there are still a lot of problems left unresolved, such as the collinearity properties among multiple variables, classification about imbalance dataset, etc.. These problems broadly exits in real-world application and especially in some disaster events. for instance, landslide dams, flooding and so on. To work out the above mentioned difficulty of SVM and significantly improve the classification performance, a boosting algorithm based on SVM and RR was presented.

The proposed novel algorithm integrates the essential features of both SVM and RR, thus the negative influence of multiple collinearity and unbalanced characteristic is reduced greatly.

In this study, an improved detection model is designed and studied to resolve the defect in traditional methods for the detection of landslide dam stability. The experimental results show that the predictive power of the CRR-SVM model is better than previous method such as DBI, discriminate analysis and logistic regression analysis.

2. Methodology

Support Vector Machine (SVM)

Support vector machine(SVM) is a promising method for both linear dataset and nonlinear dataset classification(Xie, L. and Q. Liu, 2011). The nonlinear mapping SVM method transforms the original data into higher dimensions then the Optimal Hyperplane could be considered to be a classific boundary of the dataset. Thus the key of the classification is to find the Optimal Hyperplane by searching for maximum marginal hyperplane. Given a original dataset as follows: (x_i, y), $i = 1, 2, \ldots n$, where $x_i \in R^n, y_i \in \{1, -1\}$, $i = 1, 2, \ldots n$. In this dataset, x_i represent the training set with respective classific labels $y_i \in \{1, -1\}$.

As for a linearly separable dataset, the hyperplane could be written as $W, X + b = 0$, where b is a bias, W means Weight Vector, $W = \{ w_1, w_2, \ldots, w_n$, n represents attributes' number. This hyperplane also could be expressed as follows: $w_1x_1 + w_2x_2 + w_ix_i + w_0 = 0$, $i = 1, 2, \ldots n$. Then, the weights should be adjusted so that the hyperplanes can be written as follows:

$$H_1 : w_1x_1 + w_2x_2 + \cdots w_nx_n + w_0 \geq 1 \qquad i = 1,2,\ldots n \text{ for } y_i = 1 \tag{1}$$

$$H_2 : w_1x_1 + w_2x_2 + \cdots w_nx_n + w_0 \leq -1 \qquad i = 1,2,\ldots n \text{ for } y_i = -1 \tag{2}$$

Combining the two inequalities of Equations (1) and(2), then get

$$y_i(w_1x_1 + w_2x_2 + \cdots w_nx_n + w_0) \geq 1 \qquad i = 1,2,\ldots n \tag{3}$$

The training data which fall on H_1 or H_2 hyperplanes and satisfy Equation (3) could be called SV(Support vectors). These support vectors are equally close to MMH(Maximum Marginal Hyperplane). Furthermore, based on the maximal margin $\frac{2}{\|W\|}$ and Lagrangian formulation mentioned, the MMH could be written as follows:

$$d(X^T) = < w \cdot x > + b = \sum_{i \in SV} y_i \alpha_i X_i X^T + b \tag{4}$$

Where y_i is the classific label of X_i, X^T means a test tuple, a_i and b_0 are parameters. While, considering the non-perfectly separable case, introducing ξ_i to be the slack variable could minimize the errors. What's more, SVM use another math trick to solve the classific problem of nonlinear separable dataset. It happens that the training tuples appear as scalar product $\phi(X_i).\phi(X_j)$. Instead of calculating the scalar product complicated, SVM method applies the kernel function $K(X_i, X_j)$ on the original dataset to reduce computation complexity. That is

$$K(X_i, X_j) = \phi(X_i) \cdot \phi(X_j) \tag{5}$$

Therefore, the main idea of Support Vector Machine (SVM) algorithm is as follows:

$$\min_{w,b,\xi} \frac{1}{2} w^{'} w + c \sum_{i=1}^{n} \xi_i \tag{6}$$

Subject to

$$y_i (w^{'} \phi(x_i) + b) \geq 1 - \xi_i, \ \xi_i \geq 0 \tag{7}$$

Where $b, c \in R, w, \phi(x_i) \in R^m$, $\phi : R^n \to R^m$. Thus, for any test data $x_i \in R^n$, $y_i \in \{1, -1\}$,

$$d(x_i) = \left[w^{'} \phi(x_i) + b - (1 - \xi_i) \right]$$

$$y_i = \begin{cases} +1, & if \quad d(x_i) \geq 0 \\ -1, & if \quad d(x_i) < 0 \end{cases} \tag{8}$$

Ridge Regression(RR)

Ridge Regression (RR) is a partial prediction method which have be used widely to solve the problem of multiple collinearity among the independent variables (Li and Niu, 2013). In order to describe the Ridge Regression method, given a dataset $S = \{x_i, y_i\}_{i=i}^{l}$, $x \in R^n, y \in R$, then multiple linear regression model could be defined as follows:

$$y = f(x) = X\beta + e \tag{9}$$

Where y means observations, β represents regression coefficient, X is about attribute matrix, e is random variable and follows multivariate normal with the mean vector 0 and variance-covariance matrix $\delta^2 I_n$. The Least Squares Estimate of β could be written as $\hat{\beta} = C^{-1} X^{'} y$, $C = X^{'} X$,. Thus, it can safely conclude that the estimate of β is greatly depend on C. Especially, if C is ill conditioned $(X^{'} X \approx 0)$, the Least Squares Estimate of would be likely to cause several errors. Hoerl and Kennard(1970) proposed a useful method using $C_{(k)} = C + kI_p (k \geq 0)$ to instead of C to settle this problem. According to the above discussed, the estimated β as follows:

$$\hat{\beta}_{(k)} = (C + kI_p)^{-1} X^{'} y \quad k > 0 \tag{10}$$

This is the core idea of ridge regression estimate, where k is generally named Ridge Parameter or Biasing (Orsenigo and Vercellis, 2012; Kibria et al., 2012).

The improved detection model: Combinational Ridge Regression- Support Vector Machine(CRR-SVM)

The previous research about landslide dam stability ignored the unbalance feature of the dataset and collinearity between variables. The traditional study about improving of ridge regression analysis are all considered RR as a way to develop non-linear model more effective. Aiming to effectively predict these small non-linear samples, this study proposed a new approach called CRR-SVM based on the combination of RR and SVM to address the classific detection puzzle. $d(x_i)$ represents the signed distance between x_i and the hyperplane and it could be provided by SVM method. Then consider $d(x_i)$ to be the input variables to construct another ridge regression model to predict the classific labels. As for a given sample dataset $(x_{i1}, x_{i2}, \cdots x_{in}, y_i), i = 1,2,\cdots n$, $x_i = (x_{i1}, x_{i2}, \cdots x_{in}) \in R^n, y_i \in \{1,-1\}$, above SVM method provides the signed distance $d(x_i) = \left[w^{'}\phi(x_i) + b - (1-\xi_i)\right]$ Then, construct another working dataset $(d(x_1), y), i = 1,2,\cdots n$, Where $d(x_i) \in R, y \in \{1,-1\}$.

Further, the usual RR model is changed as follows:

$$y = f(d(x)) = d(X)\beta + e \tag{11}$$

Similar to the general ridge regression model, The Least Squares Estimate of β could be written as $\hat{\beta} = C^{-1}d(X)^{'}y$, where $C = d(X)^{'}d(X)$, if $d(X)^{'}d(X) \approx 0$. Then,

$$\hat{\beta}_{(k)} = (C + kI_p)^{-1}d(X)^{'}y \quad k > 0 \tag{12}$$

$$\hat{y} = \frac{\beta_{(k)}}{(C + kI_p)^{-1}d(X)^{'}} = \hat{\beta}_{(k)}(C + kI_p)^{-2}\left[d(X)^{'}\right]^{-1} \tag{13}$$

Next define

$$\bar{y} = \quad mean \quad \hat{y} \tag{14}$$

Finally,

$$y_i = \begin{cases} +1, & if \quad \hat{y} \geq \bar{y} \\ -1, & if \quad \hat{y} < \bar{y} \end{cases} \tag{15}$$

In summary, the main idea of new combination algorithm based on SVM and RR could be described as follows:

- Identify reasonable kernel function (Gaussian Function);
- Carry out SVM algorithm to transform x_i into $d(X)$;
- Consider $(d(x_1), y), i = 1,2,\cdots n$ as a new dataset to construct a RR model based on partial Least Square Technique (LST);
- Identify reasonable *k*, and estimate $\hat{\beta}$ by Equation. (12);
- Classify y_i on the basis of Equation. (15).

Figure. 1 shows the structure graph of the above proposed CRR-SVM model.

3. Results and Discussion

Dataset

In order to verify the efficiency and practicality of the above new boosting model, this study chose a real-world dataset about landslide dams' stability to verify the improved model's performance. Tabata et al. (2002) studied about 79 Japanese landslide dams and documented 16 characteristics of the dams. Based on this record Dong et al.(2009) statistically analyzed the key elements of landslide dam stability and constructed a discriminate model utilizing the known dominant variables. According to his study, the key variables affecting landslide dam stability included catchment area(A), dam height(H) and dam volume(V). Further, Dong et al.(2011) proved the significance of these three variables once again and built another logistic regression model for the classification of landslide dam stability(into stable group and unstable group). Therefore, this study continues to

introduce these dominant variables to classific landslide dam stability.

In order to compare the performance of the proposed combination detection model with the traditional method, a worldwide data record of landslide dams (84 landslide dames documented by Ermini and Casagli, 2003) is used to be the training dataset(Table.1). As shown in Table 1, aiming to make a reasonable classification, the stable landslide dams were assigned to value 1, and the unstable landslide dams were assigned to value -1.

Although Tabata investigated a total of 79 landslide dams in Japan, there were only about 50 dams completely reported about the catchment area(A), dam height(H) and dam volume(V). In addition, there were still 13 dams among these 50 landslide dams well documented by Ermini and Casagli. So the rest 37 landslide dams in Japan was taken as our target dataset(Table.2). Similar to the above Table 1, each dam including in this dataset has three variables and a label(stable=1,unstable=-1). Different methods could be compared by dams' label prediction accuracy in this target dataset..

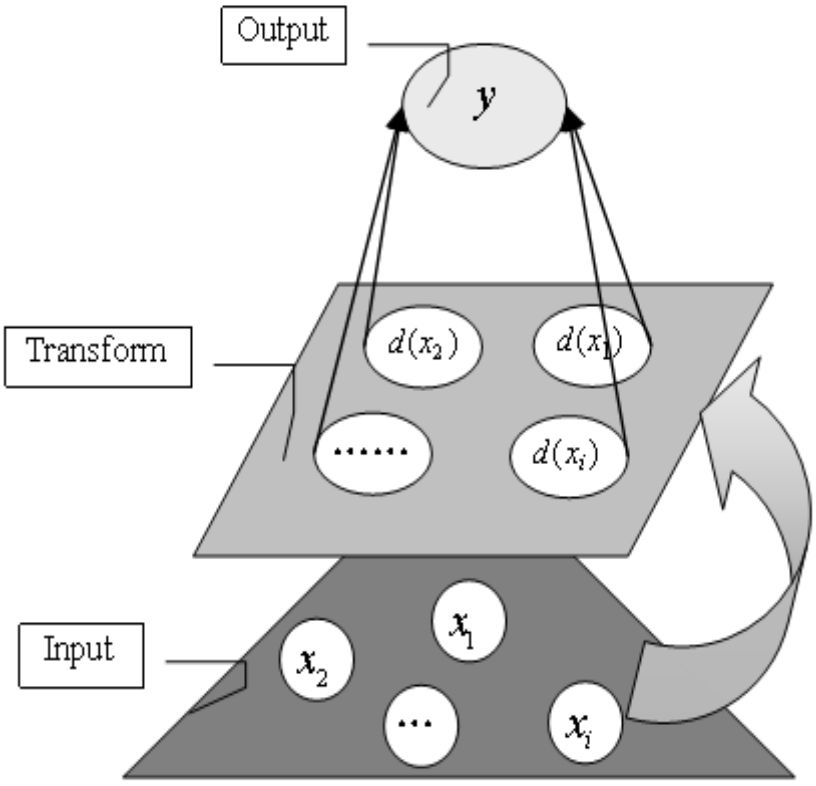

Figure 1.The structure graph of Combinational Ridge Regression- Support Vector Machine(CRR-SVM) model

Catachment Area ($10^6 m^2$)	Dam Height (m^3)	Dam Volume ($10^3 m^3$)	Class (Stable = 1 Unstable = -1)
1111	100	30000	-1
68	110	24000	-1
.....	.....	.....	.....
14.5	20	22	1

Table 1. The three variables and class of training set(84 landslide dames documented by Ermini et al.)

Results and discussion

Previous research about this area identified three dominant geomorphic variables including catchment area (A), dam height (H) and dam volume (V) effecting the dam's stability obviously, however, there was serious collinearity between dam height (H) and dam volume (V), what's more, landslide dams are typical nonlinear phenomenon, which could largely effect the predictive power of traditional logistic regression analysis result. It is shown by tests that the novel algorithm successfully combine two methods, which is better than one method only, thus inadequacy is avoided.

According to the new model above mentioned, this work depended on the software LIVSVM to carry out SVM algorithm for transforming original input variables and then made use of statistical tool SPSS to run ridge regression analysis. It is should be noted that regarding regarding the nonlinear transformation by SVM, the most important factors are Kernel Function and parameters. Considering the characteristics of dataset and the general performance of each Kernel Function(including Linear Kernel Function, Polynomial Kernel Function, Radial Basic Function) on the nonlinear mapping capacity, Radial Basic Function(RBF) is more suitable as the rational Kernel Function of our transformation. However, aiming to satisfy the input requirement, the original data should be standardized at first. Conveniently, LIBSVM package provides a useful available tool SVM-SCALE to do this.

Catachment Area ($10^6 m^2$)	Dam Height (m^3)	Dam Volume ($10^3 m^3$)	Class (Stable = 1 Unstable = -1)
5	20	1500	1
15	15	400	1
.....	.....	.....	.....
147	90	18000	-1

Table 2. The three variables and class of target set(37 landslide dams documented by Tabata)

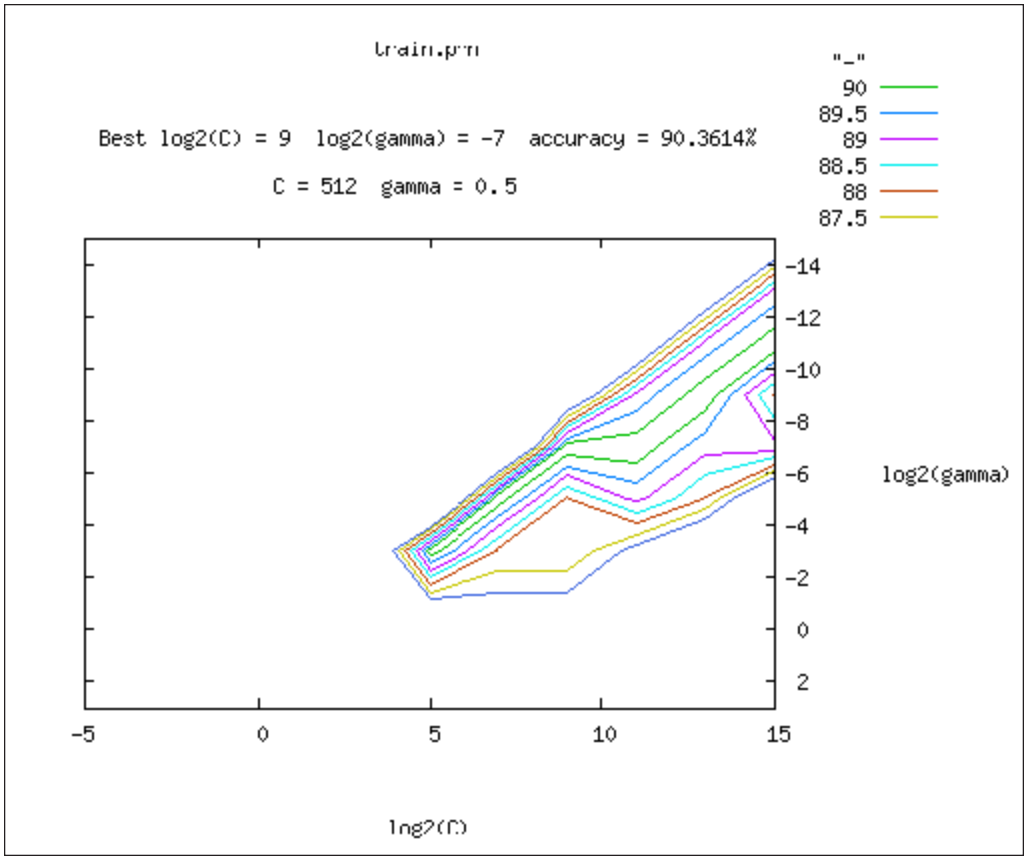

Figure 2. The graph of optimize parameters for Support Vector Machine

What's more, as for how to find the optimal parameters for SVM, there still isn't a best way recognized all over the world. So far, according to previous research, k-fold Cross Validation is the most common way to search for the optimal parameter values of c and ³. At first, the certain parameter values were given by LIBSVM package program automatically. As Figure.2 shows, Using k-fold Cross Validation methods, this research took the target dataset as original dataset to get the classical accuracy under the first given parameter values of c and γ. After 7-fold Cross Validation, the accuracy reached to be the highest value, 90.34%. So this parameter group was the best choice. In this research, the optimal penalty was $c = 510$ and the optimal parameter was $\gamma = 0.5$.

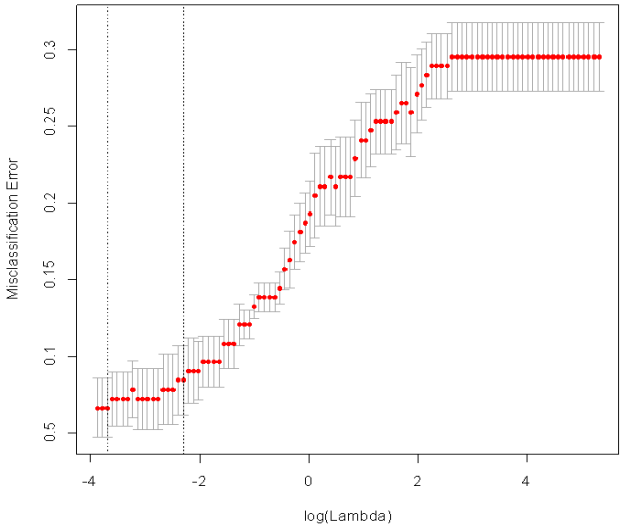

Figure 3. The estimated graph of Ridge Regression

Supported by the theory of Ridge Regression, statistic software R was used to perform landslide dam stability by input the distance between variables and their hyperplane in SVM. As is shown in Fig. 3, when $\log(\lambda)$ stabilized around -3.70, the boosting model achieve the best accuracy 93.0%.

Then compared the classific ability between this new method and previous approaches according to the same target dataset. Experiment result showed that the overall detection power(successful prediction rate) of the new approach CRR-SVM model is more effective than the traditional method. Compared with the DBI, discriminate analysis and logistic regression analysis which have been proven powerful by the same dataset, the above proposed new algorithm in this study performs better. As is shown in Table.3, the simple DBI method only got 64.90% accuracy. Discriminate analysis and Logistic regression analysis performance a little better but still less than satisfactory- at 70.10% and 76.90% respectively. In comparison, Combinational Ridge Regression-Support Vector Machine got the highest accuracy 87.95%.

Classifier	**Accuracy**
Dimensionless Blockage Index	0.6490
Discriminate analysis	0.7010
Logistic regression analysis	0.7690
Combinational Ridge Regression-Support Vector Machine	0.8795

Table 3. Comparison of four classifiers' accuracy

4. Conclusion

Rapid assessment about landslide dam stability is a crucial challenge for rescue and disaster mitigation. However, the real-world landslide dam dataset is a small sample and has obvious collinearity and imbalance features, which might limit its stability detection. Thus, this study proposed a combination model based on SVM and RR to quantitatively evaluate landslide dam stability. According to previous studies, by selection the most significant variables influencing landslide dam stability(A,H,V).

The performance of proposed CRR-SVM model and the traditional method(DBI, discriminate analysis and logistic regression analysis) is compared by the training set (84 worldwide landslide dams)and target set(37 Japanese landslide dams). Experiment result shows that the proposed detection model is useful and performs better than previous research approaches.

Besides, the above new detection method in this research can be used to classify not only the data of landslide dams stability but also other small sample. Considering this novel algorithm has combined the advantage of both SVM and RR, it will have particularly perfect performance in imbalance, multiple collinearity, even high dimension sample data. However, its application is restricted to small sample, thus, some further improving method to strengthen the generalization ability of CRR-SVM should be studied in the future.

In summary, the proposed model could be used as an important rapid prediction tool to make decisions about landslide dam hazard mitigation, particularly in a limited decision time. What's more, the method developed herein could be used to predict other potential disasters; this is also what should be done in the future. As for the CRR-SVM model itself, its application is restricted to small sample; thus we would consider searching for some further improving measures to strengthen its generalization ability.

References

[1] Ermini, L., Casagli, N. (2003). Prediction of the behaviour of landslide dams using a geomorphological dimensionless index. *Earth Surf. Process Landforms*, 28. 31-47.

[2] Schuster, R. L., Costa, J. E. (1986). A Perspective on Landslide Dams. In: Landslide Dams: Processes, Risk and Mitigation. *American Society of Civil Engineers*, USA., 1-20.

[3] Casagli, N., L. Ermini and G. Rosati, (2003). Determining grain size distribution of the material composing landslide dams in the Northern Apennines: Sampling and processing methods. *Eng. Geol*., 69, 83-97.

[4] Korup, O. (2005). Geomorphic hazard assessment of landslide dams in South Westland, New Zealand: Fundamental problems and approaches. *Geomorphology*, 66,168-188.

[5] Tabata, S., Mizuyama, T., Inoue, K. (2002). Natural Landslide Dams Hazards. Kokonshoin, Tokyo, *Japan*, 45-57.

[6] Dong, J. J., Tung, Y. H., Chen, C. C., Liao, J. J., Pan, Y. W. (2009). Discriminant analysis of the geomorphic characteristics and stability of landslide dams. Geomorphology, 110, 162-171.

[7] Dong, J.J., Tung, Y.H., Chen, C.C., Liao, J. J., Pan, Y. W. (2011). Logistic regression model for predicting the failure probability of a landslide dam. *Geol*., 117, 52-61.

[8] Hoerl, A., Kennard, R.W. (1970). Ridge regressionÿbiased estimation for non-orthogonal problems.*Technometrica*, 12, 55-88.

[9] Xie, L., Q. Liu. (2011). Integrated binary-class classification algorithm based on logistic and SVM. *Comput. Eng. Appl*., 47 (29) 149-150.

[10] Li, G., Niu, P. (2013). An enhanced extreme learning machine based on ridge regression for regression. *Neural Comput. Appl*., 22, 803-810.

[11] Orsenigo, C., Vercellis, C. (2012). Kernel ridge regression for out-of-sample mapping in supervised manifold learning. *Expert Syst. Appl*., 39, 7757-7762.

[12] Kibria, B.M.G., Mansson, K., Shukur, G. (2012). Performance of some logistic ridge regression estimators. *Comput. Econ.,* 40, 401-414.