Investigation on the Effect of Data Imbalance on Prediction of Liquefaction

Javad Sadoghi Yazdi[1]; Farzin Kalantary[2]; and Hadi Sadoghi Yazdi[3]

Abstract: Data imbalance causes learning bias in class identification techniques. A major cause for limited success in the prediction of liquefaction potential by various pattern recognition techniques is because of a liquefaction to nonliquefaction data class imbalance. It is suggested to use a support vector data description (SVDD) strategy to compensate the minority data. SVDD is used to generate virtual data points for the minority class bearing the same characteristics as the nonvirtual samples. Then an adaptive neuro-fuzzy inference system (ANFIS) classifier is employed to determine the liquefaction threshold. The ANFIS predictions are then examined by evaluating the coefficient of determination (COD) and comparing it with the Bayesian updating method. It is shown that for the liquefied data the approach is as efficient as the Bayesian method, but great improvement in the recognition rates of the nonliquefied data have been achieved. **DOI: 10.1061/(ASCE)GM.1943-5622.0000217.** *© 2013 American Society of Civil Engineers.*

CE Database subject headings: Soil liquefaction; Data analysis; Predictions.

Author keywords: Liquefaction; Data imbalance; Support vector data description; Adaptive neuro-fuzzy inference system.

Introduction

Determination of liquefaction potential of soils is of major concern and an essential criterion in the design process of civil engineering projects. Over the past 30 years, many researchers have endeavored to present various methods for the prediction of liquefaction potential of soils.

Among the situ tests, many researchers have adapted cone penetration test (CPT) results as the basis for the evaluation of liquefaction potential because of the superior nature of the test method (Juang et al. 2003; Youd and Idriss 2001). Also, the estimation of elastic constants based on CPT and standard penetration test (SPT) data have been extended (Weiher and Davis 2004).

More intricate approaches based on constitutive modeling (Mróz et al. 2003), artificial neural networks (ANN), fuzzy logic, and probabilistic analyses have been introduced. Moss et al. proposed liquefaction models that use the Bayesian updating method for CPT data (Moss et al. 2006). The relative state parameter index, (ξ_R), is used for probabilistic correlation between laboratory and field liquefaction potentials (Jafarian et al. 2010). Evolutionary polynomial regression (EPR) is used for the evaluation of liquefaction potential (Rezania et al. 2010) and later on is developed into an evolutionary-based approach for the assessment of earthquake-induced soil liquefaction and lateral displacement and presents a formulation in three-dimensional space of cyclic stress ratio, cone penetration, and effective overburden for the prediction of liquefaction potential (Rezania et al. 2011). A support vector machine was developed for use as an alternative deterministic and probabilistic empirical liquefaction model (Oommen et al. 2010). However, the issue of class imbalance is still an obstacle for all pattern recognition techniques. It is therefore proposed to use a support vector data description to define a class sphere and thus determine outliers in the first instance and then use the Monte Carlo technique to randomly compensate the minority class. A full description of the support vector data description (SVDD) is provided by Tax and Duin (1999, 2004).

Having managed the issue of class imbalance, an adaptive neuro-fuzzy inference system (ANFIS) is then used as the identification technique for the determination of the liquefaction threshold. *MATLAB 7.12.0.635 (R2011a)* toolboxes were used for ANFIS.

ANFIS is adapted using CPT-based liquefaction case histories compiled by Moss et al. (2006). The CPT database adapted here has 182 case histories of which 139 are from liquefied sites and 43 are from nonliquefied sites. This database falls within the category of imbalanced data sets because the ratio of liquefied to nonliquefied instances is more than 3. The issue of data imbalance has been recognized and an attempt to alleviate its negative effects by using Bayesian updating optimization has been presented. (Cetin et al. 2002).

Methodology

The basic steps in this methodology include feeding the CPT data into the SVDD to produce various descriptions of data ranges by applying different data region descriptions, and then for each of the determined minority class data spheres, the appropriate upsampling is carried out. The ANFIS classifier is then employed to determine the optimum data description providing the best possible recognition rate.

The liquefaction data shown in Fig. 1(a) are fed into the SVDD and a data region defined by model description parameters [width of Gaussian kernel (σ) and penalty coefficient (C) that is a constant, which determines the trade-off between the hypersphere volume and outliers] is obtained [Fig. 1(b)]. The nonliquefaction model is similarly developed [Figs. 1(c) and (d)].

The SVDD encloses both liquefaction and nonliquefaction regions and thereby detects outliers. Each SVDD for liquefaction

[1] Postgraduate Student, Civil Engineering Dept., K. N. Toosi Univ. of Technology, 19697 Tehran, Iran. E-mail: javad.sadoghiyazdi@mymail.unisa.edu.au

[2] Assistant Professor, Civil Engineering Dept., K. N. Toosi Univ. of Technology, 19697 Tehran, Iran. E-mail: fz_kalantary@kntu.ac.ir

[3] Associate Professor, Computer Dept., Ferdowsi Univ. of Mashhad, 9177948974 Mashhad, Iran (corresponding author). E-mail: h-sadoghi@um.ac.ir

Note. This manuscript was submitted on May 17, 2011; approved on March 7, 2012; published online on March 12, 2012. Discussion period open until January 1, 2014; separate discussions must be submitted for individual papers. This technical note is part of the ***International Journal of Geomechanics***, Vol. 13, No. 4, August 1, 2013. ©ASCE, ISSN 1532-3641/2013/4-463–466/$25.00.

and nonliquefaction can be used for the determination of the status of the input sample relative to the obtained enclosed region; inside, outside, or on-boundary is the status of the input samples. The status is reported, respectively, with negative, positive, or zero values.

Support Vector Data Description-Based Data Generation

In view of the fact that the ratio of liquefied sample points to nonliquefied samples is greater than 3, the imbalance between the different data classes will have an adverse effect on the pattern recognition procedure. Therefore, in the nonliquefied class region identified by the SVDD, data are generated (Fig. 1).

Both Monte Carlo and the SVDD models are jointly used to generate the data needed to remove the imbalance. A probability density function is generated using Monte Carlo for initial data generation in accordance with the determined center and the width of the minority class region. The result is shown in Fig. 2.

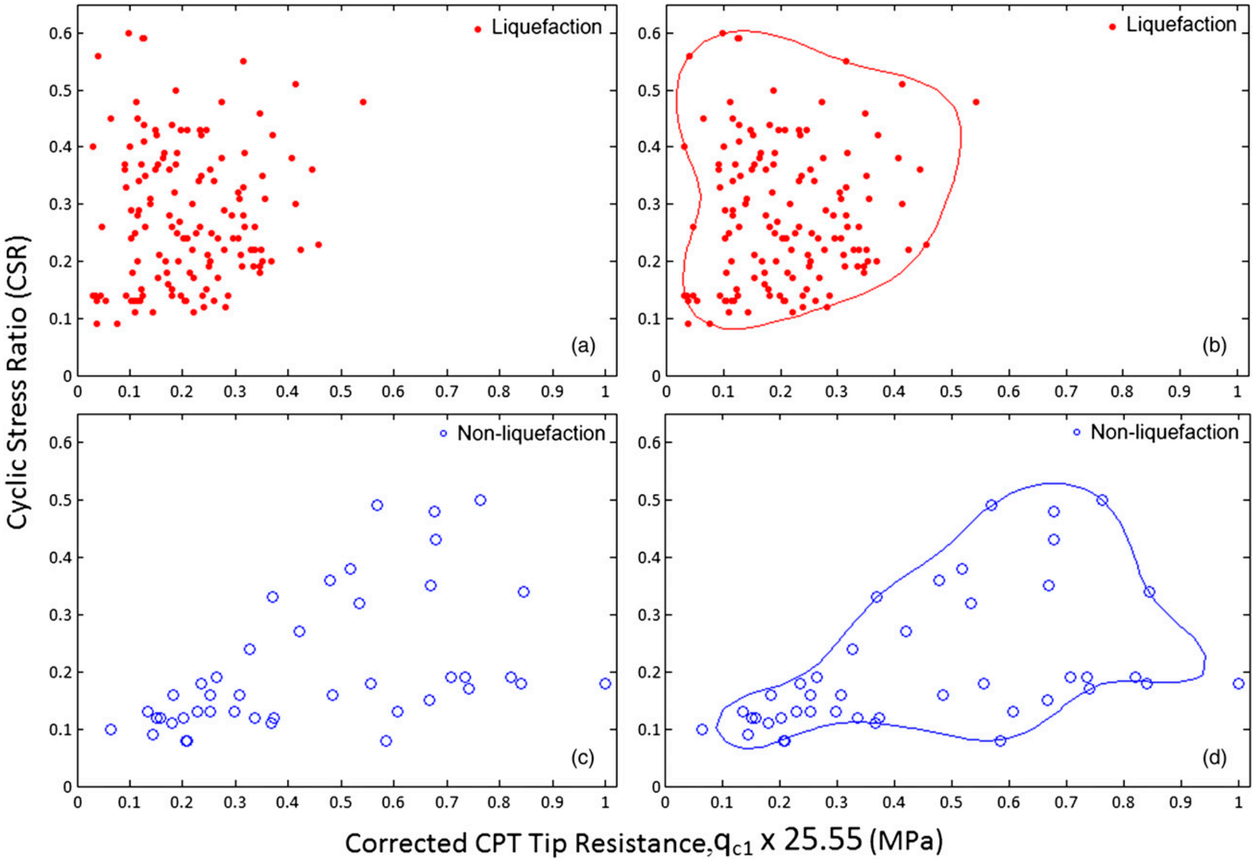

Fig. 1. (a) Liquefied data; (b) surface obtained using SVDD for liquefaction data; (c) nonliquefied data; (d) obtained surface using SVDD for nonliquefaction data

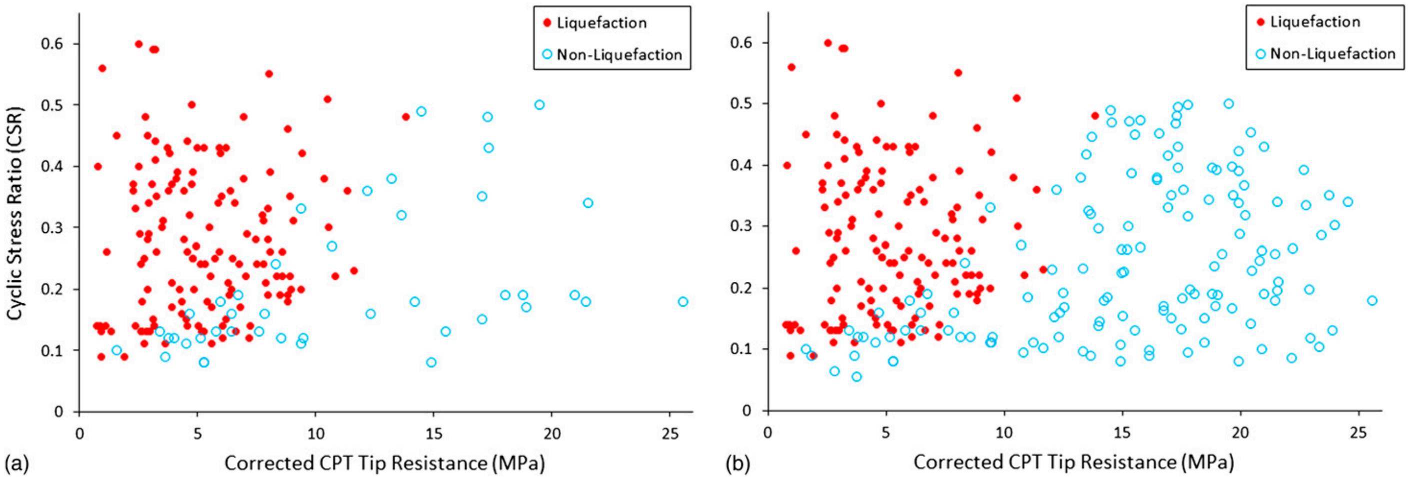

Fig. 2. (a) CPT-based case histories in $CSR - q_{c1}$ space; (b) upsample CPT data

Classification Using Adaptive Neuro-Fuzzy Inference System

An ANFIS classifier is then used to predict soil liquefaction. A K-fold cross-validation method is used for training and testing of the model. In K-fold cross-validation, the data are randomly split up into K partitions and then (K-1) folds are used for training and the remaining fold is used for validation. This process is repeated K times, leaving one different fold for evaluation each time. The ability of each model to predict is estimated by calculating the errors on each test instance of each K-fold. The advantage of the K-fold cross validation is that all the of examples in the data set are eventually used for both training and validation, yet for each example in the data set, training and validation are implemented independently (Oommen et al. 2010).

Table 1. Comparison of Average of Recognition Rate for Two Gaussian Kernels Widths σ ($\sigma = 0.15$ and 0.25) and Four C Values ($C = 0.05, 0.15, 0.25,$ and 1)

SVDD parameter	Train	Test
$C = 0.05$, $\sigma = 0.15$	91.65	90.69
$C = 0.05$, $\sigma = 0.25$	92.13	90.53
$C = 0.15$, $\sigma = 0.15$	92.48	90.61
$C = 0.15$, $\sigma = 0.25$	92.17	90.97
$C = 0.25$, $\sigma = 0.15$	92.52	91.97
$C = 0.25$, $\sigma = 0.15$	92.26	90.52
$C = 0.5$, $\sigma = 0.15$	92.21	90.96
$C = 0.5$, $\sigma = 0.25$	92.60	91.01
$C = 1$, $\sigma = 0.15$	92.79	91.54
$C = 1$, $\sigma = 0.25$	92.55	91.05

Table 2. Values of a_{1i}, b_{1i}, a_{2i}, b_{2i}, a_i^*, and b_i^*

I	a_{1i}	b_{1i}	a_{2i}	b_{2i}	a_i^*	b_i^*	c_i^*
1	2.628	6.33	0.0562	0.21	0.0481	−6.72	2.187
2	2.628	17.0745	0.0562	0.38	−0.0262	−0.954	−0.167
3	2.628	4.8	0.0562	0.37	0.0066	−0.299	1.119
4	2.628	16.597	0.0562	0.1986	0.0209	−1.675	−1.041
5	2.628	6.45	0.0562	0.13	−0.2454	11.55	−0.465
6	2.628	2.66	0.0562	0.13	−0.2709	16.09	−0.902
7	2.628	22.84	0.0562	0.225	0.00049	0.6018	−1.141
8	2.628	16.382	0.0562	0.5161	−0.2084	2.202	1.692

Recognition Rate

At this stage the recognition rates produced by ANFIS are compared for various SVDD parameters. It should be noted that except for σ = 0.05 values, which give discrete and multiple segment data boundaries, the other values have been tried to determine the best recognition rate. ANFIS was run 10 times for two Gaussian kernels widths σ (σ = 0.15 and 0.25) and four values of C ($C = 0.05, 0.15, 0.25,$ and 1), and the mean values of training and test procedures are evaluated. The outcome is shown in Table 1.

Based on the previous results, $C = 0.25$ and $\sigma = 0.15$ are chosen.

Mathematical Definition of Threshold

The methodology of defining the liquefaction threshold (f) by ANFIS classifier is described subsequently

$$f(q_{c,1}, CSR) = \sum_{i=1}^{8} \omega_i z_i \Big/ \sum_{i=1}^{8} \omega_i \tag{1}$$

where ω_i and z_i are obtained as follows:

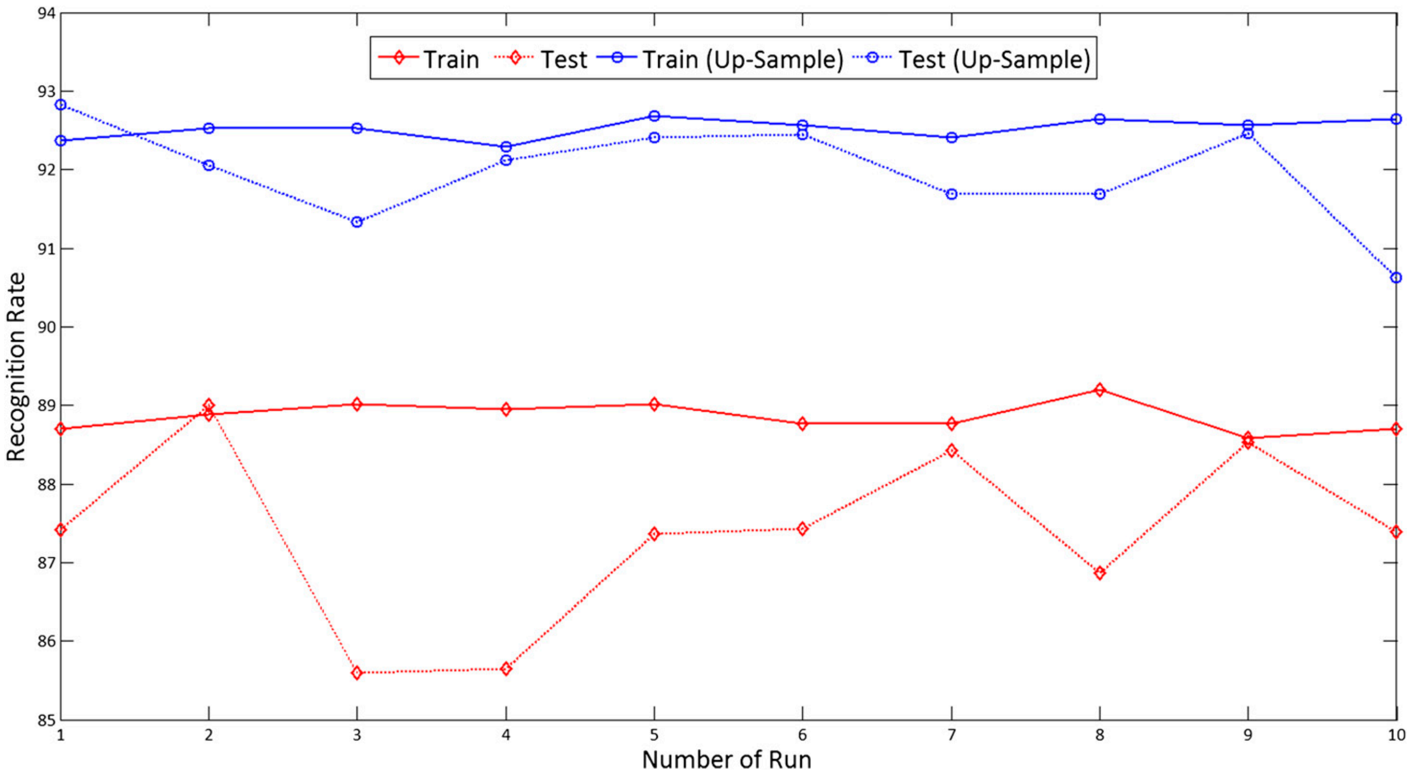

Fig. 3. Recognition rate for testing and training imbalance and balance data

Table 3. COD Value for Bayesian Updating and ANFIS Upsample Approach

Approach		Data set of Moss et al. (2006) COD of liquefied data	 COD of nonliquefied data
Moss et al. (2006)	$\mathrm{TH_L} = 5\%$	1	0.75
	$\mathrm{TH_L} = 15\%$	0.9997	0.7732
ANFIS upsample		0.9966	0.9183

$$\omega_i\left(q_{c,1}, CSR\right) = \exp\left(-\frac{\|q_{c,1} - b_{1i}\|}{a_{1i}}\right) \times \exp\left(-\frac{\|CSR - b_{2i}\|}{a_{2i}}\right) \tag{2}$$

$$z_i\left(q_{c,1}, CSR\right) = a_i^* \times q_{c,1} + b_i^* \times CSR + c_i^* \tag{3}$$

where a_{1i}, b_{1i}, a_{2i}, b_{2i}, a_i^*, b_i^*, and c_i^* are identified in Table 2

$$\begin{array}{ll} f\left(q_{c,1}, CSR\right) > 0, & \text{Liquefaction occurs} \\ f\left(q_{c,1}, CSR\right) < 0, & \text{Nonliquefaction occurs} \end{array} \tag{4}$$

The average recognition rate obtained by the 10-fold cross-validation method for each run is shown in Fig. 3. For example, in the first run, the average recognition rates of train and test data for imbalance data are 87.42% and 88.7%, respectively; whereas for upsampled data, the average recognition rates for training and test data increases to 92.36% and 92.83%, respectively.

Model Validation

To evaluate the performance of the proposed classifier and develop a quantitative basis for comparison with other methods, a number of metrics are utilized. The level of accuracy for the constructed models in each generation is evaluated based on coefficient of determination (COD) as the fitness function. The COD function used in this study is

$$\mathrm{COD} = 1 - \frac{\sum_N \left(Y_a - Y_p\right)^2}{\sum_N \left[Y_a - (1/N)\sum_N Y_a\right]^2} \tag{5}$$

where Y_a = actual is output value;Y_p = predicted value, and N = number of data points on which the COD is computed. The result is shown in Table 3.

From the previous results it appears that COD is biased against ordinary nonliquefied data, and by compensating the minority class the bias is reduced. It is evident that the proposed technique is as efficient as the method proposed by Moss et al. (2006) for the liquefied data, but it is much better for determining nonliquefied data because of the up-sampling procedure introduced here. It must also be noted that the classifier is applied to both the actual data set and the compensated data set. Hence, the previous results are directly comparable.

Summary and Conclusions

Liquefaction in soil is one of the major causes of concern in geotechnical engineering. The cone penetration test has proven to be an effective tool in the characterization of subsurface conditions and the analysis of different aspects of soil behavior, including estimating the potential for liquefaction at a specific site.

The main scope of this study is to implement an adaptive neuro-fuzzy inference system for the prediction of liquefaction threshold based on CPT upsampled data. For the identification of liquefaction and nonliquefaction regions, a support vector data description method with suitable parameters (C and σ) was used.

An ANFIS classifier was used to predict soil liquefaction. For training and testing the data model, K-fold cross-validation was used. It has been shown that by calculating the COD of the data that the ANFIS classifier is as efficient as the Bayesian approach proposed by Moss et al. (2006) for the liquefied class, but provided much better results for the nonliquefied data. In general it is shown that upsampling has a positive bearing on the recognition rates of an ANFIS classifier by about 4%.

References

Cetin, K. O., Kiureghian, A. D., and Seed, R. B. (2002). "Probabilistic models for the initiation of seismic soil liquefaction." *Struct. Saf.*, 24(1), 67–82.

Jafarian, Y., Abdollahi, A. S., Vakili, R., and Baziar, M. H. (2010). "Probabilistic correlation between laboratory and field liquefaction potentials using relative state parameter index (ξ_R)." *Soil. Dyn. Earthquake Eng.*, 30(10), 1061–1072.

Juang, C. H., Ynan, H., Lee, D. H., and Lin, P. S. (2003). "Simplified cone penetration test-based method for evaluating liquefaction resistance of soils." *J. Geotech. Geoenviron. Eng.*, 129(1), 66–80.

MATLAB 7.12.0.635 (R2011a) [Computer software]. Adelaide, South Australia, Australia, MathWorks, Inc.

Moss, R., Seed, R. B., Kayen, R. E., Stewart, J. P., Kiureghian, A. D., and Cetin, K. O. (2006). "CPT-based probabilistic and deterministic assessment of in situ seismic soil liquefaction potential." *J. Geotech. Geoenviron. Eng.*, 132(8), 1032–1051.

Mróz, Z., Boukpeti, N., and Drescher, A. (2003). "Constitutive model for static liquefaction." *Int. J. Geomech.*, 3(2), 133–144.

Oommen, T., Baise, L. G., and Vogel, R. (2010). "Validation and application of empirical liquefaction model." *J. Geotech. Geoenviron. Eng.*, 136(12), 1618–1633.

Rezania, M., Faramarzi, A., and Javadi, A. A. (2011). "An evolutionary based approach for assessment of earthquake-induced soil liquefaction and lateral displacement." *Eng. Appl. Artif. Intell.*, 24(1), 142–153.

Rezania, M., Javadi, A. A., and Giustolisi, O. (2010). "Evaluation of liquefaction potential based on CPT results using evolutionary polynomial regression." *Comput. Geotech.*, 37(1–2), 82–92.

Tax, D. M. J., and Duin, R. P. W. (1999). "Support vector domain description." *Pattern Recogni- t. Lett.*, 20, 1191–1199.

Tax, D. M. J., and Duin, R. P. W. (2004). "Support vector data description." *Mach. Learn.*, 54(1), 45–66.

Weiher, B., and Davis, R. (2004). "Correlation of elastic constants with penetration resistance in sandy soils." *Int. J. Geomech.*, 4(4), 319–329.

Youd, T. L., and Idriss, I. M. (2001). "Liquefaction resistance of soils: Summary report from the 1996 NCEER and 1998 NCEER/NSF workshops on evaluation of liquefaction resistance of soils." *J. Geotech. Geoenviron. Eng.*, 127(10), 817–833.