Computers & Geosciences 44 (2012) 10–23

Contents lists available at SciVerse ScienceDirect

Computers & Geosciences

journal homepage: www.elsevier.com/locate/cageo

Prediction of liquefaction potential based on CPT up-sampling

Javad Sadoghi Yazdi [a,*], Farzin Kalantary [a], Hadi Sadoghi Yazdi [b,c]

[a] Civil Engineering Department, K.N.Toosi University of Technology, Tehran, Iran
[b] Computer Department, Ferdowsi University of Mashhad, Mashhad, Iran
[c] Center of Excellence on Soft Computing and Intelligent Information Processing, Ferdowsi University of Mashhad, Mashhad, Iran

ARTICLE INFO

Article history:
Received 9 August 2011
Received in revised form
10 February 2012
Accepted 26 March 2012
Available online 5 April 2012

Keywords:
Soil liquefaction
Cone penetration test
Support Vector Data Description
Adaptive neuro-fuzzy inference system
Up sampling

ABSTRACT

Cone penetration test data has been widely used for determination of the threshold of seismically induced soil liquefaction. However, possible inaccuracies in the collected data from case histories as well as natural variability of parameters and other uncertainties associated with natural phenomenon have yet prohibited a conclusive definition for this threshold.

Various classification techniques have been used to define the most reliable correlations. However, available liquefied to non-liquefied data imbalance has caused learning bias to the majority class in the learning model of the pattern recognition systems. This has adversely affected the outcome of such approaches and in order to overcome this problem Support Vector Data Description (SVDD) strategy is employed to "up sample" the minority data. In other words SVDD, which is robust against noisy samples, is used to generate virtual data points for the minority class, bearing the same characteristics as the non-virtual samples. In order to specify the most appropriate data range a sphere boundary around the main body of the data are sought through an optimization process. The data inside the obtained boundary are the target data and the ones outside it are the outliers or so-called "noise", to be neglected. This procedure reduces the issue of class intermixture in the fringe zone and produces relatively well defined class that then is fed into the Adaptive Neuro-Fuzzy Inference System (ANFIS) classifier for determination of liquefaction potential. The predictions are then examined to evaluate the reliability and validation of the overall technique and compared with other prediction methods using confusion matrix. It is shown that the overall accuracy of the proposed technique is higher than all previously proposed methods and only equal to the Support Vector Machine (SVM) technique. Furthermore an improvement in the F-score of the non-liquefied data recognition has been achieved in relation to all previously proposed methods.

1. Introduction

Seismically induced liquefaction in saturated soils is a phenomenon in which soil loses much of its strength or stiffness due to rising pore water pressure for a generally short period of time but nevertheless long enough for it to cause ground failure. Determination of liquefaction potential of soils has become a major concern and an essential criterion in the design process of the civil engineering projects. Over the past 30 years, many researchers have endeavored to present various methods for prediction of liquefaction potential of soils.

Amongst in situ tests, many researchers have adapted Cone Penetration Test (CPT) results as the basis for evaluation of liquefaction potential of the test method (e.g. Juang et al., 2003; Youd et al., 2001).

Many researchers have developed charts or correlations for liquefaction threshold based on a measure of soil resistance to liquefaction (presented in the form of normalized cone tip resistance $q_{c,1}$) against a measure of seismically induced shear stress, (cast in the context of Cyclic Stress Ratio). CSR is a function of the earthquake magnitude, peak surface acceleration, the total and effective overburden stress, and the depth of the source bed. The above mentioned parameters are determined using the following equations:

$$CSR = 0.65 r_d \left(\frac{\sigma_{v0}}{\sigma'_{v0}}\right)\left(\frac{a_{\max}}{g}\right) \quad (1)$$

$$q_{c,1} = \left(\frac{p_c}{\sigma'_v}\right)^c q_c \quad (2)$$

where c is explained in details in (Moss et al., 2006).

A number of other researchers have considered a different representation of input parameters (Rezania et al., 2011). These parameters include the seismic cyclic stress ratio adjusted to an earthquake magnitude of 7.5 ($CSR_{7.5}$). Therefore for each case

* Corresponding author. Tel.: +98 9158306949; fax: +98 21 77245305 7.
E-mail address: j.sadoghi@yahoo.com (J. Sadoghi Yazdi).

0098-3004/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.cageo.2012.03.025

Nomenclature

a_{max}	maximum horizontal ground acceleration
r_d	shear stress reduction factor
CSR	cyclic stress ratio
q_c	measured cone tip resistance
$q_{c,1}$	normalized cone tip resistance
e	center of hyper-sphere
R	radius of hyper-sphere
x_i	input sample
$g(x)$	mapping function
ξ_i	slack variable
σ	width of Gaussian kernel
C	penalty coefficient
DBR	data belonging ratio
$\overline{c}$	center of Gaussian membership
$\overline{\sigma}$	standard deviation of cluster

history the above parameters need to be adjusted to the earthquake magnitude of 7.5 through the use of magnitude scaling factor (MSF) given by the following equations:

$$q_{c1N} = \frac{q_c/100}{(\sigma'_v/100)^{0.5}} \tag{3}$$

and $CSR_{7.5}$ was calculated as:

$$CSR_{7.5} = 0.65\left(\frac{\sigma_{v0}}{\sigma'_{v0}}\right)\left(\frac{a_{\max}}{g}\right)\left(\frac{r_d}{MSF}\right). \tag{4}$$

MSF was calculated as:

$$MSF = \left(\frac{10^{2.24}}{M_w^{2.56}}\right) = \left(\frac{M_w}{7.5}\right)^{-2.56} \tag{5}$$

where a_{max} is the maximum horizontal acceleration caused by earthquake; g is the acceleration of gravity; q_c is the measured cone tip resistance (kPa); P_a is the reference stress (1 atm is the 101.325 kPa); σ_{v0} and σ'_{v0} are total and effective vertical overburden stresses, respectively (kPa); r_d is the shear stress reduction factor and calculated as follows:

$$r_d = 1.0-0.00765z \text{ if } z \le 9.15\text{ m } r_d = 1.174-0.0267z$$
$$r_d = 1.174-0.0267z \text{ if } 9.15 < z \le 23\text{m} \tag{6}$$

The MSF obtained from Eq. (5) represents the lower bound of the range of MSF values recommended by the NCEER workshop (Juang et al., 2003; Youd et al., 2001). Eqs. (5) and (6) are commonly used by geotechnical researchers, although many other formulae have been proposed for calculating r_d and MSF (Juang et al., 2003).

More intricate approaches based on Artificial Neural Networks (ANN), probabilistic analyses have also been introduced recently. A summary is presented in the Table 1.

The above mentioned approaches have all achieved various degrees of success in prediction of liquefaction and each has contributed to a better understanding of the system classification process. The issue of data imbalance was initially noted by Cetin et al. (2002) and Moss et al. (2006). However, a comprehensive treatment of data imbalance as well as sampling bias has recently been presented by Oommen et al. (2010). Maximum Likelihood Logistic Regression (MLLR) has been used in this recent article to show the effect of sampling bias and it has been proven that when sampling bias is reduced the predicted probability approaches the actual probability irrespective of data imbalance.

In the present paper a different technique is employed to reduce the effect of data imbalance as well as isolating the out of range data points. The technique is called Support Vector Data Description (SVDD) which has proven its capabilities in other fields of science (Liu et al., 2011). Having removed the class imbalance then Adaptive Neuro-Fuzzy Inference System (ANFIS) is used as the classifier for determination liquefaction threshold.

It is understood that in addition to up-sampling (leading to removal of data imbalance and sampling bias), density functions have also an important role in improving predictions. However, due to limitations this issue is not treated here.

In the following section initially a review of class imbalance problems are presented, followed by definitions of the basic concepts of SVDD and ANFIS. Then in Section 3 details of liquefaction potential prediction is set out: introduction of CPT data, modeling procedure using SVDD, SVDD-based up-sampling procedure, system identification procedure using ANFIS and model validation and comparison. Finally discussion and conclusion is presented.

2. Background knowledge

2.1. Review of class imbalance problem

Imbalanced data are common in many machine learning applications. In an imbalanced data set, the number of instances in at least one class is significantly higher or lower than that in other classes. Consequently, when classification models with imbalanced data are developed; most classifiers are subjected to an unequal number of instances in each class, thus failing to construct an effective model. The class imbalance problem is encountered in real-world applications of machine learning and results in a classifier's suboptimal performance. This goes back to the place it is occurring; which is the data. The problem is that there are some datasets that have an imbalance between the

Table 1
Illustrative list of different applications having presented for soil liquefaction analysis.

Description of the work	Studied structure	Authors
Constitutive model for static liquefaction	Constitutive modeling	Mroz et al. (2003)
Probabilistic models for the initiation of seismic soil liquefaction	Probabilistic approach	Cetin et al. (2002)
CPT-Based probabilistic and deterministic assessment of in situ seismic soil liquefaction potential	Probabilistic approach	Moss et al. (2006)
Simplified cone penetration test-based method for evaluating liquefaction resistance of soils	Artificial intelligence	Juang et al. (2003)
An evolutionary based approach for assessment of earthquake-induced soil liquefaction and lateral displacement	Swarm intelligence	Rezania et al. (2011)
Evaluation of liquefaction potential based on CPT results using evolutionary polynomial regression	Swarm intelligence	Rezania et al. (2010)
Validation and application of empirical liquefaction model (SVM)	Statistical pattern recognition	Oommen et al. (2010)

numbers of instances in different classes, i.e. the number of instances in some classes is more than other classes. This imbalance could occur with different degrees; but usually a high degree is noteworthy (He and Garcia, 2009).

Class imbalance is relevant in some valuable datasets such as medical diagnosis, fraud detection, oil spill detection, mine classification etc. Imbalance may be intrinsic or extrinsic (He and Garcia, 2009). Intrinsic imbalance is related to the nature of the data but extrinsic occurs due to external interferences such as low storage and/or time. Class imbalance is studied both in two class and multi class classification. In two class classification the classes with less data are called the minority class and the other class is the majority class.

Previous algorithms for handling class imbalance could be categorized into two main groups. The first group focuses on modifying the classification method to recover the imbalance problem; modifications of SVM (Batuwita and Palade, 2010) and C4.5 decision tree (Quan et al., 2006) could be placed in this category. On the other hand, the second group of algorithms processes the data to reduce the imbalance. This category of algorithms could be considered as a preprocessing step before the classification which is the main process. The main method used for preprocessing is sampling. Sampling algorithms follow two different approaches, under sampling and oversampling. Under-sampling methods aim to decrease the size of the majority class. Easy Ensemble (Liu et al., 2006), Near-Miss (Zhang and Mani, 2003) and RUSBoost (Seiffert et al., 2010) are some good algorithms.

The second approach for preprocessing using data sampling, is oversampling. On the contrary to under sampling, oversampling algorithms tend to increase data in the minority class. Oversampling methods could be categorized into two more specific groups which are synthetic and non-synthetic. In the first category synthetic non-existent data are added to the original data in the minority class, whereas in the second category data in the minority class is replicated. Synthetic oversampling has been studied more thoroughly than the non-synthetic one. Chawla et al. (2002) proposed SMOTE; which is a synthetic oversampling method based on generating data in the neighbors of data using k-NN with a random distance. This algorithm has gained significant success in different applications. Different improvements have been proposed for SMOTE; Han et al. (2005) propose Borderline SMOTE, CE-SMOTE was proposed by Chen et al. (2010), the author of SMOTE proposed SMOTE Boost (Chawla et al., 2003) which is a combination of SMOTE and the boosting procedure.

Non-synthetic oversampling is another category of methods in which data are replicated, that is, no synthetic data are generated. Two algorithms can be defined in this category; Random Oversampling (RO) by Drummond and Holte (2003) and Cluster-Based Sampling (CBS) (Jo and Japkowicz, 2004). In RO data in the minority class is replicated randomly to a specific ratio. The second algorithm CBS, uses clustering techniques for oversampling. CBS clusters the minority and majority class using a specific number of clusters (k). Using the cluster information the data in minority and majority classes are balanced equally. One issue to consider is that sampling methods aim to alleviate the class imbalance problem in supervised classification algorithms. Other classifiers such as SVM (Bae et al., 2010), Random Forest (Gu et al., 2007) and ANN (Bae et al., 2010) have also made use of various sampling methods. A comprehensive review on sampling bias may also be found in Anderson and Gonzalez (2011).

The CPT database has 182 case histories of which 139 are from liquefied sites and 43 are from non-liquefied sites. This database which has been compiled by Moss et al. (2006) falls within the category of imbalanced datasets since the ratio of liquefied to non-liquefied instances is over 3. Moss et al. (2006) noticed this imbalance and by using Bayesian updating optimization attempted to overcome this deficiency; they modified the likelihood function by a weighting factor $W_{\text{non-liquefied}}/W_{\text{liquefied}}=1.5$ based on (Cetin et al., 2002) and the consensus of an expert panel specifically set up for reviewing the CPT dataset.

2.2. Support vector data description (SVDD)

Support vector data description is a data description method that can give the target dataset a spherically shaped description and be used for detection of outliers. In recent years, the problem of data description or one-class classification has received much attention (Tax and Duin, 1999). In domain description the task is to give a description of a training set of objects and to detect which (new) objects resemble this training set. This description should cover the class of objects represented by the training set, and ideally should reject all other possible objects in the object space. SVDD was first presented by Tax and Duin and later further developed with extensions and a more thorough treatment (Tax and Duin, 2004). The SVDD has found a wide range of applications: SVDD has been used for detection of outlier or otherwise known as uncharacteristic data points in a data set, as well as detection of an anomaly (Liu et al., 2011). It is proposed here to use SVDD for special classification problems, where one class is severely under sampled, while the other class(es) are well-sampled.

The basic idea of SVDD is to map the original normal training data both nonlinearly and implicitly into a potentially much higher inner product space (feature space), and to search for a hyper sphere with minimal volume containing most of the mapped training data. A new object, which is subjected to the same mapping, is recognized as a target if its image lies inside the hyper sphere; otherwise, it is an outlier. Several key variables such as radius and center of the hyper sphere are involved in the search of the hyper sphere.

The liquefaction/non-liquefaction data classification which is presented in the following section is a clear illustration of SVDD concept and procedure.

Let $x_i(i=1,\ldots,n)$ be p-dimensional training samples belonging to one class. We consider approximating the class region by the minimum hyper-sphere with center $e=(e_1,e_2,\ldots,e_p)^T$ and radius R in high dimensional feature space (HDS), excluding the outliers. This goal is formulated as a constrained convex optimization problem

$$\min_{R,e,\xi} R^2+C\sum_{i=1}^{n}\xi_i$$
$$\text{Subject to } \begin{cases} \|g(x_i)-e\|^2 \le R^2+\xi_i, & i=1,\ldots,n \\ \xi_i \ge 0, & i=1,\ldots,n \end{cases} \tag{7}$$

where $g(x)$ is the mapping function that maps x into a high dimension space (HDS), $\xi=(\xi_1,\ldots,\xi_n)^T$ and ξ_i is the slack variable of ith training sample and C is a constant which determines the trade-off between the hyper-sphere volume and outliers. The Lagrangian dual form of (7) is as follows:

$$\max_{\delta,\gamma} L(R,e,\xi,\alpha,\gamma)$$
$$\text{Subject to } \alpha_i,\gamma_i \ge 0, \quad i=1,\ldots,n \tag{8}$$

where $\alpha=(\alpha_1,\ldots,\alpha_n)^T$, $\gamma=(\gamma_1,\ldots,\gamma_n)^T$ and

$$L(R,e,\xi,\alpha,\gamma)=\inf\Big\{R^2+C\textstyle\sum_{i=1}^{n}\xi_i-\sum_{i=1}^{n}\alpha_i(R^2+\xi_i-g(x_i)^Tg(x_i) +2e^Tg(x_i)-e^Te)-\sum_{i=1}^{n}\gamma_i\xi_i\Big\} \tag{9}$$

For the optimal solution, the following conditions are satisfied

$$\frac{\partial L}{\partial R}=0\rightarrow\sum_{i=1}^{n}\alpha_i=1 \tag{10}$$

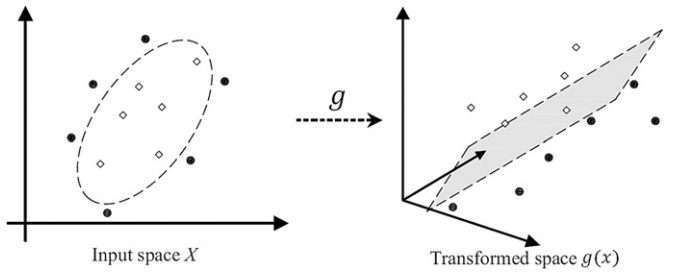


Fig. 1. Mapping of dataset X by g into a higher dimensional space.

$$\frac{\partial L}{\partial e}=0\rightarrow e=\sum_{i=1}^{n}\alpha_i g(x_i) \tag{11}$$

$$\frac{\partial L}{\partial \xi}=0\rightarrow \alpha_i=C-\gamma_i,\quad i=1,\ldots,n \tag{12}$$

$$\alpha_i(\|g(x_i-e)\|^2-R^2-\xi_i)=0,\quad i=1,\ldots,n \tag{13}$$

$$\gamma_i\xi_i=0,\quad i=1,\ldots,n \tag{14}$$

Using the above conditions, $L(R,e,\xi,\alpha,\gamma)$ is transformed to

$$\sum_{i=1}^{n}\alpha_i K(x_i,x_i)-\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j K(x_i,x_j) \tag{15}$$

where $K(x_i,x_j)=g(X)^T g(x_j)$is the kernel, and the training vectors$x_i's$ are mapped into a higher dimensional space by function g, as shown in Fig. 1.

From (8) and (12) we have $0\le\alpha_i\le C$. So, the Lagrangian dual form of (7) can be restated as follows:

$$\max_{\alpha}\sum_{i=1}^{n}\alpha_i K(x_i,x_i)-\sum_{i=1}^{n}\sum_{j=1}^{n}K(x_i,x_j)$$
$$\text{Subject to}\ \begin{cases}\sum_{i=1}^{n}\alpha_i=1,\\ 0\le\alpha_i\le C,\quad i=1,\ldots,n\end{cases} \tag{16}$$

This is a conventional quadratic program and can be solved easily. From (7), $\|g(x_i)-e\|^2=K(x_i,x_i)-2\sum_{j=1}^{n}\alpha_j K(x_i,x_j)+\sum_{j=1}^{n}\sum_{k=1}^{n}\alpha_i\alpha_k K(x_j,x_k)$ and from (13) if $\alpha_i>0, K(x_i,x_i)-2\sum_{j=1}^{n}\alpha_j K(x_i,x_j)+\sum_{j=1}^{n}\sum_{k=1}^{n}\alpha_j\alpha_k K(x_j,x_k)=R^2+\xi_i$.

From (12) if $\alpha_i<C$, $\gamma_i>0$. So, from (14) we have $\xi_i=0$.So, if $0\le\alpha_i\le C$.

$$R^2=K(x_i,x_i)-2\sum_{j=1}^{n}\alpha_j K(x_i,x_j)+\sum_{j=1}^{n}\sum_{k=1}^{n}\alpha_i\alpha_k K(x_j,x_k) \tag{17}$$

Finally, the unknown datum x is inside the hyper-sphere if $\|g(x)-e\|^2\le R^2$ or equivalently if

$$K(x,x)-2\sum_{i\in sv}\alpha_i K(x,x_i)+\sum_{i\in sv}\sum_{j\in sv}\alpha_i\alpha_j K(x_i,x_j)\le R^2 \tag{18}$$

where SV the set of indices of training is samples whose $\alpha\neq 0$.

Data Belonging Ratio (DBR) is defined as:

$$DBR=R^2-\|g(x)-e\|^2 \tag{19}$$

The Gaussian kernel $K(x,y)=\exp(-\|x-y\|/\sigma^2)$ is used.

2.3. Neuro-fuzzy inference system

Recently, there has been a growing interest in combining "Artificial Neural Networks" (ANN) and "Fuzzy Interface System", and as a result; neuro-fuzzy computing techniques have evolved. Neuro-fuzzy systems are fuzzy systems, which use neural networks theory in order to determine their properties (fuzzy sets and fuzzy rules) by processing data samples (Mitra and Hayashi, 2000). Neuro-fuzzy integrates the merits of both neural networks and fuzzy systems in a complementary way to overcome their disadvantage. The fusion of neural network and fuzzy logic in neuro-fuzzy models possess both low-level learning and computational power of neural networks and advantages of high-level human like thinking of fuzzy systems. Adaptive Neuro-Fuzzy Inference System (ANFIS) model combined the neural network adaptive capabilities and the fuzzy logic qualitative approach, initially introduced by Jang (1993).

It has attained its popularity due to a broad range of useful applications in such diverse areas in recent years as optimization of fishing predictions (Nuno et al., 2005), vehicular navigation (Noureldin et al., 2007), identification of the turbine speed dynamics (Kishor et al., 2007), radio frequency power amplifier linearization (Lee and Gardner, 2006), image de-noising (Qin and Yang, 2007; Çivicioglu., 2007), prediction in cleaning with high pressure water Daoming and Jie (2006), sensor calibration (Depari et al., 2007), fetal electrocardiogram extraction from ECG signal captured from mother (Assaleh., 2007), identification of normal and glaucomatous eyes (Huang et al., 2007). All these works show that ANFIS is a good universal approximated, predictor, interpolator and estimator and demonstrate that ANFIS has the approximation capabilities of neural networks and any non-linear function of several inputs and outputs can be easily constructed with ANFIS. The summarized advantage of the ANFIS technique is listed below.

- Real-time processing of instantaneous system input and output data. This property helps the use of this technique for many operational research problems.
- Offline adaptation instead of online system-error minimization, thus easier to manage and no iterative algorithms are involved.
- System performance is not limited by the order of the function since it is not represented in polynomial format.
- Fast learning time.

In the following section a detailed description of ANFIS architecture is presented.

2.4. Adaptive neuro-fuzzy inference system (ANFIS) architecture

Neuro-fuzzy systems are fuzzy systems, which use NNs to determine their properties (fuzzy sets and fuzzy rules) by processing data samples. ANFIS has been proven to have significant results in modeling nonlinear functions. In ANFIS, the membership functions (MF) are extracted from a data set that describes the system behavior. The ANFIS learns features in the data set and adjusts the system parameters according to given error criterion. In a fused architecture, NN learning algorithms are used to determine the parameters of fuzzy inference system.

A typical architecture of ANFIS is shown in Fig. 2, in which a circle indicates a fixed node, and a square indicates an adaptive node. For simplicity, we consider two inputs x, y and one output z in the fuzzy inference system (FIS). The ANFIS used in this paper implements a first-order Sugeno fuzzy model. Among many fuzzy inference systems, the Sugeno fuzzy model is the most widely used for its high interpretability and computational efficiency, and built-in optimal and adaptive techniques. For example for a first-order Sugeno fuzzy model, a common rule set with two fuzzy if-then rules can be expressed as (20).

$$\begin{aligned}&\text{Rule 1: If } x \text{ is } A_1 \text{ and } y \text{ is } B_1, \text{ then}\\ &\quad z_1=p_1x+q_1y+r_1\\ &\text{Rule 2: If } x \text{ is } A_2 \text{ and } y \text{ is } B_2, \text{ then}\\ &\quad z_2=p_2x+q_2y+r_2\end{aligned} \tag{20}$$

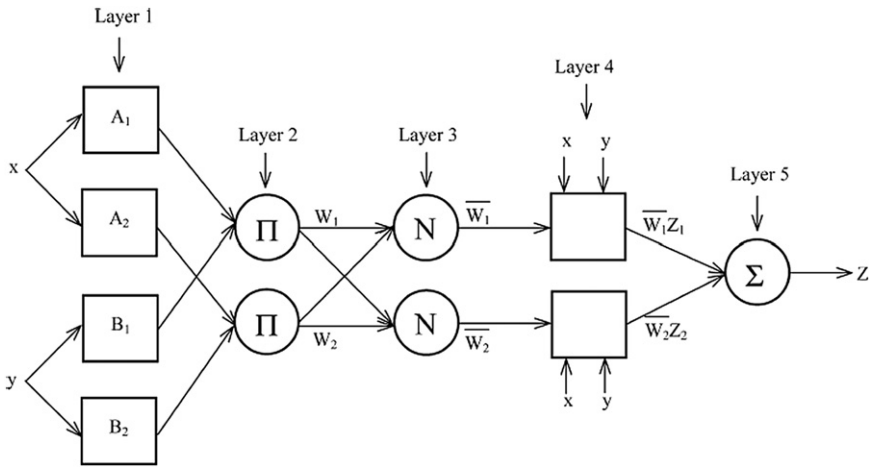


Fig. 2. ANFIS architecture (Π, *N*, Σ are defined in (23), (24), (26) respectively).

where A_i,B_i ($i=1,2$) are fuzzy sets in the antecedent, and p_i,q_i, r_i ($i=1,2$) are the design parameters that are determined during the training process. As in Fig. 2, the ANFIS consists of five layers.

Layer 1, every node i in this layer is an adaptive node with a node function

$$O_i^1 = \mu_{A_i}(x), \quad i=1,2$$
$$O_i^1 = \mu_{B_i}(y), \quad i=3,4 \tag{21}$$

where x, y are the input of node i, $\mu_{A_i}(x)$ and $\mu_{B_i}(y)$ can adopt any fuzzy membership function (MF). In this paper, Gaussian MFs are used

$$gaussian(x,\bar{c},\bar{\sigma}) = e^{-(1/2)((x-\bar{c})/\bar{\sigma})^2} \tag{22}$$

where $\bar{c}$is center of Gaussian membership function and $\bar{\sigma}$ is standard deviation of this cluster.

Layer 2, every node in the second layer represents the ring strength of a rule by multiplying the incoming signals and forwarding the product as

$$O_i^2 = \omega_i = \mu_{A_i}(x)\mu_{B_i}(y), \quad i=1,2 \tag{23}$$

Layer 3, the ith node in this layer calculates the ratio of the ith rule's ring strength to the sum of all rules ring strengths

$$O_i^3 = \overline{\omega}_i = \frac{\omega_i}{\omega_1+\omega_2}, \quad i=1,2 \tag{24}$$

where $\overline{\omega}_i$ is referred to as the normalized ring strengths.

Layer 4, the node function in this layer is represented by

$$O_i^4 = \overline{\omega}_i z_i = \overline{\omega}_i(p_i x+q_i y+r_i), \quad i=1,2 \tag{25}$$

where $\overline{\omega}_i$ is the output of layer 3, and $\{p_i,q_i,r_i\}$ are the parameter set. Parameters in this layer are referred to as the consequent parameters.

Layer 5, the single node in this layer computes the overall output as the summation of all incoming signals

$$o_i^5 = \sum_{i=1}^{2} \overline{\omega}_i z_i = \frac{\omega_1 z_1+\omega_2 z_2}{\omega_1+\omega_2} \tag{26}$$

It is seen from the ANFIS architecture that when the values of the premise parameters are fixed, the overall output can be expressed as a linear combination of the consequent parameters:

$$z = (\overline{\omega}_1 x)p_1+(\overline{\omega}_1 y)q_1+(\overline{\omega}_1)r_1+(\overline{\omega}_2 x)p_2+(\overline{\omega}_2 y)q_2+(\overline{\omega}_2)r_2 \tag{27}$$

The hybrid learning algorithm (Nuno et al., 2005; Huang et al., 2007) combining the least square method and the back propagation (BP) algorithm can be used to solve this problem. This algorithm converges much faster since it reduces the dimension of the search space of the BP algorithm. During the learning process, the premise parameters in layer 1 and the consequent parameters in layer 4 are tuned until the desired response of the FIS is achieved. The hybrid learning algorithm has a two-step process. First, while holding the premise parameters fixed, the functional signals are propagated forward to layer 4, where the consequent parameters are identified by the least square method. Second, the consequent parameters are held fixed while the error signals, the derivative of the error measure with respect to each node output, are propagated from the output end to the input end, and the premise parameters are updated by the standard BP algorithm.

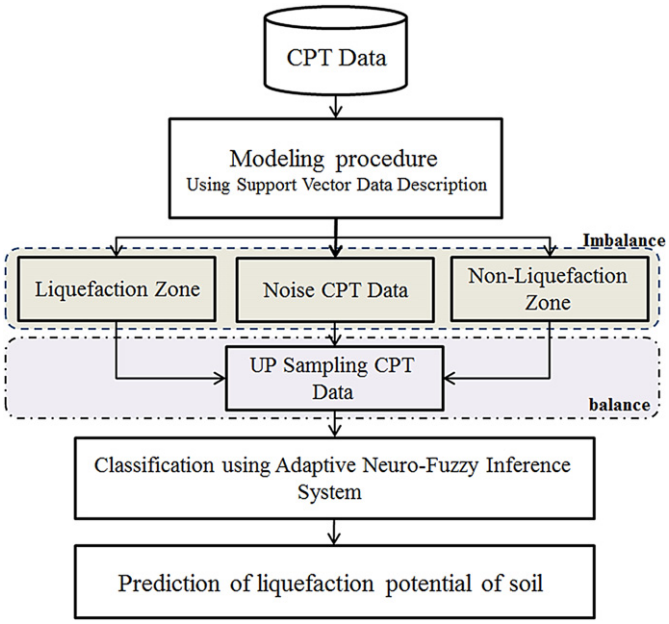


Fig. 3. The proposed structure.

3. The proposed approach

In this paper it is proposed to use Adaptive Neuro Fuzzy Inference System (ANFIS) as the identification technique for determination of the liquefaction threshold. MATLAB have been used for programming ANFIS and SVDD using a modified code,

originally proposed by Tax (Tax and Duin, 1999, 2004). ANFIS is trained using CPT based liquefaction case histories. However, before feeding the raw data into the class identification procedure, a number of steps must be taken to remove the data imbalance as well as isolating the so-called "noise" or outlier data sample methodologically. The organization chart for the proposed approach is presented in Fig. 3.

The process includes feeding the CPT data into the SVDD to produce various descriptions of data ranges by applying different data region description parameters (σ, C). For each of the determined minority class data sphere, the appropriate "up sampling" is carried out and then the ANFIS classifier is employed to determine the optimum data description providing the best possible recognition rate.

3.1. CPT data

The data used in this study is adapted from reference (Moss et al., 2006) and include 139 liquefied and 43 non-liquefied records from 18 different earthquakes spanning over four decades. The earthquakes included are; 1964 Niigata, 1968 Inangahua, 1975 Haicheng, 1976 Tangshan, 1977 Vrancea, 1979 Imperial Valley, 1980 Mexicali, 1981 Westmorland, 1983 Nihonkai-Chubu, 1983 Borah Peak, 1987 Elmore Ranch, 1987 Superstition Hills, 1987 Edgecumbre, 1989 Loma Prieta, 1994 Northridge, 1995 Hyogoken-Nambu (Kobe), 1999 Kocaeli, and 1999 Chi-Chi earthquakes.

Moss et al. (2006) have presented the above data in (CSR, $q_{c,1}$) space and have shown that similar to all previous works, class intersection exists. This class intersection may be attributed to uncertainties associated with physical measurements and other unknown factors and shall prohibit definition of a sharp and clear cut threshold. Recently Rezania et al. (2011), introduced a three dimensional space ($CSR_{7.5}$, q_{c1N} and σ'_v) mainly to alleviate the class intersection issue (Rezania et al., 2011). However, feeding the above data into this newly proposed space reveals that even this attempt did not totally remove the class intersection either as is evident from different views of this space (Fig. 4). Hence in this study it is decided to revert to the (CSR, $q_{c,1}$) space and the issue of class imbalance is thus treated in this space.

The data in the minority class (non-liquefied records) is "up sampled" by over three folds to overcome the imbalance. The best data description has been achieved by a trial and error process. The influencing parameters in the SVDD and the procedure by which parameter tuning is achieved are described in the next sections.

3.2. Modeling procedure using SVDD

The modeling procedure is demonstrated graphically in Fig. 5. The liquefaction data shown in Fig. 5(a) is fed into SVDD (according Section 2.2) and a data region defined by model description parameters (σ, C) is obtained (Fig. 5(b). The non-liquefaction model is similarly developed (Fig. 5(c and d)).

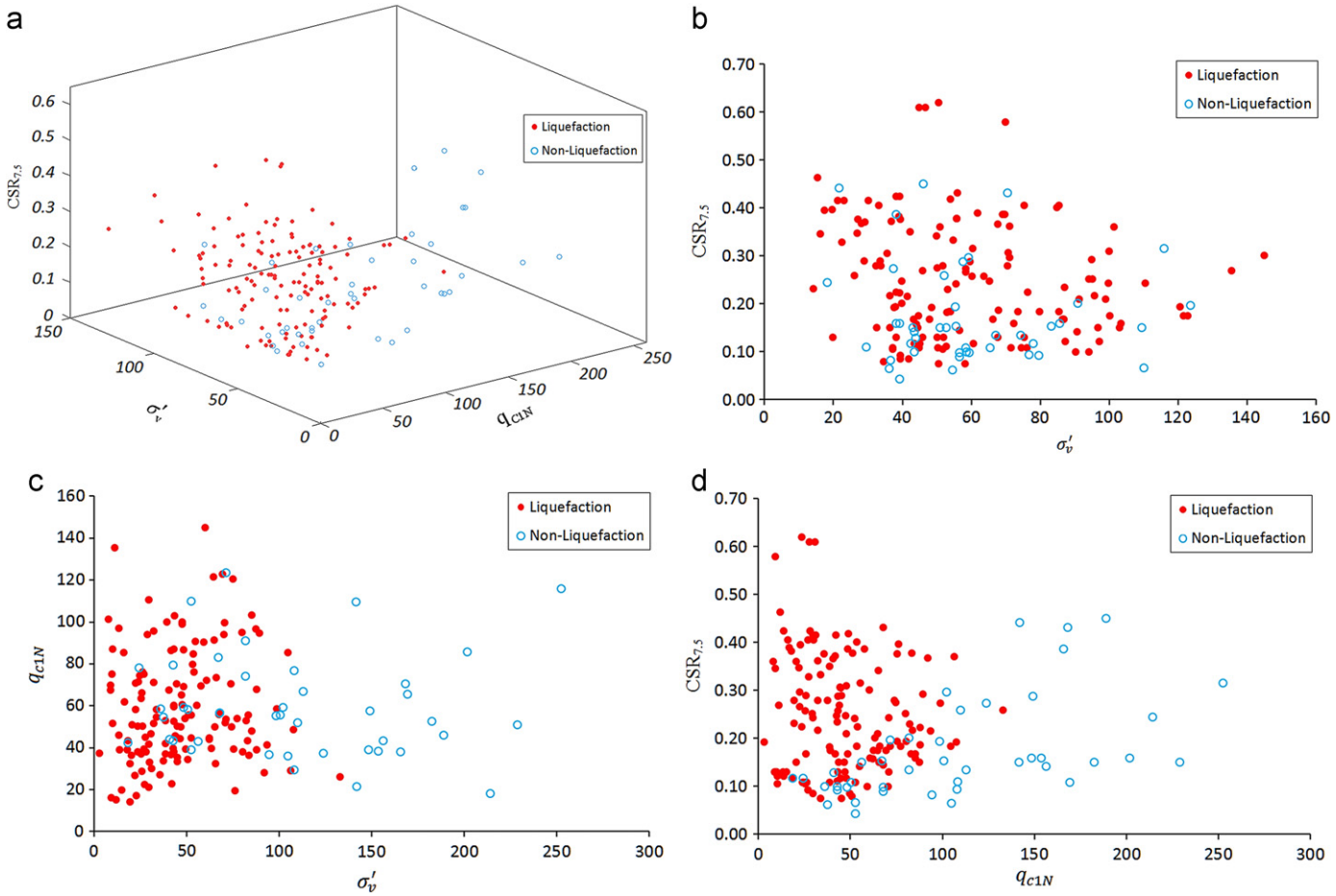


Fig. 4. Data Scatter of CPT based case histories in ($CSR_{7.5}$,q_{c1N},σ'_v) space. (a) View in third dimension ($CSR_{7.5}$,q_{c1N},σ'_v), (b) view in (σ'_v,$CSR_{7.5}$), (c) view in (σ'_v,q_{c1N}), (d) view in (q_{c1N},$CSR_{7.5}$).

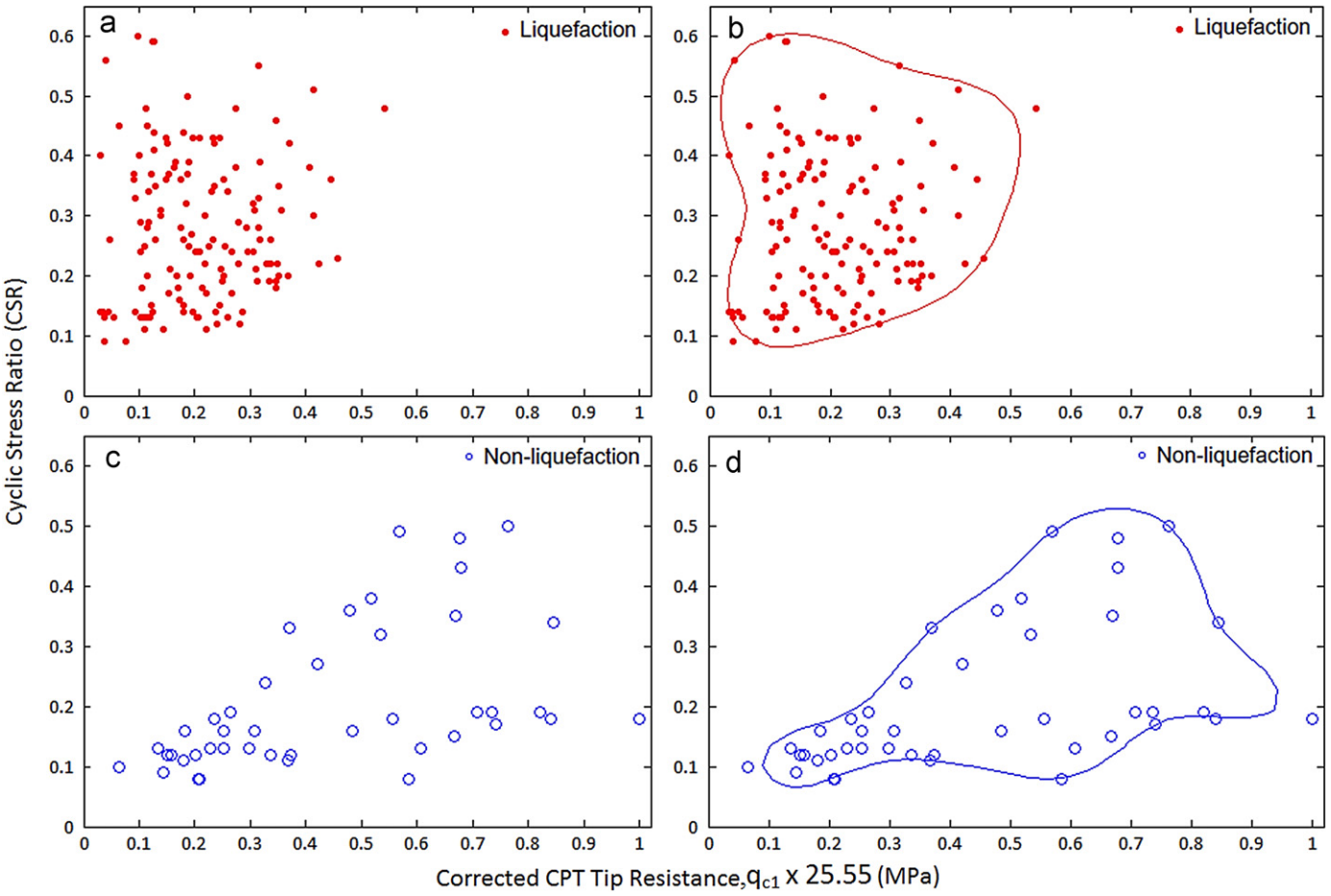


Fig. 5. (a) Liquefied data, (b)obtained surface using SVDD for liquefaction data, (c) non-liquefied data, (d) obtained surface using SVDD for nonliquefaction data.

The SVDD encloses liquefaction/non-liquefaction regions and thereby detects outliers. Each SVDD for liquefaction and non-liquefaction can be used for determination of status of input sample relative to the obtained enclosed region; inside/outside/on-boundary is status of input samples. The status is reported respectively with negative/positive/zero values (more details of SVDD operation appears in Section 3.3).

Fig. 5(a) liquefied data (b) obtained surface using SVDD for liquefaction data (c) non-liquefied data (d) obtained surface using SVDD for non-liquefaction data.

3.3. Discussion on SVDD parameters

There are two parameters σ, C (width parameter, penalty coefficient) in the SVDD which influences the outreach and the extent of the data domain. The influence of the two parameters to describe the data classification with Gaussian kernel function is shown in Fig. 6. Gaussian kernels with different widths σ ($\sigma=0.05, 0.15, 0.25$) and different penalty values C are combined in these figures. It must be noted that the penalty values for the two data classes are assumed to be the same (i.e. $C_1=C_2=0.05, 0.15, 0.5, 1$)

It is evident that for small values of the width parameter "σ", which represents weaker and more limited correlations between neighboring data, multiple isolated regions are detected, whereas by increasing its value a unified data zone with wider data class sphere is described.

The influence of the penalty coefficient "C" on inclusion of outliers is also presented in these figures. The smaller the value of "C", the tighter the region becomes by reducing the weight of outlier data. It is noteworthy that for a constant width parameter, the penalty coefficient determines the extent of data range and thus the noise data. The optimized description is determined by comparison of recognition rates obtained through ANFIS classifier. In other words, the strength of recognition pattern by ANFIS is set as a criterion for tuning the SVDD parameters.

3.4. SVDD-based up sampling CPT data

In view of the fact that the ratio of liquefied sample points to non-liquefied samples is over 3, the imbalance between different data classes will affect over pattern recognition procedure. Therefore, in the non-liquefied class region identified by the SVDD (as shown in Fig. 5(d)) data are generated.

3.5. Data generation procedure

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 to the determined center and the width of the minority class region. Then all SVDD models are used for acceptance or rejection of the generated data. In order to

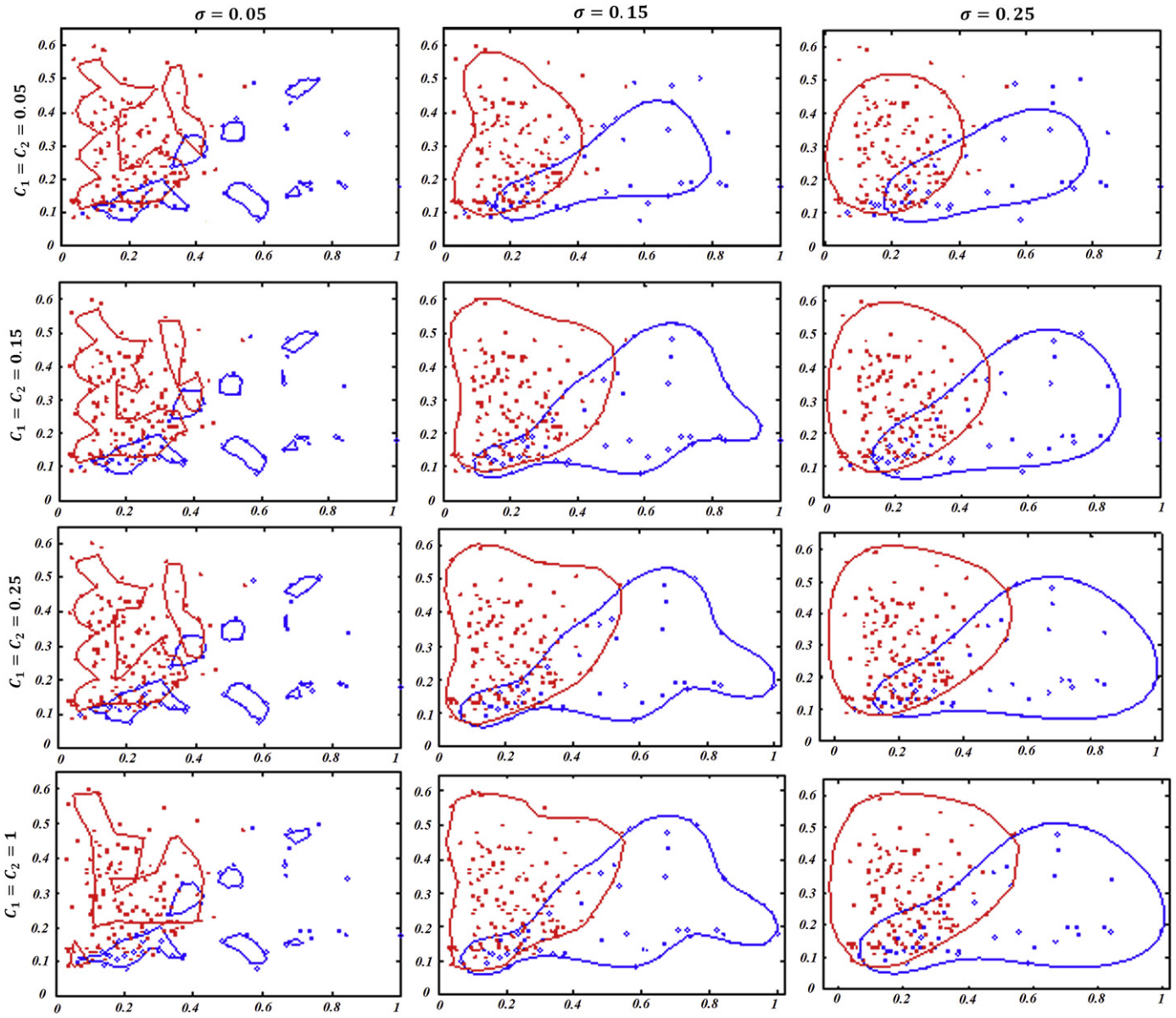


Fig. 6. Influence of different parameters σ, C_i (i=1, 2) on the SVDD with Gaussian kernel function. Three Gaussian kernels widths σ (σ=0.05, 0.15, 0.25) and different values for C_i (C_i =0.05, 0.15, 0.25, 1) are combined. Training data set contains two classes of objects size for liquefaction and non-liquefaction is 139 and 43 respectively.

clarify this procedure, an example of this is provided for σ=0.15 and C=0.15 in Fig. 7. It is evident that three zones exist; a liquefaction zone (L_Z), a non-liquefaction zone (NL_Z) and a fringe zone which includes intersection of the two sets ($L_Z \cap NL_Z$).

A sample point within the region boundary has a value less than zero, according to Eq. (18), DBR of a data point on the boundary is zero and for a data point outside the boundary this ratio is greater than zero. Data points A and B have percentage belonging ratios of (3.7%) and ($-$2.93%) respectively. Therefore, the data point A is an outlier and thus neglected for the up-sampling process.

Having defined the data region, the data generation is carried out on a Monte Carlo bases in the minority class region (i.e. NL_Z) excluding the fringe zone NL_Z-($NL_Z \cap L_Z$). The belnging ratio for the generated data must be less than zero for zone NL_Z and greater than zero for zones L_Z. Fig. 8 shows the up-sampled data.

3.6. Classification using ANFIS

In this study, ANFIS classifier (which was discussed in Section 2.3) is used to predict soil liquefaction. K-fold cross-validation 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 (Fig. 9). The ability of each model to predict is estimated by calculating errors on each test instances of each K fold. The advantage of K-fold cross validation is that all the 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).

In this study ANFIS main parameters ($\overline{c}$,$\overline{\sigma}$) have been determined to maximize generality. Ten folds were used in the K-fold cross-validation. The results of training and testing based on up-sampled data are shown in Table 2.

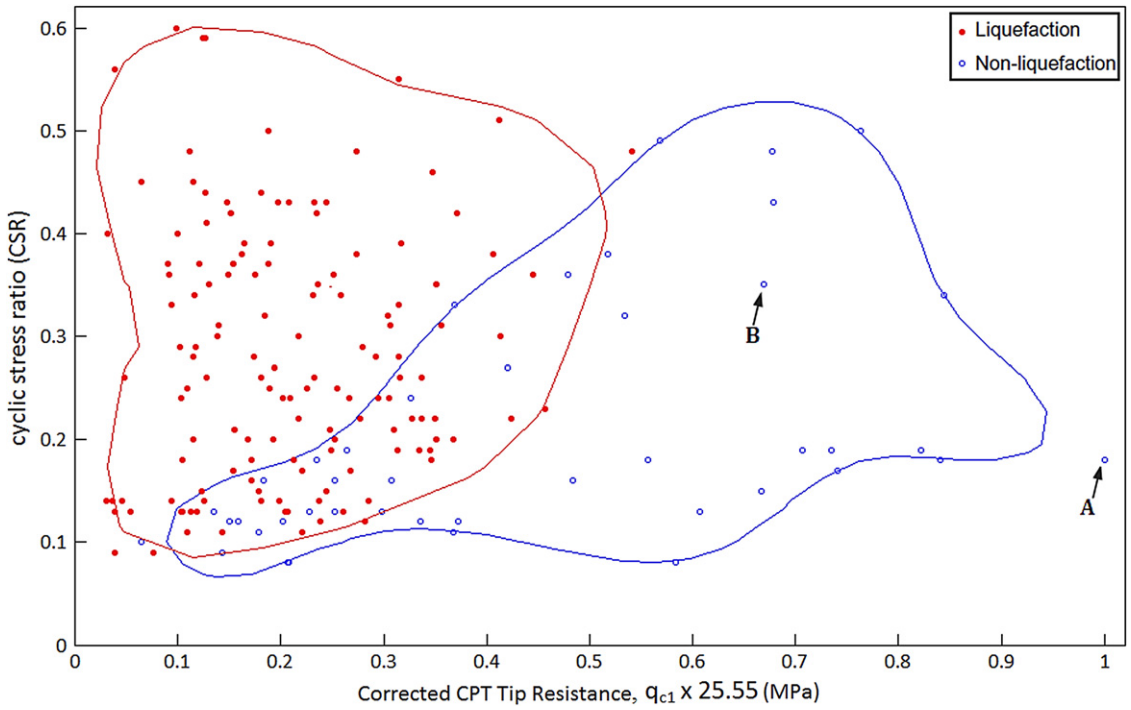


Fig. 7. Liquefaction, non-liquefaction and interference zone.

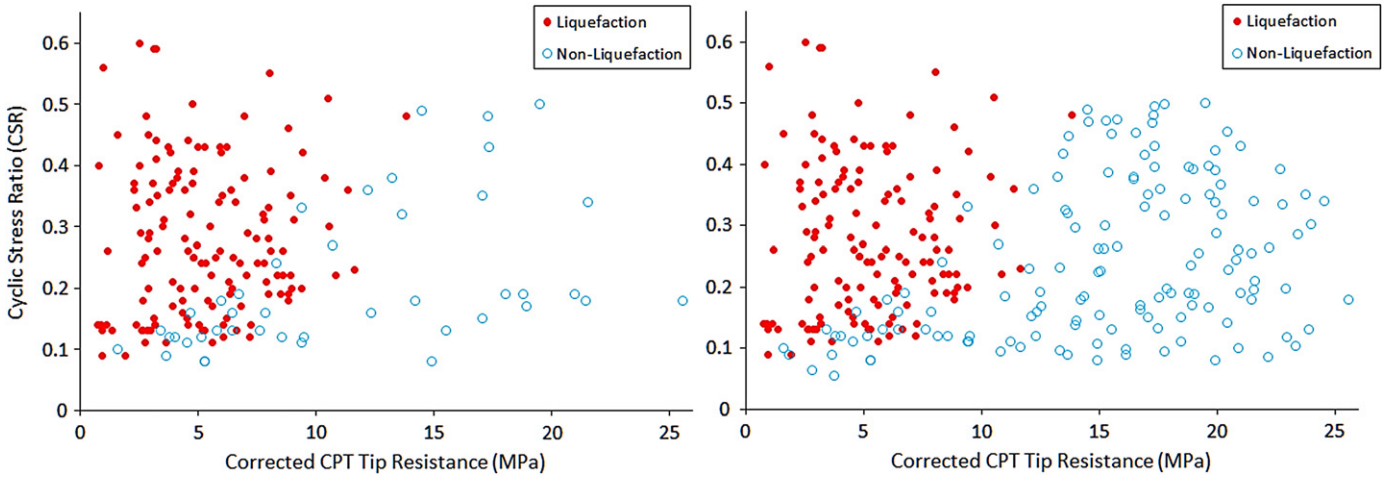


Fig. 8. (a) CPT based case histories in $CSR-q_{c1}$ space and (b) up sample CPT Data.

3.7. Recognition rate

At this stage the recognition rates produced by ANFIS are compared for various SVDD parameters. It should be noted that except for $\sigma=0.05$ values which give a discrete and multiple segment data boundaries, the other values have been tried for determination of best recognition rate. ANFIS is run ten times for two Gaussian kernels widths σ ($\sigma=0.15$, 0.25) and four values C ($C=0.05$, 0.15, 0.25, 1) as described in the previous section and the mean values of train and test procedure is evaluated. The outcome is shown in Fig. 10.

3.8. Parameter tuning procedure

Based on the values of recognition rates obtained by ANFIS classifier, it can be noted that the most accurate predictions are obtained using $C=0.25$, $\sigma=0.15$ and $C=1$, $\sigma=0.15$ which are not much different, and except for the higher test values for the first combination they can be assumed to produce the results with the same kind of accuracy. Here the former combination is used since it gives a slightly better test values.

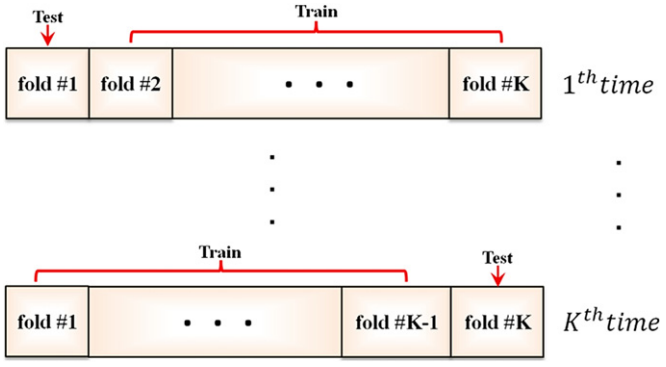
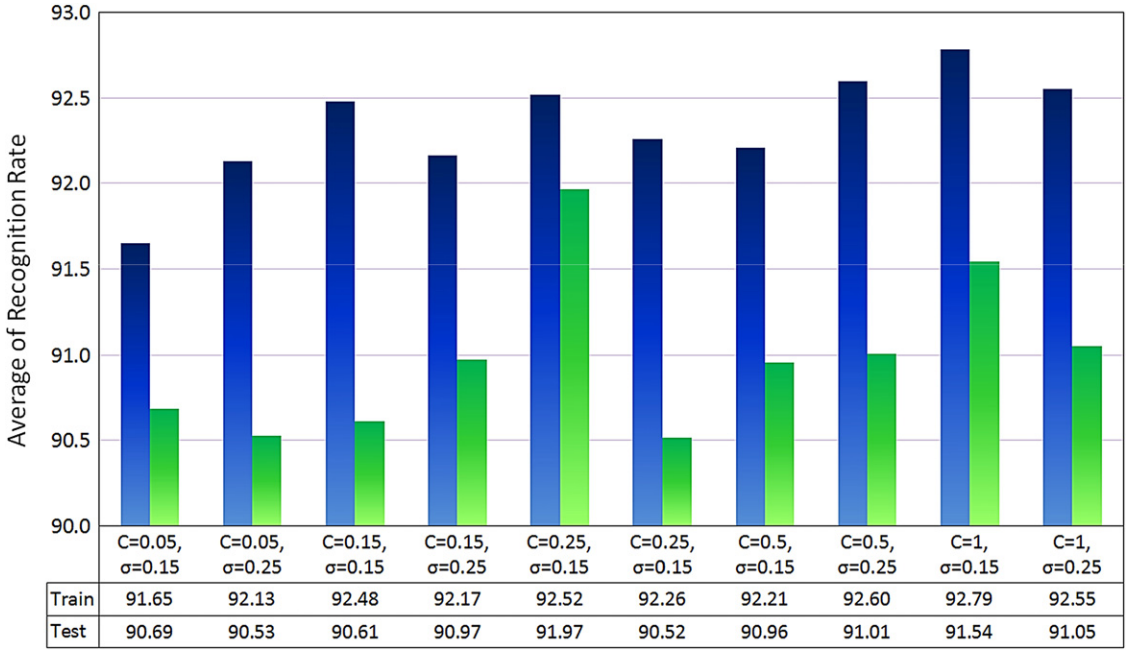


Fig. 9. K-fold cross-validation method for training and testing CPT data.

Table 2
Recognition rate for a single run (C=0.15, σ=0.15).

Fold number	1	2	3	4	5	6	7	8	9	10
Train	92	92.8	92	92.83	91.6	92.4	92	92.43	92.4	92.4
Test	96.43	85.71	89.29	88.89	100	89.29	92.86	88.89	89.29	89.29



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

Fig. 10. Comparison of average of recognition rate for two Gaussian kernels widths σ (σ=0.15, 0.25) and four C values (C =0.05, 0.15, 0.25, 1).

3.9. *Mathematical definition of the threshold*

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

$$f(q_{c,1},CSR)=\sum_{i=1}^{8}\omega_i z_i/\sum_{i=1}^{8}\omega_i \tag{28}$$

where ω_i and z_i are obtained as follows,

$$\omega_i(q_{c,1},CSR)=\exp\left(-\frac{\|q_{c,1}-b_{1i}\|}{a_{1i}}\right)\exp\left(-\frac{\|CSR-b_{2i}\|}{a_{2i}}\right) \tag{29}$$

$$z_i(q_{c,1},CSR)=a_i^* q_{c,1}+b_i^* CSR+c_i^* \tag{30}$$

where a_{1i}, b_{1i}, a_{2i}, b_{2i}, a_i^*, b_i^* and $c_i^*$are as following Table 3.

$$\begin{cases} f(q_{c,1},CSR)>0, & \text{Liquefaction occurs} \\ f(q_{c,1},CSR)<0, & \text{Non-Liquefaction occurs} \end{cases} \tag{31}$$

In order to demonstrate the effect of data imbalanced on ANFIS predictions, the prediction of ANFIS classifier based on both sets of imbalanced as well as up-sampled (or balanced) data sets are presented in Fig. 11.

The average recognition rate obtained by 10-fold cross-validation method for each run is shown in Fig. 12. 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 up-sampled data the average recognition rates for train and test data increase to 92.36% and 92.83% respectively.

Table 3
Value of a_{1i}, b_{1i}, a_{2i}, b_{2i}, a_i^*, b_i^*, c_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

4. Model validation

In order to evaluate the performance of the proposed classifier and develop a quantitative basis for comparison with other methods, a number of metrics are utilized. These include "overall accuracy", "precision", "recall" and "F-score". These metrics can be computed from the elements of a "confusion matrix", where each column of the matrix represents the instances in a predicted class, while each row represents the instances in an observed class (Table 3).

The instances where liquefaction has or has not occurred and they have correctly been predicted (i.e. row and column indices equal) are denoted as "True-Positive (TP)" and "True-Negative (TN)". These indices form the diagonal of the matrix where the classifier has proven effective in its recognition task. Whereas the off-diagonal indices indicate the instances of misclassification; that is where liquefaction has or has not occurred but the classifier has predicted

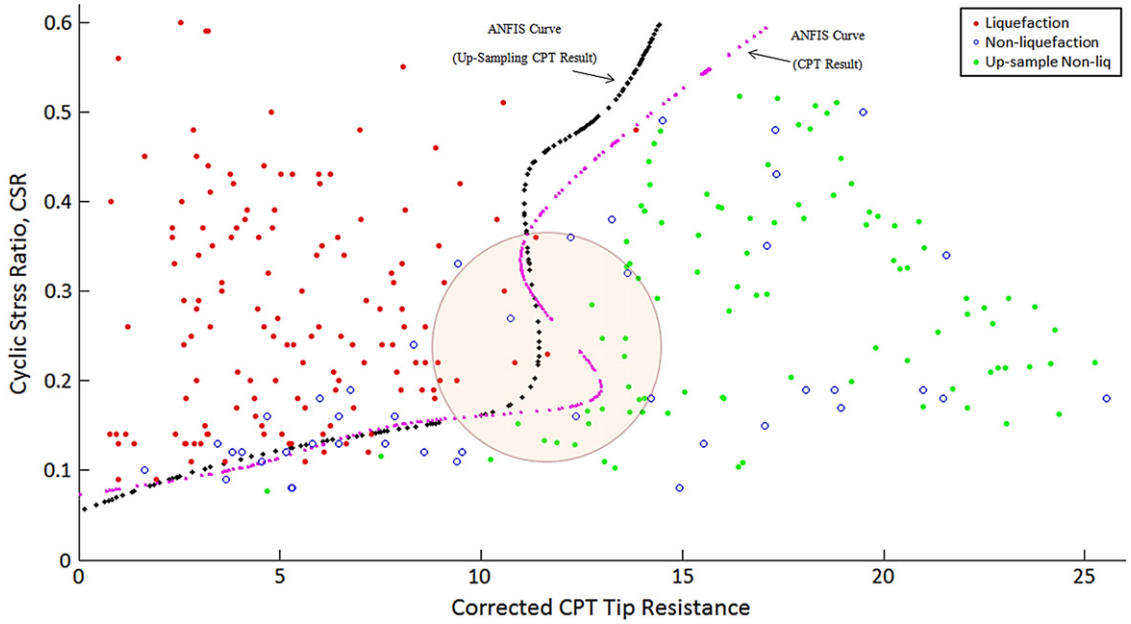


Fig. 11. ANFIS Curve for prediction of liquefaction based imbalance and balanced data.

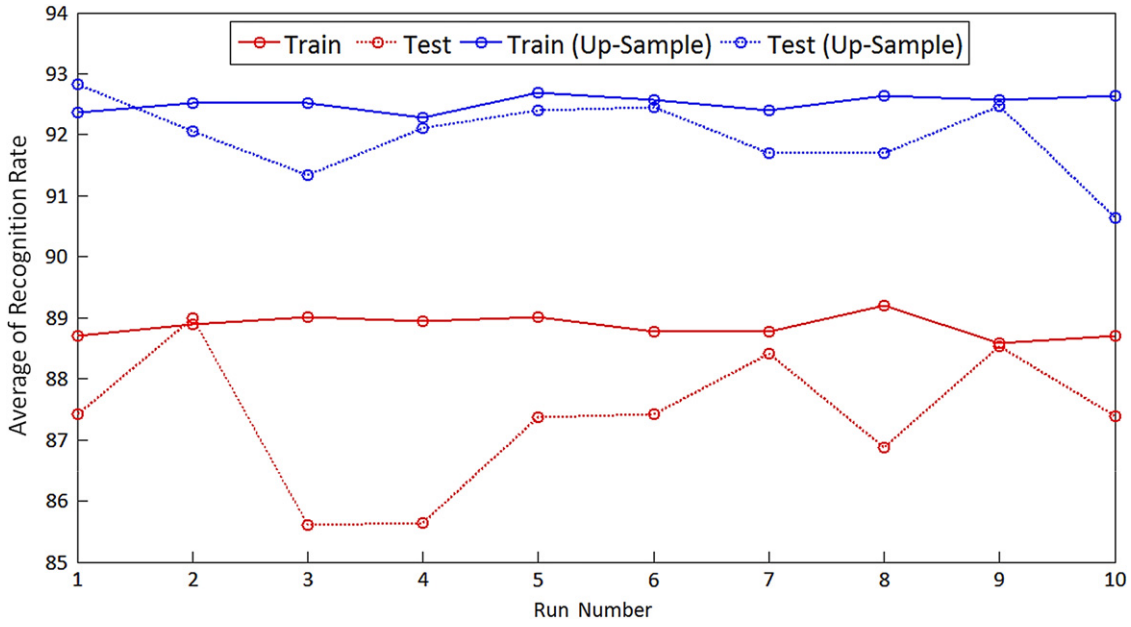


Fig. 12. Recognition rate for testing and training imbalance and balance data.

the converse. They are denoted are "False-Negative (FN)" and "False-Positive (FP)".

The accuracy of the classifier may then be defined as:

$$\text{Overall accuracy} = (TP+TN)/(TP+TN+FP+FN) \tag{32}$$

Overall accuracy is an overall measure of the capability of the classifier to predict the correct result. However, it is not necessarily an accuracy measure of each class individually. Precision and Recall are better measures of classifier performance in each class and are defined as:

$$\text{Precision} = p = TP/(TP+FP) \tag{33}$$

$$\text{Recall} = R = TP/(TP+FN) \tag{34}$$

These two metrics are especially useful where class imbalance exists and in this article the improvement achieved by SVDD in removal of the class imbalance are shown using these two metrics.

Finally, *F*-score is a weighted harmonic mean of precision and recall and combines the two measures to give a single evaluation metric.

$$F_\beta = (1+\beta^2)(P.R)/(\beta^2.P+R) \tag{35}$$

where β=measure of the importance of recall to precision and can be defined by the user for a specific project.

The above metrics have been evaluated using the dataset presented by Moss et al. (2006) in the approaches proposed by different researchers listed below in Table 4.

Before examining the performance of each approach it must be noted that Rezania et al. (2010), introduced three sub-categories for different soil types and thus the above metrics have been

calculated separately for each soil type and in order to form a common basis for comparison weighted averages of the metrics have been calculated according to the following formula:

$$OA_{EPR} = \frac{data_{SP}}{data_{total}} OA_{SP} + \frac{data_{SM}}{data_{total}} OA_{SM} + \frac{data_{SM-ML}}{data_{total}} OA_{SM-ML} \quad (36)$$

The overall performance of the SVDD up-sampled ANFIS technique is evidently better than many of the previously proposed approaches and is partially equaled by SVM (see discussion below).

In order to examine the effect of sampling bias on the overall accuracy and F-score of both data class, various ratios of non-liquefied to liquefied data ranging 0.5–2 were also tested. The results are presented in Fig. 13.

It is evident that optimum predictions are obtained when sampling bias approaches unity (Table 5).

Table 4
Confusion matrix.

Observed	Predicted	
	Yes	No
Yes	True-positive	False-negative
No	False-positive	True-negative

5. Discussion

The following points can be deduced from the above results:

- As shown in Fig. 12 the recognition rates of up-sampled CPT data have increased on average by about 4%. Furthermore, recognition rates are more consistent during different runs (see Fig. 12), which indicate a more stable identification procedure.
- The overall accuracy of the proposed method ranks higher than most of the others methods and is only equal to the SVM.
- Due to the improvement achieved by up sampling, F-score of the minority class (i.e. non-liquefaction) is the highest value achieved to date.
- The ratio of Precision to Recall is a measure of centrality of threshold in the fringe (intersection) zone. That is, the closer the ratio is to unity the more central the threshold is in the fringe zone. According to the results obtained by the up-sampled ANFIS, both Precision and Recall of liquefied and non-liquefied data are very close to each other, whereas the SVM has not been as successful, especially in the minority (non-liquefied) class. This proves the enhanced efficiency gained by up-sampling.
- As can be seen from Fig. 11, up-sampling has prevented over-fitting of the data and thus has caused a more general description. Imbalanced data had caused ANFIS to try to fit

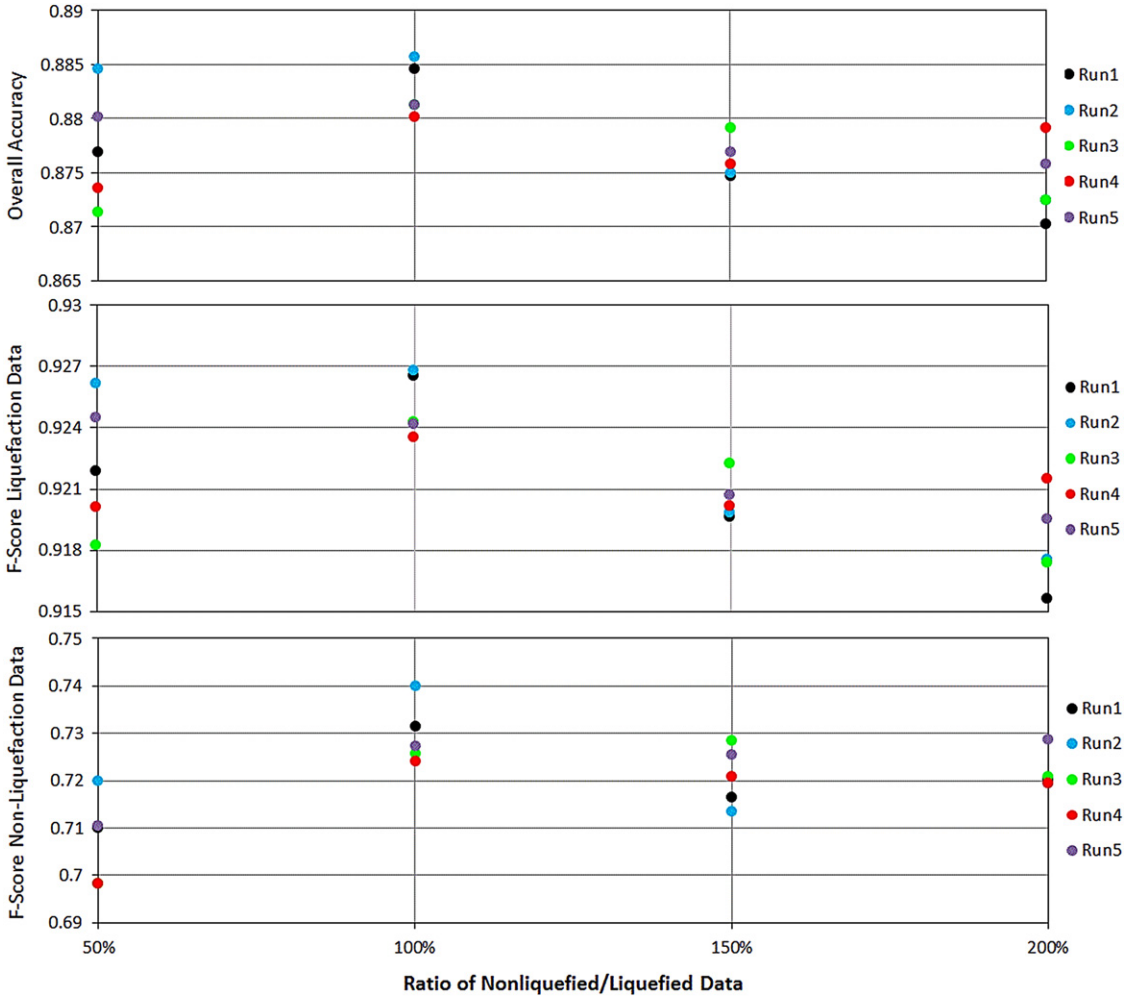


Fig. 13. Effect of non-liquefied/liquefied data ratio on OA and F-Score for five consecutive run (cluster center's range=0.5).

Table 5
Various estimates of the predictive performance of the CPT-Based deterministic models: (1) OA and (2) recall, precision, and F-score for both liquefaction and non-liquefaction occurrences.

Approach		Data set of Moss et al. 2006						
		OA	Liquefied			Non-liquefied		
			R	*P*	F-score	*R*	*P*	F-score
Youd et al. 2001	Simplified procedure	0.846	0.877	0.917	0.897	0.744	0.653	0.695
Moss et al. 2006	TH_L=0.15	0.879	0.985	0.872	0.925	0.534	0.92	0.676
	TH_L=0.5	0.857	0.913	0.9	0.907	0.674	0.7	0.69
Oommen et al 2010	SVM	0.89	0.978	0.888	0.931	0.604	0.896	0.722
Rezania et al. 2010	EPR (SP)	0.9	1	0.889	0.941	0.5	1	0.667
	EPR (SM)	0.84	0.939	0.869	0.903	0.462	0.667	0.545
	EPR (SM-ML)	0.556	0.294	1	0.455	1	0.455	0.625
	EPR (weighted average)	0.808	0.854	0.892	0.843	0.548	0.69	0.577
Rezania et al. 2011	EPR (three dimensional space)	0.841	0.878	0.91	0.894	0.721	0.646	0.681
$ANFIS_{up\text{-}sample}$		0.89	0.942	0.916	0.926	0.721	0.795	0.756

the description onto the minority class, whereas un-sampled data has led to a more general description.

6. 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 characterization of subsurface conditions and analysis of different aspects of soil behavior, comprising estimating the potential for liquefaction at a specific site.

The CPT database used in this study has 182 case histories of which 139 are from liquefied sites and 43 are from non-liquefied sites. The ratio of the data in the two classes indicates that serious data imbalance exists. The main scope of this study is to implement Adaptive Neuro-Fuzzy Inference System for the prediction of liquefaction threshold based on CPT Up-Sampled data. For identification of liquefaction and non-liquefaction regions, Support Vector Data Description method with suitable parameters (C and σ) has been used.

ANFIS classifier was used to predict soil liquefaction. For training and testing data model, K-fold cross-validation was used. It is shown that up-sampling has a positive bearing on recognition rates of ANFIS classifier by about 4%.

Furthermore, the performance of the overall technique has been compared against other newly proposed method. Certain metrics that exist in predictive analytics have been used as measures of classifier accuracy and generality. These have been invoked and calculated to form a basis of comparison. It is shown that the proposed approach has the highest overall accuracy equal only to SVM method simultaneously with generality in both classes of data.

Our future work shall focus on the development of kernel density estimation for inclusion of risk analysis.

References

Assaleh, K., 2007. Extraction of fetal electrocardiogram using adaptive neuro-fuzzy inference systems. IEEE Transactions on Biomedical Engineering 54 (1), 59–68.

Anderson, A.P., Gonzalez Jr, I., 2011. Species-specific tuning increases robustness to sampling bias in models of species distributions: an implementation with Maxent. Ecological Modelling 222 (15), 2796–2811.

Bae, M.H., Wu, T., Pan, R., 2010. Mix-ratio sampling: classifying multiclass imbalanced mouse brain images using support vector machine. Expert Systems with Applications 37 (7), 4955–4965.

Batuwita, R., Palade, V., 2010. FSVM-CIL: fuzzy support vector machines for class imbalance learning. IEEE Transactions on Fuzzy Systems 18 (3), 558–571.

Cetin, K.O., Kiureghian, A.D., Seed, R.B., 2002. Probabilistic models for the initiation of seismic soil liquefaction. Structural Safety, 67–82.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research 16, 321–357.

Chawla, N.V., Lazarevic, A., Hall, L.O., Bowyer, K.W., 2003. SMOTE Boost: improving prediction of the minority class in boosting. Proceedings of Seventh European Conference. Principles and Practice of Knowledge Discovery in Databases, 107–119.

Çivicioglu, P., 2007. Using uncorrupted neighborhoods of the pixels for impulsive noise suppression with ANFIS. IEEE Transactions on Image Processing 16 (3), 759–773.

Chen, S.I., Guo, G., Chen, L., 2010. A new over-sampling method based on cluster ensembles. Proceedings of International Conference on Advanced Information Networking and Applications Workshops, 599–604.

Daoming, G., Jie, C., 2006. ANFIS for high-pressure water jet cleaning prediction. Surface and Coatings Technology 201 (3–4), 1629–1634.

Depari, A., Marioli, A.D., Taroni, A., 2007. Application of an ANFIS algorithm to sensor data processing. IEEE Transactions on Instrumentation and Measurement 56 (1), 75–79.

Drummond, C., Holte, R.C., 2003. C4.5, Class imbalance, and cost sensitivity: why under-sampling beats over-sampling. ICML'2003 Workshop on Learning from Imbalanced Datasets II.

Gu, J., Zhou, Y., Zuo, X., 2007. Making class bias useful: a strategy of learning from imbalanced data. Lecture Notes in Computer Science, 287–295.

Han, H., Wang, W.Y., Mao, B.H. 2005.Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. Proceedings of International Conference on Intelligent Computing, 878–887.

He, H., Garcia, E.A., 2009. Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering 21 (9), 1263–1284.

Huang, M.L., Chen, H.Y., Huang, J.J., 2007. Glaucoma detection using adaptive neuro-fuzzy inference system. Expert Systems with Applications 32 (2), 458–468.

Jang, J.S.R., 1993. ANFIS: adaptive-network-based fuzzy inference system. IEEE Transaction on System Man and Cybernet 23 (5), 665–685.

Juang, C.H., Yuan, H., Lee, D.H., Lin, P.S., 2003. Simplified cone penetration test-based method for evaluating liquefaction resistance of soils. Journal of Geotechnology and Geoenvironmental Engineering 129 (1), 66–80.

Jo, T., Japkowicz, N., 2004. Class imbalances versus small disjuncts. ACM SIGKDD Explorations Newsletter 6 (1), 40–49.

Kishor, N., Singh, S.P., Raghuvanshic, A.S., 2007. Adaptive intelligent hydro turbine speed identification with water and random load disturbances. Engineering Applications of Artificial Intelligence 20 (6), 795–808.

Lee, K.C., Gardner, P., 2006. Adaptive neuro-fuzzy inference system (ANFIS) digital predistorter for rf power amplifier linearization. IEEE Transactions on Vehicular Technology 55 (1), 43–51.

Liu, X.Y., Wu, J., Zhou, Z.H., 2006.Exploratory under Sampling for Class Imbalance Learning. Proceedings of the International Conference on Data Mining, 965–969.

Liu, Y.H., Liu, Y.C., Chen, Y.Z., 2011. High-speed inline defect detection for TFT-LCD array process using a novel support vector data description 38 (5), 6222–6231.

Mitra, S., Hayashi, Y., 2000. Neuro-fuzzy rule generation: survey in soft computing framework. IEEE Transactions on Neural Networks, 748–768.
Moss, R., Seed, R.B., Kayen, R.E., Stewart, J.P., Kiureghian, A.D., Cetin, K.O., 2006. CPT-based probabilistic and deterministic assessment of in situ seismic soil liquefaction potential. Journal of Geotechnology and Geoenvironmental Engineering 132 (8), 1032–1051.
Nuno, A.I., Arcay, B., Cotos, J.M., Varela, J., 2005. Optimization of fishing predictions by means of artificial neural networks, ANFIS, functional networks and remote sensing images. Expert Systems with Applications 29 (2), 356–363.
Noureldin, A., El-Shafie, A.M., Tahab, R., 2007. Optimizing neuro-fuzzy modules for data fusion of vehicular navigation systems using temporal cross-validation. Engineering Applications of Artificial Intelligence 20 (1), 49–61.
Oommen, T., Baise, L.G., Vogel, R., 2010. Validation and Application of Empirical Liquefaction Model. Journal of Geotechnology and Geoenvironmental Engineering 136 (12), 1618–1633.
Oommen, T., Baise, L.G., Vogel, R., 2010. Sampling bias and class imbalance in maximum likelihood logistic regression. Mathematical Geosciences 43 (1), 99–120.
Quan, Z., Gang, L.G., Jun, C.W., Jun, W., Fu, C.S., 2006. Using an improved C4.5 for imbalanced dataset of intrusion. International Conference on Privacy, 15–19.
Qin, H., Yang, S.X., 2007. Adaptive neuro-fuzzy inference systems based approach to nonlinear noise cancellation for images. Fuzzy Sets and Systems 158 (10), 1036–1063.
Rezania, M., Javadi, A.A., Giustolisi, O., 2010. Computers and Geotechnics 37 (1–2), 82–92.
Rezania, M., Faramarzi, A., Javadi, A.A., 2011. An evolutionary based approach for assessment of earthquake-induced soil liquefaction and lateral displacemen. Engineering Applications of Artificial Intelligence 24 (1), 142–153.
Seiffert, C., Khoshgoftaar, T.M., Hulse, J.V., Napolitano, A., 2010. RUSBoost: a hybrid approach to alleviating class imbalance. IEEE Transactions on Systems, Man and Cybernetics-PART A: Systems And Humans 40 (1), 185–197.
Tax, D.M.J., Duin, R.P.W., 1999. Support vector domain description. Pattern Recognition Letters, 1191–1199.
Tax, D.M.J., Duin, R.P.W., 2004. Support vector data description. Machine Learning 54 (1), 45–66.
Youd, T.L., Idriss, I.M., Andrus, R.D., Arango, I., Castro, G., Christian, J.T., et al., 2001. Liquefaction resistance of soils: summary report from the 1996 NCEER and 1998 NCEER/NSF workshops on evaluation of liquefaction resistance of soils. Journal of Geotechnology and Geoenvironmental Engineering ASCE 127 (10), 817–833.
Zhang, J., Mani, I., 2003. K-NN approach to unbalanced data distributions: a case study involving information extraction. Proceedings of International Conference on Machine Learning, ICML.