SAFETY RISK EVALUATIONS OF DEEP FOUNDATION CONSTRUCTION SCHEMES BASED ON IMBALANCED DATA SETS

Peisong GONG [1], Haixiang GUO[2*], Yuanyue HUANG[1], Shengyu GUO[1]

[1]*School of Economics and Management, China University of Geosciences (WUHAN), Wuhan, China*
[2]*Mineral Resource Strategy and Policy Research Center, China University of Geosciences (WUHAN), Wuhan, China*

Abstract. Safety risk evaluations of deep foundation construction schemes are important to ensure safety. However, the amount of knowledge on these evaluations is large, and the historical data of deep foundation engineering is imbalanced. Some adverse factors influence the quality and efficiency of evaluations using traditional manual evaluation tools. Machine learning guarantees the quality of imbalanced data classifications. In this study, three strategies are proposed to improve the classification accuracy of imbalanced data sets. First, data set information redundancy is reduced using a binary particle swarm optimization algorithm. Then, a classification algorithm is modified using an Adaboost-enhanced support vector machine classifier. Finally, a new classification evaluation standard, namely, the area under the ROC curve, is adopted to ensure the classifier to be impartial to the minority. A transverse comparison experiment using multiple classification algorithms shows that the proposed integrated classification algorithm can overcome difficulties associated with correctly classifying minority samples in imbalanced data sets. The algorithm can also improve construction safety management evaluations, relieve the pressure from the lack of experienced experts accompanying rapid infrastructure construction, and facilitate knowledge reuse in the field of architecture, engineering, and construction.

Keywords: safety risk evaluation, construction scheme, deep foundation, imbalanced data set, ensemble learning algorithm, machine learning.

Introduction

Even though safety accidents are rare, deep foundation construction safety risks are generally high (Eskesen et al., 2004). Deep foundation construction schemes include all key information about construction, such as construction methods, schedules, and safeguard measures. As such, most construction safety technology risks can be identified and evaluated at the construction scheme stage. Safety risk evaluation systems of deep foundation construction schemes are therefore decided on the basis of construction schemes before construction starts to allow for a detailed evaluation of safety management issues that can arise during subsequent construction, optimize the allocation of related resources, and improve safety management quality and efficiency (Zhou & Zhang, 2011).

Safety risk reports of deep foundation construction scheme usually involve the establishment of an expert committee. Committee members study construction schemes and then offer their opinions on safety risk reviews on the basis of their own experiences. Safety review reports, which list safety risks that can occur, associated risk levels, and corresponding countermeasures, are then produced on the basis of collective decisions.

An analysis on major deep foundation construction safety accidents revealed that a common cause of accidents is an unreasonable construction design passed by safety risk evaluation committees (Ding et al., 2012). In the field of architecture, engineering, and construction (AEC), works, such as scheme evaluations, are based on related knowledge, which is enormous and barely grasped entirely by human brains (Zhong et al., 2012). Explicit and tacit knowledge are usually used in scheme evaluations. Code compliance checking primarily depends on explicit knowledge, whereas risk evaluations mainly depend on tacit knowledge. Finishing these works manually is time consuming because of the need for enormous knowledge and the influence of some other uncertain factors, such as the lack of experienced experts, the physical and mental state of the experts, the working arrangements, and work

*Corresponding author. E-mail: *15101286@qq.com*

attitudes, on the quality (e.g. Tan et al., 2010; Zhong et al., 2012, 2018; Luo & Gong, 2015). Computer-assisted code compliance checking has been researched and used in practice for many years to improve the quality and efficiency of the work (e.g. Yang & Xu, 2004; Tan et al., 2010; Zhong et al., 2012; Preidel & Borrmann, 2015; Zhong et al., 2018). Many researchers have focused on computer-assisted construction risk identification and evaluation (e.g. Ding et al., 2012; Zhou et al., 2013; Wang & Chen, 2017; Liu et al., 2018). As mentioned earlier, safety risk evaluation mainly depends on tacit knowledge, which is as enormous as explicit knowledge and is difficult to express using computers. Experienced experts accompanying the rapid development of infrastructure construction are also lacking. Thus, theoretical models that express tacit knowledge by using computers and that can improve the quality and efficiency of the safety risk evaluation of deep foundation construction schemes and facilitate knowledge reuse in the AEC field should be developed.

Safety risk evaluations of deep foundation schemes require a large amount of information. If a safety risk is associated with m individual risk factors, then the ith risk factor has a n_i possible value; therefore, that safety risk corresponds to $\prod_{i=1}^{m} n_i$ combination of individual risk factors, with each combination of risk factors representing a possible working condition. If all risk factors and safety risk values are scientifically reasonable, then all combinations of the $\prod_{i=1}^{m} n_i$ risk factors include all the safety risk review rules. Therefore, if a safety risk is associated with six risk factors and each risk factor has at least 2–5 possible values (discrete values), then the combination of risk factor values for that safety risk can easily exceed 1000. Therefore, knowledge on the safety risk evaluation include thousands of review rules.

The three major engineering deep excavation methods are open cut, cut and cover, and undercutting. Each method has many associated construction safety risks, and knowledge on safe risk evaluations of deep foundation construction schemes is enormous. Therefore, human experts cannot consider all information. Furthermore, human experts have different knowledge backgrounds and work experiences. As such, they have different and sometimes contradicting views on the same safety risks. Manual reviews are also affected by uncertain factors, such as the physical and mental state of human experts, working arrangements, and work attitudes. All these factors can significantly affect the evaluation quality. Given these problems, current safety risk evaluations of deep foundation construction schemes are generally incomplete. This condition compromises the construction safety management.

Given these difficulties regarding the quality and accuracy of safety risk evaluations of deep foundation construction schemes, significant research on improvement methods is available on the theoretical level.

Safety risk evaluations based on design documents have always been an important part of construction safety management. Seo and Choi (2008) studied the factors related to safety risks in underground engineering design documents and proposed a safety risk evaluation method based on design documents. Ding et al. (2012) proposed a method to identify metro construction safety risks automatically on the basis of engineering drawings and developed a metro construction safety risk identification system based on engineering drawings. Cao (2014) proposed a safety risk evaluation method for the design phase and verified the significant role of safety risk evaluations at the design stage for the risk control of two bridge construction projects.

Some risk evaluation methods, such as analytic hierarchy process (AHP), fuzzy sets, Bayesian networks, and fault tree analysis, have achieved good results. Sun et al. (2008) used the Delphi method to identify the safety risks in the construction of Olympic venues and established an AHP model to complete the risk evaluation. Zhou and Zhang (2011) improved a Bayesian network using a fault tree method and established an engineering technology risk deep foundation evaluation model with a fuzzy comprehensive method for evaluating multiple risks. Gao et al. (2013) used a fuzzy comprehensive evaluation method to conduct safety risk evaluations of construction worker behavior, from which a preliminary construction safety management system study was conducted. Pinto (2014) developed a safety risk evaluation model based on fuzzy sets to evaluate the risks of personal injuries. Zheng and Ma (2014) used a fuzzy AHP to conduct a comprehensive evaluation of construction safety risks associated with the Changchun Metro Line 1. Zhang and Ma (2014) integrated a fuzzy AHP and CIM model to evaluate the safety risks of the Xi'an Metro Line 4 construction. Park et al. (2015) collected data on safety accidents from the past 10 years, studied accident causes, and established an index system set to evaluate the safety risks for each construction process. This study achieved good results in practical project application. Sansakorn and An (2015) used fuzzy inference techniques to evaluate the safety risks of construction projects. Al-Anbari et al. (2015) analyzed the safety risks of an Oman construction project by consulting 40 experts in the field to identify the influencing factors, namely, 11 safety risks and 8 health risks. These factors were then evaluated using traditional risk analysis methods and health and safety risk analysis methods. The results proved that the latter gave more accurate and reliable evaluation results than the former. Wang et al. (2016) used a Bayesian belief network to conduct an evaluation of tunnel construction risks. Patel and Jha (2017) established a construction safety risk evaluation index system and evaluated construction project safety by using a fuzzy AHP. Liu et al. (2018) proposed a systematic method by integrating exploratory factor analysis and structural equation modeling. Zhang et al. (2019) proposed a real-time safety risk identification model to ensure the metro construction site safety of the metro.

Although the above-mentioned studies have achieved good results, none of these studies have scientifically man-

aged or maximized knowledge on safety risk evaluations of previous deep foundation construction schemes successfully. Most of these studies have used index systems or deep foundation construction evaluation safety risk models and have assessed these systems and models using comprehensive evaluation algorithms. However, several problems are observed with these evaluation models. First, these models have poor flexibility; specifically, once an evaluation index system is determined, updating it with new evaluation requirements is difficult, and the cost of such changes is high. Second, some models require experts for each evaluation. This precondition not only brings difficulty in overcoming the subjective factors but also increases the complexity in evaluation implementation. The evaluation becomes onerous because of the multiple evaluations needed for the safety risks of different projects or the different stages in these projects. Third, the accumulated knowledge from evaluations is ineffectively managed and cannot be reused.

Therefore, machine learning is a feasible method to complete safety reviews and resolve the aforementioned problems because it can manage massive safety risk knowledge. Machine learning can scientifically manage the large amount of valuable historical data and can compensate for the theoretical flaws in traditional safety risk evaluation methods. The developments in construction technology have introduced new engineering methods, which in turn present new safety risks. First, machine learning can quickly grasp and manage all related information. As such, the complete safety risk evaluation knowledge is considered when machine learning is applied to safety risk evaluations. Second, machine learning is unaffected by uncertain factors; the evaluation quality can be guaranteed as long as safety risk knowledge is included completely. Third, machine learning is efficient because it needs less time and constant reviews. In general, machine learning not only can compensate for the theoretical defects in current safety risk evaluation methods but also can overcome the practical difficulties of artificial evaluations.

Some studies have been conducted on using computers to manage knowledge on safety risks of deep foundation construction schemes. For example, Zhong and Li (2015) developed an ontology model to integrate construction knowledge and risk for deep foundation construction. Using a semantic reasoning method, the construction process safety risk knowledge was acquired quickly, and deep foundation construction safety risks were efficiently identified. This study provided a good basis for risk evaluations. Zhang et al. (2017) established a technical regulation case base for deep foundation construction using a weighted k-nearest neighbor (KNN) algorithm, which was found to improve the safety risk evaluation quality substantially.

However, although these studies have effectively established knowledge bases on safe risks of deep foundation construction, the problems associated with combining specific evaluation algorithms and complete actual evaluations and directly using knowledge bases for the safety risk

evaluations of specific projects have still not been resolved. Therefore, the current study aims to make existing knowledge bases effective for practical applications. To this end, machine learning algorithms based on existing canonical data sets (imbalanced data sets) are developed to achieve high-quality automatic safety risk evaluations. High-risk construction incidents rarely occur. Thus, knowledge data sets on the safety risk evaluation of deep foundation construction schemes are typically imbalanced. As a result, ensuring machine learning accuracy is difficult. Ensemble-based classifiers, also named multiple classifier systems (Krawczyk & Schaefer, 2013), improve the performance of a single classifier by combining several base classifiers that outperform every independent one (López et al., 2013). Tao et al. (2019a) proposed a novel self-adaptive cost weight-based support vector machine (SVM) cost-sensitive ensemble for imbalanced data classification. This method has better generalization performance in terms of G-Mean and F-Measure than other existing imbalanced data set classification techniques. Tao et al. (2019b) proposed a new over-sampling technique that uses the real-value negative selection procedure to generate artificial minority data with no requirement of actual minority data available. The proposed hybrid approach can achieve better performance in terms of G-Mean and F-Measure evaluation metrics than other existing imbalanced data set classification techniques. Classifier ensembles have become a popular solution method for class imbalance problems. These ensembles overcome the difficulties inherent in traditional evaluation methods and improve the quality and efficiency of safety risk reviews of deep foundation construction schemes.

The remainder of the paper is organized as follows. In Section 1, knowledge on deep foundation construction safety reviews is sourced and standardized, and a corresponding information system is established. In Section 2, the integrated classification algorithm for imbalanced data sets is introduced. In Section 3, a specific risk is used to compare the proposed algorithm with other popular algorithms for verifying the effectiveness of the proposed algorithm.

1. Knowledge on safety risk evaluation of deep foundation construction schemes

Previous deep foundation construction projects have resulted in a large amount of valuable safety risk management information. If computer science is used to manage this safety risk knowledge, then the quality and efficiency of deep foundation construction safety reviews can be significantly improved.

1.1. Expression and source for the knowledge on safety risk evaluation of deep foundation construction schemes

The knowledge expressions of safety risk evaluation should be first standardized to enable their learning by using computers. The knowledge on safety risk evalua-

tion of deep foundation construction schemes can be expressed as a series of IF–THEN rules; in other words, for any specific safety risk, the evaluation knowledge can be expressed as "if the… condition occurs, then the risk occurrence level is …". Therefore, in theory, the safety risk evaluation knowledge can be expressed as an information system.

In the information system equation, $S = (U, C \cup D, V, f)$, $U = (x_1, x_2, \cdots x_n)$ is a domain space, C is a conditional attribute set, D is a decision attribute set, and $V = \bigcup_{a \in C \cup D} V_a$, where V_a is the value range for attribute a and $f : U \times A \to V$ is an information function; if $\forall x \in U, a \in C \cup D$, then $f(x, a) \in V_a$.

Previous daily management of deep foundation engineering constructions has resulted in a massive amount of construction management information, with the safety risk review knowledge embedded deep in the original information. The original information includes all kinds of extraneous daily construction management information, such as construction diaries, supervision diaries, monthly supervision reports, safety patrol records, safety situation analysis, regular meetings, and meeting minutes as well as related technical data, such as the Case Compilation of Safety Accidents in Deep Foundation Construction (hereinafter referred to as the Compilation of Cases, which is compiled from past safety accidents and lessons learned).

Many construction and supervision diaries have no precise information about construction progress and management methods. Therefore, the original information is insufficient to build the safety risk evaluation knowledge, and the information needs to be obtained by other ways. Some full-time engineers must be sent to construction sites to record information, such as construction progress and safety management information, in detail in accordance with the consult contract. In this manner, the precise information of safety risk information can be obtained. For instance, in accordance with the construction progress, engineers can record different values of the ratio of penetration depth to excavation depth and other key safety risk factors and different safety management information that can infer safety risk occurrence probability levels (Table 9). The main information that engineers record in the construction sites is usually transferred into two forms. The day-to-day recorded information is imported into the early warning construction safety system, and the high-risk event information and accident information is gathered into the Case Compilation of Safety Accidents in Deep Foundation Construction.

The following is a brief description of the process required to identify the safety risk factors from the daily construction data Safety Patrol Records and the Compilation of Cases.

1.1.1. Obtained from safety situation analysis

In the early warning construction safety system (an information system that works as a safety management platform for all the units involved in the deep foundation construction project), safety management engineers and experts use their knowledge and experience to analyze and evaluate site safety management and make suggestions on the basis of the actual construction and the construction schemes. Then, they submit the safety situation analysis (compiled from the evaluation result and suggestions) on the system daily. Therefore, the safety situation analysis in the construction safety early warning system is specific to the risk evaluation knowledge of the safety engineers and experts. However, in general, only professionals with relevant knowledge can read, understand, or use this knowledge. Therefore, given that the specific information needed by the machine learning is implied in the safety situation analyses, it needs to be manually processed before it can be used.

Safety engineers and experts gather to investigate an independent project. The hidden risks can be identified and the risk factors can be obtained by combining and analyzing the project construction scheme information from the detailed analyses of the daily safety situation over the entire construction process. The safety risk evaluation knowledge can be determined by integrating all projects in the construction safety early warning system. However, all this preliminary work needs to be completed by professionals because safety risk evaluation knowledge is difficult to extract from safety situation analysis texts accurately.

Figure 1 shows the logical extraction process of safety risk evaluation knowledge. The results of the extraction are expressed as a combination of safety risks and risk factors. Figure 2 shows a safety situation analysis obtained from the safety early warning system for a specific project on a certain day. From the construction scheme and project safety situation analysis, the safety risk and its risk factors can be identified. The information in the first column in Table 1 is extracted from Figure 2. For example, the safety analysis on the first line in Table 1 requires that construction units excavate the earth following the excavation plan. On the basis of the professional knowledge and experience of the safety engineer, if the excavation plan is not strictly applied, then a risk of landslides exists in the foundation pit. The construction scheme information related to this risk is soil quality and excavation height.
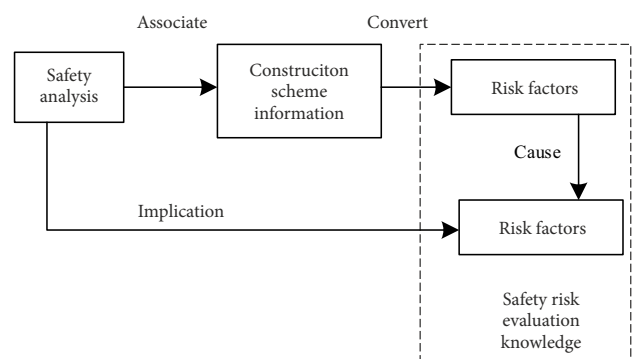

Figure 1. Diagram of logical extraction process of safety risk evaluation knowledge

Recommendations for safety evaluation

According to the on-site safety risk inspection and monitoring data analysis, my project group gives the following suggestions:

1. At present, earthwork is being excavated; the construction unit is requested to carry out grading and excavation in strict accordance with the excavation plan, and the piles of dikes shall be cleared and transported promptly;

2. The excavations of slope protection on both sides for advice, timely supporting, and timely carry out precipitation work, ensure the water level in the excavation face, pay close attention to digging the soil under the stability, timely reporting of the monitoring unit monitoring data;

3. The construction unit should be urged to strengthen the inspections on the wall to the ground, especially the joints, and promptly handle any signs of seepage or leakage;

4. The construction unit should be asked to strengthen the monitoring of the water supply pipeline and the piers of light rail in the small mileage of the foundation pit;

5. The construction unit should be asked to pay attention to the stability inspection of the scaffolding to ensure the stability of the stiffness of the scaffold;

6. The construction unit is urged to strengthen the on-site temporary electricity, mechanical hoisting and other safety management;

7. Rain in the next few days, the construction unit should be asked to pay attention to the slope protection, drainage of water in the foundation pit in time, and pay attention to the construction measures of rainy days.

Figure 2. Example of a safety situation analysis (source: Wuhan metro construction safety early warning system)

Table 1. Risk factors extracted form a safety situation analysis

Safety analysis	Safety risk	Project construction scheme information	Risk factors
At present, the earthwork is being excavated, and the construction units are requested to be excavated in strict accordance with the layered subsections of the excavation plan	Soil landslide in foundation pit	Soil texture of foundation ditch and delamination height of excavation	Soil texture at the bottom of foundation ditch and layered height of earthwork excavation
The pit soil heap must be removed in a timely manner. Sloping soil should be retained on both sides of the excavation. Supports should be set up in time	Deformation of enclosure structure	The underground continuous wall is used as the surrounding support structure (e.g., wall thickness, depth, and depth ratio) using steel support	Surrounding load for the foundation pit, type of exterior-protected construction, depth ratio of the exterior-protected construction, rock entry for the exterior-protected construction, and support style
The construction unit should be urged to strengthen the inspections on the wall to the ground, especially the joints, and promptly handle any signs of seepage or leakage	Sand gushing and influx in the enclosure structure	The underground continuous wall is used as the surrounding support structure (e.g., wall thickness, depth, rock entry condition, and depth ratio)	Type of exterior-protected construction, rock entry condition of exterior-protected construction, depth ratio of exterior-protected construction, and groundwater
The construction unit should be asked to strengthen the monitoring of the water supply pipeline at the end of the small mileage of the foundation pit	Peripheral pipeline damage	Groundwater and surrounding pipeline situation	Groundwater situation, types of pipeline, pipeline material, and distance between pipeline and foundation pit
The construction unit should be asked to strengthen the monitoring of the pier of the light rail at the end of the foundation pit.	Destruction of surrounding buildings	Bridge pier foundation of light rail (e.g., type, embedded depth, rock entry, and distance from foundation pit)	Type of foundation, embedded depth of foundations, rock entry of foundations, and distance from foundation pit

Therefore, the safety risk evaluation knowledge is extracted following the safety situation analysis. In other words, the soil landslides in the foundation pit are related to two risk factors (construction scheme information), namely, the soil texture at the bottom of foundation ditch and the delamination excavation height. The combination of the two risk factors determines the risk evaluation results. The formal expression for the safety risk evaluation knowledge is shown in Table 6.

1.1.2. Obtained from the Compilation of Cases

For long deep foundation construction, builders make a detailed summary of high-risk events and safety accidents that can occur during construction. They collate and summarize the process, cause, treatment, and experience of these events or accidents and then compile the Case Compilation of Safety Accidents in Deep foundation Construction. This book focuses on the clues associated with the safety risk factors. We can determine the information associated with the hydrogeology, construction design, surrounding environment, and climate, analyze the risk events described in the case, summarize the accident causes, and identify the associated risk factors in each case depending on the general project situation.

For example, the Case Compilation of Safety analyzes in detail the risk of an excessive settlement of the surrounding buildings during the deep foundation construction of a metro station and summarizes the main risks.

The direct cause for any excessive settlement of the surrounding buildings is the settlement of the surrounding surfaces. As such, the risk factors associated with the surface subsidence risk directly in the text describing the cause of the risk can be generalized. The descriptions of the risk causes are listed in the Case Compilation of Safety, with the added risk factors behind the risk factor clues being given in the parentheses:

- Stratum consolidation settlement (soil settlement caused by artificial dewatering of the foundation pit). The underground continuous wall of the station has not entered the rock (the rock entry condition of the enclosure structure), and underground confined water is not cut off (condition of groundwater). Dewatering in the foundation pit and groundwater level inside and outside the foundation pit is decreased at the same time. This condition leads to an increase in pumping time and quantities. In addition, a certain water-head exists between the inside and outside of the foundation pit (water head inside and outside the foundation pit). This condition causes soil destruction, especially of the fine sand layer (soil texture of foundation ditch) under the water flow action field. Dewatering causes water loss and soil erosion under the building foundation, a decrease in the soil water content around the building, an increase in the effective stress between the soil particles, and a downward movement of the soil mass. These phenomena cause buildings to sink.
- Time-space effect (exposure of the foundation pit for a long period). A long time is needed for the foundation pit excavation due to the influence of dewatering time. This condition causes a deformation in the continuous concrete wall (deformation of exterior-protected foundation pit construction) and an increased settlement of the surrounding soil.
- Settlement of surrounding soil caused by the deformation of in the continuous concrete wall (deformation of exterior-protected construction of the foundation pit). The shape of the station foundation pit is irregular, and the steel supports (support type and erection situation) are not set in time. This situation results in a deformation in the continuous concrete wall to the inside of the foundation pit because of the soil and water actions. In other words, the soil surrounding the foundation pit settles. This settlement eventually leads to further building settlement.

Table 2. List of risk factors obtained from the Compilation of Cases

Safety risk	Risk factors
Surrounding ground surface settlement	Artificial precipitation
	Foundation ditch soil texture
	Rock entry condition of the enclosure structure
	Groundwater
	Water-head inside and outside the foundation pit
	Deformation of envelop enclosure
	Support type and erection situation

By analyzing the above-mentioned reasons, the main risk factors for the risk of surrounding ground surface subsidence are obtained and shown in Table 2.

The safety risk evaluation result (decision attribute set D) is obtained from the relevant construction norms and the relevant daily construction management measures.

The risk level is determined depending on the probability of risk occurrence and the associated risk losses. The method for determining the risk level is stipulated in the Code for Risk Management of Underground Works in Urban Rail Transit (GB50652-2011). The evaluation of the risk level in this specification is shown in Table 3.

The risk occurrence probability is divided into five evaluation grades: frequent, possible, occasional, rare, and impossible. The risk level is then determined from the possibility of the risk occurrence and the extent of the possible losses. Different principles and measures are available to cope with the risks depending on the risk evaluation level, as shown in Table 4.

In safety risk evaluation knowledge information systems, the class label is the risk possibility level (decision attribute set D). However, the definitions for the risk possibility grade are inoperable because the risk probability cannot be measured accurately; however, it can be indirectly determined using the daily construction management data.

In practical situations, depending on the risk level, construction site managers apply different management measures, which are given in daily construction documents, such as construction or supervision diaries. These risk management measures can be used as a clue to indirectly determine the probability levels for the various

Table 3. Risk level determination table
(source: Code for Risk Management of Underground Works in Urban Rail Transit (GB50652-2011))

Degree of possibility		Degree of damage				
		A: Calamitous	B: Very serious	C: Serious	D: To be considered	E: Negligible
> 0.1	Frequent	I	I	I	II	III
0.01–0.1	Possible	I	I	II	III	III
0.001–0.01	Occasionally	I	II	III	III	IV
0.0001–0.001	Rare	II	III	III	IV	IV
< 0.001	Impossible	III	III	IV	IV	IV

Table 4. Risk management principle table
(source: Code for Risk Management of Underground Works in Urban Rail Transit (GB50652-2011))

Rank	Acceptance criteria	Principles of disposal	Control scheme	Response Department
I	Unacceptable	Risk control measures must be taken to reduce risk, and the risk level must be reduced to acceptable or unwilling levels	Risk early warning and emergency treatment plan should be prepared, or the scheme should be revised and adjusted	Government departments and the parties to the construction of the project
II	Unwilling to accept	Risk management should be implemented to reduce the risk, and the cost required to reduce the risk should not be higher than the loss after the risk	Risk prevention and monitoring should be implemented, and the risk management measures to be supported should be specified	
III	Acceptable	Risk management should be implemented, and risk treatment measures should be taken	Daily management and monitoring should be strengthened	All parties to the project construction
IV	Negligible	Risk management can be implemented	Routine examination and examination can be performed	

Table 5. Risk occurrence probability level judgment standard

Rank	Possibility	Management measure characteristics
1	Frequent	All units should reduce the possibility of common measures to handle the risk.
2	Possible	Reinspection and improvement of the construction scheme, focus on the daily work, and close monitoring of all accident symptoms.
3	Occasional	Field investigation is strengthened to focus on the risk.
4	Rare	Not a major concern; management limited to routine general inspections.
5	Impossible	Not listed in the risk list; little possibility of the occurrence and completely negligible.

risk occurrences indirectly. Referring to the provisions for the early warning level management measures in the Standard for Construction Safety Evaluation of Metro Engineering (GB50715-2011, 2011) that correspond to the five risk probability grades in Introduction of Code for Risk Management of Underground Works in Urban Rail Transit (GB50652-2011, 2011), the management measure characteristics that correspond to the probability grades for each risk can be defined, as shown in Table 5.

From the evaluation terms in Table 5, the risk probability grade at the time of each risk factor combination can be determined on the basis of actual past situations and the management measures taken at that time.

1.2. Characteristics of safety risk evaluation knowledge of deep foundation construction

By sorting and analyzing the deep foundation construction management data with a specific deep foundation construction safety risk as the object, rich safety risk evaluation knowledge can be extracted and expressed as an information system, as shown in Table 6.

A typical drawback in information systems, such as in Table 6, is the small sample size for the high-risk decision-making attributes. High-risk construction incidents rarely occur, and safety accidents occur even less frequently. As such, the majority of recorded incidents is low-risk or general risk incidents (impossible, rare, or occasional). Therefore, the maximum data imbalance ratio is generally

Table 6. Information system of safety risk evaluation of deep foundation construction

CA1	CA 2	......	CA n	DA
V_1^{ca1}	V_1^{ca2}	......	V_1^{can}	Frequent
V_2^{ca1}	V_1^{ca2}	......	V_3^{can}	Possible
......	......	......	......	......

Note: CA is a conditional attribute, DA is a decision attribute, and V_1^{ca1} is the first attribute value for the conditional attribute CA 1.

above 30 (maximum imbalance ratio = maximum sample number/least sample number). Therefore, the specific safety risk evaluation information system has significant data imbalance characteristics. However, the purpose of construction safety risk evaluations is to focus on the accuracy of the high-risk grade evaluations with a small sample size.

Therefore, the information system of safety risk evaluation knowledge of deep foundation construction is a typical imbalanced data set because the lack of high-risk samples creates difficulties in ensuring machine learning accuracy. Consequently, traditional classification algorithms cannot guarantee classification accuracy. The relationships between imbalanced proportions and classification performance have been previously studied; furthermore, although the relationships between the imbalanced proportion and the classification performance have not

been accurately given, 1:35 or even 1:10 has been found to cause classifier failure in some cases (Weiss & Provost, 2003). A typical KNN algorithm selects the K samples closest to the classification object and votes in accordance with the class labels on the K samples to decide on the category for the objects to be classified. Therefore, the classification results from the KNN algorithm are significantly biased toward the majority. The current study focuses on a machine learning method for safety risk evaluation of deep foundation construction schemes based on an imbalanced data set.

2. Risk evaluation based on improved ensemble classification algorithm

The essence of the theory for the evaluation of safety risks based on historical data is classification. For specific classifications, available mature classification methods include decision tree, Bayesian, artificial neural networks, KNN, and SVM. Unfortunately, these traditional classification methods show poor performance on imbalanced data sets (Li et al., 2016a).

However, the minority of imbalanced data sets has had significant focused research attention. For safety concerns, classification errors of minority often will pay high cost. The minority data in the deep foundation construction safety risk evaluation knowledge information system are related to high risk grades. As such, any classification errors can cause significant safety management risks and profound consequences for the construction.

Fruitful research has been conducted on the classification of imbalanced data sets. In general, three strategies have been used to improve the classification of imbalanced data sets (Li et al., 2016b). The first strategy requires that the data set condition attributes be decreased to reduce data redundancy and improve classification accuracy. The second strategy requires a transformation of the data set and the algorithm, for which a sampling method is used to transform the data sets and balance the number of samples to improve the data set (Chawla et al., 2002; C. Y. Lee & Z. J. Lee, 2012). Common methods for the algorithmic transformation are ensemble learning, cost-sensitive learning (Chawla et al., 2003; Lesser et al., 2006), and one-class classifier (Zheng et al., 2006); among which, ensemble learning methods have been the most widely used. Typical integrated learning methods include the boosting algorithm (Schapire, 1990) and the bagging algorithm (Breiman, 1996). Ensemble learning algorithms have been significantly improved on the basis of the two algorithms in recent years (Galar et al., 2012; Yang et al., 2008). The third strategy is the adoption of a new classification evaluation criterion to prevent the classifier being biased toward the majority.

Although the aforementioned strategies can improve the classification of the imbalanced data set, using a single strategy effect has a relatively poor performance than combination strategy effect. Therefore, the present study

attempts to integrate several strategies for proposing an integrated learning method based on the safety risk evaluation knowledge of deep foundation construction imbalanced data set. This algorithm uses the feature selection in the first strategy to reduce the condition attributes of the data set. Thereafter, the second strategy based on an integrated AdaBoost and SVM algorithm is adopted to improve the classifications. Then, the third strategy is used to evaluate the effect of a classifier based on the classifier evaluation criteria tailored for multi classification objectively. Finally, the complete integrated algorithm proposed in Section 2.4 is an integrated binary particle swarm optimization (BPSO), AdaBoost, and SVM algorithm. The evaluation index from the third strategy is then used to evaluate the effect of the integrated algorithm.

2.1. Feature selection

Treating minority samples in imbalanced data sets as noise is easy. As such, the classification results are unsatisfactory. Therefore, feature selection is necessary before the formal evaluation to reduce the redundant features and noise in the data set and thus prevent the risk of the minority samples being regarded as noise for improving the classification effect of minority groups (Chuang et al., 2008).

As genetic algorithm (GA) is widely used in feature selection, other methods are also used frequently because of their own advantages. Filter and wrapper feature selection methods are most regularly used in feature selection (Li et al., 2016b). BPSO is a wrapper feature selection method. Particle swarm optimization (PSO), which was devised by Kennedy and Eberhart (1995), is a widely used stochastic evolutionary algorithm for solving optimization problems. Unlike other evolutionary algorithms (e.g., GA and DE), PSO does not contain crossover and mutation operations. Thus, the complexity of the model is decreased.

PSO has been a common method for feature selection. In 1995, Kennedy and Eberhart proposed a swarm-based intelligent evolutionary algorithm PSO algorithm (Hassan et al., 2010), which can solve feature selection problem in a continuous real number space, on the basis of the behavioral characteristics of bird predation. However, given that many problems are associated with the binary discrete space in practice, Kennedy and Eberhart further proposed the BPSO algorithm in 1997 (Kennedy & Eberhart, 1997).

The idea is specified as follows. Each particle is regarded as an n-dimensional (n characteristic number) 01 vector, with each character in the vector representing a feature. A value of "1" indicates that the feature is an important feature and should be selected, and a value of "0" indicates that a feature is not an important feature and should be abandoned. A number of 01 vectors are randomly generated. The standard PSO algorithm is used for the iterative calculation to determine the optimal 01 vector. The result of the attribute selection is obtained. The fitness value of the particle is AUC_{area} as the standard, which is further described in Section 2.4.

2.2. AdaBoost+SVM classification algorithm

2.2.1. AdaBoost series classifiers

The Boosting algorithm uses a combined learning method to train a weak classifier to be a strong classifier (Khan et al., 2014). The most widely used algorithm of this type is the AdaBoost algorithm proposed by Schapire and Freund in 1996 (Sun et al., 2007). The basic idea of the AdaBoost algorithm is specified as follows. AdaBoost selects a weak classifier and sets the initial weight $1/m$ for each sample of the training set (m is the number of training set samples). Then, the classifier is used several times, and the sample weight is updated in accordance with the training result. The updating principle increases the weight of the samples that fail in the training to ensure that the classifier focuses on learning the failed samples for the subsequent classification calculation. After the completion of each classification computation, a classification prediction function is generated, and each prediction function is given a corresponding weight depending on its classification accuracy. Notably, a high precision indicates a great weight. In this manner, a prediction function sequence can be obtained after a number of classification calculations. A weighted voting method is applied to the prediction function sequence to obtain the final prediction result, which is the classification calculation result. Therefore, the AdaBoost algorithm can transform a weak classifier into a strong classifier by multiple iterations.

The basic AdaBoost algorithm can only handle binary classification problems. On this basis, Schapire (1990) proposed a series of derivative algorithms in which the AdaBoost.M1 algorithm can handle multiple classification problems (Cao et al., 2013). The process for the AdaBoost. M1 algorithm is shown in Table 7.

2.2.2. SVM

SVM was formally proposed in 1995 (Cortes & Vapnik, 1995) and demonstrated superior text categorization performance (Joachims, 1998). However, the support vector concept emerged as early as the 1960s. It provided a solid theoretical basis for SVM. SVM is an excellent small sample learning method with good classification performance because its core is maximizing the classification margin. The final SVM decision function is only determined by a few support vectors. The complexity of the calculation depends on the number of support vectors rather than the sample space dimensions. Thus, this method effectively avoids the curse of dimensionality. SVM has good robustness because it determines the final result using only a few support vectors.

Nonetheless, the SVM calculation is complex because it has a high time complexity that reaches $O\left(N_{sv}^3 + N \times N_{sv}^2 + D \times N \times N_{sv}\right)$ (N: sample dimension; N_{sv}: number of support vectors) when calculating large samples. In addition, the basic SVM algorithm can only support binary classification problems.

However, given the data set sample space in this study is not large (only 432 samples in all), the calculation is not complicated. Although the data set in this study is a multi-classification problem, the SVM can be transformed

Table 7. AdaBoost.M1 algorithm process (Weiss & Provost, 2003)

AdaBoost.M1
Input: Training set $D = \left\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_i) \cdots (\boldsymbol{x}_m, y_k)\right\}$, $i = 1, 2, \cdots, m$, $y_i \in Y$, $Y = \left\{C_1, C_2, \cdots, C_k\right\}$ as the classification label; Base learning algorithm $\mathcal{L}$; Number of iterations T.
Process: 1. Initial weighting of samples.: $\mathcal{D}_1(i) = 1/m$. 2. for $t = 1, 2, \cdots, T$ do 3. $h_t = \mathcal{L}(D, \mathcal{D}_t)$ Training classifier h_t from training set D based on distributed $\mathcal{D}_t$, estimate the error of h_t. 4. $\epsilon_t = \sum_{i=1}^{m} \mathcal{D}_t(i)\left[y_i \neq h_t(x_i)\right]$; Calculate the error for the classification results ϵ_t. 5. if $\epsilon_t > 0.5$ then if $\epsilon_t > 0.5$ then stop the iteration calculation. 6. $T \leftarrow t - 1$ 7. continue 8. end if 9. $\beta_t = \dfrac{\epsilon_t}{1 - \epsilon_t}$ Set the parameter β_t. 10. for $i = 1, 2, \cdots, m$ 11. $\mathcal{D}_{t+1}(i) = \mathcal{D}_t(i)\beta_t^{1 - \left[y_i \neq h_t(x_i)\right]}$ Update the weight of the sample. 12. end for i 13. end for T
Output: Final classification results: $H(x) = \arg\max_{y \in Y}\left(\sum_{t=1}^{T} \ln\left(\dfrac{1}{\beta_t}\right)\left[y_p \neq h_t(s_p)\right]\right)$

into multi-classifier using one vs one (OVO) or one vs all (OVA) strategies (Fernández et al., 2013). This procedure increases the computing complexity. However, considering that the sample space of the data set is not large, the computing time is completely acceptable; therefore, this study chooses SVM as the base classifier for the integrated classification algorithm.

2.3. Evaluation index for the classification results

In ensemble learning, a certain classification evaluation standard must be used to control the calculation process. However, for imbalanced data sets, common evaluation indexes often cannot reflect the classifier performance of risk evaluation data sets in which low risk and general risk samples account for the majority. If all feature combinations are judged as low or general risk by the classifier, even if the classification accuracy exceeds more than 90%, then no actual application value exists for the classifier.

The receiver operating characteristic (ROC) curve is a comprehensive classifier evaluation index (S. Zhang & H. X. Zhang, 2011). The ROC curve is based on the basic evaluation index for the machine learning–obfuscation matrix. The ROC curve takes FPR $\left(FP/\left(TN+FP\right)\right)$ as the transverse axis and TPR $\left(TP/\left(TP+FN\right)\right)$ as the vertical axis. By adjusting the classifier threshold, a set of (FPR, TPR) in the coordinate system can be derived, and these points can be connected to determine the ROC curve. The classifier threshold is the probability output of the classifier; a sample of the test set belongs to the probability of a certain category. In general, the area under the ROC curve (AUC) is used as the criterion to evaluate the classifier performance. Notably, a great AUC value indicates improved performance. However, basic ROC curves can only evaluate binary classification problems.

For a multi-classification problem, all data set categories can be combined into two groups (C_c^2 groups) to obtain C_c^2 ROC curves and the same number of AUC values; this way draws all AUC values into a polar coordinate system to evaluate the classifier performance from the graphic area of each AUC point (Everson & Fieldsend, 2006).

For example, if a data set has three categories with the category labels 1, 2, and 3, then any combination of two categories can generate a ROC curve and derive an AUC value. Therefore, from the three categories, three values are derived, namely, "$AUC_{1,2}$", "$AUC_{1,3}$", and "$AUC_{2,3}$".

After the AUC value for any two categories is determined, a traditional practice is to find the average value for evaluating the performance of the classifier. However, if the two AUC values have an opposite difference and the amplitude is roughly equal, then the mean value is stable. This condition is not conducive to comparing the advantages and disadvantages of multiple classifiers. In other words, even if the two classifiers have quite different specific classification performance, the average index is the same. Therefore, all AUC values are plotted in this study into polar coordinates. Thereafter, each AUC point

is connected to form a closed area. Then, the classifier performance is evaluated in accordance with the size of the area to avoid any shortcomings in the traditional methods. The area is named AUC_{area}. Figure 3 shows the location of the AUC points obtained for the polar coordinates from a data set that has three categories. The angle of each point is evenly allocated in accordance with the number of AUC points across 360°.

Equation (1) is used to calculate the area of each point. For convenience, the original AUC_{area} value is treated as normalized to restrict the value of the evaluation index to a [0,1] interval. If all r_i in Eqn (1) are 1, then the $AUC_{areamax}$ is obtained. Accordingly, the normalized AUC_{area} value is obtained from $AUC_{area}/AUC_{areamax}$ (Eqn (2)). The normalized AUC_{area} values are used later in this article.

$$AUC_{area}=\frac{1}{2}\sin\left(\frac{2\pi}{q}\right)\left(\left(\sum_{i=1}^{q-1}r_i\times r_{i+1}\right)+\left(r_q\times r_1\right)\right),\quad(1)$$

where $q=C_c^2$ is the number of AUC points and C is the number of categories.

$$AUC_{area}=\frac{\frac{1}{2}\sin\left(\frac{2\pi}{q}\right)\left(\left(\sum_{i=1}^{q-1}r_i\times r_{i+1}\right)+\left(r_q\times r_1\right)\right)}{\frac{1}{2}\sin\left(\frac{2\pi}{q}\right)\times q}=$$

$$\left(\left(\sum_{i=1}^{q-1}r_i\times r_{i+1}\right)+\left(r_q\times r_1\right)\right)/q.\quad(2)$$

AdaBoost and SVM are used as classifiers. The corresponding probability output matrix from the AdaBoost and the SVM is set to ensure that the probability of the test sample belongs to each category for obtaining the AUC_{area} value for the evaluation classification. As mentioned earlier, given the basic SVM can only handle a two-classification problem, an OVO policy is used to enable the SVM to handle the multiple classification problem.
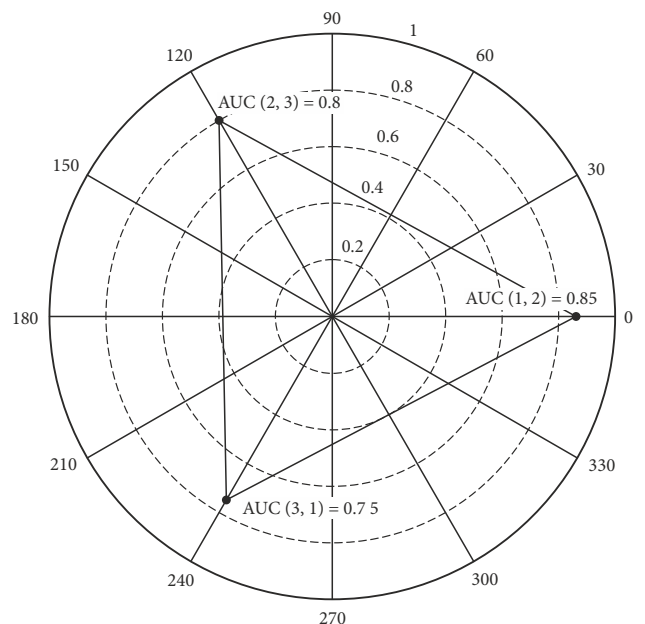

Figure 3. AUC polar diagram for three categories

We suppose that a data set has C classes. Using the OVO strategy, C_c^2 two-classification results. For a sample S_i and a category C_j, sample S_i belongs to class C_j with a probability of $\left[C_j / C_c^2 \right]$. Therefore, the SVM classification calculation can generate a probabilistic output matrix $svm_{score(i,j)}$ every time. Therefore, the probability of element matrix sample S_i belonging to the C_j class is $svm\_score(i,j) = \left[C_{ij} / C_c^2 \right]$. If the number of AdaBoost.M1 iterations is T, then ϵ_t is the classification SVM error rate when each iteration is calculated. $Ada\_score(i,j)$ corresponding to $svm_{score(i,j)}$ is preferred to the weighted $svm\_score(i,j)$ obtained T times. $Ada_{score}(i,j) = \sum_{t=1}^{T} \left(\left(1-\epsilon_t\right) / \sum_{t=1}^{T} \left(1-\epsilon_t\right) \right) \times svm\_score(i,j)$. $\left(1-\epsilon_t\right) / \sum_{t=1}^{T} \left(1-\epsilon_t\right)$ is the weight of the

SVM classifier for the t iteration. After the output probability of all the samples in each category is obtained, the ROC curve and the AUC value for any two categories are obtained. After all the C_c^2 AUC values are obtained, the standard AUC_{area} value can be determined using Eqn (2).

2.4. Improved ensemble classification algorithm for the BPSO+AdaBoost+SVM

The BPSO+AdaBoost+SVM ensemble learning algorithm uses the BPSO to extract the features and selects a subset of samples for the AdaBoost+SVM, after which they are classified and calculated. After the classification results are obtained, the AUC_{area} values are calculated as the fitness. The specific process for the entire algorithm is shown in Table 8.

Table 8. BPSO+AdaBoost+SVM algorithmic process

Integrated classification algorithm BPSO+AdaBoost+SVM
Input: Training set $D = \left\{ (x_1, y_1), (x_2, y_1) \cdots (x_m, y_m) \right\}$, $y_1 \in Y$ as the classification label;   Base learning algorithm $\mathcal{L}$;   Number of iterations T.
Process: 1. Randomly generate the initialized particles   2. while (Failure to reach the maximum number of iterations or failure to meet the stop criteria)   3. for $i = 1$ to N, N: Number of population   4. Extract sample subset *data* based on the feature subset extracted by particle i.   5. for $t_1 = 1$ to T_1, T_1: number of iterations for the particle swarm   6. $\mathcal{D}_1(i) = 1/m$ // start AdaBoost.M1 Initialize the weight of the sample   7. for $t = 1$ to T 8. $h_t : data_i \rightarrow \{C_1, C_2, \cdots, C_k\}$, Probability output matrix svm_score_i requires the SVM to obtain the classifier generated by the iteration calculation and the corresponding probability output matrix.   9. $\epsilon_t = \sum_{i=1}^{m} \mathcal{D}_t(i)\left[y_i \neq h_t(x_i) \right]$; Calculate the error for the classification results ϵ_t.   10. $\beta_t = \epsilon_t / \left(1-\epsilon_t\right)$ Set the parameter β_t.   11. $\mathcal{D}_{t+1}(i) = \mathcal{D}_t(i)\beta_t^{1-\left[y_i \neq h_t(x_i) \right]}$ Update the weight of the sample   12. end for t 13. $H(x) = \arg\max_{y \in Y} \left(\sum_{t=1}^{T} \ln\left(1/\beta_t\right)\left[y_p \neq h_t(s_p) \right] \right)$ Calculate the sequence of classifiers   14. $Ada_{score} = \sum_{t=1}^{T} \left(\left(1-\epsilon_t\right) / \left(\sum_{t=1}^{T} \left(1-\epsilon_t\right) \right) \right) \times svm\_score$ Calculate the probability output matrix for AdaBoost   15. Calculate all the AUC values, draw the polar diagram, and calculate the AUC_{area}, $fitness(x_i) = AUC_{area}$ 16. if $fitness(x_i) > fitness(pbest_i)$ // update the individual optimal solutions in the particle swarm   17. then $pbest_i = x_i$ 18. if $fitness(x_i) > fitness(gbest)$ 19. then $gbest = x_i$ 20. Update the speed and displacement of particle i 21. end for t_1 22. end for i 23. end while
Output: the optimal attribute subset S_{best} and classification results, the corresponding optimal polar diagrams, Each AUC value and the optimal AUC_{area}.

3. Empirical analyses

In this section, the safety risk "water and sand gushing in the foundation pit bottom" is taken as an example, and the proposed integrated classification algorithm is used to classify and calculate the knowledge information system risk.

3.1. Basic data

The knowledge information system for the risk evaluation of water and sand gushing in foundation pit bottom is taken from previous deep foundation construction pro-

ject management data, as shown in Table 9. Table 9 has 432 records (only part is listed). The attribute values in Table 9 are preconditioned to ensure that the analysis and calculation are accurate. The comparisons between the preconditioned attribute values and the original attribute values are shown in Table 10. In Table 10, all insecure attribute values are large and all secure attribute values are small. The attribute range is controlled within [0,1], and the equivalent value for the ratio of the penetration depth to the excavation depth for the supporting structure of a sample in an information system is K_i, with the original value being K_i'. K_{min}' and with K_{max}' are the minimum and

Table 9. Knowledge Information system for the risk evaluation of water and sand gushing in the foundation pit bottom

Number	Soil texture at the bottom of the foundation ditch	Groundwater	Ratio of penetration depth to excavation depth	Water-head	Foundation reinforcement	Entering-rock	Risk occurrence probability level
1	1	1	0.85	0.50	1	1	1
2	1	1	0.98	0.85	1	0	0.75
3	1	1	0.73	0.72	0	1	0.75
4	1	1	0.98	0.58	0	0	0.5
5	1	1	0.86	0.43	1	1	1
…	…	…	…	…	…	…	…
107	1	0	0.28	0.22	0	1	0
108	1	0	0.09	0.19	0	0	0
109	0.66	1	0.76	0.94	1	1	1
110	0.66	1	0.79	0.86	1	0	0.75
111	0.66	1	0.86	0.92	0	1	0.75
…	…	…	…	…	…	…	…
324	0.33	0	0.05	0.05	0	0	0
325	0	1	0.91	0.50	1	1	0.75
326	0	1	0.99	0.57	1	0	0.5
327	0	1	0.84	1.00	0	1	0.5
328	0	1	0.89	0.52	0	0	0.5
…	…	…	…	…	…	…	…

Table 10. Attribute value preprocessing control table

Attribute	Domain value				
Soil texture at the bottom of foundation ditch	Silt soil	Arenosol	Cohesive soil	Stratum	
	1	0.66	0.33	0	
Groundwater	Confined water	Unconfined water	Without water		
	1	0.5	0		
Ratio of penetration depth to excavation depth	$K_i = 1 - (K_i' - K_{min}')/K_{max}'$				
Water head inside and outside the foundation pit	$W_i = (W_i' - W_{min}')/W_{max}'$				
Foundation pit bottom reinforcement	nil	exist			
	1	0			
Rock entry condition of the enclosure structure	not entering-rock	entering-rock			
	1	0			
Risk occurrence probability level	Frequent	Possible	Occasionally	Rare	Impossible
	1	0.75	0.5	0.25	0

maximum values for the penetration depth to excavation depth ratio of the supporting structure in this information system. The definitions for the relative parameters for the water head inside and outside the foundation pit are similar.

The statistical information for the samples shown in Table 9 is shown in Table 11. Therefore, the knowledge information system for the risk evaluation of "water and sand gushing in the foundation pit bottom" is a multi-class (5 classification) imbalanced data set (the imbalance ratio is 36.8).

3.2. Analysis of the experimental results

The data set shown in Table 9 has 432 samples. The training set and test set composition is shown in Table 12. Four experiments are performed. The first experiment used the algorithm proposed in this study, and the others respectively used the AdaBoost+SVM algorithm, the KNN algorithm, and the AdaBoost+KNN algorithm. Furthermore, the relevant parameters in BPSO are the same as the corresponding parameters set in Kennedy and Eberhart

(1997), where inertia weight is set to 0.729 and learning factors c_1 and c_2 are both set to 1.49445. We set k of KNN equals to 5 for those data sets that the rarest class contains at least 5 samples; otherwise, K is set to be equal with the number of samples of the rarest class. The number of samples of the rarest class is 5 shown in Table 11. Therefore, the K of KNN equals to 5. In addition, MATLAB libsvm is used to realize SVM as base classifier. Other unexplained Numbers are default values.

In this study, four evaluation indexes are used to calculate the effectiveness and superiority of the integrated classification algorithm proposed in Section 2. The accuracy rates for every classification are given, and the results of the classification calculations are shown in Table 13.

- Experiment one used the integrated classification algorithm proposed in this study. The overall accuracy was found to be very high; in particular, the correct classification rate for the minority class (1 grade category) reached 100%.
- Considering that the characteristic data set dimensions shown in Table 9 were low, experiment two

Table 11. Sample statistics

Risk probability level	Sample number	Maximum imbalance ratio
Impossible (0)	148	
Rare (0.25)	184	Maximum imbalance ratio = (Maximum sample number) / (Minimum sample number) = 184/5=36.8
Occasionally (0.5)	65	
Possible (0.75)	30	
Frequent (1)	5	

Table 12. Training set and test set sample distribution

Risk level	Training set	Test set	Total
0	111	37	148
0.25	142	42	184
0.5	49	16	65
0.75	21	9	30
1	3	2	5
	326	106	432

Table 13. Analysis of the results for the classification calculation

	Experiment 1 BPSO+AdaBoost+SVM	Experiment 2 AdaBoost +SVM	Experiment 3 KNN	Experiment 4 AdaBoost +KNN
A_{cc}	0.8962	0.9680	0.81132	0.9057
AUC_{area}	0.7971	0.8079	0.3085	0.7616
GM	0.9052	0.7891	0.4945	0.8221
FV	0.8892	0.7881	0.3269	0.7888
Classification accuracy for the various categories				
Overall classification accuracy	96/106	99/106	86/106	86/106
Accuracy rate for the 0 grade classification	36/37	37/37	36/37	37/37
Accuracy rate for the 0.25 grade classification	39/42	41/42	42/42	42/42
Accuracy rate for the 0.5 grade classification	11/16	12/16	7/16	13/16
Accuracy rate for the 0.75 grade classification	8/9	8/9	1/9	3/9
Accuracy rate for the 1 grade classification	2/2	1/2	0/2	1/2

Note: $A_{CC} = (TP+TN)/(TN+FP+FN+TP)$, $GM = \sqrt{sensitivity \cdot specificity}$, $FM = (2 \cdot sensitivity \cdot precision)/(sensitivity+precision)$, $sensitivity = TP/(TP+FN)$, $specificity = TN/(FP+TN)$, $precision = TP/(FP+TP)$, FN = False Negative, FP = False Positive, TN = True Negative, TP = True Positive.

attempted to remove the feature selection process and adopted the AdaBoost+SVM algorithm directly. The overall accuracy of this method was higher than that of the experiment one; however, the accuracy of the most critical minority classification (1 grade category) was lower. A comparison of experiment one with experiment two showed that the learning of the minority samples was improved given that experiment one used the feature selection method to process the data set in Table 9. However, as some valuable information was lost for most other samples, the classification accuracy of the majority was lower in experiment one than in experiment two, which did not use feature selection. From an application perspective, the loss of classification errors in the low risk categories was considerably smaller than in the high risk categories. Therefore, the algorithm used in experiment one had more practical value than the algorithm used in experiment two.

- In experiment three, a traditional KNN single classifier was used. The calculation results showed that overall the classification accuracy was lower; however, the classification accuracy for the 0.25 and 0 grades, which had the largest number of samples, was the highest. As the number of samples decreased, the classification accuracy in the other categories fell sharply, which was in line with the traditional bias toward larger samples associated with the KNN classifier. This result proved that the KNN classifier was unsuitable for imbalanced data sets.

- In experiment four, the AdaBoost+KNN algorithm was used. As the performance of the KNN classifier was enhanced with AdaBoost ensemble algorithm, the classification accuracy of KNN classifier was significantly improved. Therefore, the AdaBoost ensemble algorithm was very powerful in improving weak classifiers. However, the classification effect on the minority groups was still unable to meet the application requirements due to the weaknesses in the KNN classifier.

This study conducted an integrated algorithmic experiment that used SVM as the base classifier, AdaBoost as the integrated framework, and BPSO for feature selection to ensure that the machine learning effect of the imbalanced data set was suitable for practical applications.

Conclusions

This study analyzed the practical difficulties of safety risk evaluations of deep foundation construction schemes to improve the safety of deep foundation construction and proposed that machine learning could be effective for manual analyses of safety risk evaluations of comprehensive construction schemes. Compared with previous studies that ignored historical data, the safety risk evaluation knowledge information system was developed from historical data, and a machine learning method was applied to extract the relevant knowledge.

Furthermore, accurate minority class classifications were needed for practical applications because the safety risk evaluation knowledge information system was a typical imbalanced data set. As such, traditional classifiers could not be used. Focusing on the characteristics of the deep foundation safety risk construction knowledge information system, an integrated classification algorithm, the BPSO+AdaBoost+SVM, was proposed to improve the minority classification accuracy. Compared with other related algorithms, the developed algorithm is suitable for the machine learning of the developed information system because it can effectively classify the minority samples.

The ensemble classification algorithm proposed in this study can assist in promoting the use of intelligent evaluations of safety risk evaluations of deep foundation construction schemes. The developed algorithm does not involve subjective factors. However, it emphasizes the full utilization and management of historical data. It also is completely dependent on previous safety management knowledge and experience. Therefore, the quality of safety risk evaluations is high when the historical data are complete. On this basis, the collection and consolidation of relevant daily safety management data is very important. This research improves the quality and efficiency of safety risk evaluations of deep foundation construction schemes. This enhancement in turn improves safety risk management of deep foundation construction.

Further research can be focused on the collection and sorting process for deep foundation construction safety risk management information. A high quality database can also be established for knowledge on safety risk evaluations of deep foundation construction schemes. Furthermore, with the developed algorithm at the core, a computer program can be developed to automate the safety risk evaluations of deep foundation construction schemes. This way can improve accuracy and save time and money associated with manual evaluations.

Author contributions

Haixiang Guo conceived the study and were responsible for the design and development of the data analysis. Peisong Gong, Haixiang Guo and Yuanyue Huang were re-

sponsible for data collection and model construction. Yuanyue Huang and Shengyu Guo were responsible for data interpretation. Peisong Gong wrote the draft of the article.

Disclosure statement

Authors confirm that all of the content, figures (charts, photographs, etc.), and tables in the submitted manuscript work are original work created by the authors and no any competing financial, professional, or personal interests from other parties.

References

Al-Anbari, S., Khalina, A., Alnuaimi, A., Normariah, A., & Yahya, A. (2015). Risk assessment of safety and health (RASH) for building construction. *Process Safety and Environmental Protection*, *94*, 149–158.
https://doi.org/10.1016/j.psep.2015.01.009

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140. https://doi.org/10.1007/BF00058655

Cao, H. M. (2014). Research on the risk assessment for the construction safety in the planning and design stages of bridge engineering. *Advanced Materials Research*, *998–999*, 1678–1681.
https://doi.org/10.4028/www.scientific.net/AMR.998-999.1678

Cao, Y., Miao, Q.-G., Liu, J.-C., & Gao, L. (2013). Advance and prospects of AdaBoost algorithm. *Acta Automatica Sinica*, *39*(6), 745–758.
https://doi.org/10.1016/S1874-1029(13)60052-X

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*(3), 321–357.
https://doi.org/10.1613/jair.953

Chawla, N. V., Lazarevic, A., Hall, L. O., & Bowyer, K. W. (2003). SMOTEBoost: Improving prediction of the minority class in boosting. Paper presented at the *Knowledge Discovery in Databases: PKDD 2003* (pp. 107–119). Springer.
https://doi.org/10.1007/978-3-540-39804-2_12

Chuang, L. Y., Chang, H. W., Tu, C. J., & Yang, C. H. (2008). Improved binary PSO for feature selection using gene expression data. *Computational Biology and Chemistry*, *32*(1), 29–38. https://doi.org/10.1016/j.compbiolchem.2007.09.005

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.
https://doi.org/10.1023/A:1022627411411

Eskesen, S. D., Tengborg, P., Kampmann, J., & Holst Veicherts, T. (2004). Guidelines for tunnelling risk management: International Tunnelling Association, Working Group No. 2. *Tunnelling and Underground Space Technology*, *19*(3), 217–237.
https://doi.org/10.1016/j.tust.2004.01.001

Ding, L. Y., Yu, H. L., Li, H., Zhou, C., Wu, X. G., & Yu, M. H. (2012). Safety risk identification system for metro construction on the basis of construction drawings. *Automation in Construction*, *27*, 120–137.
https://doi.org/10.1016/j.autcon.2012.05.010

Everson, R. M., & Fieldsend, J. E. (2006). Multi-class ROC analysis from a multi-objective optimisation perspective. *Pattern Recognition Letters*, *27*(8), 918–927.
https://doi.org/10.1016/j.patrec.2005.10.016

Fernández, A., López, V., Galar, M., del Jesus, M. J., & Herrera, F. (2013). Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches. *Knowledge-Based Systems*, *42*, 97–110.
https://doi.org/10.1016/j.knosys.2013.01.018

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., & Herrera, F. (2012). A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *42*(4), 463–484.
https://doi.org/10.1109/TSMCC.2011.2161285

Gao, P., Liu, X., & Tong, R. P. (2013). Risk assessment and preliminary study of safety management system on construction works. *Applied Mechanics and Materials*, *368–370*, 1917–1921.
https://doi.org/10.4028/www.scientific.net/AMM.368-370.1917

GB50652-2011. *Code for risk management of underground works in urban rail transit.*

GB50715-2011. *Standard for construction safety evaluation of metro engineering.*

Hassan, M. R., Ramamohanarao, K., Karmakar, C., Hossain, M. M., & Bailey, J. (2010). A novel scalable multi-class ROC for effective visualization and computation. In *Advances in Knowledge Discovery and Data Mining* (pp. 107–120). Springer.
https://doi.org/10.1007/978-3-642-13657-3_14

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *Machine Learning: ECML-98* (pp.137–142). Springer.
https://doi.org/10.1007/BFb0026683

Kennedy, J., & Eberhart, R. C. (1995). Particle swarm optimization. In *Proceedings of ICNN'95 – International Conference on Neural Networks*. Perth, WA, Australia. IEEE.
https://doi.org/10.1109/ICNN.1995.488968

Kennedy, J., & Eberhart, R. C. (1997). A discrete binary version of the particle swarm algorithm. In *1997 IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation*. Orlando, FL, USA. IEEE.
https://doi.org/10.1109/ICSMC.1997.637339

Krawczyk, B., & Schaefer, G., (2013). An improved ensemble approach for imbalanced classification problems. In *2013 IEEE 8th International Symposium on Applied Computational Intelligence and Informatics* (*SACI*). Timisoara, Romania. IEEE.
https://doi.org/10.1109/SACI.2013.6609011

Lee, C. Y., & Lee, Z. J. (2012). A novel algorithm applied to classify unbalanced data. *Applied Soft Computing*, *12*(8), 2481–2485. https://doi.org/10.1016/j.asoc.2012.03.051

Lesser, V., Durfee, E., & Corkill, D. (2006). Trends in cooperative distributed problem solving. *IEEE Transactions on Knowledge & Data Engineering*, *18*, 63–77.

Liu, W., Zhao, T., Zhou, W., & Tang, J. (2018). Safety risk factors of metro tunnel construction in China: an integrated study with EFA and SEM. *Safety Science*, *105*, 98–113.
https://doi.org/10.1016/j.ssci.2018.01.009

Li, Y. J., Guo, H. X., Liu, X., Li, Y. N., & Li, J. L. (2016a). Adapted ensemble classification algorithm based on multiple classifier system and feature selection for classifying multi-class imbalanced data. *Knowledge-Based Systems*, *94*, 88–104.
https://doi.org/10.1016/j.knosys.2015.11.013

Li, Y. J., Guo, H. X., Li, Y. N., & Liu, X. (2016b). A boosting based ensemble learning algorithm in imbalanced data classification. *System Engineering – Theory & Practice*, *36*(1), 189–199.

López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics. *Information Sciences*, *250*(11), 113–141.
https://doi.org/10.1016/j.ins.2013.07.007

Luo, H., & Gong, P. (2015). A BIM-based code compliance checking process of deep foundation construction plans. *Journal of Intelligent & Robotic Systems*, *79*(3–4), 549–576. https://doi.org/10.1007/s10846-014-0120-z

Khan, M. N., Ksantini, R., Ahmad, S. I., & Guan, L. (2014). Covariance-guided one-class support vector machine. *Pattern Recognition*, *47*(6), 2165–2177. https://doi.org/10.1016/j.patcog.2014.01.004

Park, J., Park, S., & Oh, T., (2015). The development of a web-based construction safety management information system to improve risk assessment. *KSCE Journal of Civil Engineering*, *19*(3), 528–537. https://doi.org/10.1007/s12205-014-0664-2

Patel, D. A., & Jha, K. N. (2017). Developing a process to evaluate construction project safety hazard index using the possibility approach in India. *Journal of Construction Engineering and Management*, *143*(1), 04016081. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001205

Pinto, A. (2014). QRAM a qualitative occupational safety risk assessment model for the construction industry that incorporate uncertainties by the use of fuzzy sets. *Safety Science*, *63*, 57–76. https://doi.org/10.1016/j.ssci.2013.10.019

Preidel, C., & Borrmann, A. (2015). Automated code compliance checking based on a visual language and building information modeling. In *Proceedings of the International Symposium on Automation and Robotics in Construction (ISARC)* (Vol. 32). IAARC Publications. https://doi.org/10.22260/ISARC2015/0033

Sansakorn, P., & An, M. (2015). Development of risk assessment and occupational safety management model for building construction projects. *International Journal of Civil and Environmental Engineering*, *9*(9), 1248–1255.

Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning*, *5*(2), 197–227. https://doi.org/10.1007/BF00116037

Seo, J. W., & Choi, H. (2008). Risk-based safety impact assessment methodology for underground construction projects in Korea. *Journal of Construction Engineering and Management*, *134*(1), 72–81. https://doi.org/10.1061/(ASCE)0733-9364(2008)134:1(72)

Sun, Y., Fang, D., Wang, S., Dai, M., & Lv, X., (2008). Safety risk identification and assessment for Beijing Olympic Venues construction. *Journal of Management in Engineering*, *24*(1), 40–47. https://doi.org/10.1061/(ASCE)0742-597X(2008)24:1(40)

Sun, Y., Kamel, M. S., Wong, A. K. C., & Wang, Y. (2007). Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition*, *40*(12), 3358–3378. https://doi.org/10.1016/j.patcog.2007.04.009

Tan, X., Hammad, A., & Fazio, P. (2010). Automated code compliance checking for building envelope design. *Journal of Computing in Civil Engineering*, *24*(2), 203–211. https://doi.org/10.1061/(ASCE)0887-3801(2010)24:2(203)

Tao, X., Li, Q., Guo, W., Ren, C., Li, C., Liu, R., & Zou, J. (2019a). Self-adaptive cost weights-based support vector machine cost-sensitive ensemble for imbalanced data classification. *Information Sciences*, *487*, 31–56. https://doi.org/10.1016/j.ins.2019.02.062

Tao, X., Li, Q., Ren, C., Guo, W., Li, C., He, Q., & Zou, J. (2019b). Real-value negative selection over-sampling for imbalanced data set learning. *Expert Systems with Applications*, *129*, 118–134. https://doi.org/10.1016/j.eswa.2019.04.011

Wang, F., Ding, L., Love, P. E. D., & Edwards, D. J. (2016). Modeling tunnel construction risk dynamics: Addressing the production versus protection problem. *Safety Science*, *87*, 101–115. https://doi.org/10.1016/j.ssci.2016.01.014

Wang, Z. Z., & Chen, C. (2017). Fuzzy comprehensive Bayesian network-based safety risk assessment for metro construction projects. *Tunnelling and Underground Space Technology*, *70*, 330–342. https://doi.org/10.1016/j.tust.2017.09.012

Weiss, G. M., & Provost, F. (2003). Learning when training data are costly: The effect of class distribution on tree induction. *Journal of Artificial Intelligence Research*, *19*, 315–354. https://doi.org/10.1613/jair.1199

Yang, M., Yin, J. M., & Ji, G. L., (2008). Classification methods on imbalance data: A survey. *Journal of Nanjing Normal University* (*Engineering and Technology Edition*), *8*(4), 7–12.

Yang, Q. Z., & Xu, X. (2004). Design knowledge modeling and software implementation for building code compliance checking. *Building and Environment*, *39*(6), 689–698. https://doi.org/10.1016/j.buildenv.2003.12.004

Zhang, B. X., & Ma, F. H. (2014). Metro construction safety risk assessment of Xi'an based on CIM model. *Applied Mechanics and Materials*, *638–640*, 804–808. https://doi.org/10.4028/www.scientific.net/AMM.638-640.804

Zhang, S., Shang, C., Wang, C., Song, R., & Wang, X. (2019). Real-time safety risk identification model during metro construction adjacent to buildings. *Journal of Construction Engineering and Management*, *145*(6), 04019034. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001657

Zhang, S., & Zhang, H. X. (2011). Modified KNN algorithm for multi-label learning. *Application Research of Computers*, *28*(12), 4445–4450.

Zhang, Y., Ding, L., & Love, P. E. D. (2017). Planning of deep foundation construction technical specifications using improved case-based reasoning with weighted k-nearest neighbors. *Journal of Computing in Civil Engineering*, *31*(5), 04017029. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000682

Zheng, E. H., Li, P., & Song, Z. H. (2006). Cost sensitive support vector machines. *Control and Decision*, *21*(4), 473–476.

Zheng, X., Ma, F. H. (2014). Metro construction safety risk assessment based on the fuzzy AHP and the comprehensive evaluation method. *Applied Mechanics and Materials*, *580–583*, 1243–1248. https://doi.org/10.4028/www.scientific.net/AMM.580-583.1243

Zhong, B. T., Ding, L. Y., Luo, H. B., Zhou, Y., Hu, Y. Z., & Hu, H. M. (2012). Ontology-based semantic modeling of regulation constraint for automated construction quality compliance checking. *Automation in Construction*, *28*, 58–70. https://doi.org/10.1016/j.autcon.2012.06.006

Zhong, B., Gan, C., Luo, H., & Xing, X. (2018). Ontology-based framework for building environmental monitoring and compliance checking under BIM environment. *Building and Environment*, *141*, 127–142. https://doi.org/10.1016/j.buildenv.2018.05.046

Zhong, B., & Li, Y. (2015). An ontological and semantic approach for the construction risk inferring and application. *Journal of Intelligent & Robotic Systems*, *79*(3), 449–463. https://doi.org/10.1007/s10846-014-0107-9

Zhou, H.-b., & Zhang, H. (2011). Risk assessment methodology for a deep foundation pit construction project in Shanghai, China. *Journal of Construction Engineering and Management*, *137*(12), 1185–1194. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000391

Zhou, Y., Ding, L. Y., & Chen, L. J. (2013). Application of 4D visualization technology for safety management in metro construction. *Automation in Construction*, *34*, 25–36. https://doi.org/10.1016/j.autcon.2012.10.011