Journal of Geochemical Exploration 165 (2016) 35–48

Contents lists available at ScienceDirect

Journal of Geochemical Exploration

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

A geochemical modeling to predict the different concentrations of REE and their hidden patterns using several supervised learning methods: Choghart iron deposit, bafq, Iran

S. Zaremotlagh, A. Hezarkhani *

Department of Mining & Metallurgical Engineering, Amirkabir University of Technology (Tehran Polytechnic), Hafez Ave. No. 424, Tehran, Iran

ARTICLE INFO

Article history:
Received 25 August 2015
Revised 3 February 2016
Accepted 7 February 2016
Available online 11 February 2016

Keywords:
Bafq mining district
Supervised learning
Geochemical rules
Rocks compositions
Decision tree

ABSTRACT

The Bafq mining district hosts some Kiruna-type iron oxide-apatite (IOA) deposits which are commonly formed as a result of the multistage interaction of hydrothermal-magmatic processes within the Early Cambrian volcano-sedimentary sequence. Rare earth elements (REEs) are potentially concentrated under different physicochemical conditions in IOA deposits. Choghart orebody is one of the main magnetite-apatite deposits in the region. A wide range of primary and secondary geological events have affected the Choghart deposit, causing the behavior of REEs to vary in different zones. This study proposes a suitable exploration technique using various classification methods to identify the different concentrations of REE and their hidden patterns. To provide the required data, a systematic lithogeochemical sampling was performed in the north and NE of the Choghart orebody. The REE contents of samples were transformed into discrete values as distinct classes based on the results of clustering analysis. All possible combinations of features, being the geographical location and the major oxide composition of samples, were selected as subsets of predictors in every classification method. For each REE, 455 prediction models were constructed using these predictors. The performances of the classification methods were evaluated by error criteria with regard to all cases. Having the least amount of errors, the decision tree method was selected as the most suitable method. Based on decision tree results, the best subsets of predictors were chosen for each element. The existence of a significant relationship between the distribution patterns of each REE and the related predictors was assessed by its prediction errors. The assessment illustrated that some REE are reasonably predictable, and others are too irregular to be modeled. The extracted classification rules describe the geochemical relationships among the most important factors influencing the different concentrations of REE in the Choghart orebody. These results can be extended to other similar deposits to predetermine some REE-enriched zones based on major elements analysis. Merely by employing such techniques in REE exploration projects, a great savings in time and cost will be affected.

1. Introduction

Rare earth elements (REEs) are a famous group of chemically similar elements commonly associated to each other in the Earth's crust. These elements have been usually split into the Light-Rare-Earth elements (LREE contain La, Ce, Pr, Nd, Pm, Sm, and Eu) and Heavy-Rare-Earth elements (HREE contain Gd, Tb, Dy, Ho, Er, Tm, Yb, Lu, Y and Sc) (Emsbo et al., 2015; Jha, 2014; Long et al., 2012; Simandl, 2014). The exceptional properties of REEs make them essential for hi-tech manufacturing in today's society (Massari and Ruberti, 2013; Stegen, 2015).

Although there are several valuable deposits in a few countries in the world, the major production of REE minerals, concentrates, and metals are monopolized by China (Jaireth et al., 2014; Jha, 2014).

The REE behavior is principally defined based on their mineral/melt partition coefficients, temperature, pressure, oxygen fugacity, ionic radius and charge of the element. Despite the similarity in behavior, the REE can be partially fractionated from each other by several petrological and mineralogical processes (Henderson, 1984, 1996; Jaireth et al., 2014; Rollinson, 1993). Mobilization and redistribution of REE may occur during the weathering and alteration processes and adjusts the REE distribution patterns in these systems. (Baioumy et al., 2014; Cole et al., 2014; Ehya, 2012; Foley and Ayuso, 2013; Küpeli, 2010; Schacht et al., 2010; Shikazono et al., 2008). So different geological processes and thermodynamic conditions specify the distribution of REEs in various environments, each with its unique pattern. Therefore, the REEs are known as important geochemical tracers for a wide range of geological processes and their abundances, ratios, isotope compositions, and normalized patterns are the important criteria for geochemical exploration studies (Berger et al., 2014; Cole et al., 2014; Tsay et al., 2014).

The REEs are mainly concentrated in specific types of rocks and deposits. In addition, they are potentially known as an important by-

* Corresponding author.
E-mail addresses: s_zarem@aut.ac.ir, zare_eng@yahoo.com (S. Zaremotlagh), Ardehez@aut.ac.ir (A. Hezarkhani).

http://dx.doi.org/10.1016/j.gexplo.2016.02.001
0375-6742/

product of Iron oxide-apatite (IOA) deposits (Simandl, 2014). The massive IOA deposits occur with various shapes, sizes, and mineral grades all over the world (e.g. Kiruna of Sweden and Bafq mining district of Iran). Although many researchers, e.g., (Borrok et al., 1998; Daliran, 1990, 2002; Daliran et al., 2010; Foerster and Jafarzadeh, 1994; Groves et al., 2010; Hildebrand, 1986; Nystroem and Henriquez, 1994; Rajabi et al., 2015; Stosch et al., 2011), have studied in these districts and proposed sundry descriptive or genetic models for them, there are many unsolved problems about the origin and metallogenesis of IOA deposits yet.

REE distribution patterns are commonly varied and complicated in the IOA deposits. Therefore, the main objective of this paper is to present an innovative technique for studying different concentrations of REE in the Choghart orebody, an important IOA deposit in Central Iran, in order to identify their patterns and separate zones. For this purpose, several parametric and nonparametric classification methods including discriminant analysis, Naïve Bayes, and decision tree were applied to reliably model the influence of various rock compositions on the spatial distribution patterns of REE. The performances of these methods were evaluated by illustrating the best prediction models. In addition, some classification rules are introduced as the geochemical factors effectively influencing different concentrations of REE. These rules can be used to map various zones. The results of this study can be extended to other similar deposits to predetermine enriched zones of some REEs which are critical for detailed geochemical exploration.

2. Bafq mining district

The The Bafq mining district, also known as the Zarigan-Chahmir basin, is located in Central Iranian microcontinent. This district hosts many mineral deposits, chiefly the Kiruna-type iron oxide-apatite (IOA) and thus is one of the main metallogenic provinces in Iran. The most important IOA deposits (e.g., Choghart, Chahgaz, Esfordi, Sechahun, Mishdovan and North Anumaly) are located in the western and NW portions of the Zarigan-Chahmir basin as shown in Fig. 1. (Bonyadi et al., 2011; Daliran et al., 2009; Ghanbari et al., 2014; Sabet-Mobarhan-Talab et al., 2015; Stosch et al., 2011).

2.1. Regional geological and tectonic setting

The intruding of granitic plutons into the Precambrian sequence and formation of felsic to intermediate volcanic and volcano-sedimentary rocks happened at the period of the Early Cambrian. This sequence is composed of an unmetamorphosed series which includes interlayered micro-conglomerates, sandstones, black siltstones and shales, dolomites and dolomitic limestones, mafic to felsic volcanic rocks, volcanoclastic beds and tuffaceous shales (Foerster and Jafarzadeh, 1994; Ramezani and Tucker, 2003; Samani, 1988).

Central Iranian microcontinent includes three main structural zones which are Lut, Tabas, and Yazd blocks as depicted in Fig. 1. The Bafq mining district locates in the central of Posht-e-Badam Block. This block separates the Tabas and Yazd blocks with regional-scale faults (Alavi, 1991). The subduction of Proto-Tethys oceanic crust under Central Iranian microcontinent and the consequences of continental arc and back-arc control the evolution of the Bafq mining district. Several tectonic models have been considered in relation to this region (Alavi, 1991; Bagheri and Stampfli, 2008; Rajabi et al., 2012, 2015; Ramezani and Tucker, 2003; Samani, 1988).

2.2. Mineralization, alteration, and geochemistry

It have been claimed that mineralization and felsic magmatism were simultaneous. The closely temporal and spatial relations among the iron oxide-apatite (IOA) deposits in addition to the Early Cambrian felsic volcanic rocks confirm these opinion (Daliran, 1990; Daliran et al., 2009, 2010). Some synsedimentary attributes also illustrate that the mineralization might be partially simultaneous with the sedimentation in the basin (Aftabi et al., 2009; Daliran, 2002; Daliran et al., 2009; Stosch et al., 2011).

According to some researchers (Jami et al., 2007; Moore and Modabberi, 2003), the most of the Iron oxide-apatite and apatite-rich deposits are epigenetic. Interaction of multistage hydrothermal-magmatic processes within the Early Cambrian volcano-sedimentary sequence caused various mineralization styles. So most likely, hydrothermal fluids are a significant factor in the evolution of some of these deposits (Daliran, 2002; Jami et al., 2007, 2009; Torab and Lehmann, 2007). These fluids may be connected to the arc calk-alkaline magmatism (Daliran et al., 2009; Jami et al., 2007; Stosch et al., 2011). These deposits are sometimes found within alkali alteration zones with secondary Na- and K-feldspars at the regional scale. It would point toward a genetic relation between mineralization processes and alkali metasomatism (Stosch et al., 2011).

The Choghart and some REE-bearing IOA deposits occur within unmetamorphosed welded rhyolitic to rhyodacitic tuffs and volcano-sedimentary rocks (Daliran, 2002; Stosch et al., 2011). The intrusive rocks being predominantly syenite and secondarily pyroxenite, gabbro and granite are enclosed by alkali rhyolites. The orebody and its country rocks are cut by several diabasic dikes. The entire complex is surrounded by the plain comprised of Quaternary formations and recent alluvium. The shape of main orebody at Choghart deposit is approximately vertical, discordant and pipe-shaped body plunging 73°NNW (Moore and Modabberi, 2003).

The main primary ore mineral located at the bottom of the Choghart orebody is massive magnetite. Hematite is the second abundant mineral mostly created from a secondary source. Hydrous iron oxide and goethite vanish with distance from surface quickly so depth of the oxidation zone is about 150 m. The minor minerals such as apatite, pyrite, tremolite, actinolite, calcite, talc, quartz, monazite, davidite and allanite have been identified throughout the orebody (Moore and Modabberi, 2003).

The magnetite-apatite is the most important ore type in this orebody (Foerster and Jafarzadeh, 1994). The primary magnetite crystals vary in size from fine-grained to coarse-grained which have occasionally thin exsolved ilmenite. Despite the crystals usually display particular inner intergrowths for crystallization of melt, some of which reveal indication of recrystallization. Apatite is the most plentiful gangue that occurs in the form of two distinct generations. The older of which is simultaneous with that the creation of iron oxide and shows euhedral crystals closely intergrown with magnetite whereas the newer ones appears as subhedral to anhedral crystals in lenses, dikes, and veinlets cutting the magnetite-apatite ore (Moore and Modabberi, 2003).

The Early Cambrian igneous rocks of the Bafq mining district have a bimodal nature. The chondrite-normalized REE patterns display significant variation from LREE to HREE with no considerable Eu anomalies for basaltic rocks and also show obvious enrichment in the LREE with important negative Eu anomalies for the rhyolitic domes (Rajabi et al., 2015).

The REE enrichment is intensely associated with the formation of phosphate minerals in many IOA deposits. However bastnaesite, parasite, and allanite are so significant in some locations (Oreskes and Einaudi, 1990). Edfelt (2007) explained there are few complications in the phosphate-REE relationship in some Kiruna district. Hence the relationship between REE and phosphate minerals in such deposits should be more understood. In these deposits, appetites characteristically comprise 2000–6000 ppm REE (Frietsch, 1982; Frietsch and Perdahl, 1995). Daliran (2002) claimed Bafq district apatites contain up to 1.75 wt.% REE.

Some researches present that post-depositional REE leaching could be happened in apatite in which the inclusions of monazite and xenotime may be seen (Bonyadi et al., 2011; Stosch et al., 2011; Torab and Lehmann, 2007). The U–Pb dating of monazite inclusions in apatite demonstrates that the REE redistribution in apatite might be happen frequently throughout hydrothermal process several million years after the formation of the IOA deposits (Stosch et al., 2011).

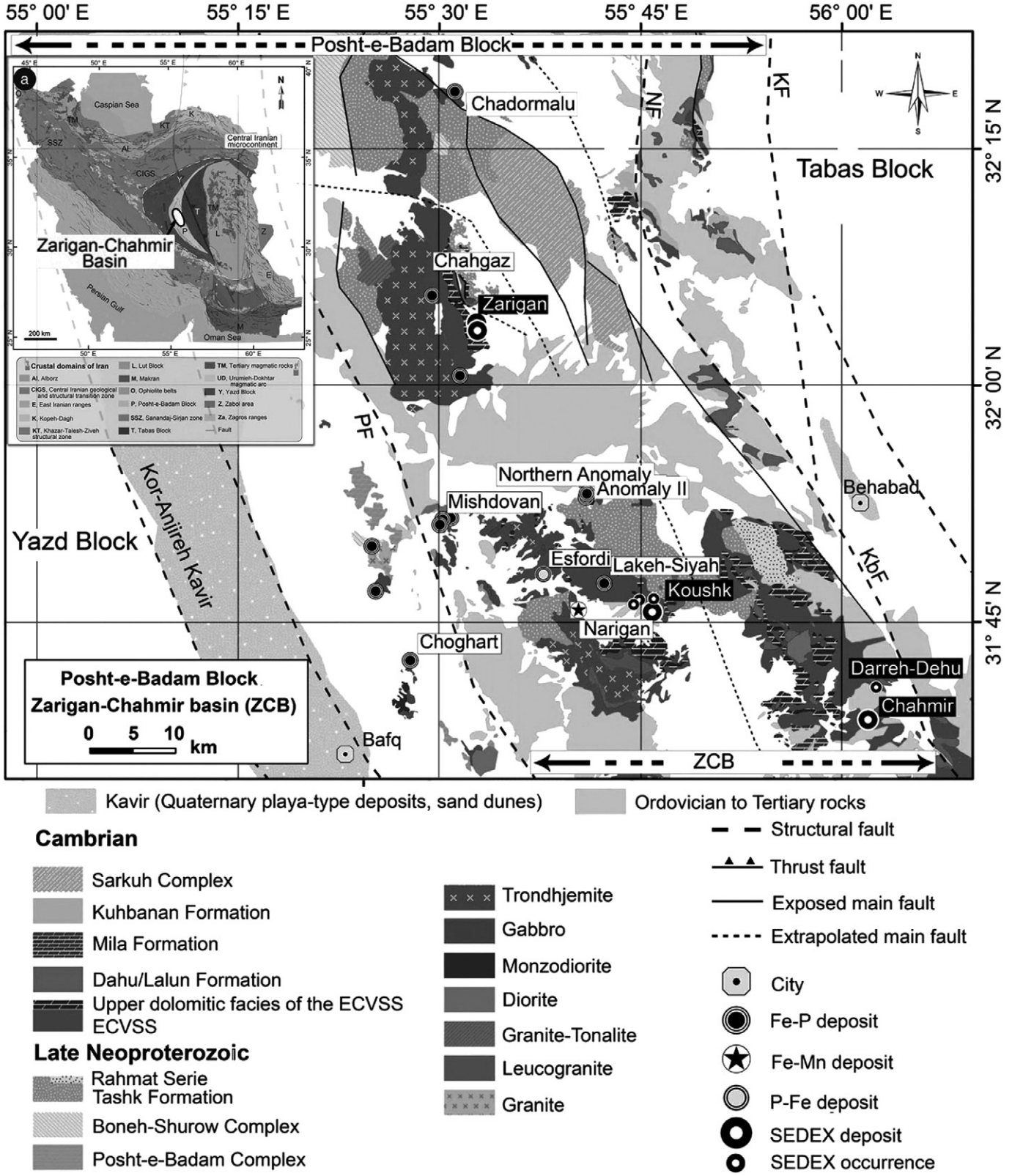

Fig. 1. Geological map of the Zarigan-Chahmir basin, showing the location of basin's deposits. CF: Chapedony Fault, KbF: Kuhbanan Fault, KF: Kalmard Fault, NF: Naeini Fault, PF: Posht-e-Badam Fault (Rajabi et al., 2015). a) Structural map of Iran (modified after Aghanabati, 1998) and location of the Zarigan-Chahmir basin in the Posht-e-BadamBlock (Rajabi et al., 2015).

3. Methodology and materials

Geoscientists are faced with a huge amount of data repeatedly collected in various projects. Examining such data is significant to discover its hidden knowledge. Data mining techniques such as classification and clustering are applied to obtain a clearer comprehension of the data and solve some geological problems. A large number of classification and clustering techniques have been proposed to solve several geological problems. Classification is the process of finding a model describing and distinguishing data classes. Clustering analyzes data objects

without using class labels and clusters them based on the principle of maximizing the intra-group similarity and minimizing the inter-group similarity. A large number of classification and clustering techniques have been proposed to solve several geological problems, e.g., (Abbaszadeh et al., 2013; Astel et al., 2014; Eggenkamp and Marques, 2013; Li et al., 2012; Li and Anderson-Sprecher, 2006; Lourenço et al., 2010; Nazarpour et al., 2015; Peh and Halamić, 2010; Sadeghi et al., 2013, 2015; Tahmasebi and Hezarkhani, 2012a, 2012b).

3.1. Clustering methods

Clustering related to unsupervised learning methods can discover previously unknown groups within the data. It groups a set of objects into various clusters so that the objects within a cluster have high similarity, but are very dissimilar to objects in other clusters. The dissimilarity measure is computed by the feature values describing the object. Several clustering methods can be used to understand data distribution or as a preprocessing step for other data mining algorithms. In this study, the k-means algorithm was employed to prepare the geochemical dataset for use in classification models.

The k-means algorithm known as one of the most common clustering methods operates in the following manner (Han, Kamber et al. 2012). The k-means is a centroid-based technique that distributes the n objects of a dataset D into k clusters, $C_1, \ldots, C_k$, that is, $C_i \subset D$ and $C_i \cap C_j = \varnothing$; for $1 \le i, j \le k$. It uses the centroid of a cluster, c_i, defined by the mean or medoid of the objects allocated to this cluster. The difference between an object $p \in C_i$ and c_i is measured by their Euclidean distance that is denoted by $dist(p, c_i)$. The quality of cluster C_i can be calculated by the sum of squared error between all objects in C_i and the centroid c_i as Eq. (1)

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} dist(p, c_i)^2 \tag{1}$$

3.2. Classification methods

The supervised learning method is commonly characterized by a two-step process. Firstly, a classification model is built based on previous data; secondly, the model's accuracy is determined. If it is acceptable, the model can be used to classify new data (Dougherty, 2013; Han et al., 2012). Supervised learning consists of parametric and non-parametric methods in statistical classification. Parametric methods such as Naïve Bayes and discriminant analysis need probability distributions to estimate a representation of the classes. On the other side, non-parametric methods must be used when the probability distributions are not known. In these situations, the density functions are estimated or the probabilities are bypassed and directly build decision boundaries on the basis of training data (Dougherty, 2013). The decision tree methodology is a nonparametric inductive learning technique.

3.2.1. Discriminant analysis

Discriminant analysis (DA) can be a generally robust and powerful method, if fundamental statistical assumptions are confirmed. These assumptions are multivariate normal distributions for independent variables in each class and homogeneity of variance–covariance structures for the different classes (Li and Anderson-Sprecher, 2006). Nevertheless, sufficient instances in each class in addition to fairly few predictors (five or fewer) makes it a proper classification technique in sundry applications (Tabachnick and Fidell, 1996).

Since the numbers of predictors considered in our problem are much less than the numbers of observations, this classifier can be robust. Here, the linear and quadratic functions were taken in discriminant analysis according to the statistical characteristics of the analyzed dataset. In the linear discriminant analysis (LDA) instances have similar covariance structures with different means for each class while quadratic discriminant analysis (QDA) provides conditions allowing different covariance matrices.

The LDA known as a generalization of Fisher's linear discriminant (Fisher, 1936) is vastly applied in artificial intelligent and data mining problems, e.g. (Cox and Wang, 2014; Duda et al., 2012; Huang and Guan, 2015; Imani and Ghassemian, 2015). Here is a review of Fisher discriminant analysis (Bishop, 2006) applied to the classification problem is briefly explained as follow. The LDA divides a D-dimensional input region into decision regions by the best (D-1)-dimensional hyper-planes being linear functions of input vector x. A discriminate function takes an input vector x (an instance) and assigns it to one of K Classes symbolized C_K. A simple linear discriminant function for two-class problem offered by:

$$y(x) = w^T x + w_0 \tag{2}$$

where w is a weight vector and w_0 is a bias that sometimes named threshold. The input vector x is allocated to class C_1 if $y(x) > 0$, and to C_2 else. The related decision boundary is specified by $y(x) = 0$.

Consider there are N_1 points of class C_1 and N_2 points of class C_2, so that the mean vectors of the two classes can be given by:

$$M_1 = \frac{1}{N_1} \sum_{n \in C_1} X_n, \quad M_2 = \frac{1}{N_2} \sum_{n \in C_2} X_n \tag{3}$$

The mean of projected data for class C_1 and C_2 are calculated by $m_1 = W^T M_1$ and $m_2 = W^T M_2$. Although a line joining means of two classes to each other properly separates them, they have somewhat overlapped for the instances projected onto the line. Fisher suggested a criterion which not only reduce the overlap but also maximize the separation simultaneously. This criterion, known as the objective function in Fisher discriminant analysis, is computed by Eq. (4).

$$J(W) = \frac{W^T S_B W}{W^T S_W W} \tag{4}$$

where S_B is the between-class scatter matrix and S_W is the within-class matrix.

In order to improve separated classes, this objective function should be maximized. The optimal W weight vector which classifies the instances with the maximum between-class variance and minimum within-class variance will be given by solving Eq. (5).

$$S_W^{-1} S_B W - J(W) W = 0 \tag{5}$$

The extension of LDA for over two classes is discussed in a similar manner (Fukunaga, 1990).

QDA is suggested as an alternative to avoid the tendency of cases for assigning to classes with higher variance due to greater posterior probability that happens because of the disruption of homogeneity assumption (Tabachnick and Fidell, 1996). If the assumption of a shared covariance matrix is relaxed, each class-conditional density $p(X|C_K)$ will have its own covariance matrix. Accordingly quadratic functions of X are obtained.

3.2.2. Naïve Bayes

Naïve Bayes (NB) is widely used for classification of high-dimensional data. In this simple form of Bayesian network, the values of the features are assumed to be conditionally independent of one another (Hernández-González et al., 2013; Wu et al., 2015). The class-conditional independence between features is generally optimistic assumption in various classification tasks (Webb et al., 2012). As long as this assumption is satisfied, the NB classifier estimates better parameters for classification and applies less training data than many other classifiers. However, this classifier may usually act well practically even if independence assumption is invalid.

With regard to the characteristics of data used in this study, two Gaussian and kernel distributions were considered separately to determine their effectiveness. The former is suitable when features have normal distributions in each class, and the latter is suitable when the distribution of features is skewed or has multiple modes.

In short, the Naïve Bayes classifier works as follows (Han et al., 2012):

Given an instance,*X*, the NB classifier predicts that *X* belongs to class C_i for which $P(C_i|X)$ is maximized. The class C_i is called the maximum posteriori hypothesis according to Bayes' theorem as Eq. (6)

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \tag{6}$$

To reduce the volume of calculations in estimating $P(X|C_i)$ for high-dimensional datasets, the Naïve assumption of class-conditional independence is considered. Thus,

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) = P(x_1|C_i) \times P(x_2|C_i) \times \ldots \times P(x_n|C_i) \tag{7}$$

where x_k refers to the value of attribute A_k for instance X. Above simple probabilities in Eq. (7) can be easily estimated from the training dataset. It is usually assumed that the continuous variable have a Gaussian distribution with a mean μ and standard deviation σ, defined by Eqs. (8) and (9).

$$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{(x-\mu)^2}{2\sigma^2}} \tag{8}$$

$$P(x_k|C_i) = g\left(x_k, \mu_{C_i}, \sigma_{C_i}\right) \tag{9}$$

where μ_{C_i} and σ_{C_i} indicates mean and standard deviation of the values of attribute A_k for training instances of class C_i. To predict the class label of X, $P(X|C_i)P(C_i)$ is calculated for each class C_i. The classifier predicts that the class label of instance *X* is the class C_i if and only if Eq. (10) is satisfied.

$$P(X|C_i)P(C_i) > P(X|C_j)P(C_j) \text{ for } 1 \le j \le m, j \ne i \tag{10}$$

3.2.3. Decision tree

Decision tree (DT) is employed to predict the response variables as a function of the predictors. It uses training data and creates a model that predicts a goal class based on input attributes for an unobserved test instance (Dougherty, 2013; Quinlan, 1993). The researchers of this study wanted to use the decision tree to extract rules defining the relationship between predictors and to illustrate how a sample belongs to a specific class.

The decision tree includes advantages such as robustness, being simple to understand, and easy to implement. Moreover, it requires little prior knowledge and is applicable on large and noisy datasets (Farid et al., 2014). This method has been applied progressively by geoscientists in several classification tasks (Akkaş et al., 2015; Chasmer et al., 2014; Li et al., 2013; Pradhan, 2013; Shi, 2014; Xue et al., 2015).

Decision tree algorithms (e.g., ID3, C4.5, and CART) are considered for classification. These algorithms adopt a greedy approach in which decision trees are built in a top-down recursive manner. To create a decision tree, all data first are collected in a root node and then divided into relatively more homogenous branch nodes until achieving leaf nodes as following steps. (1) Check all possible splits on each attribute in all input data, (2) choice the best attribute separating the instances into distinct classes based on attribute selection measures and enforce it, (3) repeat mentioned steps for the two child nodes recursively, and (4) stop splitting in a node when it contains just instances of one class or satisfy some predefined conditions. Finally, labels of the leaf nodes are specified by their allocated statistical information (Pal and Mather, 2003).

Attribute selection measures commonly comprise information gain, gain ratio, and Gini index which are used in different algorithms. A review of Gini index giving reasonably good results in our study is explained as follows (Han et al., 2012). The notation used in this fact is summarized in Table 1.

The Gini index evaluates the impurity of D as Eq. (11) and it considers a binary split for every attribute.

$$Gini(D) = 1 - \sum_{i=1}^{k} p_i^2 \tag{11}$$

where p_i is estimated by $|C_{i,,D}|/|D|$. If a binary split on *A* partitions *D* into D_1 and D_2, the Gini index of D is defined by Eq. (12)

$$Gini_A(D) = \frac{|D_1|}{|D|} \text{Gini}(D_1) + \frac{|D_2|}{|D|} \text{Gini}(D_2) \tag{12}$$

Each of the possible binary splits is considered for every attribute. The reduction in impurity incurring by a binary split on an attribute *A* is explained by Eq. (13).

$$\Delta Gini(A) = Gini(D) - Gini_A(D) \tag{13}$$

The attribute with minimum Gini index maximizes the reduction in impurity and is chosen as the best splitting variable at each node.

Many branches in a decision tree commonly indicate noise or outliers in the training data. The decision tree can be improved by controlling the leafiness. Since a leafy tree tends to overtrain, it has high predictive power on the training set, but not on the test set.

Pruning is an alternative method that optimizes the tree leafiness. It is the process of reducing a tree by turning some branch nodes into leaf nodes and removing the leaf nodes under the original branches. Such method usually uses statistical measures to eliminate the least reliable branches. Although giving a higher resubstitution error, the optimal pruned tree tends to be less complex and acts usually faster and better at correctly classifying independent test data. Hence it's generally faster and more efficient to classify test data than unpruned trees (Dougherty, 2013; Han et al., 2012).

A simple classification tree which predicts classes based on a set of rules by two predictors, P_1 and P_2, is shown in Fig. 2. Predicting is started at the top node. The first decision is whether P_1 is smaller than 0.5. If so, then follow the left branch. Here the tree asks if P_2 is smaller than 0.5. If so, then follow the left branch, else follow the right branch to show that the tree classifies the instance into class C_1 or C_2 respectively. Otherwise, if P_1 exceeds 0.5, then follow the right branch and the instance is classified to class C_1.

Table 1
Description of the symbols generally used in decision tree algorithm.

Symbols	Description
D	A training dataset of class-labeled instances
D_j	A subset of a training dataset
X	A set of instances with n-dimensional attribute
A	An attribute
k	The number of distinct classes
C_i	A class label for i = 1, … , k
$C_{i,D}$	The set of instances of class C_i in D
$\|D\|$	The number of instances in *D*
$\|C_{i,D}\|$	The number of instances in $C_{i,D}$
P_i	The probability that a instances in *D* belongs to class C_i

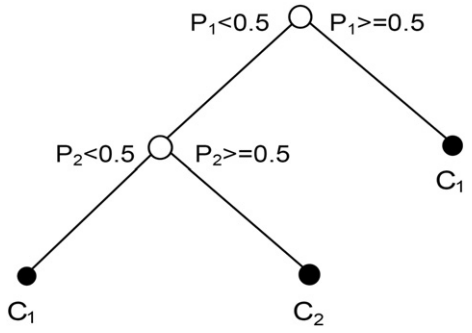

Fig. 2. A set of rules for estimating the different classes by a simplified decision tree.

4. Discussion and results

4.1. Sampling and data preparation

In the current study, an exploration program was followed to recognize geochemical patterns related to REE occurrences based on whole rock compositions. Detection and interpretation of geochemical anomalies in this study were converted to a comprehensible classification procedure where background and anomalous samples belonged to distinct classes.

For this propose, a systematic sampling program was designed to detect the geochemical variability of the Choghart orebody with an adequate level of confidence. The sampling pattern covers the bottom parts of ore preserved from the effects of weathering to its surrounding host rocks frequently affected by hydrothermal processes as shown in Fig. 3. A sufficient number of representative rock samples (N = 113) were taken from the surface of the pit. The collected samples were cleaned in deionized water, dried at room temperature, crushed, and powdered using standard procedures. The prepared samples were analyzed for major element oxides consisting of TiO_2, Fe_2O_3, Al_2O_3, SiO_2, CaO, MgO, SO_3, P_2O_5, K_2O, MnO, Na_2O, and L.O.I. using the X-ray fluorescence method (XRF) and for REE using inductively coupled plasma mass spectrometry (ICP-MS).

Because they originated from heterogeneous sources, the geochemical datasets contain noisy, missing, and inconsistent data. Such low-quality data will lead to undesirable effects in data processing results. Therefore, preprocessing methods should be used on the data to increase its quality and improve the performance of classification methods. Some data preprocessing techniques such as data cleaning, data outlier detection, and data transformation were used on the chemical analytical results in this study. The basic statistical parameters of the prepared geochemical dataset are given in Table 2 and the histograms of the LREE and the HREE are shown in Fig. 4.

Some classification models were employed to predict discrete values of REE concentrations based on some continuous predictors. To do this, the REE content values had to be transformed into discrete values properly using the k-means clustering algorithm. The clustering results were fairly modified using the mean and standard deviation of continuous values to identify distinct classes of REE concentrations as very low, low, medium or high.

4.2. Geochemistry of REEs in the choghart orebody

The Choghart deposit exhibit a wide variation in mineralogy, rocks texture, compositions, and hydrothermal alteration degrees. According to the geological characteristics, at least four main types of formation are recognized in the Choghart orebody: albitophire zone (host rock),

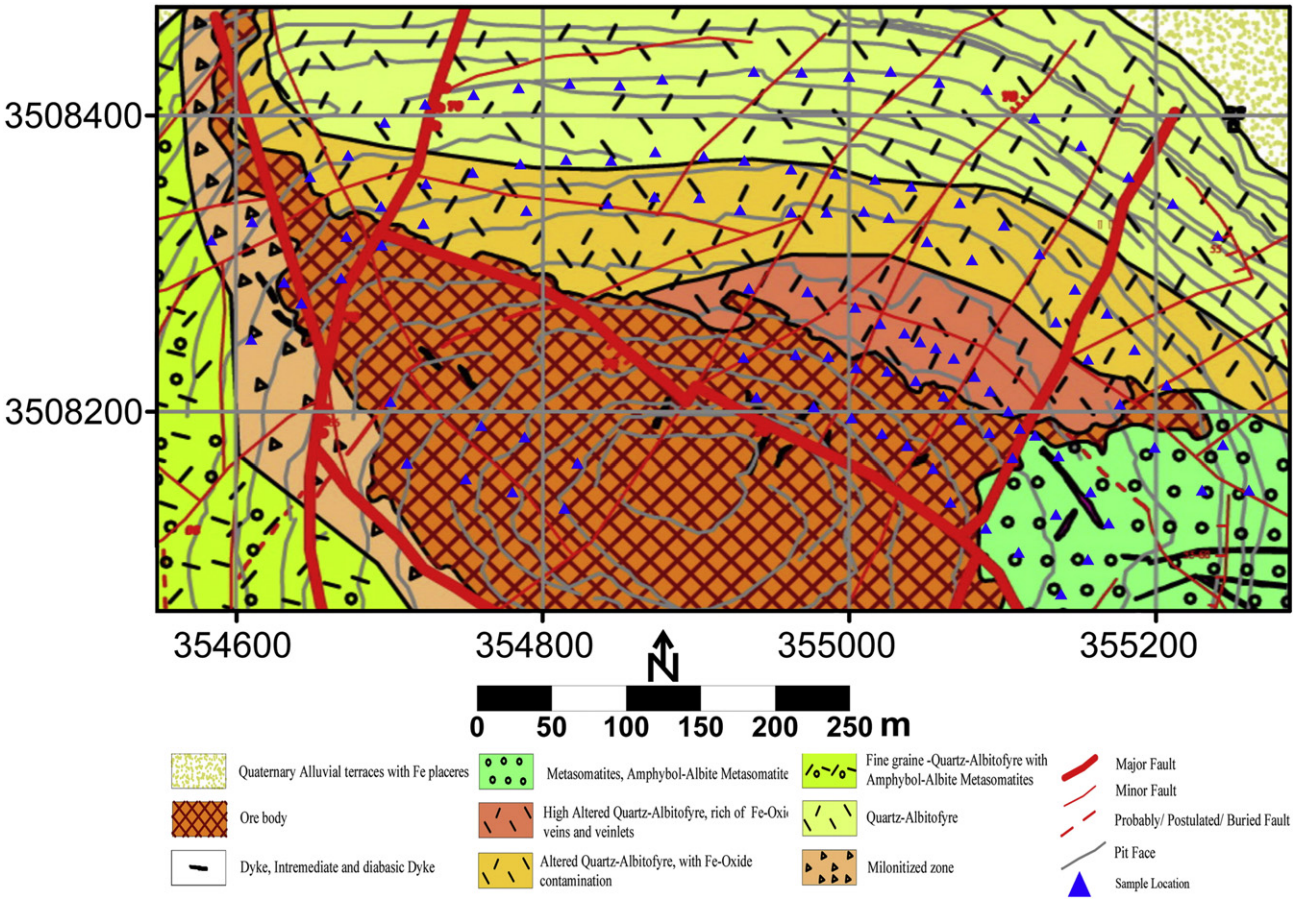

Fig. 3. Simplified geological map (modified after Zahed and Seidy (2012)) showing sample locations in the Choghart orebody.

Table 2
The basic statistics of geochemical data (major element oxides and REE contents) in Choghart orebody.

Compositions	No.	Minimum	Maximum	Mean	Std. Deviation	Median	Skewness	Kurtosis
(wt.%)								
TiO_2	113	0.02	3.56	0.67	0.88	0.2	1.69	1.75
Fe_2O_3	113	0.38	93.56	21.47	29.71	5	1.29	0.02
Al_2O_3	113	0	18.49	6.88	6.18	7	0.07	−1.79
SiO_2	113	0	81.2	43.58	28.4	46.48	−0.23	−1.66
CaO	113	0.14	51.15	10.53	15.11	3.19	1.52	0.76
MgO	113	0.23	13.7	3.89	3.22	3.41	0.95	0.48
SO_3	113	0	2.92	0.08	0.3	0.01	8.05	74.47
P_2O_5	113	0.01	45.76	6.29	11.88	0.04	1.85	2.28
K_2O	113	0	7	0.66	0.94	0.32	3.69	19.86
MnO	113	0	0.21	0.04	0.05	0.02	1.39	1.35
Na_2O	113	0	9.53	3.49	3.61	1.05	0.33	−1.78
L.O.I	113	0	17.95	2.42	2.53	1.87	3.43	16.1
(ppm)								
La	113	6.54	2728.4	454.44	730.28	100.89	1.91	2.43
Ce	113	30.43	5290.59	732.97	1310.31	165.79	2.33	4.28
Pr	113	0.01	476.65	59.89	98.78	18	2.61	6.66
Nd	113	0.01	1971.63	306.98	483.29	78.09	2.03	3.23
Sm	113	0.01	432.05	52.71	86.94	18.66	2.87	8.03
Eu	113	0.44	3050.23	151.92	534.77	13.45	4.15	17.19
Gd	113	3.24	483.54	117.72	85.63	89.65	1.79	3.97
Tb	113	0.27	573.76	34.83	105.63	6.46	4.23	17.09
Dy	113	0.01	270.49	26.36	47.31	11.28	3.31	11.9
Ho	113	0.16	140.78	6.81	14.61	3.27	7.28	64.05
Er	113	0.01	93.98	13.22	17.98	6.59	2.94	8.85
Tm	113	0.12	91.97	7.64	16.61	2.04	3.6	13.19
Yb	113	0.45	183.07	14.3	20.91	7.09	5.26	38.04
Lu	113	0.18	245	4.78	22.86	2.45	10.54	111.77
Y	113	3.91	6624.52	1355.16	1323.67	1002.08	2.18	5.4
Sc	108	12.04	66.2	32.47	11.61	31.17	0.4	−0.07
Nb	113	20.05	4975.41	2314.71	852.42	2323.54	0.35	0.92
Σ LREE	113	91.81	11,844.61	1758.9	2904.11	477.91	2.15	3.48
Σ HREE	113	22.83	776.9	225.66	168.94	148.58	1.33	1.19
Σ REE	113	190.3	12,353.65	1984.56	2993.75	695.74	2.16	3.52
Σ REE + Y	113	585.11	14,543.9	3339.72	3756.71	1638.45	1.85	2.09

iron oxide zone (ore), metasomatic zone, and high phosphate zone. Moreover, a subtype of high phosphate and iron oxide zone is recognized. Magnetite is the major ore mineral, hematite is typically created from a secondary source, and apatite is the most abundant gangue in this orebody.

The fluid inclusion study illustrated that the formations of apatite in Choghart occur in two mineralization episodes. The first is associated with a magmatic origin in high-pressure conditions (up to 2 Kb) and a temperature close to 600 °C. The second also follows the same mechanism with the difference that it was composed at a temperature lower than 500 to 300 °C. The inclusions of monazite and xenotime were also observed within apatite types I and II in the Choghart orebody.

The average contents of LREE, HREE, and Y in various zones are stated in Table 3. These values suggest the main carriers might be different for each of these groups in each zone. Phosphate mineralization can be a critical factor controlling REEs concentration in rocks so that its total average content may rise to 1.5 wt%. In addition significant concentrations for REE have been observed in transitional zone between ores and gangue in the NE of orebody. The chondrite-normalized REE patterns generally demonstrate significant variation from LREE to HREE with negative Eu anomalies in different zones. These patterns might be due to localized crystallization of LREE minerals. These diagrams also show that the concentration of REE is much higher in the phosphate zone than in the metasomatic, iron oxide, and albitophyre zones, respectively.

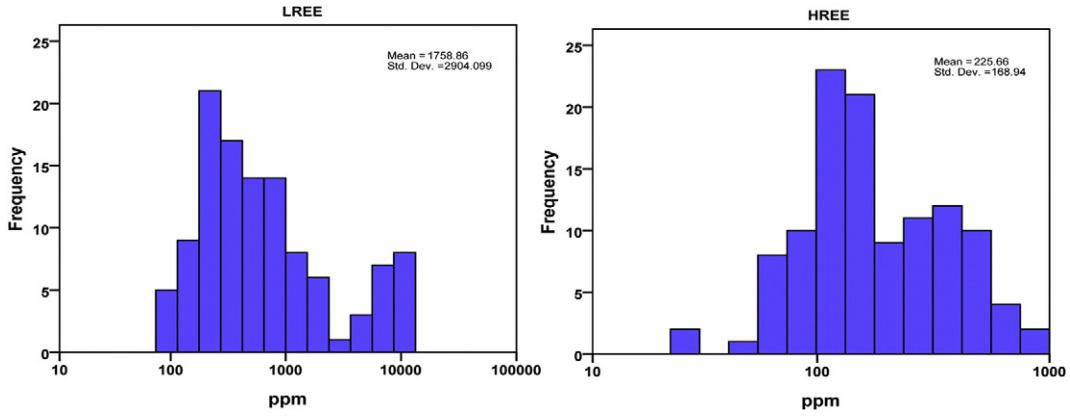

Fig. 4. Histograms of LREE and HREE content of samples in the Choghart orebody.

Table 3
The average contents of REEs in various zones of the Choghart orebody.

Compositions (ppm)	Iron oxide	High phosphate	Metasomatic	Albitophire
LREE	731.1	3277.9	450.5	310.9
HREE	371.9	241.1	147.1	148.2
Y	562.7	1898.5	1173.1	1107.3

4.3. Quantifying the influence of rock compositions on the spatial distribution patterns of REEs

Primary geochemical characteristics were influenced by secondary processes, such as multi-stage mineralization, alteration, and weathering in Choghart deposits. The wide range of primary and secondary processes might cause redistributions of REE in different zones. Hence, REE distribution patterns are supposed to be complicated.

The researchers in this study believe that the rocks could effectively record geochemical information related to genesis and formation conditions of deposits with variations in their chemical compositions. According to the statistical characteristics presented in Table 2, diverse classification methods such as discriminant analysis (with linear and quadratic functions), Naïve Bayes (normal and kernel distributions), and decision tree were used separately to predict distinct classes of REE concentrations based on their indicator predictors.

Classification methods are generally evaluated and compared by the criteria of misclassification rates (Hastie et al., 2009). These criteria were calculated based on resubstitution and cross-validation methods in all possible states of predictors in the current survey.

Resubstitution error is the difference between predicted values and the corresponding real values through all training data. A high amount of resubstitution error clearly states a bad classification model; however, a low amount of it is not a suitable guarantee to predict new datasets.

If enough data exists, a validation set can be set aside to evaluate the performance of a classifier model, but inadequate data makes this impossible in the present study. In order to solve this problem, the k-fold cross-validation method was used. This method randomly divides the instances into k separate parts with the same size. The k defines the number of training and test sets in cross-validation. The classification function uses the existing training subsets to fit a suitable model and predicts class labels for the different test subset in every fold. The number of misclassifications was calculated between the predicted value and corresponding real value in a test set, and the overall misclassification rate was returned through all test sets.

The predictive performance of a decision tree was also measured by a cost of tree. The cost of each node is the ratio of bad classified instances in that node. The cost of the tree is equal to the overall sum of the cost of each leaf node multiplied by occurred probability (Zadrozny et al., 2003). The cost values were calculated for remaining subtrees after each pruning step.

The suitable cost values obtained by k-fold cross-validation were applied effectively to evaluate predictive performance of the decision tree classifier in this paper. To achieve the optimal subtree, the cross-validation costs were plotted versus the number of terminal nodes. The smallest tree was specified with one standard error of the minimum cost of subtree as shown in Fig. 5.

The inputs of the classification methods, called indicator predictors, are N-dimensional vectors of attributes including the geographical location of samples (X, Y, and Z) and their major element oxide contents. According to some considerations and desired goals, three predictors are optimally chosen for all classification models. Therefore, all possible 3-combinations of 15 predictors, denoted by $\binom{15}{3}$, are considered as subsets of predictors to identify the behavior of each REE.

For each classification method, 455 prediction models were constructed using these subsets of predictors; then the best of them causing the minimum misclassification rate based on cross-validation were selected as the main indicators. Some REE having at least one minimum error less than 20% and their related results are shown in Table 4. For example, the first row of this table indicates the subset of predictors $\{Z, TiO_2, P_2O_5\}$ causing the minimum cross-validation error (17.70) to predict different classes of La concentrations by linear discriminant analysis (LDA). If the same predictors are similarly used by other methods, their related cross-validation errors will be QDA = 15.04, GNB = 15.93, KNB = 15.04, and DT = 8.85.

For each classifier, the minimum misclassification rate was introduced as the best prediction errors as shown in Figs. 6 and 7. The final model constructed by the best selected predictors could be reasonably applied to the new data to predict their REE classes.

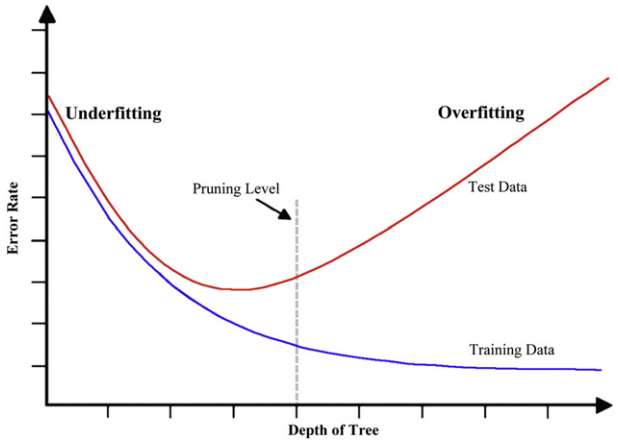

Fig. 5. A typical example of determining the optimal pruning level to avoid overfitting effects in a decision tree (Dougherty, 2013).

Table 4
The best combinations of predictors based on minimum cross-validation (Only those REE are shown which have at least one minimum error less than 20%. The minimum error for each classification method is indicated in bold.).

REE	1st Predictor	2nd Predictor	3rd Predictor	LDA*	QDA*	GNB*	KNB*	DT*
La	**Z**	**TiO_2**	**P_2O_5**	**17.70**	15.04	15.93	15.04	8.85
	Z	**P_2O_5**	**Na_2O**	21.24	**10.62**	17.70	14.16	9.73
	Y	**Z**	**P_2O_5**	20.35	15.04	**13.27**	17.70	10.62
	Y	**P_2O_5**	**Na_2O**	23.01	14.16	19.47	**13.27**	11.50
	Z	**SiO_2**	**P_2O_5**	23.89	15.04	16.81	15.93	**7.96**
Ce	**CaO**	**SO_3**	**L.O.I**	**17.70**	23.01	23.01	19.47	22.12
	Z	**CaO**	**P_2O_5**	20.35	**14.16**	22.12	22.12	18.58
	Y	**SiO_2**	**P_2O_5**	31.86	20.35	**15.93**	23.01	18.58
	X	**CaO**	**SO_3**	23.89	22.12	23.01	**16.81**	22.12
	Y	**Z**	**CaO**	28.32	20.35	21.24	23.89	**15.04**
Nd	**CaO**	**SO_3**	**L.O.I**	**19.47**	22.12	25.66	19.47	22.12
	X	**P_2O_5**	**L.O.I**	35.40	**18.58**	22.12	26.55	24.78
	Fe_2O_3	**P_2O_5**	**Na_2O**	30.97	29.20	**17.70**	29.20	19.47
	CaO	**MgO**	**L.O.I**	29.20	23.01	23.89	**16.81**	26.55
	Z	**TiO_2**	**P_2O_5**	27.43	23.89	24.78	33.63	**15.04**
Sm	**Fe_2O_3**	**Al_2O_3**	**CaO**	**31.86**	69.03	32.74	36.28	37.17
	X	**Fe_2O_3**	**MgO**	67.26	**29.20**	28.32	30.09	44.25
	TiO_2	**SiO_2**	**MgO**	36.28	50.44	**21.24**	30.97	30.09
	Z	**SiO_2**	**MgO**	42.48	44.25	23.01	**21.24**	35.40
	X	**TiO_2**	**SiO_2**	38.05	35.40	27.43	29.20	**18.58**
Eu	**Z**	**P_2O_5**	**MnO**	**20.35**	13.27	19.47	28.32	8.85
	Z	**CaO**	**MnO**	21.24	**11.50**	13.27	24.78	9.73
	Z	**CaO**	**MnO**	21.24	11.50	**13.27**	24.78	9.73
	Z	**Fe_2O_3**	**L.O.I**	36.28	30.97	29.20	**17.70**	15.04
	Z	**TiO_2**	**Al_2O_3**	34.51	29.20	27.43	20.35	**6.19**
Ho	**X**	**Fe_2O_3**	**P_2O_5**	**30.97**	40.71	19.47	24.78	17.70
	X	**Al_2O_3**	**MgO**	43.36	**25.66**	19.47	22.12	18.58
	X	**Fe_2O_3**	**MgO**	52.21	29.20	**13.27**	15.93	23.01
	X	**TiO_2**	**SiO_2**	36.28	27.43	25.66	**9.73**	20.35
	Z	**SiO_2**	**Na_2O**	46.02	47.79	36.28	26.55	**9.73**

* LDA: Linear Discriminant Analysis, QDA: Quadratic Discriminant Analysis, GNB: Gaussian Naïve Bayes, KNB: Kernel Naïve Bayes, and DT: Decision Tree.

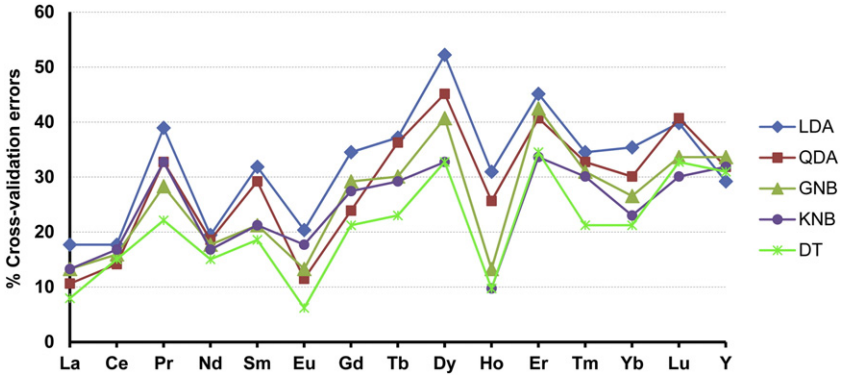

Fig. 6. The best prediction errors for predicting concentrations of REE by the cross-validation method.

These line charts show that the least amount of prediction errors occurred mostly in the decision tree. Other methods, such as kernel Naïve Bayes classification, Gaussian Naïve Bayes classification, quadratic discriminant analysis, and linear discriminant analysis, are ordered by taking the least amount of prediction errors, respectively. Therefore, the decision tree classifier was selected as a more suitable method than the others for predicting different distribution patterns of REEs in the Choghart orebody.

It is desirable to have both training accuracy and good generalization ability in the decision tree results. To do this, a new parameter was defined as a combination of the resubstitution error, cross-validation error, and cost of the tree. This parameter was calculated for all possible subsets of predictors and sorted in ascending order for each REE. The three best subsets of predictors from the head of each list were selected as shown in Fig. 8.

With regard to the least amount of prediction errors in the diagram, some REE (La, Ce, Nd, Sm, Eu, and Ho) are reasonably predictable, others (Pr, Gd, Tb, Tm, and Yb) are partly predictable, and the others (Dy, Er, Lu, and Y) are hardly predictable when their respective predictors are changed in different parts of the orebody.

The prediction of some elements, especially in the first group, is more meaningful than the others because either source of magma would have a high content of them or secondary evolution process might modify their distributions. Most of them, except the Ho, are known as LREE which are generally more enriched in different zones of Choghart orebody. Moreover, these elements have more ability to form complexes rather than others in hydrothermal systems; thus, they might be mobilized in the Choghart deposit and would extend their distributions.

The other elements almost include the HREE and exhibit more irregular behaviors in this region. Previous studies and analytical results imply there is no evidence for extended enriched zones of these elements. However, some elements might be concentrated locally in accessory minerals such as allanite, monazite, xenotime and apatite which would not break down during metasomatism and alteration. Therefore, it is likely that there is no meaningful relationship between distribution patterns of these elements and using predictors in the Choghart orebody. The irregular distribution of some REEs might be related to more complex geological features which might not affect rock compositions at an acceptable level.

Among the stated subsets in Fig. 8, a number of predictors were selected based on the greater relative frequencies. These predictors are presented in a stacked column chart as in Fig. 9. Each column of the chart includes the participation rate of the most important influencing factors to predict distinct classes of each REE.

These factors related to chemical composition of minerals that occur in the same type of ore deposit. Most rock-forming minerals, with the exception of plagioclase, have more enrichment in LREE than HREE in different zones of Choghart orebody. The replacement of REE in particular minerals is restricted due to their different ionic radius. Allanite hosts the larger LREE and zircon host the smaller HREE. Apatite and sphene indicates any priority for the LRRE, HREE, or the MREE. Among the major REE-bearing minerals phosphates, monazite incorporates REE from La to Gd and xenotime incorporates them from Tb to Lu preferentially (Ni et al., 1995; Rollinson, 1993). Hence the presence of each of these minerals might cause different behavior patterns of each element in the region.

The last column of the chart exhibits the overall participation rates of factors in the process of predicting the geochemical behavior of REEs. The most important features as Z, Fe_2O_3, SiO_2, P_2O_5, and Na_2O can be considered related to depth factor in addition to iron oxide, metasomatic, phosphate, and albitophire zones. They are so important that their changes can effectively determine the different concentrations of most of the REEs in the orebody. These results can be confirmed geologically

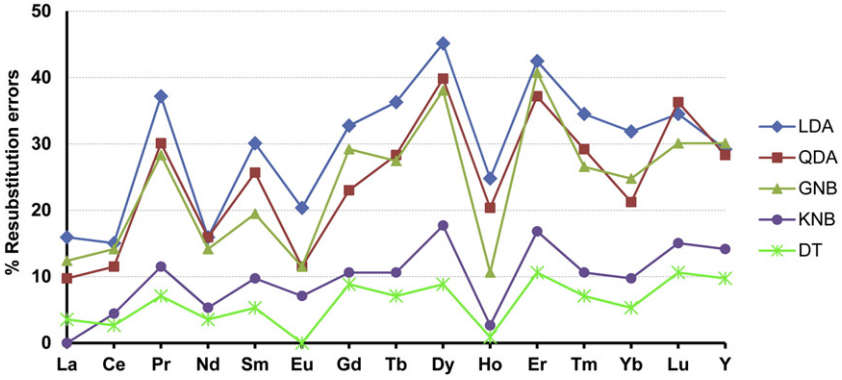

Fig. 7. The best prediction errors for predicting concentrations of REE by the resubstitution method.

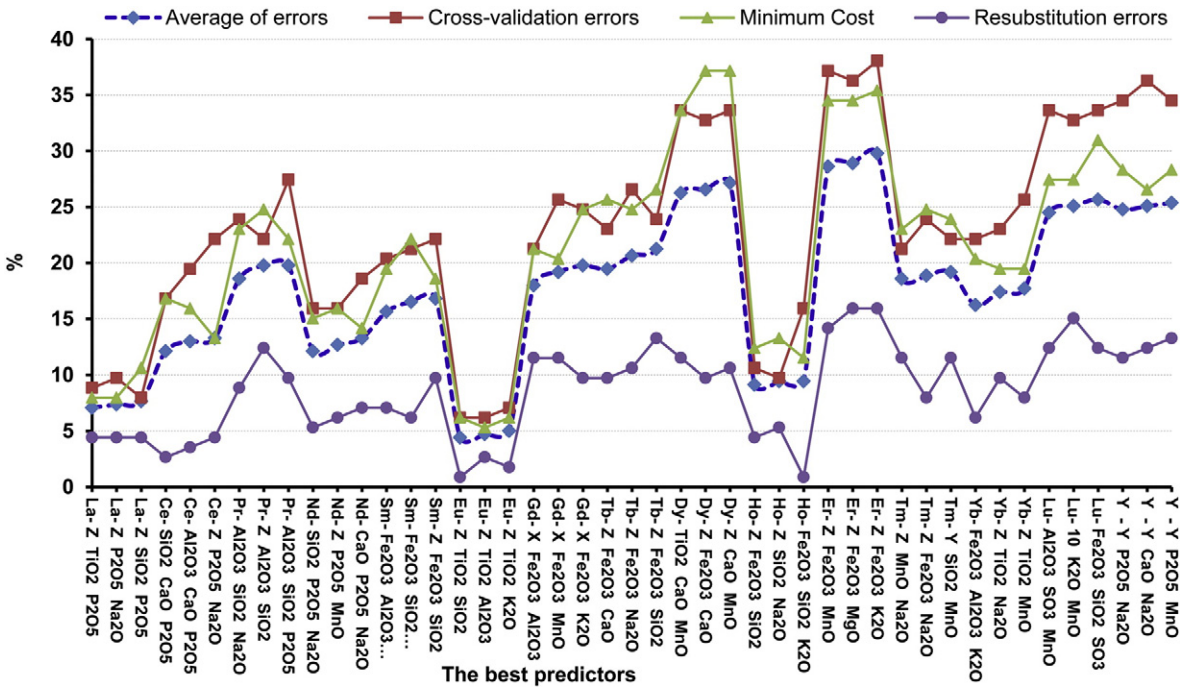

Fig. 8. The best subsets of each REE predictor according to the decision tree results.

by the main geochemical zones in the Choghart orebody mentioned in previous section.

The decision tree was pruned by replacing a whole subtree with a leaf node until optimal level is reached in order to prevent the effects of leafiness, overfitting, and complexity. This work caused the pruned trees to exhibit a better performance than the unpruned ones to predict REE distribution patterns. Some pruned trees related to La, Ce, Nd, and Eu, for example, are shown in Fig. 10.

After training and pruning the decision trees, several classification rules were extracted to describe the relationship among the most important factors influencing the different concentrations of REE. These rules are related to the geochemical conditions dominating in the Choghart orebody. Because the aim of exploration projects is to identify REE-rich zones, the geochemical rules associated to class 'High' are stated in Table 5.

The values of the cost and the estimated probability determine the validity of these rules. The cost explains the ability of the best pruned tree to predict different classes of REE concentrations, and the estimated probability measures the purity of a leaf node in a pruned tree and defined as the proportion of samples correctly placed in this node. Both

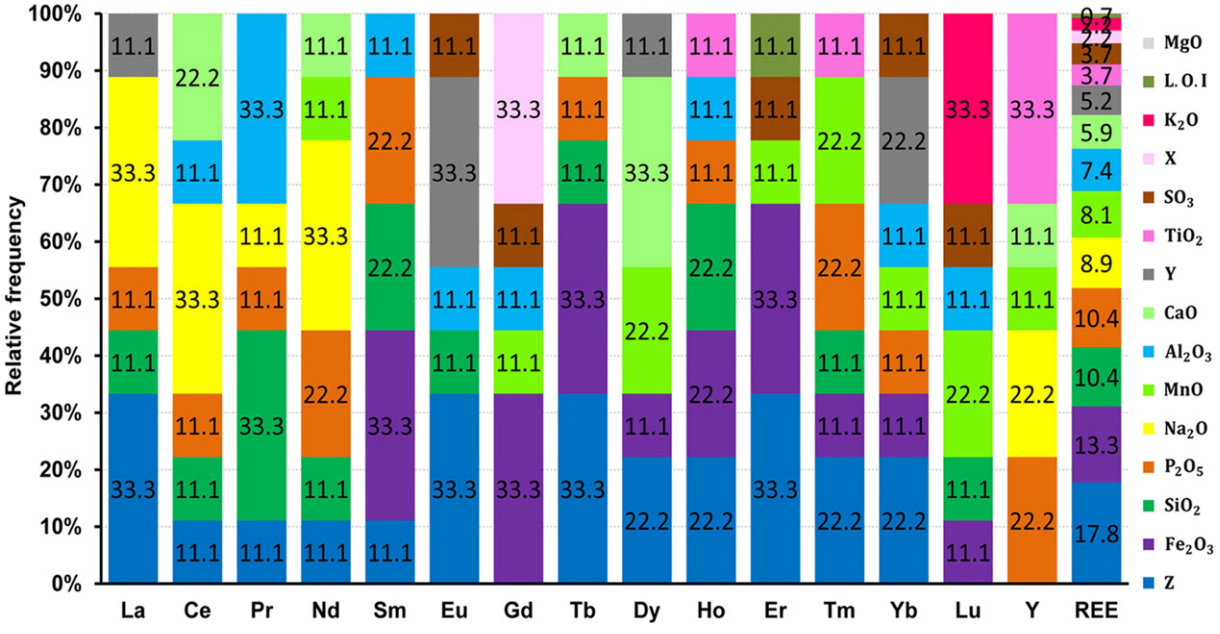

Fig. 9. The most important influencing factors and their participation rate percentages for predicting the distribution patterns of REE.

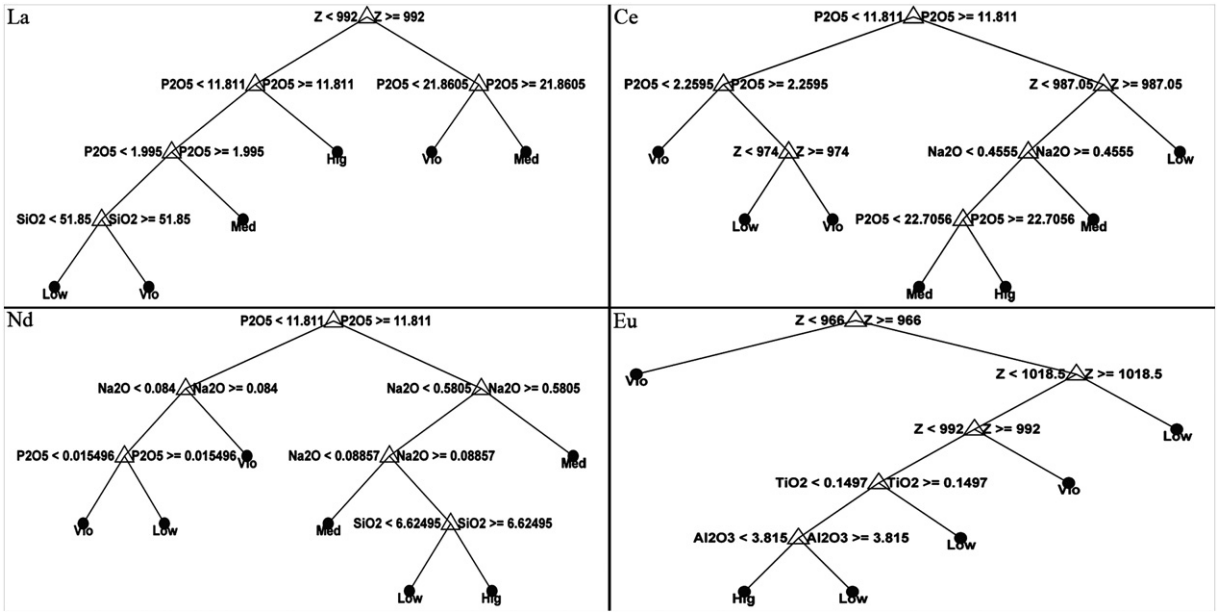

Fig. 10. Pruned trees and related rules in order to predict the distribution patterns of La, Ce, Nd, and Eu.

of these values are between zero and one. The lower amounts of cost along with the higher amounts of estimated probability represent more accurate geochemical rules.

To clarify, the entries related to Ce prediction in Table 5 are stated as follows: The best subsets of predictors are {Z, P_2O_5, Na_2O}, {SiO_2, CaO, P_2O_5}, and {Al_2O_3, CaO, P_2O_5}. The costs of the pruned trees for these subsets are equal to 0.115, 0.150, and 0.168, respectively. For each pruned tree, one rule leading to a high Ce concentration is extracted. A new sample will be classified to class 'High' if it satisfies one of the following rules:

- Z<987.05 and Na_2O<0.46 and P_2O_5> = 22.71
- P_2O_5> = 11.81 and SiO_2> = 6.38 and CaO> = 36.64

Table 5
Extracted geochemical rules to predict high concentrations of REE in the Choghart orebody (The most accurate geochemical rules are indicated in bold.).

REE	Predictors*	Cost of pruned tree	The rule related to REE prediction as class 'High'**	Estimated probability for class 'High'
	03 04 11	0.071	**If Z < 992 & P_2O_5 > = 11.81**	1
La	03 11 14	0.080	**If Z < 992 & P_2O_5 > = 11.81**	1
	03 07 11	0.071	**If Z < 992 & P_2O_5 > = 11.81**	1
	03 11 14	0.115	**If Z < 987 & Na_2O < 0.46 & P_2O_5 > = 22.71**	1
Ce	07 08 11	0.150	If P_2O_5 > = 11.81 & SiO_2 > = 6.38 & CaO > = 36.64	0.833
	06 08 11	0.168	If P_2O_5 > = 11.81 & P_2O_5 < = 34.55 & CaO > = 36.64 & Al_2O_3 > = 0.12	0.833
	06 07 11	0.195	If P2O5 > = 22.71 & SiO2 > = 19.06	1
Pr	06 07 14	0.212	If Al_2O_3 < 12.8565 & SiO_2 > = 19.06 & SiO_2 < 26.21 & Na_2O > = 0.16	1
	03 06 07	0.212	If Al_2O_3 < 12.8565 & SiO_2 > = 19.06 & SiO_2 < 26.21 & Z > = 966.4	1
	07 11 14	0.150	**If P_2O_5 > = 11.81 & Na_2O > = 0.09 & Na_2O < 0.58 & SiO_2 > = 6.62**	0.909
Nd	08 11 14	0.133	If P_2O_5 > = 11.81 & Na_2O > = 0.09 & Na_2O < 0.58	0.839
	03 11 13	0.177	If P_2O_5 > = 11.81 & MnO > = 0.03	0.833
	05 07 14	0.150	**If SiO_2 < 44.00 & Na_2O > = 0.15 & Fe_2O_3 > = 1.77 & Fe_2O_3 < 6.02**	1
Sm	05 06 14	0.186	**If Al_2O_3 < 6.75 & Na_2O > = 0.15 & Fe_2O_3 > = 1.77 & Fe_2O_3 < 6.02**	1
	03 05 07	0.177	If SiO_2 < 44.00 & Z > = 973.45 & Z < 982.55	0.889
	03 04 06	0.053	**If Z > = 966 & Z < 992 & TiO_2 < 0.15 & Al_2O_3 < 3.82**	1
Eu	03 04 07	0.053	**If Z > = 966 & Z < 992 & TiO_2 < 0.15 & SiO_2 < 43.19**	1
	03 04 12	0.062	**If Z > = 966 & Z < 992 & TiO_2 < 0.15 & K_2O < 0.15**	1
	01 05 06	0.195	**If Al_2O_3 < 9.8 & Fe_2O_3 < 3.84 & X < 5133.5**	0.857
Gd	01 05 13	0.177	If MnO > = 0.006 & Fe_2O_3 < 3.84	0.667
	01 05 12	0.212	If Fe_2O_3 < 3.84 & X > = 5035 & X < 5125.5	0.667
	03 05 08	0.195	If Fe_2O_3 > = 75.93 & Z < 976.1	0.625
Tb	03 05 07	0.186	If Fe_2O_3 > = 75.93 & Z < 976.1	0.625
	03 05 14	0.186	If Fe_2O_3 > = 75.93 & Z < 976.1	0.625
	04 08 13	0.283	If CaO > = 37.5 & MnO > = 0.025	1
Dy	03 05 08	0.336	If CaO > = 37.5 & Z < 987.05	1
	03 08 13	0.345	If CaO > = 37.5 & Z < 987.05	1
	03 05 13	0.319	If Z > = 982.55 & Fe_2O_3 > = 58.66	1
Er	03 05 12	0.319	If Z > = 982.55 & Fe_2O_3 > = 58.66	1
	03 05 09	0.319	If Z > = 982.55 & Fe_2O_3 > = 58.66	1
	02 11 13	0.257	If P_2O_5 > = 22.71 & MnO < 0.034	0.875
Y	02 11 14	0.265	If P_2O_5 > = 22.71 & Na_2O > = 0.54	1
	02 08 14	0.248	If Y > = 8234.5 & CaO > = 26.39 & Na_2O > = 0.54	1

* 01: *X*, 02: *Y*, 03: *Z*, 04: *TiO_2*, 05: *Fe_2O_3*, 06: *Al_2O_3*, 07: *SiO_2*, 08: *CaO*, 09: *MgO*, 10: *SO_3*, 11: *P_2O_5*, 12: *K_2O*, 13: *MnO*, 14: *Na_2O*, 15: *L.O.I.*
** Major oxides are expressed *in weight percentages and X, Y and Z* are expressed *in meters.*

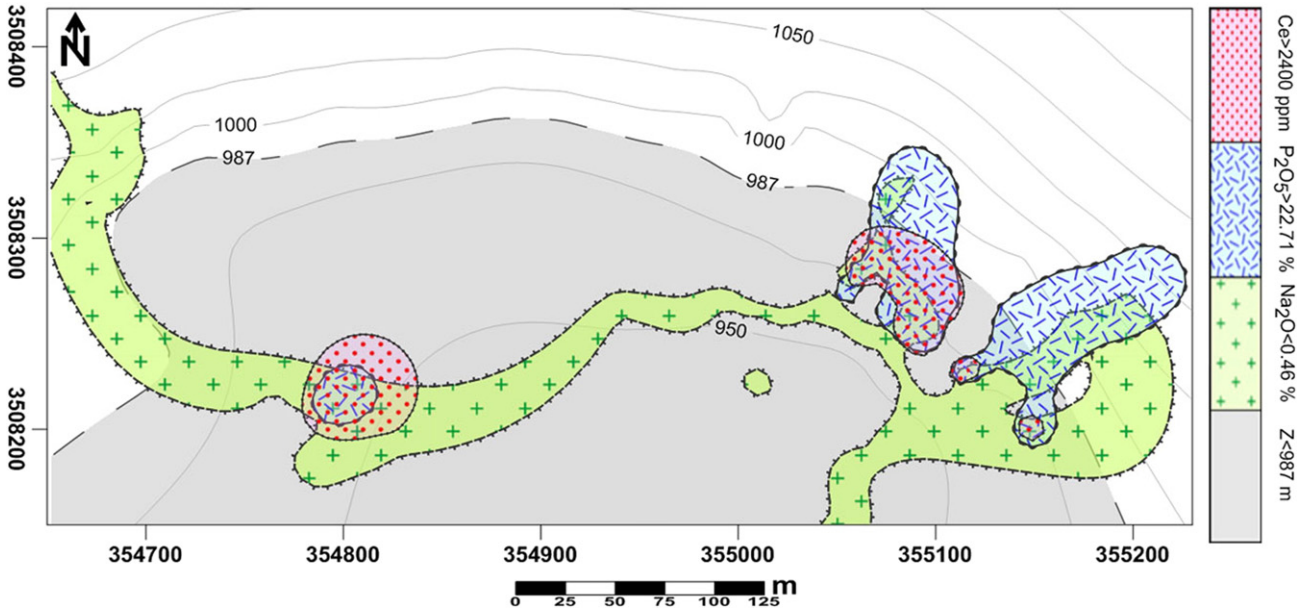

Fig. 11. Intersection of several regions bounded by extracted rules to identify target area representing Ce-enriched zones.

- $P_2O_5 >= 11.81$ and $P_2O_5 <= 34.55$ and $CaO >= 36.64$ and $Al_2O_3 >= 0.12$

The estimated probabilities of these rules are 1, 0.833, and 0.833, respectively. Therefore, the first rules having lower costs and higher estimated probability rates are introduced as the best geochemical rule for prediction of Ce-enriched zones in the Choghart orebody in this paper. According to the results presented in Table 5, the significant features P_2O_5 and CaO are representative of phosphate minerals. They control REEs concentrations and modify their patterns in related zones. The features Na_2O and Al_2O_3, related to albitophire zones which are depleted of REEs. The variations of depth factor ($966 < Z < 992$) illustrate that the high concentrations of REEs are mostly limited to a certain depth of orebody. This range refers to transitional zone between ores and gangue in the NE of orebody which also confirms the validity of the model used. Moreover, it will be important to give more attention to the concentration of REEs in those formations in which the mentioned geochemical rules are established.

Data visualization is an effective technique used to see patterns, recognize trends, and identify anomalies. Thus, a specific geochemical map related to the rules above is shown in Fig. 11. In such a case, the target area is considered to be the intersection of several regions, each of which is bounded by one of the conditions. According to this figure, the mentioned area and Ce-enriched zones overlap significantly. As a result, it is possible to propose the geochemical maps representing variation of major oxides as target zones. These zones will be candidates for more detailed exploration studies which require accurate and expensive analytical methods to determine the amount of rare earth elements.

The method employed in this paper can be similarly extended in Bafq mining district to predetermine possible enriched zones of REEs. In addition, it is even useful to discover spatial distribution patterns of REEs in other deposits. By using the findings of this study in REE exploration projects, a great savings in time and cost will be affected.

5. Conclusion

Multistage hydrothermal-magmatic processes played a significant role in the evolution of Choghart IOA deposit which might give rise to the REE redistribution. The concentrations of REE in the different zones of orebody could be reflected by the whole-rock chemical compositions. In this study, a systematic lithogeochemical sampling was performed in the Choghart orebody and the samples were analyzed for the major-oxides and the rare earth elements. To identify the different distribution patterns of REE, the extensive computational experiments were conducted on the dataset with five different classification methods. By comparing these methods in terms of the prediction error, the decision tree was selected as the most suitable method in order to achieve a better understanding of conceptual relationships between the data. The decision tree results showed that, the distribution patterns of most LREE were reasonably predictable based on changes in their respective predictors. The geochemical characteristics are due to either high LREE content in magma source or their modified distributions during alteration and metasomatism. In contrast, the HREE exhibited the irregular distribution patterns and could not be reasonably predicted by the predictors. This behavior may be related to other complex events not affecting the major element composition of rocks considerably. However, some HREE might be concentrated locally in accessory minerals with no evidence of extended enriched zones. A number of factors having the greatest impact on the REE prediction were sorted as: Z (depth), Fe_2O_3, SiO_2, P_2O_5, Na_2O, MnO, Al_2O_3 and CaO. These factors are related to the main geochemical zones known in the Choghart orebody. Moreover, the important geochemical rules were proposed to identify the different concentrations of REE and illustrate their behavior. The recent rise in global prices of REE makes it feasible and profitable to explore giant IOA deposits. Hence, the results of this study could be extended to similar REE geochemical exploration projects and also could result in a considerable savings in both time and cost.

Acknowledgements

The authors express their gratitude to the managing director of the ICIOC (Iranian Central Iron Ore Company), Eng. Askari for his support and assistance. We would like to thank Eng. Rahimipour and Eng. Akhigan provided access to the Choghart Mine, sampling and support during field work, geological data, and other things. We wish to thank Eng. Zahed, Eng. Shekarian, Eng. Najafi, Eng. Ghanbarpour and others for their useful advices that helped our study. The authors would like to thank the editor and anonymous reviewers for their constructive comments that greatly improved the manuscript.

References

Abbaszadeh, M., Hezarkhani, A., Soltani-Mohammadi, S., 2013. An SVM-based machine learning method for the separation of alteration zones in sungun porphyry copper deposit. Chem. Erde-Geochem. 73, 545–554.

Aftabi, A., Mohseni, S., Babeki, A., Azaraien, H., 2009. Fluid inclusion and stable isotope study of the esfordi apatite-magnetite deposit, Central Iran- a discussion. Econ. Geol. 104, 137–139.

Akkaş, E., Akin, L., Çubukçu, H.E., Artuner, H., 2015. Application of decision tree algorithm for classification and identification of natural minerals using SEM–EDS. Comput. Geosci. 80, 38–48.

Alavi, M., 1991. Tectonic Map of the Middle East: Tehran, Scale 1:5,000,000, Geological Survey of Iran.

Astel, A., Michalski, R., Łyko, A., Jabłońska-Czapla, M., Bigus, K., Szopa, S., Kwiecińska, A., 2014. Characterization of bottled mineral waters marketed in Poland using hierarchical cluster analysis. J. Geochem. Explor. 143, 136–145.

Bagheri, S., Stampfli, G.M., 2008. The anarak, jandaq and posht-e-badam metamorphic complexes in Central Iran: new geological data, relationships and tectonic implications. Tectonophysics 451, 123–155.

Baioumy, H.M., Ahmed, A.H., Khedr, M.Z., 2014. A mixed hydrogenous and hydrothermal origin of the bahariya iron ores, Egypt: evidences from the trace and rare earth element geochemistry. J. Geochem. Explor. 146, 149–162.

Berger, A., Janots, E., Gnos, E., Frei, R., Bernier, F., 2014. Rare earth element mineralogy and geochemistry in a laterite profile from Madagascar. Appl. Geochem. 41, 218–228.

Bishop, C.M., 2006. Pattern recognition and machine learning. Springer, New York.

Bonyadi, Z., Davidson, G.J., Mehrabi, B., Meffre, S., Ghazban, F., 2011. Significance of apatite REE depletion and monazite inclusions in the brecciated Se–chahun iron oxide–apatite deposit, bafq district, Iran: insights from paragenesis and geochemistry. Chem. Geol. 281, 253–269.

Borrok, D.M., Kelser, S.E., Boer, R.H., Essene, E.J., 1998. The vergenoeg magnetite-fluorite deposit, South Africa; support for a hydrothermal model for massive iron oxide deposits. Econ. Geol. 93, 564–586.

Chasmer, L., Hopkinson, C., Veness, T., Quinton, W., Baltzer, J., 2014. A decision-tree classification for low-lying complex land cover types within the zone of discontinuous permafrost. Remote Sens. Environ. 143, 73–84.

Cole, C.S., James, R.H., Connelly, D.P., Hathorne, E.C., 2014. Rare earth elements as indicators of hydrothermal processes within the East Scotia subduction zone system. Geochim. Cosmochim. Acta 140, 20–38.

Cox, R.A., Wang, G.W.-Y., 2014. Predicting the US bank failure: A discriminant analysis. Econ. Anal. Policy 44, 202–211.

Daliran, F., 1990. The Magnetite-Apatite Deposit of Mishdovan, East Central Iran: an Alkali Rhyolite Hosted," Kiruna Type" Occurrence in the Infracambrian Bafg Metallotect (Mineralogic, Petrographic and Geochemical Study of the Ores and the Host Rocks). Ruprecht-Karls-Universität.

Daliran, F., 2002. Kiruna-type iron oxide-apatite ores and apatitites of the bafq district, Iran, with an emphasis on the REE geochemistry of their apatites. Hydrotherm. Iron Oxide Copper-gold Relat. Depos. 2, 303–320.

Daliran, F., Stosch, H.-G., Williams, P.J., 2009. A review of the Early Cambrian magmatic and metasomatic events and their bearing on the genesis of the Fe oxide-REE-apetite deposits (IOA) of the Bafq district, Iran. In: Williams, et al. (Eds.), Smart Science for Exploration and Mining: Proceedings of the 10th Biennial SGA Meeting, Townsville, Australia 17th-20th August 2009.

Daliran, F., Stosch, H., Williams, P., Jamali, H., Dorri, M., 2010. Early Cambrian iron oxide-apatite-REE (U) deposits of the Bafq District, east-Central Iran. exploring for iron oxide copper–gold deposits: Canada and global analogues. Geol. Assoc. Can. Short Course Notes 20, 143–155.

Dougherty, G., 2013. Pattern Recognition and Classification: an Introduction. Springer Science & Business Media.

Duda, R.O., Hart, P.E., Stork, D.G., 2012. Pattern Classification. John Wiley & Sons.

Edfelt, A., 2007. The Tjårrojåkka Apatite-Iron and Cu (−Au) Deposits, Northern Sweden-Products of one Ore Forming Event. Luleå University of Technology.

Eggenkamp, H.G.M., Marques, J.M., 2013. A comparison of mineral water classification techniques: occurrence and distribution of different water types in Portugal (including Madeira and the Azores). J. Geochem. Explor. 132, 125–139.

Ehya, F., 2012. Variation of mineralizing fluids and fractionation of REE during the emplacement of the vein-type fluorite deposit at bozijan, Markazi Province, Iran. J. Geochem. Explor. 112, 93–106.

Emsbo, P., McLaughlin, P.I., Breit, G.N., du Bray, E.A., Koenig, A.E., 2015. Rare earth elements in sedimentary phosphate deposits: solution to the global REE crisis? Gondwana Res. 27, 776–785.

Farid, D.M., Zhang, L., Rahman, C.M., Hossain, M., Strachan, R., 2014. Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks. Expert Syst. Appl. 41, 1937–1946.

Fisher, R., 1936. The use of multiple measurements in taxonomic problems. Annual Eugenics 7 (part II). John Wiley, NY, pp. 179–188 (Reprinted in Contributions to Mathematical statistics, 1950).

Foerster, H., Jafarzadeh, A., 1994. The bafq mining district in Central Iran; a highly mineralized infracambrian volcanic field. Econ. Geol. 89, 1697–1721.

Foley, N.K., Ayuso, R.A., 2013. Rare earth element mobility in high-alumina altered metavolcanic deposits, South Carolina, USA. J. Geochem. Explor. 133, 50–67.

Frietsch, R., 1982. On the chemical composition of the ore breccia at luossavaara, Northern Sweden. Mineral. Deposita 17, 239–243.

Frietsch, R., Perdahl, J.-A., 1995. Rare earth elements in apatite and magnetite in Kiruna-type iron ores and some other iron ore types. Ore Geol. Rev. 9, 489–510.

Fukunaga, K., 1990. Introduction to Statistical Pattern Recognition. Academic press.

Ghanbari, Y., Hezarkhani, A., Ataei, M., Pazand, K., 2014. Mineral potential mapping for rare earth elements mineralization with AHP method in the Kerman-kashmar tectonic zone, Central Iran. J. Geol. Soc. India 83, 457–465.

Groves, D.I., Bierlein, F.P., Meinert, L.D., Hitzman, M.W., 2010. Iron oxide copper-gold (IOCG) deposits through earth history: implications for origin, lithospheric setting, and distinction from other epigenetic iron oxide deposits. Econ. Geol. 105, 641–654.

Han, J., Kamber, M., Pei, J., 2012. Data Mining, Southeast Asia Edition: Concepts and Techniques. Morgan kaufmann.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning- Data Mining, Inference, and Prediction. second ed. Springer-Verlag, New York.

Henderson, P., 1984. Rare earth element geochemistry. In: Henderson, P. (Ed.), Developments in Geochemistry. Elsevier.

Henderson, P., 1996. Rare Earth Minerals: Chemistry, Origin and Ore Deposits. Springer Science & Business Media.

Hernández-González, J., Inza, I., Lozano, J.A., 2013. Learning bayesian network classifiers from label proportions. Pattern Recogn. 46, 3425–3440.

Hildebrand, R.S., 1986. Kiruna-type deposits; their origin and relationship to intermediate subvolcanic plutons in the great bear magmatic zone, Northwest Canada. Econ. Geol. 81, 640–659.

Huang, Y., Guan, Y., 2015. On the linear discriminant analysis for large number of classes. Eng. Appl. Artif. Intell. 43, 15–26.

Imani, M., Ghassemian, H., 2015. Feature space discriminant analysis for hyperspectral data feature reduction. ISPRS J. Photogramm. Remote Sens. 102, 1–13.

Jaireth, S., Hoatson, D.M., Miezitis, Y., 2014. Geological setting and resources of the major rare-earth-element deposits in Australia. Ore Geol. Rev. 62, 72–128.

Jami, M., Dunlop, A.C., Cohen, D.R., 2007. Fluid inclusion and stable isotope study of the esfordi apatite-magnetite deposit, Central Iran. Econ. Geol. 102, 1111–1128.

Jami, M., Dunlop, A.C., Cohen, D.R., 2009. Fluid inclusion and stable isotope study of the esfordi apatite-magnetite deposit, Central Iran—A reply. Econ. Geol. 104, 140–143.

Jha, A., 2014. Rare Earth Materials: Properties and Applications. CRC Press.

Küpeli, Ş., 2010. Trace and rare-earth element behaviors during alteration and mineralization in the attepe iron deposits (Feke-Adana, Southern Turkey). J. Geochem. Explor. 105, 51–74.

Li, Y., Anderson-Sprecher, R., 2006. Facies identification from well logs: A comparison of discriminant analysis and naïve Bayes classifier. J. Pet. Sci. Eng. 53, 149–157.

Li, J., Wang, Y., Xie, X., Su, C., 2012. Hierarchical cluster analysis of arsenic and fluoride enrichments in groundwater from the Datong basin, Northern China. J. Geochem. Explor. 118, 77–89.

Li, X., Chan, C.W., Nguyen, H.H., 2013. Application of the neural decision tree approach for prediction of petroleum production. J. Pet. Sci. Eng. 104, 11–16.

Long, K., Van Gosen, B., Foley, N., Cordier, D., 2012. The principal rare earth elements deposits of the United States: A summary of domestic deposits and a global perspective. In: Sinding-Larsen, R., Wellmer, F.-W. (Eds.), Non-Renewable Resource Issues. Springer, Netherlands, pp. 131–155.

Lourenço, C., Ribeiro, L., Cruz, J., 2010. Classification of natural mineral and spring bottled waters of Portugal using principal component analysis. J. Geochem. Explor. 107, 362–372.

Massari, S., Ruberti, M., 2013. Rare earth elements as critical raw materials: focus on international markets and future strategies. Res. Policy 38, 36–43.

Moore, F., Modabberi, S., 2003. Origin of choghart iron oxide deposit, bafq mining district, Central Iran: new isotopic and geochemical evidence. J. Sci. Islam. Rep. Iran 14, 259–270.

Nazarpour, A., Sadeghi, B., Sadeghi, M., 2015. Application of fractal models to characterization and evaluation of vertical distribution of geochemical data in zarshuran gold deposit, NW Iran. J. Geochem. Explor. 148, 60–70.

Ni, Y., Hughes, J.M., Mariano, A.N., 1995. Crystal chemistry of the monazite and xenotime structures. Am. Mineral. 80, 21–26.

Nystroem, J.O., Henriquez, F., 1994. Magmatic features of iron ores of the Kiruna type in Chile and Sweden; ore textures and magnetite geochemistry. Econ. Geol. 89, 820–839.

Oreskes, N., Einaudi, M.T., 1990. Origin of rare earth element-enriched hematite breccias at the Olympic dam Cu-U-Au-Ag deposit, Roxby downs, South Australia. Econ. Geol. 85, 1–28.

Pal, M., Mather, P.M., 2003. An assessment of the effectiveness of decision tree methods for land cover classification. Remote Sens. Environ. 86, 554–565.

Peh, Z., Halamić, J., 2010. Discriminant function model as a tool for classification of stratigraphically undefined radiolarian cherts in ophiolite zones. J. Geochem. Explor. 107, 30–38.

Pradhan, B., 2013. A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS. Comput. Geosci. 51, 350–365.

Quinlan, J.R., 1993. C4. 5: Programs for Machine Learning. Morgan Kaufmann.

Rajabi, A., Rastad, E., Alfonso, P., Canet, C., 2012. Geology, ore facies and sulphur isotopes of the koushk vent-proximal sedimentary-exhalative deposit, posht-e-badam block, Central Iran. Int. Geol. Rev. 54, 1635–1648.

Rajabi, A., Canet, C., Rastad, E., Alfonso, P., 2015. Basin evolution and stratigraphic correlation of sedimentary-exhalative Zn–Pb deposits of the early Cambrian zarigan–Chahmir Basin, Central Iran. Ore Geol. Rev. 64, 328–353.

Ramezani, J., Tucker, R.D., 2003. The saghand region, Central Iran: U–Pb geochronology, petrogenesis and implications for gondwana tectonics. Am. J. Sci. 303, 622–665.

Rollinson, H.R., 1993. Using Geochemical Data: Evaluation, Presentation, Interpretation. Longman, Harlow, p. 352.

Sabet-Mobarhan-Talab, A., Alinia, F., Ghannadpour, S.-S., Hezarkhani, A., 2015. Geology, geochemistry, and some genetic discussion of the chador-malu iron oxide-apatite deposit, Bafq District, Central Iran. Arab. J. Geosci. 1–20.

Sadeghi, M., Morris, G.A., Carranza, E.J.M., Ladenberger, A., Andersson, M., 2013. Rare earth element distribution and mineralization in Sweden: An application of principal component analysis to FOREGS soil geochemistry. J. Geochem. Explor. 133, 160–175.
Sadeghi, B., Madani, N., Carranza, E.J.M., 2015. Combination of geostatistical simulation and fractal modeling for mineral resource classification. J. Geochem. Explor. 149, 59–73.
Samani, B.A., 1988. Metallogeny of the Precambrian in Iran. Precambrian Res. 39, 85–106.
Schacht, U., Wallmann, K., Kutterolf, S., 2010. The influence of volcanic ash alteration on the REE composition of marine pore waters. J. Geochem. Explor. 106, 176–187.
Shi, G., 2014. Chapter 5 - decision trees. In: Shi, G. (Ed.), Data Mining and Knowledge Discovery for Geoscientists. Elsevier, Oxford, pp. 111–138.
Shikazono, N., Ogawa, Y., Utada, M., Ishiyama, D., Mizuta, T., Ishikawa, N., Kubota, Y., 2008. Geochemical behavior of rare earth elements in hydrothermally altered rocks of the Kuroko mining area, Japan. J. Geochem. Explor. 98, 65–79.
Simandl, G., 2014. Geology and market-dependent significance of rare earth element resources. Mineral. Deposita 49, 889–904.
Stegen, K.S., 2015. Heavy rare earths, permanent magnets, and renewable energies: An imminent crisis. Energy Policy 79, 1–8.
Stosch, H.-G., Romer, R.L., Daliran, F., Rhede, D., 2011. Uranium–lead ages of apatite from iron oxide ores of the Bafq District, East-Central Iran. Mineral. Deposita 46, 9–21.
Tabachnick, B., Fidell, L., 1996. Using Multivariate Statistics. Harper Collins College Publishers, New York.
Tahmasebi, P., Hezarkhani, A., 2012a. A fast and independent architecture of artificial neural network for permeability prediction. J. Pet. Sci. Eng. 86–87, 118–126.
Tahmasebi, P., Hezarkhani, A., 2012b. A hybrid neural networks-fuzzy logic-genetic algorithm for grade estimation. Comput. Geosci. 42, 18–27.
Torab, F., Lehmann, B., 2007. Magnetite-apatite deposits of the bafq district, Central Iran: apatite geochemistry and monazite geochronology. Mineral. Mag. 71, 347–363.
Tsay, A., Zajacz, Z., Sanchez-Valle, C., 2014. Efficient mobilization and fractionation of rare-earth elements by aqueous fluids upon slab dehydration. Earth Planet. Sci. Lett. 398, 101–112.
Webb, G.I., Boughton, J.R., Zheng, F., Ting, K.M., Salem, H., 2012. Learning by extrapolation from marginal to full-multivariate probability distributions: decreasingly naive Bayesian classification. Mach. Learn. 86, 233–272.
Wu, J., Pan, S., Zhu, X., Cai, Z., Zhang, P., Zhang, C., 2015. Self-adaptive attribute weighting for naive Bayes classification. Expert Syst. Appl. 42, 1487–1502.
Xue, D., Pang, F., Meng, F., Wang, Z., Wu, W., 2015. Decision-tree-model identification of nitrate pollution activities in groundwater: A combination of a dual isotope approach and chemical ions. J. Contam. Hydrol. 180, 25–33.
Zadrozny, B., Langford, J., Abe, N., 2003. Cost-sensitive learning by cost-proportionate example weighting, Data Mining, 2003. ICDM 2003. Third IEEE International Conference on. IEEE, pp. 435–442.
Zahed, A., Seidy, M., 2012. Coghart mine geology map (Unpublished Report). Iran Central Iron Ore Co (ICIOC).