Available online at www.sciencedirect.com

ScienceDirect

Expert Systems with Applications 36 (2009) 102–110

Expert Systems with Applications

www.elsevier.com/locate/eswa

Application of rough set and decision tree for characterization of premonitory factors of low seismic activity

Iftikhar U. Sikder *, Toshinori Munakata [1]

Department of Computer and Information Science, Cleveland State University, Cleveland, OH 44115, USA

Abstract

This paper presents a machine learning approach to characterizing premonitory factors of earthquake. The characteristic asymmetric distribution of seismic events and sampling limitations make it difficult to apply the conventional statistical predictive techniques. The paper shows that inductive machine learning techniques such as rough set theory and decision tree (C4.5 algorithm) allows developing knowledge representation structure of seismic activity in term of meaningful decision rules involving premonitory descriptors such as space–time distribution of radon concentration and environmental variables. The both techniques identify significant premonitory variables and rank attributes using information theoretic measures, e.g., entropy and frequency of occurrence in reducts. The cross-validation based on "leave-one-out" method shows that although the overall predictive and discriminatory performance of decision tree is to some extent better than rough set, the difference is not statistically significant.

Keywords: Rough set theory; Decision tree; Earthquake prediction; Machine learning

1. Introduction

Identifying the premonitory factors for earthquake is a difficult task (Biagi, Ermini, Kingsley, Khatkevich, & Gordeev, 2001). Classical statistical techniques fall short of complying with stringent constraints and implicit assumptions to be used in predicting premonitory factors for earthquake. In recent years, researchers have identified the need for the implementation of advanced statistical methods in earthquake data evaluation (Cuomo et al., 2000). Firstly, this is due to the fact that high seismic activity is relatively a rare phenomenon; therefore, it is difficult to collect statistically significant number of samples to derive conclusive prediction. Secondly, the classical statistical methods often include *a priori* assumptions on the sample distribution, e.g., discriminant analysis requires assumptions of within group variances (Browne, Duntsch, & Gediga, 1998) – a difficult condition to be satisfied for seismic classes. It also requires assumptions of normal distribution of continuous attributes and comparable number of objects in decision classes. It should be noted that frequently used time series methods for earthquake prediction such as AR model or GRACH model also require stationarity assumption (Mega et al., 2003; Telesca, Lapenna, & Macchiato, 2005) of condition variable. Thirdly, it is often difficult to interpret the output of the statistical results. For example, discriminant analysis generates the final result in a form of discriminant functions, which aggregate the input information in a non-transparent way (Flinkman et al., 2000).

In recent years, a wide variety of machine learning and knowledge discovery techniques have been used for rule induction in many different disciplines including environmental science (Dmeroski, 2002). These techniques include commonly used data mining tools such as neural network (Fu, 1999), decision tree (Quinlan, 1992), and rough sets (Pawlak & Slowinski, 1994). A major advantage of using such techniques is that they are mostly data driven, non-parametric and less restrictive in *a priori* assumptions.

* Corresponding author. Tel.: +1 216 687 4758; fax: +1 216 687 5448. *E-mail addresses:* i.sikder@csuohio.edu (I.U. Sikder).

[1] Tel.: +1 216 687 3684; fax: +1 216 687 5448, E-mail address: munakata@grail.cba.csuohio.edu.

0957-4174/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2007.09.032

For example, decision tree is better suited for non-normal and non-homogeneous dataset. The rough set is usually preferred for its "non-invasive" approach because it does not require any distribution assumption (Düntsch & Gediga, 1998). Unlike fuzzy set theory or statistical analysis, a unique advantage of rough set is that it does not rely on other model assumption or external parameter. It solely utilizes the structure of the given data.

Given asymmetrical distribution of the seismic levels, the rough set and decision tree approaches are the appropriate tools because of their non-parametric stance. Rough set and decision tree are often considered to have better knowledge representation structure in term of deriving meaningful decision rules (Daubie, Levecq, & Meskens, 2002). The extracted rules are easily interpretable allowing complex relationships to be represented in an intuitive and comprehensible manner. The rules establish a relationship between descriptions of objects by attributes and their assignment to specific class. Moreover, the rules can be used for the classification of new objects (Krusinska, Slowinski, & Stefanowski, 1992). Rough set and decision tree eliminates superfluous or redundant attributes to determine significant attributes for classification. It has been shown that under special circumstances, when the distribution of objects in the boundary region of rough set is equally probable, the criteria for selecting dominant attributes is a special case of ID3 (Wong, Ziarko, & Ye, 1986). Rough set has been compared with other techniques. This include comparison of performance of rough set and discriminant analysis (Browne et al., 1998; Dimitras, Slowinski, Susmaga, & Zopounidis, 1999; Krusinska et al., 1992), logistic regression (Dimitras et al., 1999), neural network (Jelonek, Krawiec, & Slowinski, 1995; Szczuka, 1998), ordinal statistical methods (Teghem & Benjelloun, 1992) and other statistical methods (Wong et al., 1986). Rough set has also been compared with decision tree. The comprehensive comparison of rough set and decision tree (ID3) can be found in (Beynon & Peel, 2001; Daubie et al., 2002; Mak & Munakata, 2002).

Among premonitory factors, variation of radon concentration and other terrestrial gases are considered important indicators associated with seismic events (Belayaev, 2001; Fleischer & Mogro-Campero, 1981; King, 1985; Magro-Campero, Fleischer, & Likes, 1980; Takahashi, 2003). Traditionally, regression methods have been used to predict radon concentration in soil gas on the basis of environmental data. Zmazek, Todorovski, Dzeroski, Vaupotic, and Kobal (2003) reports that model trees (MT) – a variant of decision tree outperform other regression methods like linear regression (LR) and instance based regression (IB)(Aha & Kibler, 1991) in predicting radon concentration from meteorological data. Teghem and Charlet (1992) searched for premonitory factors for earthquakes by emphasizing radon concentration and gas geochemistry by using rough set model. Using the data set provided by Teghem and Benjelloun (1992), Düntsch and Gediga (1997) established a procedure to evaluate the validity of prediction based on the approximation quality of attributes of rough set dependency analysis. Based on the results of their statistical evaluation procedure the casualness of a prediction can be found to ensure that the prediction is not based on only a few (casual) observations.

While a large body of literature exists on identifying premonitory factors for earthquake and the application of inductive machine learning techniques, surprisingly the integration of the two approaches has been inadequately addressed. In this paper we investigate the overall performance of rough set and decision tree to characterize the premonitory factors responsible for low seismic activity. In particular, we investigate the predictive and discriminatory performance of the two models. We used the data set reported in (Teghem & Charlet, 1992) which includes weekly measure of low seismic activity and associated radon concentration measure and climatic condition at different location in Belgium. The classical rough set theory and C4.5 – a modified version of ID3 for decision tree building algorithm were used to compare the overall performance in the context of earthquake prediction.

2. The rough set methods

The main idea that seismic activity can be modeled by means of classical rough set theory presupposes the granularity of observations and the resulting indiscernibility relations. Introduced by Pawlak and Slowinski (1994), rough sets is relatively new machine learning technique to deal with inexact, uncertain or vague knowledge. Rough can be used when the sample size is small and the distribution of the data deviates significantly from the multivariate normal distribution (Stefanowski, 1992). Rough set was introduced to characterize a concept in terms of elementary sets in an approximation space. The underlying assumption in rough set is the notion that information can be associated with every object in the universe. A set of object are indiscernible or indistinguishable from each other based on their attribute value or characteristic behavior. The concept of indiscernibility or information granulation is at the heart of rough set theory. A finer granulation means more definable concept. Granularity is expressed by partition and their associated equivalence relations on the sets of objects, also called *indiscernibility relations*. Indiscernibility leads to the concept of boundary-line cases, which means that some elements can be classified to the concepts or its complements with the available information, and thus forms the boundary-line cases. The object that belongs to a set with certainty is called *lower approximation* while *upper approximation* contains all objects that may possibly belong to the set (see Fig. 1).

The knowledge system in rough set can be represented in the form $(U, C \cup \{d\})$, where $d \notin C$ is the earthquake intensity class which, for example may represents the decision to change or unchanged land use in a given location and U is the closed universe which consists of non-empty finite set of objects and C is a non-empty finite set of attributes such that $c : U \to V_c$ for every $c \in C, V_c$ is a value of attribute

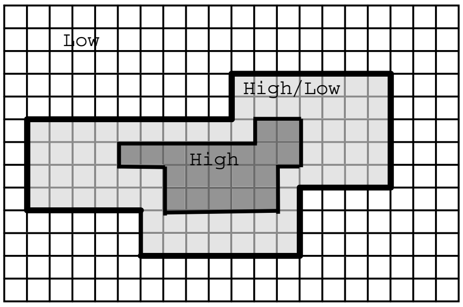

Fig. 1. Approximating the set of seismic activity class generated by indiscernibility relation. The lower approximation set (represented by dark gray area) indicates observations that can be classified, with certainty, as the high seismic activity on basis of attribute knowledge. The boundary region represents the uncertain cases where prediction is ambiguous.

c. For $P \subseteq C$ the granule of knowledge about a decision with respect to indiscernibility relation can be represented as

$$\mathrm{Ind}(P) = \{(x, x') \in U^2 | \forall c \in Pc(x) = c(x')\} \tag{1}$$

Thus, the instances x and x' are indiscernible from each other if $(x, x') \in \mathrm{Ind}(P)$ relations and decision about changing or unchanging a given location is approximated by lower and upper approximation of decision concept as follows

$$P\underline{X} = \{x \in U | \mathrm{Ind}(x) \subseteq X\} \tag{2}$$

$$P\overline{X} = \{x \in U | \mathrm{Ind}(x) \cap X \neq \emptyset\} \tag{3}$$

An effort to mapping rough set decision classes with earthquake intensity class induces two important partitions in a sample space, one that is induced by the set of all conditioning attributes (consisting of atoms) and another the partition induced by the earthquake intensity class. Hence, the earthquake prediction can be formally stated as

$$f : C \rightarrow R$$

Here C, the domain of f, is the conditional attributes or premonitory factors. The set R consists of different degrees or levels of earthquake classes. The overall objective is to determine the appropriate f from a given real world subset of information Train $\subset C \times R$ The goal of application of rough set theory is to generate set of universal decision rules in the form of if $a_{i1} = v_1$ and $a_{i2} = v_2$ and…then decision $= d$. Using such rules it is possible to classify new, unclassified instances. Given a training set, each object can be used to derive rule from its characteristic attribute values. However, since such rule would be very specific and could only be applied to that specific object, therefore, data mining techniques strive to produce universal or minimal covering decision rules that would have small number of descriptors so that they could be used to classify of many different objects. The advantage of rough set is that it is inherently data driven and "non-invasive" (Düntsch & Gediga, 1998). Unlike fuzzy set theory or statistical analysis, a unique advantage of rough set is that it does not rely on other model assumptions or external parameters. It utilizes the structure of the given data. The numerical value of imprecision or membership function is not required to be inferred, rather approximations are derived objectively from the given training set itself.

The data representation in the rough set framework is achieved by an information system, which is a pair $\mathrm{A} = (U, A)$, where U is a non-empty, finite set called the universe and A – a non-empty, finite set of attributes, i.e., $a : U \rightarrow V_a$ for $a \in A$, where V_a is the *value set* of attribute a. A *decision table* is an information system of the form $\mathrm{A} = (U, A \cup \{d\})$, where $d \notin A$ is the *decision attribute* or class labels where we can assume a set V_d of values of the decision d is equal to $\{1, \ldots, r(d)\}$. Decision d determines the partition $\{X_1, \ldots, X_{r(d)}\}$ of the universe U, where $\mathrm{X}_k = \{x \in U : d(x) = k\}$ for $1 \leqslant k \leqslant r(d)$.This system can be generalized as the decision system $\mathrm{A} = U, A, d_{\mathrm{D}}(x)$, where $d_{\mathrm{D}}(x) = (d_1(x), \ldots, d_k(x))$ for $x \in U$ (Polkowski & Skowron, 1998). Such decision table is equivalent to the training sample in machine learning used to induce concept approximation in machine learning (Mitchell, 1997).

For an information system $\mathrm{A} = (U, A)$, any $B \subseteq A$ is associated with an equivalence relation $\mathrm{IND}_A(B)$ (also called as *B-indiscernibility relation*, its classes are denoted by $[x]_B$.) defined by: $\mathrm{IND}(B) = \{(x, x') \in U^2 : \text{for every } a \in B,\ a(x) = a(x')\}$. Objects x, x' are indiscernible by attributes B. Given a set of attributes $B \subseteq A$ and $X \subseteq U$ a set of objects, we can approximate X by constructing the *B-lower* and *B-upper approximations* of X, $\underline{B}X$ and $\overline{B}X$, respectively, where $\{\underline{B}X = \{x \in U : [x]_B \subseteq X\}$ and $\{\overline{B}X \in U : [x]_B \cap X \neq \emptyset\}$. The set $BN_B(X) = \overline{B}X - \underline{B}X$ represents the *B-boundary* of X. The accuracy of approximation is measured by $\alpha_B = \frac{|\underline{B}(X)|}{|\overline{B}(X)|}$, where $0 \leqslant \alpha_B \leqslant 1$. A set is rough if $\alpha_B(X) < 1$ (i.e., X is vague with respect to B).

In addition to identifying indiscernibility relationship and equivalence classes, classical machine learning equivalent to feature reduction in rough set is the concept of *reduct*. Reducts allows one to decide whether some of the attributes are redundant or superfluous with respect to decision class. Hence, reducts are all the subset of attributes that are minimal, i.e., that do not include any dispensable attribute. Extraction of reducts requires construction of $n \times n$ matrix (c_{ij}), called the *discernibility matrix*of an information system such that $c_{ij} = \{a \in \mathrm{A} : a(x_i) \neq a(x_j)\}$ for $i, j = 1, \ldots, n$. While, a *discernibility function* f_{A} for an information system is a boolean function defined by

$$f_A(\bar{a}_1, \ldots, \bar{a}_m) = \wedge\{\vee_{\bar{c}_{ij}} : 1 \leqslant j < i \leqslant n, c_{ij} \neq \emptyset\}$$

where $\bar{c}_{ij} = \{\bar{a} : a \in c_{ij}\}$. It has been shown that the set of all prime implicants of f_{A} determines the set of all *reducts* (Skowron & Rauszer, 1992). Computing all possible reducts is a non-trivial task. While computing prime implicants is an NP-Hard (Skowron & Grzymalla-Busse, 1994) problem, it is possible to use heuristic algorithm (e.g., genetic algorithm) (Wroblewski, 1995) or dynamic reducts (Bazan, Skowron, & Synak, 1994) to generate a computationally efficient set of minimal attributes.

From machine learning perspective, decision rules derived from lower approximation represents certain rules while rules extracted from upper approximation corresponds to possible rules. Since the objective is to derive *minimal* decision rule, reduct serves this purpose by providing minimal number of descriptor in the conditional part. Once the reduct have been computed deriving decision rule is a simple task of laying the reducts over the original decision table and mapping the associated values. Such rules derived from the training set can be used to classify new instances for which the decision classes are unknown. However, it is likely that more than one rule may fire to decide a class for a new object. In that case strategies are to be adopted (e.g., standard voting) to resolve conflict among candidate rules recognizing same object (Polkowski & Skowron, 1998). Instead of using a fixed strategy, it has been shown that rough set methods can be used to learn from data the strategy for conflict resolving between decision rules (Szczuka, 1999).

3. Dataset description

We used the dataset reported in Teghem and Benjelloun (1992) which includes time series data involving radon concentrations in soils for eight point of measurements with different geo-climatic environment. The data set includes premonitory factors for earthquakes consisting of seismic activity on 155 records of weekly measures of the Richter scale and associated radon concentration measured at eight different locations (attributes C1–C8) and seven measures of climatic factors (attributes C9–C15). At each site a mean of measured rates of radon emanation is estimated. C1 and C2 belong to the first set of sites, while site C3–C8 belong to the second one. The condition attribute variables – the radon concentration at different sites and climatic variables are classified into nominal category from numeric values using histogram binning algorithm under normal distribution assumption $N(\mu,\sigma)$. Assuming normal distribution $N(\mu,\sigma)$ for each site, the data was discretized as follows:

$$N(-\infty,\mu-\sigma[;]\mu-\sigma,\mu-\sigma/3[;]\mu-\sigma/3,\mu+\sigma/3[;]\mu + \sigma/3,\mu+\sigma[;]\mu+\sigma,+\infty)$$

numbered from 1 to 5, respectively. The conditional attributes C9–C15 correspond to climatic attributes atmospheric pressure (C9), sun period (C10), air temperature (C11), relative humidity (C12), and rainfall C13. The decision table consists of two different levels of seismic activity, where $D^{(1)} = f(x,D) = R \leqslant 1.5$ and $D^{(2)} = f(x,D) = R > 1.5$. The cardinality $|D^{(2)}| = 8$ and $|D^{(1)}| = 147$ show a large asymmetry between the two equivalence classes $D^{(1)}$ and $D^{(2)}$.

4. Application of rough set

4.1. Representation of decision table

The first step in rough set application is the development of decision table. As discussed previously, we have used the dataset reported in (Teghem & Charlet, 1992). The decision table includes 155 objects or samples reflecting periodic measure of Radon and climatic variables as well as seismic measure. For each record 15 conditional attributes are registered. The decision variable represents two different degree of seismic levels, where $D^{(1)} = f(x,D) = R \leqslant 1.5$ and $D^{(2)} = f(x,D) = R > 1.5$. Since the conditional and decision attributes are in nominal scale discretization was not necessary.

4.2. Approximation of decision space

In the second step, approximation of object's classification is evaluated. This includes construction of approximation of each decision class with respect to all the condition attributes. Since we are more interested in predicting $D^{(2)}$ rather than $D^{(1)}$ it is sufficient to analyze the former case. The analysis shows that when all the attributes are used the lower approximations $\{\underline{B}D^2 = \{x \in U : [x]_B \subseteq D^2\}$ is the identical set as the upper approximation $\overline{B}D^2 = \{x \in U : [x]_B \subseteq X \neq \emptyset\}$ Therefore, the set is B-definable as $\underline{B}D^2 = \overline{B}D^2$. The quality of approximation, accuracy, and entropy measures are equal to 1.

4.3. Reduction of attributes

The extraction of reducts from data involves construction of minimal subset of attributes ensuring the same quality of sorting as that of all attributes. Using discernibility matrix algorithm (Skowron & Rauszer, 1992) for searching reducts 440 reducts were found with cardinality ranging from 4 to 7. The intersection of all reducts or *core* was found null, i.e., there exist no common attribute or indispensable attribute. We used a methodology proposed in (Slowinski, Zopounidis, & Dimtras, 1997) to manually evaluate reducts. Table 1 illustrates some examples of reducts obtained. The second column shows the predicted loss on quality of classification if such attribute is removed.

Table 2 shows the frequency of individual attribute occurring in all 440 reducts. Evidently, there is no common attribute occurring in all the reducts. The most frequent attribute is C2 (radon concentration at site 2) occurring 224 times (i.e., 50.91%) among the set of reducts.

It should be noted that the total occurrence of the environmental attributes in the reducts is 6.8% more than the attributes related to radon concentration. This could be

Table 1
Some examples of reducts and the predicted quality loss when attributes are removed (for all reducts the quality of sorting = 1)

Reducts	Quality loss
{C2, C5, C6, C7, C9}	{0.026, 0.058, 0.013, 0.013, 0.026}
{C2, C3, C5, C7, C9}	{0.013, 0.013, 0.058, 0.039, 0.013}
{C1, C2, C4, C7, C8}	{0.026, 0.071, 0.019, 0.039, 0.026}
{C2, C5, C7, C9, C11}	{0.026, 0.039, 0.013, 0.032, 0.013}
{C1, C2, C4, C5}	{0.052, 0.142, 0.052, 0.084}

Table 2
Condition attributes and their frequency in reducts

Attributes in reducts	%Frequency
C2	50.91
C11	47.50
C9	42.73
C10	42.50
C12	37.95
C8	37.50
C5	34.32
C7	32.50
C6	31.36
C1	30.45
C13	29.77
C3	29.77
C4	27.27
C14	25.00
C15	24.77

explained by the fact that majority of objects in the decision table is related to low seismic activity, i.e., $D^{(1)}$. As the number of objects related to $D^{(2)}$ is extremely few, the influence of radon concentration alone is insufficient to produce major change in the quality of sorting, even though radon concentration may have strong association with $D^{(2)}$.

4.4. Decision rules

Rule extraction is a relatively straightforward procedure. Reducts are used to generate decision rules from the decision table. The objective is to generate basic minimal covering rules or minimal number of possibly shortest rules covering all the cases. The classical LEM2 algorithm (Grzymala-Busse & Than, 1992; Stefanowski, 1998) was used to derive minimal set of rules covering all the objects from learning set. The algorithm generates 15 minimum covering rules. Four rules are generated that predict class $D^{(2)}$. These rules are

Rule 1: $(\mathrm{C1}=2)$ & $(\mathrm{C2}=4)$ & $(\mathrm{C5}=2)$ & (C10 in $\{2,1\}) \Rightarrow (\mathrm{Class}=D^{(2)})$ [4, 4, 50.00%, 100.00%]

Rule 2: $(\mathrm{C1}=1)$ & $(\mathrm{C2}=4)$ & $(\mathrm{C5}=5) \Rightarrow (\mathrm{Class}=D^{(2)})$ [2, 2, 25.00%, 100.00%]

Rule 3: $(\mathrm{C2}=5)$ & $(\mathrm{C9}=2)$ & $(\mathrm{C13}=1)$ & $(\mathrm{C15}=1) \Rightarrow (\mathrm{Class}=D^{(2)})$ [1, 1, 12.50%, 100.00%]

Rule 4: $(\mathrm{C2}=3)$ & $(\mathrm{C8}=1)$ & $(\mathrm{C11}=5) \Rightarrow (\mathrm{Class}=D^{(2)})$ [1, 1, 12.50%, 100.00%]

The support and strength for rule 1 is 4 and the relative strength and level of discrimination is 50% and 100%, respectively. The level of discrimination indicates the ratio of the number of covered positive objects to the number of all objects covered by the rule. The remaining 11 rules predict class $D^{(1)}$. Some examples of such rules are

Rule 5: (C3 in $\{1,5\}) \Rightarrow (\mathrm{Class}=D^{(1)})$ [37, 37, 25.17%, 100.00%]

Rule 6: (C2 in $\{2,1\}) \Rightarrow (\mathrm{Class}=D^{(1)})$ [71, 71, 48.30%, 100.00%]

Rule7: $(\mathrm{C3}=3)$ & (C11 in $\{3,2,5\}) \Rightarrow (\mathrm{Class}=D^{(1)})$ [28, 28, 19.05%, 100.00%]

Since the quality of approximation is equal to unity, there are no inconsistencies in the rules. All the rules extracted are deterministic. However, the quality of those rules (in terms of support and strength) which predicts class $D^{(1)}$ is higher than $D^{(2)}$. It should be noted that the average number of descriptors in the rules required for predicting class $D^{(2)}$ is almost twice as the number required for class $D^{(1)}$.

4.5. Validation and implementation

The predictive performance of the rules derived is tested for new instances using cross-validation method. As the number of objects related to $D^{(2)}$ is significantly fewer, we used "*Leave-one-out*" cross-validation method, because of its advantage of using maximum possible number of training instances for learning phase. In this method each instance in the data set is left out in turn and the remaining instances are used for training the tree. The correctness of prediction is judged with respect to the remaining instance. This process is repeated for possible sample instance. The final error estimate is measured by the averaging each iteration result. This method of cross-validation is particularly appropriate for the given dataset for several reasons. First, it allows using maximum possible number of training instances for learning tree. Secondly, since the number instances of decision variable reflecting high seismic activity is very small, 10-fold stratified cross-validation or percentage split method would not have produced sufficient number of training set of high seismic activity for the learning tree. Thirdly, the procedure does not require any random sampling, i.e., it is essentially deterministic. In other words, repeated experiments should produce the same result. However, since it exhaustively calculates the entire learning procedure n number of times (n being the total number of instances in the dataset), computational cost is very high. In our case, since there are only 155 sample points, the computational requirement is not very demanding. Using basic minimal covering rule and majority threshold of 15% (i.e., an object is assigned to a decision class if the class collects as many vote as 15%) the new instances are classified. The result shows that the overall accuracy is 88.39% with 7.74% objects misclassified. 3.87% objects (originally$D^{(1)}$) remains unclassified. The kappa statistic was found 0.316. The area under ROC curve is computed to be 0.60. Table 3 summarizes the main statistic.

Table 3
Summary statistic of rough set theory

Seismic class	True positive rate	False positive rate	Sensitivity	Specificity	F-measure
$D^{(2)}$	0.25	0.042	0.25	0.96	0.25
$D^{(1)}$	0.996	0.75	0.96	0.25	0.96

5. Decision tree methods

Using recursive partitioning of the data set, classification or prediction in decision tree proceeds in top down induction approach of divide-and-conquer algorithm. Partitioning is done at each node by selecting a variable and a split which guarantees highest reduction of entropy or maximum information gain. The entropy values for information of a set of training examples $\Re$ are

$$E(\Re) = \sum_{i=1}^{m} p_i \log(1/p_i) = -\sum_{i=1}^{m} p_i \log p_i$$

where, p_i is the ratio of the class $\Re_i$ in the set of example $\Re$. When class $\Re_i$ partitions set $\Re$ using attribute A_j, the value of information $E(A_j)$ is $E(A_j) = \sum_{i=1}^{|X_j|} w_i * E(\Re'_i)$, where $\Re'_i$ is the lower level of example with partition based on attribute A_j, and w_i is the ratio pf the example in $\Re'_i$ to the number of example in $\Re$. Thus, information gain achieved by using attribute A_j in decision tree from example $\Re$ is

$$\mathrm{Gain}(A_j) = E(\Re) - E(A_j).$$

A popular tree building algorithm is Quinlan's ID3 (*Iterative Dichotomizer 3*) (Quinlan, 1986, 1992). The tree building process starts by selecting an attribute to place at the root node and at each succeeding level the subsets generated by preceding levels are further partitioned until it reaches a relatively homogenous terminal node or leaf node consisting of majority of the examples in a single class. The condition attributes that induces most amount of entropy reduction and information gain are placed closer to the root node. The so called homogeneity is a subject to predefined threshold and the node is labeled to a class having maximum frequency at that node. An extension of ID3 includes Quinlan's C4.5 and C5 which model both discrete and continuous variables (Quinlan, 1992). Additional modification includes handling missing value, pruning of decision tree, and rule derivation. Another variant of decision tree is CART (Breiman, Friedman, Olshen, & Stone, 1984), which use diversity index (Gini Index) to decide the best split.

6. Results of decision tree

We used C4.5 learning scheme implemented as J48 class in Weka (Witten & Frank, 2005) – a machine learning workbench which includes a framework in the form of Java class library. Initially, we evaluate the worth of an attribute by measuring the information gain ratio with respect to the class. The result is shown in Table 4. Attributes are then ranked by their individual evaluations by using in conjunction with attribute evaluators such as ReliefF, GainRatio, and Entropy, etc.

The ranking of attributes shows that concentration at site 1 and 2 are clearly the most important attributes followed by environmental attributes such as sun period, concentration at site 5 and atmospheric pressure. However, these attributes should not be considered in terms of seismic predictive value. Fig. 2 shows the unpruned tree resulting from C4.5 classifier. The tree consists of 16 nodes with 13 terminal leaves. Clearly, the tree shows the importance of non-climatic variables. The initial split is made at site 1 and then at site 2. The first split on the C1 causes 70.3% of instances to be classified a class $D^{(1)}$. The leaf resulting from split $f(x, \mathrm{C1} = 1)$ consists of 22 instances, of which 3 instances having $D^{(2)}$are misclassified (represented as 1(22/3.0)). At the second level, the splits are on C2 where the leaf associated with split $f(x, \mathrm{C2} = 3)$ has 4 instances classified as $D^{(1)}$. One instance is misclassified as $D^{(1)}$. At this level, 20.6% of instances are assigned final class. The remaining 9.03% instances are finally classified at third level following a split on C5. At this level, the condition $f(x, \mathrm{C2} = 3)$ generates a leaf with class $D^{(2)}$ wherein 3 out of 4 instances are correctly classified as $D^{(2)}$. The tree does not produce any unclassified instance.

Table 4
Average ranking of attributes with respect to information gain ratio, entropy and ReliefF

Average rank	Attributes	Average merit
1	C1	.064
2	C2	.056
3	C10	.043
4	C5	.037
5	C9	.036
6	C6	.032
7	C3	.027
8	C4	.026
9	C12	.025
10	C13	.019
11	C8	.016
12	C7	.012
13	C14	.007
14	C11	004
15	C15	0

The predictive performance of the tree is measured using "*Leave-one-out*" stratified cross-validation method.

The tree classifier correctly classified 93.55% instances with only 10 instances misclassified (6.45%). Since we are more interested in predicting higher seismic activity than lower activity, the overall accuracy of the model does not adequately reflect the model performance. Although the accuracy is high, out of 8 instances of having

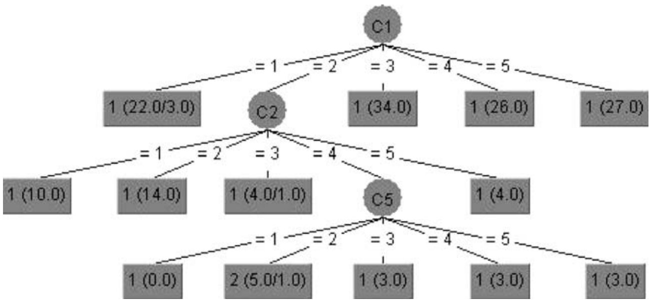

Fig. 2. Decision tree generated by C4.5 algorithm.

Table 5
Summary statistic of C4.5 algorithm

Seismic class	True positive rate	False positive rate	Sensitivity	Specificity	F-measure
$D^{(2)}$	0.375	0.034	0.375	0.966	0.375
$D^{(1)}$	0.966	0.625	0.966	0.375	0.966

$D^{(2)} = f(x, D) = R > 1.5|D^{(2)}| = 8$ only 3 instances are correctly identified by the tree. Although true positive rate for $D^{(2)}$ is relatively low, the false positive rate is extremely low 0.034 indicating classifier's conservative bias towards classification to less risky class.

Table 5 summarizes the overall result of C4.5 algorithm. This is also reflected in the specificity score of $D^{(2)}$which is considerably higher (0.966). The product of sensitivity and specificity, which is often used as overall measure, is equal to 0.362 reflecting moderate predictive performance of the decision tree. The area under ROC curve (AUC) is calculated to be 0.626. The kappa statistic is 0.34. The AUC and kappa score show that the predictive performance of the tree is somewhat in the range of low to moderate.

7. Comparison of rough set and decision tree results

Unlike black box models such as neural network, rough set and J48 provide transparent method for inductive learning from data. This is especially important when the seismic activity need to be understood in term of its causal structure involving multiple factors and their interactions. Thus, the rules generated from these machine learning techniques could provide further insight into the complex dynamics of seismic activity.

While both methods provide algorithm for evaluating conditioning attributes, their inherent significance is entirely different. In decision tree the main objective of attribute evaluation is based on information theoretic value viz. attributes' information gain (i.e., entropy reduction). The ranking of individual attribute in Table 4 are made chiefly by using parameters such as, GainRatio, Entropy, ReliefF, etc. While the concept of reduct in rough set is based on elimination of superfluous or redundant attributes in a decision table. The focus is to identify minimal set of attributes that preserve the indiscernibility relation. Hence, ranking of the attributes in Tables 4 and 2 reflect two distinct characteristics. For example, among the top five attributes in the two tables only C2 and C9 are appears as common attribute, which means that they have high information gain value as well as high quality of sorting.

In total 13 decision rules can be extracted from the decision tree generated by J48 algorithm. Only one rule predicts $D^{(2)}$ which is "if $(C1 = 2)$ & $(C2 = 4)$ & $(C5 = 2) \Rightarrow$ $(\text{Class} = D^{(2)})$". This rule is similar to the one derived in rough set except the additional descriptor C10 included in the antecedent. While J48 offers only one rule for$D^{(2)}$, rough set generates four distinct rules $D^{(2)}$. Although rough set offers extensive and explanatory rules, there is higher degree of compactness and compression of knowledge in C4.5-classifier's rule.

Table 6
Comparison of predictive and discriminatory performance

	Overall accuracy (%)	Kappa coefficient	Area under ROC curve (AUC)	Unclassified (%)
Decision tree (C4.5)	93.55	0.34	0.626	0
Rough set	88.39	0.32	0.60	3.87

Table 6 shows the predictive accuracy and discriminatory performances. The overall accuracy is lower in rough set because a number of instances were left unclassified. However, the kappa coefficients do not vary significantly. In fact, the difference is not statistically significant. Kappa statistic shows improvements over random assignment. The expected value of kappa for a random predictor is zero. Hence, the predicted classes are beyond chance agreement for both classifiers. The area under ROC curve (AUC) provide a measure of discrimination, i.e., how well the classifier is able to discriminate objects in decision class $D^{(2)}$ from objects in decision class $D^{(1)}$. AUC is generally accepted as the best ROC-derived single-number statistic for performance assessment (Øhrn, 1999). An AUC < 0.5 is considered no discriminatory ability, while an AUC of 1 represents perfect discrimination. Although the AUC for C4.5 is slightly higher their overall discriminatory performance is somewhat in the moderate range.

8. Conclusion

While seismic activity is subject to many complex factors associated with space and time, inductive machine learning approaches like decision tree and rough set provide predictive tools to unfold hidden patterns. Both rough set and decision tree (C4.5) methods generate comprehensible and explanatory descriptions in terms of radon concentration and climatic variables of seismic patterns by means of inductive learning from time series data. Using information theoretic measures such as entropy, information gain, the relative ranking of condition attributes shows that radon emanation at site 1 and 2 are the most important factors. Clearly, the both methods identify site 1 and 2 as major indicators of elevated seismic activity. Moreover, rough set induced rules show that higher seismic activity corresponds to low values of C1 and high values of C2. The decision tree induced by C4.5 algorithm shows splits based on radon emanation at site 1, 2 and 5, respectively. While maximal frequency of occurrence in rough set reducts is at site 2, environmental factors appear also significant suggesting a strong relationship of radon emanation and associated environmental factors (Zmazek et al., 2003). The decision rules derived from rough are extensive, while

C4.5 rules are more compact and could be useful to manage large number of rules. The cross-validation based on "leave-one-out" method shows that although the overall predictive and discriminatory performance of decision tree is to some extent better than rough set, the difference is not statistically significant. A hybrid approach combining rough set and decision tree (Minz & Jain, 2005; Nguyen, 1998) could provide better predictive and discriminatory performance. By generating homogeneous patterns occurring in time series data it is possible to extract temporal templates (Synak, 2000) of seismic events where decision rules describing dependencies between temporal features can be induced by employing rough set or hybrid rule induction techniques for each template.

References

Aha, D., & Kibler, D. (1991). Instance based learning algorithms. *Machine Learning, 6*, 37–66.

Bazan, J. G., Skowron, A., & Synak, P. (1994). Dynamic reducts as a tool for extracting laws from decisions tables. Paper presented at the eigth international symposium on methodologies for intelligent systems. London, UK.

Belayaev, A. (2001). Specific Features of radon earthquake precursors. *Geochemistry International, 12*, 1245–1250.

Beynon, M., & Peel, M. (2001). Variable precision rough set theory and data discrimination: An application to corporate failure prediction. *OMEGA, 29*, 561–576.

Biagi, P. F., Ermini, A., Kingsley, A., Khatkevich, S. P., & Gordeev, Y. M. (2001). Difficulties with interpreting changes in groundwater gas content as earthquake precursors in Kamchatka. *Russian Journal of Seismology, 5*, 487–497.

Breiman, L., Friedman, J., Olshen, J., & Stone, R. (1984). *Classification and regression trees*. Chapman & Hall.

Browne, C., Duntsch, I., & Gediga, G. (1998). IRIS revisited: A comparison of discriminant and enhanced rough set data analysis. In L. Polkowski & A. Skowron (Eds.), *Rough sets in knowledge discovery 2: Applications. Case studies and software systems* (pp. 345–368). New York: Physica-Verlag.

Cuomo, V., Bello, G. D., Lapenna, V., Piscitelli, S., Telesca, L., Macchiato, M., et al. (2000). Robust statistical methods to discriminate extreme events in geoelectrical precursory signals: Implications with earthquake prediction. *Natural Hazards, 21*, 247–261.

Daubie, M., Levecq, P., & Meskens, N. (2002). A comparison of rough sets and recursive partitioning induction approaches: An application to commercial loans. *International Transactions in Operational Research, 9*, 681–694.

Dimitras, A., Slowinski, R., Susmaga, R., & Zopounidis, C. (1999). Business failure using rough sets. *European Journal of Operational Research, 114*, 263–280.

Dmeroski, S. (2002). Applications of KDD methods in environmental sciences. In W. Kloesgen & J. Zytkow (Eds.), *Handbook of data mining and knowledge discovery*. Oxford: Oxford University Press.

Düntsch, I., & Gediga, G. (1997). Statistical evaluation of rough set dependency analysis. *International Journal of Human–Computer Studies, 46*, 589–604.

Düntsch, I., & Gediga, G. (1998). Uncertainty measures of rough set prediction. *Artificial Intelligence, 106*(1), 77–107.

Fleischer, R. L., & Mogro-Campero, A. (1981). Radon transport in theearth a tool for uranium exploration and earthquake prediction. Paper presented at the 11th international SSNTD conference, 7–12 September.

Flinkman, M., Michalowski, W., Nilsson, S., Slowinski, R., Susmaga, R., & Wilk, S. (2000). Use of rough sets analysis to classify Siberian forest ecosystems according to net primary production of phytomass. *INFOR, 38*(3), 145–160.

Fu, L. M. (1999). Knowledge discovery based on neural networks. *Communications of the ACM, 42*(11), 47–50.

Grzymala-Busse, J. W., & Than, S. (1992). Reduction of instance space in machine learning from examples. Paper presented at the Proceedings of the fifth international symposium on artificial intelligence, Cancun, Mexico, December 7.

Jelonek, J., Krawiec, K., & Slowinski, R. (1995). Rough set reduction of attributes and their domains for neural networks. *Computational Intelligence, 11*, 339–347.

King, C. Y. (1985). Radon monitoring for earthquake prediction in China. *Earthquake Prediction Research, 3*, 47–68.

Krusinska, E., Slowinski, R., & Stefanowski, J. (1992). Discriminant versus rough set approach to vague data analysis. *Applied Stochastic Models and Data Analysis*(8), 43–56.

Magro-Campero, A., Fleischer, R. L., & Likes, R. S. (1980). Changes insubsurface radon concentration associated with earthquakes. *Journal of Geophysics Research, 85*, 3053–3057.

Mak, B., & Munakata, T. (2002). Rule extraction from expert heuristics: A comparative study of rough sets with neural networks and ID3. *European Journal of Operational Research, 136*, 212–229.

Mega, M. S., Allegrini, P., Grigolini, P., Latora, V., Palatella, L., Rapisarda, A., et al. (2003). Power-law time distribution of large earthquakes. *Physical Review Letters, 90*(18), 1–4.

Minz, S., & Jain, R. (2005). Refining decision tree classifiers using rough set tools. *International Journal of Hybrid Intelligent Systems, 2*(2), 133–148.

Mitchell, T. M. (1997). *Machine learning*. Portland: McGraw-Hill.

Nguyen, H. S. (1998). From Optimal Hyperplanes to Optimal Decision Trees. *Fundamenta Informaticae, 34*(1–2), 145–174.

Øhrn, A. (1999). Discernibility and rough sets in medicine: Tools and applications. Unpublished PhD Thesis, Norwegian University of Science and Technology.

Pawlak, Z., & Slowinski, R. (1994). Rough set approach to multi-attribute decision analysis. *European Journal of Operational Research, 72*, 443–459.

Polkowski, L., & Skowron, A. (1998). Rough sets in knowledge discovery: Methodology and applications. In L. Polkowski & A. Skowron (Eds.). Heidelberg: Physica-Verlag.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning, 1*, 81–106.

Quinlan, J. R. (1992). *C4.5: Programs for machine learning*. San Mateo, CA: Morgan Kaufman.

Skowron, A., & Grzymalla-Busse, J. (1994). From rough set theory to evidence theory. In R. Yager, M. Fedrizzi, & J. Kacprzyk (Eds.), *Advances in the Dempster–Shafer theory of evidence* (pp. 192–271). New York: John Wiley & Sons, Inc.

Skowron, A., & Rauszer, C. (1992). *The discernibility matrices and functions in information systems. Intelligent decision support*. Dordrecht, Netherlands: Kluwer.

Slowinski, R., Zopounidis, C., & Dimtras, A. I. (1997). Prediction of company acquisition in Greece by means of the rough set approach. *European Journal of Operational Research, 100*, 1–15.

Stefanowski, J. (1992). Rough sets theory and discriminant methods as tools for analysis of information systems – A comparative study. *Foundations of Computing and Decision Sciences, 2*, 81–98.

Stefanowski, J. (1998). On rough set based approaches to induction of decision rules in rough set in knowledge discovery. In A. Skowron, & L. Polkowski (Eds.). (Vol. 1, pp. 525–529). Physica Verlag: Heidelberg.

Synak, P. (2000). Temporal templates and analysis of time related data. *Lecture notes in computer science: Second international conference on rough sets and current trends in computing* (Vol. 2005, pp. 420–427). London, UK: Springer.

Szczuka, M. S. (1998). Rough sets and artificial neural networks. In L. Polkowski & A. Skowron (Eds.), *Rough sets in knowledge discovery 2:*

Applications, case studies and software systems (pp. 449–470). New York: Physica-Verlag.

Szczuka, M. S. (1999). Symbolic and neural network methods for classifier construction. Unpublished PhD Dissertation, Warsaw University.

Takahashi, M. (2003). Observation of Radon for earthquake prediction research. *Geochimica et Cosmochimica Acta Supplement, 67*(18), 468.

Teghem, J., & Benjelloun, M. (1992). Some experiments to compare rough sets theory and ordinal statistical methods. In R. Slowinski (Ed.), *Intelligent decision support: Handbook of applications and advances of rough set theory*. *System Theory, Knowledge Engineering and Problem Solving* (Vol. 11, pp. 267–284). Dordrecht: Kluwer Academic Publishers.

Teghem, J., & Charlet, J. M. (1992). Use of "rough sets" method to draw premonitory factors for earthquakes by emphasing gas geochemistry: The case of a low seismic activity context in Belgium. In R. Slowinski (Ed.), *Intelligent decision support: Handbook of applications and advances of rough set theory*. *System Theory, Knowledge Engineering and Problem Solving* (Vol. 11, pp. 165–179). Dordrecht: Kluwer Academic Publishers.

Telesca, L., Lapenna, V., & Macchiato, M. (2005). Multifractal fluctuations in earthquake-related geoelectrical signals. *New Journal of Physics, 7*(214).

Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques* (2nd ed.). New York: Morgan Kaufmann.

Wong, S. K. M., Ziarko, W., & Ye, R. L. (1986). Comparison of rough-set and statistical methods in inductive learning. *International Journal of Man–Machine Studies, 24*, 53–72.

Wroblewski, J. (1995). Finding minimal reducts using genetic algorithms. In P. P. Wang (Ed.), *Proceedings of the international workshop on rough sets soft computing at second annual joint conference on information sciences (JCIS'95)* (pp. 186–189). Wrightsville Beach, NC.

Zmazek, B., Todorovski, L., Dzeroski, S., Vaupotic, J., & Kobal, I. (2003). Application of decision trees to the analysis of soil radon data for earthquake prediction. *Applied Radiation and Isotopes, 58*, 697–706.