International Journal of Earth, Energy and Environmental Sciences
ISSN: 2517-942X
Vol:6, No:4, 2012

Bayesian Networks for Earthquake Magnitude Classification in a Early Warning System

G. Zazzaro, F.M. Pisano, G. Romano

***Abstract*—** During last decades, worldwide researchers dedicated efforts to develop machine-based seismic Early Warning systems, aiming at reducing the huge human losses and economic damages. The elaboration time of seismic waveforms is to be reduced in order to increase the time interval available for the activation of safety measures. This paper suggests a Data Mining model able to correctly and quickly estimate dangerousness of the running seismic event.

Several thousand seismic recordings of Japanese and Italian earthquakes were analyzed and a model was obtained by means of a Bayesian Network (BN), which was tested just over the first recordings of seismic events in order to reduce the decision time and the test results were very satisfactory.

The model was integrated within an Early Warning System prototype able to collect and elaborate data from a seismic sensor network, estimate the dangerousness of the running earthquake and take the decision of activating the warning promptly.

***Keywords*—**Bayesian Networks, Decision Support System, Magnitude Classification, Seismic Early Warning System

I. Introduction

OVER the last few decades there has been ongoing experimentation into seismic early-warning (EW) systems in several active seismic areas of the world. EW systems are operating (active) in Mexico, Japan, Taiwan, Romania and Turkey; while other systems are under development (Italy, India, California, Greece, …). Although the prediction of earthquakes is not yet practicable, current technology allows prompt identification of the onset of any dangerous seismic event. As it is well known, seismic EW concerns the capability of estimating the destructive potential of an earthquake in the seconds immediately following its generation. Such an estimation can then be used in sending out an alarm to strategic sites in order to allow activities for their securing before the arrival of the destructive seismic waves. In addition earthquake EW systems utilize the capability of modern telecommunication systems to process and transmit information faster than seismic waves propagate. When a suitable seismic sensor network is available to protect a geographical area, or a specific site, fast processing methods can be applied to locate an earthquake, calculate the event magnitude, and estimate the distribution of ground motion. A seismic network could be distributed in the epicentral area, or localized around the area to be protected, if the epicenter is unknown.

A monitoring network is composed by remote sensing stations that transmit in real-time to a central processor that provides to calculate in real-time seismic parameters such as location, origin time, magnitude.The purpose of the earthquake EW is to quickly announce people that an earthquake has occurred and inform them about the estimated seismic intensity several seconds or more before the arrival of strong tremors caused by the quake. The present paper describes part of Data Elaboration Center included in the research project “SIT_MEW – Integrated Network of broadband communication with early warning methodologies for land and emergency management in case of natural disaster”, funded by Italian MIUR (Minestry of Education, University and Research); the part described in the paper was in charge of the authors. The project aimed at developing an EW system both for site-specific and regional warning, receiving seismic waveform from a monitoring sensor network placed in the Irpinia region (southern Italy). The system was asked for efficiently taking decision upon the opportunity of alerting people and infrastructures in the area of Naples city when an earthquake originated in Irpinia, reducing the probability of missed and false alarms.With such requirements, the system was designed in order to match the pressing time constraint of assuring at least a time interval of 20 seconds for the activation of automatic safety measures (e.g. traffic lights to prevent transit on threatened structures, shutdown of fuel pipelines and dangerous tanks, isolation of hospital operating rooms, etc.) in the urban area of Naples. In order to do that, each step from the data collection to the warning decision was carefully designed to assure a limited time-consumption. In more detail, the data analysis step had to take into account that every earthquake is recorded by more than one seismic sensor, and each sensor produces at least three accelerograms, one for each coordinate axis; such complex physical phenomenon makes Data Mining (DM) worthy for application because of its ability to work with many variables and data. Actually, in some recent papers [7], DM algorithms such as Decision Trees, Clustering and Association Rules were applied to the seismic classification for post-processing analysis. This work was carried out through the realization of the Knowledge Discovery in Database (KDD) process according to the standard model process conceived by the Cross-Industry Consortium Standard Process for DM (CRISP-DM) [11]. The process is finalized to create a numerical model for seismic magnitude classification based on an appropriate selection of seismic parameters of the earthquake.This work was part of the algorithm-based analytical core of a prototype system for the application of seismic EW methods, for real-time mitigation of earthquake effects

G. Zazzaro, F.M. Pisano and G. Romano are with CIRA (Italian Aerospace Research Centre), Via Maiorise sn, 81043 Capua (CE) Italy. (phone: 00390823623558; e-mail {g.zazzaro, f.pisano, g.romano}@cira.it).

In this work Weka tool (Version 3.6.2) (Waikato Environment for Knowledge Analysis) was used ([2], [9], [12]) to carry on DM analysis, from data exploration to model evaluation.

II. Business Understanding

A. Early Warning Definition

EW is widely defined as all the actions that can be taken during the lead time of a catastrophic event. The lead time is defined as the time elapsing between the instant when the occurrence of a catastrophic event in a given place is reasonably certain and the moment it actually occurs [4]. Typical lead times are of the orders of seconds to tens of seconds for earthquakes, minutes to hours for tsunamis, etc.

B. Early Warning Principle

Tremors extend out from the seismic focus in a wave-like motion. When an earthquake occurs it releases energy in the form of waves that radiate from the earthquake source in all directions. The principle on which EW systems are based exploits the consideration that seismic waves travel with velocity less than electromagnetic signals, used to transmit the seismic information about the incoming event from the sensor networks to the elaboration centre. In addition there are two main types of seismic waves: P-waves (Primary) or initial tremors (not destructive), and S-waves (Secondary) which cause stronger tremors and damages. P-waves are compressional waves that are longitudinal in nature. S-waves are slower than P-waves and move at about half the speed of P-waves. Vertical ground motion generated by S-waves is highly damaging to the structures. An EW system is based on the different propagation velocities between P and S-waves.

TABLE I
Differences among P and S-waves and Electromagnetic waves

Traveling speed of seismic waves		
P-waves	cause rattling tremors	around 7km/s
S-waves	cause larger, more powerful tremors	around 4km/s
Electromagnetic waves	to spread the seismic alert	around 300000km/s

The time interval from the arrival of P-waves and the S-waves may be used to activate security measures: as matter of fact, the goal is to exploit the time delay of P-waves with respect to S-waves in order to forecast the effect of the latter based on the automatic elaboration of the former.

Assuming that the warning time provided by the EW system is sufficient for the activation of the protection measures, based on the predictions from the first few seconds of P-wave observation, an effective decision has to be made whether to activate the alarm or not.

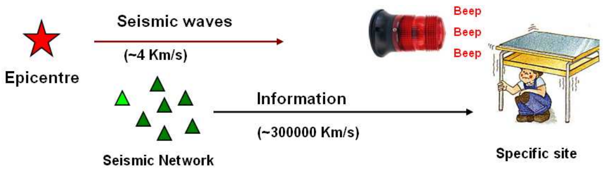


Fig. 1 EW system can save a lot of lives

Since prediction is uncertain in making this decision, false and missed alarms are possible. As a consequence a key element of an EW system is a better understanding of the parameters that play a fundamental role in this uncertainty. As a result performance-based approach to EW system design and decision models is a mandatory necessity.

A decision model is then presented to take a decision in a real-time scenario based on the expected consequences and savings coming from the decision itself.

If the magnitude threshold is exceeded, a warning signal is transmitted through an area-wide transmitter or to a monitored target site. The message contains information of the incoming event. As the event evolves, more data are available in order to confirm and increase the accuracy of the information processed starting from the incoming signals.

C. Data Mining Goals

In order to predict the level of danger of an earthquake when it is running and to distinguish dangerous from non-dangerous seismic events, a lot of data mining techniques have been applied to create a numerical model of classification.

In order to recognize a seismic event as dangerous or not while it's running many different DM techniques were applied to create a successful model which, once deployed, satisfied the strict time constraint for classification.

The problem of seismic events classification was faced by means of Bayesian Networks, whose objective is to build up a model able to classify a seismic event, represented by a set of significant parameters, as dangerous by associating the value YES to the binary target value; the classification is correct only if the magnitude of the event is higher than the chosen threshold.

The Business goal was the prediction of the magnitude of an earthquake in progress, so it was translated into a data mining goal consisting of the classification of the magnitude of seismic events. In particular, the classification is a binary classification and the target class is the magnitude threshold.

The threshold is fixed to 5, because an earthquake is considered dangerous (in Irpinia area of Italy [6], [10]) if its magnitude is greater than 5.

If the value of the expected magnitude exceeds the threshold value, then a warning message could be given.

III. Data Understanding

A. Data Source

For the present work two data sources were used:

1) JAPAN, called J from KIK database [14], is a dataset of seismic registration from Japan KIKnet.

2) RISSC, called R from RISSC database [15], is a dataset of southern Italy seismic registration from ISnet (Irpinia Seismic network).

B. Data Format

The original data format was SAC which stands for "Seismic Analysis Code". It was originally developed to analyze data in time series, especially seismic data. It is one of the most widely used data formats for data storage in the seismological research community [13].

Every SAC file contains a fixed length header section followed by one or two data sections. The header contains floating point, integer, logical, and character fields.

The following table shows some of the contents of the SAC binary data file header. For example, Name of the station, date of the seismic event, magnitude and event location. Every SAC file contains 133 fields, some defined and some undefined (set to “-12345”).

The second section of a SAC file contains the time series of the dependent variable (acceleration or velocity) related to the header, registered by a seismic sensor.

The following is a SAC file header of a Japanese earthquake of magnitude 4 occurred in 2006, February 18. This header has 32 defined fields.

TABLE II
SAC HEADER EXAMPLE

FILE: AIC0010602181621.UD.sac - 1

NPTS = 6000
B = 0.000000e+00
E = 5.999000e+01
IFTYPE = TIME SERIES FILE
LEVEN = TRUE
DELTA = 1.000000e-02
DEPMIN = -3.949165e+00
DEPMAX = 2.167225e-01
DEPMEN = -1.835354e+00
AMARKER = 9.19 (IP-0)
T0MARKER = 15.418
KZDATE = FEB 18 (049), 2006
KZTIME = 16:21:26.000
KSTNM = AIC001
STLA = 3.529440e+01
STLO = 1.367530e+02
STEL = 6.000000e+00
KEVNM = NONE
EVLA = 3.568500e+01
EVLO = 1.364210e+02
EVDP = 1.300000e+01
DIST = 5.277971e+01
AZ = 1.450985e+02
BAZ = 3.252903e+02
GCARC = 4.746583e-01
LOVROK = TRUE
USER1 = 4.100000e+00
NVHDR = 6
NWFID = 198
LPSPOL = TRUE
LCALDA = TRUE
KCMPNM = Q

The header parameters have the following meanings [13]:

TABLE III
MEANINGS OF SAC HEADER PARAMETERS

NPTS	Number of points per data component
B	Beginning value of the independent variable
E	Ending value of the independent variable
IFTYPE	Type of file
LEVEN	TRUE if data is evenly spaced
DELTA	Increment between evenly spaced samples
DEPMIN	Minimum value of dependent variable
DEPMAX	Maximum value of dependent variable
DEPMEN	Mean value of dependent variable
AMARKER	First arrival time (seconds relative to reference time) – P-wave arrival time
TOMARKER	Second arrival time (seconds relative to reference time) – S-wave arrival time
KZDATE	Alphanumeric form of GMT reference date
KZTIME	Alphanumeric form of GMT reference time
KSTNM	Station name
STLA	Station latitude
STLO	Station longitude
STEL	Station elevation
KEVNM	Event name
EVLA	Event latitude (degrees)
EVLO	Event longitude (degrees)
EVDP	Event depth below surface (meters)
DIST	Station to event distance (km)
AZ	Event to station azimuth (degrees)
BAZ	Station to event azimuth (degrees)
GCARC	Station to event great circle arc length (degrees)
LOVROK	TRUE if it is okay to overwrite this file on disk
USER1	User defined variable storage area. Magnitude event in this header
RNVHDR	Header version number
NWFID	Waveform ID
LPSPOL	TRUE if station components have a positive polarity
LCALDA	TRUE if DIST, AZ, BAZ, and GCARC are to be calculated from station and event coordinates
KCMPNM	Component name.

The other header fields are undefined.

C. Earthquake Magnitude

Usually, the SAC field number 39 called MAG stores the earthquake magnitude.

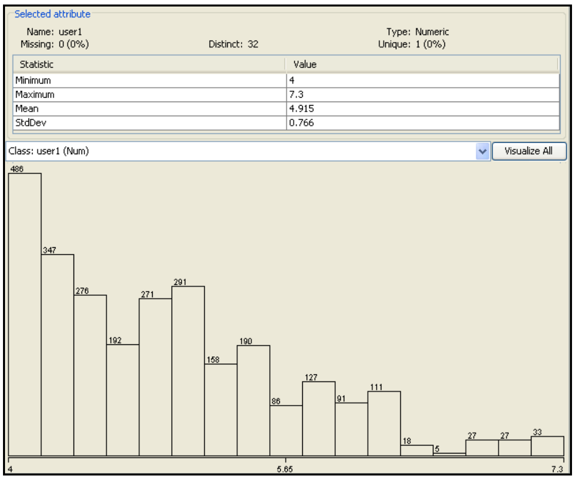
The magnitude is a parameter used by seismologists to quantify the earthquake size. The Richter magnitude scale summarizes the amount of seismic energy released by an earthquake. It is obtained by calculating the logarithm of the combined horizontal amplitude of the largest displacement from zero on a seismometer output. Measurements have no limits and can be either positive or negative [10].

D. Japan DataBase

The initial dataset consisted of 8208 files in SAC format, representing 2736 seismic events occurred in JAPAN. For each event the dataset contained three files: the first for the EW component (east-west), the second for the NS component (north-south) and the third for the UD (up-down) of the acceleration of registration of seismic events.

1. *Japan Data Exploration*

User1 attribute of SAC file header stores magnitude of seismic events whose histogram over the dataset is shown in Figure 4. The minimum value taken from this field is 4, while the maximum is 7.3, the average is 4,915 while the standard deviation is 0,766.



Selected attribute

Name: user1 Type: Numeric
Missing: 0 (0%) Distinct: 32 Unique: 1 (0%)

Statistic	Value
Minimum	4
Maximum	7.3
Mean	4.915
StdDev	0.766

Fig. 2 User1=mag statistical distribution

E. RISSC DataBase

1. *Irpinia Seismic Network*

Irpinia Seismic Network (ISNet) is a local network of strong motion and it was designed in 2002. ISNet covers an area of approximately 100 km x 70 km along Campania-Lucania Appennine chain in Irpinia and is deployed along the active fault responsible for the 1980, November 23, Mag 6.9 Campania–Lucania earthquake [4]. ISNet consists of 28 seismic stations, each of which is connected with real-time communication to a Local Control Center (LCC) that is generally located in an urban area. The six LCCs make first elaborations over the incoming data from seismic stations.

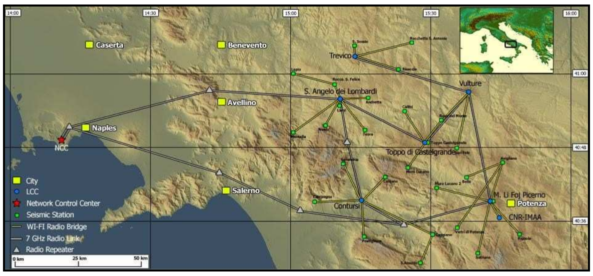


Fig. 3 Topology of communication system of Irpinia Seismic Network (ISNet) in southern Italy

2. *The Waveforms and Events Database*

RISSC (http://www.rissclab.unina.it/) keeps track of the events detected by ISNet and the relative waveforms recorded by the sensors.

After the request for permission it is possible to access the database RISSC (http://dbserver.ov.ingv.it:8080/login.jsp). This database stores objects for events, origin estimations (time and location), magnitude estimations and waveforms. A waveform object for each sensor that recorded the earthquake is also linked to the event object and stored a pointer to a SAC file.

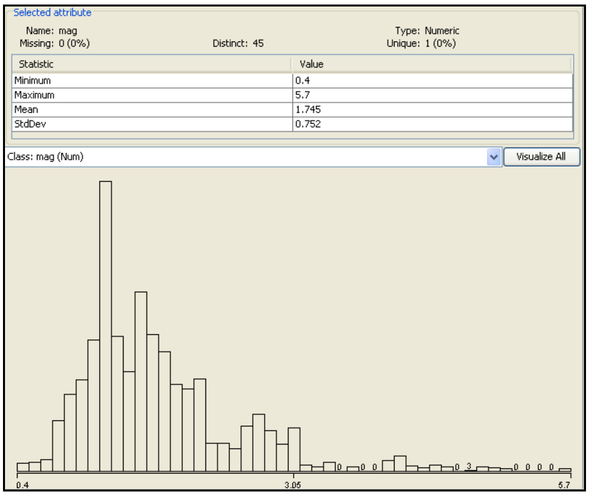
3. *Irpinia Seismic Events*

In the last years no dangerous seismic event has occurred in Irpinia, thankfully.

A total of 38763 SAC files, related to events of low magnitude occurring between 2005 and 2009, were downloaded from online RISSC database. The files described 1297 earthquakes. For each event the dataset contained three files: the first for the up-down component (0), the second for the north-south component (1) and the third for the east-west (2) of the acceleration of registration of seismic events.

4. *RISSC Data Exploration*

MAG attribute of SAC file header stores the magnitude of seismic events whose histogram over the dataset is shown in Figure 4. The minimum value recorded in this field is 0.4, while the maximum is 5.7, the average is 1.745 while the standard deviation is 0,752.



Selected attribute

Name: mag Type: Numeric
Missing: 0 (0%) Distinct: 45 Unique: 1 (0%)

Statistic	Value
Minimum	0.4
Maximum	5.7
Mean	1.745
StdDev	0.752

Fig. 4 mag statistical distribution

From data exploration analysis of the dataset it was observed that few earthquakes with magnitudes greater than 3 are related to distant earthquakes from the seismic network (dist > 200 km).

F. J + R Dataset

From data exploration step it is observed that Japan events have higher magnitude than RISSC events.

TABLE IV
COMPARISON BETWEEN DATASETS J AND R

	JAPAN	RISSC
# SAC	8208 files 2736 UD, NS and EW components	38763 files 12921 0, 1 and 2 components
Events dates	From 1996 to 2006	From 2005 to 2009
# recorded events	2736	12690

	JAPAN	RISSC
# undefined fields	102	78
# fields in final dataset	26	26
# seismic events	337	1297
Event Magnitude	In *user1* SAC header field (*mag* field is undefined) [4, 7.3]	In *mag* SAC header field [0.4, 5.7]
Increment between evenly spaced samples (Hertz)	In *delta* SAC header field 0.01 Hertz	In *delta* SAC header field 0.008 Hertz
Station to event distance (km)	In *dist* SAC header field [2.07, 59.86]	In *dist* SAC header field [0.32, 427.4]
Event depth below surface (Km)	In *evdp* SAC header field [0, 50]	In *evdp* SAC header field [0.7, 459.8]

IV. Data Preparation

A. Time-check

The number $\Delta t = t_0 - a$ is calculated for each record, where a and t_0 are contained in the SAC file headers; in particular, t_0 is the S-wave temporal marker (seconds relative to reference time), while a is the P-wave temporal marker (first arrival time – seconds relative to reference time). The check requires that $\Delta t \geq 4$ seconds [Fig. 5]. All those records, for which the time interval elapsing between the arrival of the first wave (P-wave longitudinal, no-destructive seismic waves) and the second wave (S-wave transversal, destructive seismic waves) is less than 4 seconds, are excluded from the dataset. In fact, if $\Delta t < 4$ seconds the S-wave covers the P-wave and the signal to be analyzed will be corrupt and unusable. For the sake of clarity, all seismic parameters will be calculated in the first 4 seconds of P-waves in order to reduce the time of warning. Making a recap, if the beginning of the S-wave is too close to the beginning of the P-wave, time series of P is covered by the time series of S and it is not possible to use the first 4 seconds of the initial P-wave in order to predict the trend of S-wave as expected from an Early Warning System.

The number of J + R records in the dataset that pass the time-check is 11196 corresponding to about 76% (1113 seismic events) of the original dataset.

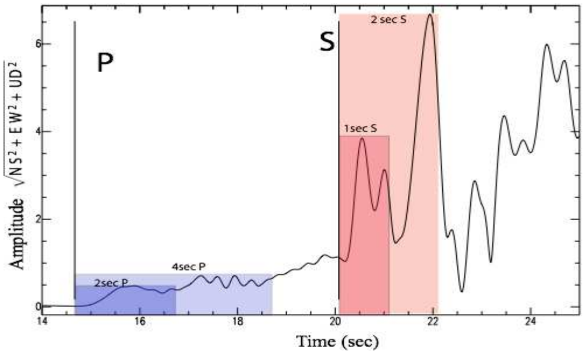


Fig. 5 Time Interval between P-wave and S-wave

B. Seismic Attributes

On kind suggestion of an expert on seismology, a number of physical indicators were selected and threshold of magnitude distinguishing events as dangerous or not was set to 5.

These parameters were extracted from the time series using an "ad hoc" developed JAVA procedure for real-time data integration.

TABLE V
Seismic derivate Attributes with Formulas and Meanings

Attribute	Description
LOG(PD)	Where PD is the module of the *peak displacement*, measured in the first 4 seconds of initial P-wave. LOG is a base-10 logarithm.
LOG(TP)	Where TP is the maximum, within 4 seconds of the initial P-wave, of the *predominant period* τ_p , of the vertical component waveform. $\tau_p(t) = \sqrt{\frac{\int_0^t v_z^2(s)ds}{\int_0^t a_z^2(s)ds}}$ Where v_z and a_z are the vertical components of speed and acceleration, and 0 is the arrival of P-wave.
LOG(TD)	Where TD is the maximum, within 4 seconds of the initial P-wave, of the *predominant period* τ_d , of the vertical component waveform. $\tau_d(t) = \sqrt{\frac{\int_0^t u_z^2(s)ds}{\int_0^t v_z^2(s)ds}}$ Where v_z and u_z are the vertical components of speed and displacement, and 0 is the arrival of P-wave.
LOG(IV2)	Where IV is the peak of speed integral, within 4 seconds of the initial P-wave. IV2 is IV to square
LOG(PD2/IV2)	PD2 is PD to square
LOG(IV2/PD)	
LOG(R/10)	Where R is calculated from the parameters *dist* and *evdp* found in the SAC file header. $R = \sqrt{dist^2 + evdp^2}$

THRESHOLD_5	IF MAG<5 THEN THRESHOLD_5 = 'NO'
	ELSE THRESHOLD_5 = 'YES'
	Where MAG is the earthquake *magnitude*.

In particular, R parameter is the Euclidean distance from recording station to earthquake hypocenter:

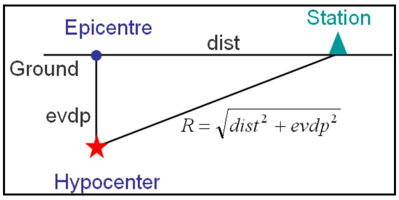


Fig. 6 R seismic parameter

TABLE VI
SEISMIC ATTRIBUTES WITH TYPES

#	Name_Attribute	Type
1	*LOG(PD)*	*Numeric*
2	*LOG(TP)*	*Numeric*
3	*LOG(TD)*	*Numeric*
4	*LOG(IV2)*	*Numeric*
5	*LOG(PD2/IV2)*	*Numeric*
6	*LOG(IV2/PD)*	*Numeric*
7	*LOG(R/10)*	*Numeric*
8	*THRESHOLD_5*	*Nominal* {YES,NO}

THRESHOLD_5 is the target attribute for all the Data Mining classification procedures.

TABLE VII
THRESHOLD_5 TARGET CLASS DISTRIBUTION

THRESHOLD_5 class	
Value	**Count**
SI	887
No	10309

From Table VII, the target class has an unbalanced distribution. In addition, for any registration of dangerous seismic event in the dataset, there are about 11.6 registrations of events that are not dangerous (with MAG<5).

C. J + R Clustering

In order to explore the complete dataset J+R and to seek possible outliers, a clustering algorithm was applied.

1. K-Means

The k-means algorithm is one of the most widely used algorithms for data clustering. Using WEKA tool, fixing k=2, k-means gave the following results:

```
=== Run information ===
Instances:  11196

Test mode: Classes to clusters evaluation on training data
=== Model and evaluation on training set ===
Cluster centroids:

      Attribute          Full Data     1          2
                         (11196)       (9147)     (2049)
      MAG                2.388         1.8298     4.88
      LOG(PD)            -0.114        0.2823     -1.8835
      LOG(TP)            -0.604        -0.5258    -0.953
      LOG(TD)            -0.1902       -0.1323    -0.4489
      LOG(IV2)           0.5868        1.2504     -2.376
      LOG(PD2/IV2)       -0.8149       -0.6858    -1.391
      LOG(IV2/PD)        0.7008        0.9681     -0.4925
      LOG(R/10)          0.774         0.7984     0.6653

Clustered Instances
1      9147 (82%)
2      2049 (18%)

Class attribute: ORIGIN

Classes to Clusters:
   1    2  <-- assigned to cluster
   0 2045 | J
 9147    4 | R

Cluster 1 <-- R       Cluster 2 <-- J
Incorrectly clustered instances :  4.0    0.0357 %
```

In particular, Cluster number 1 has 9147 (82% of full dataset) records, while Cluster 2 has 2049 (18%) records.

2. Cluster's Evaluation

Starting from the dataset consisting of 11196 records of J + R described by 10 attributes (7 seismic attributes + THRESHOLD_5 + MAG + ORIGIN) two groups called Cluster1 and Cluster2 were obtained. The above WEKA printout shows three obtained centroids: the first one for the full dataset and the others for two centroids.

Fixing Class attribute=ORIGIN, choosing Test mode="Classes to clusters evaluation on training data", WEKA showed the distribution of JAPAN and RISSC record over the two classes, represented in Table VIII:

TABLE VIII
CLUSTERING MATRIX

1	2	
0	**2045**	J
9147	**4**	R

TABLE IX
INSTANCE NUMBER 7555 OF J + R

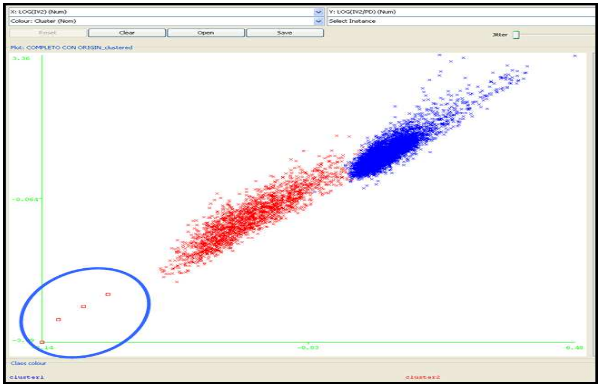
1	LOG(PD)	-3.980144
2	LOG(TP)	-0.904814
3	LOG(TD)	-0.701155
4	LOG(IV2)	-6.335259
5	LOG(PD2/IV2)	-1.625029
6	LOG(IV2/PD)	-2.355115
7	LOG(R/10)	1.105698
8	THRESHOLD_5	NO
9	MAG	2.7
10	ORIGIN	R
11	CLUSTER	cluster2

In particular the four red-highlighted records in the matrix belonged to R but they were attributed to J by the algorithm. So these four records could be outliers. In statistics, an outlier is an observation that is numerically distant from the rest of the data. These four records were removed from the dataset R + J.

The table IX shows one of the records (instance number 7555 of the original dataset).

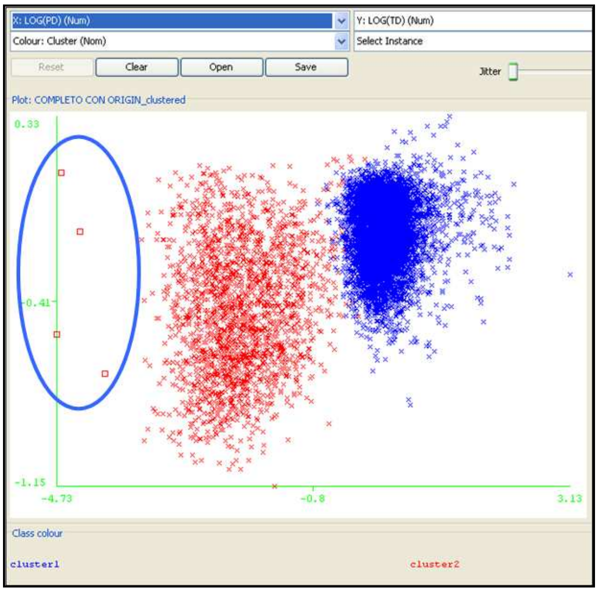
3. Cluster's Representation

In Fig. 7 and in Fig. 8 below, the clusters are represented in planes LOG(IV2),LOG(IV2/PD) and LOG(PD),LOG(TD), respectively.



X: LOG(IV2) (Num)
Y: LOG(IV2/PD) (Num)
Colour: Cluster (Nom)
Plot: COMPLETO CON ORIGIN_clustered

Fig. 7 Clusters in LOG(IV2), LOG(IV2/PD) plane



X: LOG(PD) (Num)
Y: LOG(TD) (Num)
Colour: Cluster (Nom)
Select Instance
Plot: COMPLETO CON ORIGIN_clustered
Class colour
cluster1
cluster2

Fig. 8 Clusters in LOG(PD), LOG(TD) plane

In addition Cluster1 is the blue one while Cluster2 is the red one. The outliers are circled in blue and we can see their distance from the centroids of the obtained clusters in the considered plans.

D.HoldOut Method with Stratified Remove Folds Filter

In order to carry on the modeling phase of the CRISP-DM, R and J datasets were split into a selection of subsets.

Business and Data Understanding phases showed that each earthquake was recorded by many stations within the seismic network.

To facilitate rapid prediction of the earthquake hazard that is running, it was decided to split the dataset J into two subsets IJ and NIJ, containing respectively all of the first waveforms (time-check passed) of the seismic events and the subsequent ones. Accordingly, the same splitting for the dataset R was made and the subsets IR and NIR were obtained.

Using a sequence of Stratified Remove Folds filter of Weka tool, it's possible to apply HoldOut Method [8] to obtain some subsets from original dataset J+R.

The following Fig. 9 and Fig. 10 show the splitting of J and R datasets.

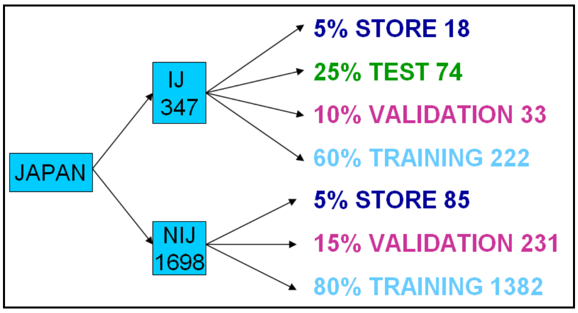


Fig. 9 Splitting of JAPAN dataset

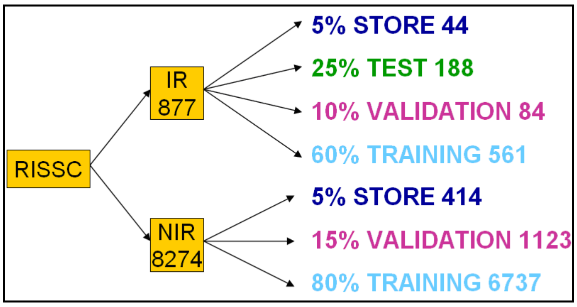


Fig. 10 Splitting of RISSC dataset

In table X, the results obtained concerning the subdivision of the original dataset are summarized. In addition the descriptions and uses of these subsets are also listed.

TABLE X
USES AND DESCRIPTIONS OF DATASETS

Dataset Name	Description	Use	Cardinality
TRAINING	First and no-first registrations from J and R datasets	To build the classificatio n models	8898
VALIDATION	First and no-first registrations from J and R datasets	To select the best models varying the parameters	1471
TEST	First registrations from J and R datasets	To test the selected models	262
STORE	First and no-first registrations from J and R datasets	For new records if necessary	561
without 4 outliers (they were in training set)			**11192**

V. MODELING

In order to classify dangerous earthquakes, various modeling techniques were selected and applied, and their parameters were calibrated to optimal values. In particular, many WEKA algorithms were applied MultilayerPerceptron for Neural Networks, J48 for Inductive Decision Trees and BayesNet for Bayesian Networks. In general all obtained classification models showed good results on the test set consisting of only the first seismic recordings. In this paper we show the results of applying an algorithm based on Bayesian Networks.

A. Bayesian Approach

As it is well known Bayesian classifiers are statistical classifiers. They can be Naïve or (Belief) Networks mainly. They can predict class membership probabilities, such as the probability that a given sample belongs to a particular class. Bayesian classification is based on Bayes theorem.

Bayesian or Belief Network (BN) used in this work is a probabilistic graphical model that represents a set of random variables and their conditional dependencies via a directed acyclic graph (DAG). BN specifies joint conditional probability distributions.

Formally a BN is defined by two components [8]:

1) DAG (defined by its topology), where each node represents a random variable and each arc represents a probabilistic dependence (if an arc is drawn from a node A to a node B, then A is a parent of B, and B is a descendent of A).
2) Conditional Probability Table (CPT) for each variable (the CPT for a variable A specifies the conditional distribution P(A│ Parents(A)), where Parents(A) is the set of parents of A).

Each node is associated with a probability function that takes as input a particular set of values for the node's parent variables and gives the probability of the variable represented by the node.

B. WEKA BayesNet

WEKA tool provides several algorithms for bayesian classification. In order to classify seismic events registrations, in this work BayesNet algorithm of WEKA was applied.

As it was already said, a BN is made up of two components: the network topology and the conditional probability tables.

WEKA BayesNet algorithm [1] let to define such components by means of the following parameters:

1) searchAlgorithm selects the method for searching network topology; we fixed it to K2.
2) Estimator selects the algorithm for calculating the conditional probability tables. We chose the SimpleEstimator algorithm.

In the next table, WEKA BayesNet parameters are summarized.

TABLE XI
BAYESNET PARAMETERS

estimator	SimpleEstimator
searchAlgorithm	K2

A lot of bayesian models have been produced by applying BayesNet algorithm, by changing the values of the next two parameters:

1) A = alpha parameter of the SimpleEstimator algorithm which sets the starting value for the calculation of conditional probability.
2) P = maxNrOfParents parameter of K2 algorithm which sets the maximum value of the number of parents of each node in the network topology.

The performances of the built models were calculated and compared.

C. Model Performance Metrics

The Data Mining models test was designed with the purpose of selecting models with high performance results of correctly recognizing dangerous seismic events (i.e. classifying magnitude) elaborating only the first registration of the earthquake, so that the whole elaboration time of the Early Warning System was appreciably reduced.

Performances of obtained models were estimated by using ad hoc metrics on the TEST dataset containing 262 records concerning all the first registration of earthquakes. Such metrics are traditionally the followings:

1) True Positive (TP) and True Negative (TN) rates.
2) False Positive (FP) and False Negative (FN) rates.
3) ROC and ROC Area (AUC).
4) Confusion Matrix.
5) Total Cost.

In particular the overall (total) cost performance metric is defined as follows:

$C = N_{TP}C(+,+) + N_{FP}C(-,+) + N_{TN}C(-,-) + N_{FN}C(+,-)$, for a binary class problem.

N_{TP} indicates the amount of positive cases correctly classified by the model, N_{FP} describes the amount of negative records misclassified as positive and so on. Moreover, $C(i, j)$ is the cost of classifying a record in the i-th class as a record of the j-th class.

The next 2x2 cost matrix was fixed following domain expert advices for evaluating the models built:

TABLE XII
COST MATRIX

	Positive	**Negative**
Positive	0	11.6
Negative	1	0

The cost C(+,-) = 11.6 of committing a false negative error was chosen taking into account the unbalanced distribution of "THRESHOLD_5" attribute target (Table VII): the cost of committing a false negative error is 11.6 times larger than the cost of committing a false alarm. In other words, failure to detect any positive example is just as bad as committing 11.6 false alarms.

D. BayesNet Applications

The algorithm parameters were calibrated based on the model performance results over the VALIDATION set. For the sake of clarity, we set P and changed A. The results of the obtained models were compared. It was selected the model whose metrics got the best values corresponding to P = 1, P = 2 and P = 3.

In the following boxes the results are shown. The first section of every box contains the testing results on VALIDATION set and the second one reports the testing results on TEST set. In addition, the topologies (DAG) of the networks (P = 1,2,3) are given.

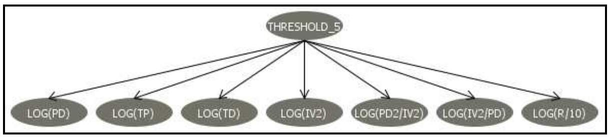


Fig. 11 Bayes Network with P=1

P=1 A=8.4

```
=== Evaluation on VALIDATION set ===

Correctly Classified Instances       1332             90.5506 %
Incorrectly Classified Instances      139              9.4494 %
Kappa statistic                        0.5577
Total Cost                           223.8
Mean absolute error                    0.0925
Root mean squared error                0.283
Relative absolute error               64.3528 %
Root relative squared error          105.8562 %
Total Number of Instances           1471

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   ROC Area   Class
0.93      0.097     0.447       0.971       YES
0.903     0.07      0.994       0.971       NO

=== Confusion Matrix ===

    a      b   <-- classified as
  106      8 |   a = YES
  131   1226 |   b = NO
--------------------------------------------------------------------------
=== Evaluation on TEST set ===

Correctly Classified Instances        219             83.5878 %
Incorrectly Classified Instances       43             16.4122 %
Kappa statistic                        0.498
Total Cost                            64.2
Mean absolute error                    0.1609
Root mean squared error                0.3893
Total Number of Instances            262

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   ROC Area   Class
0.938     0.178     0.423       0.959       YES
0.822     0.063     0.99        0.959       NO

=== Confusion Matrix ===

   a     b   <-- classified as
  30     2 |   a = YES
  41   189 |   b = NO
```

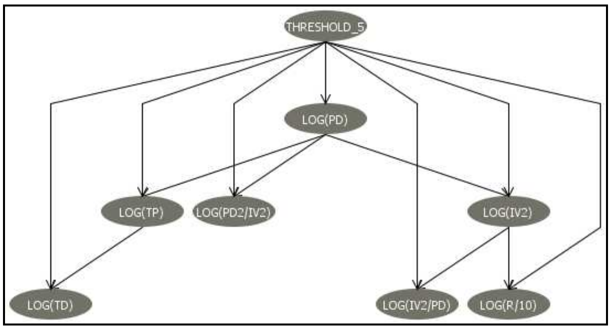


Fig. 12 Bayes Network with P=2

P=2 A=11.3

```
=== Evaluation on VALIDATION set ===

Correctly Classified Instances       1413             96.0571 %
Incorrectly Classified Instances       58              3.9429 %
Kappa statistic                        0.7027
Total Cost                           460.8
Mean absolute error                    0.0454
Root mean squared error                0.1734
Relative absolute error               31.5909 %
Root relative squared error           64.8546 %
Total Number of Instances           1471

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   ROC Area   Class
0.667     0.015     0.792       0.98        YES
0.985     0.333     0.972       0.98        NO

=== Confusion Matrix ===

   a      b   <-- classified as
  76     38 |   a = YES
  20   1337 |   b = NO
--------------------------------------------------------------------------
=== Evaluation on TEST set ===

Correctly Classified Instances        247             94.2748 %
Incorrectly Classified Instances       15              5.7252 %
Kappa statistic                        0.7618
Total Cost                            46.8
Mean absolute error                    0.0667
Root mean squared error                0.2044
Relative absolute error               36.7881 %
Root relative squared error           61.8782 %
Total Number of Instances            262

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   ROC Area   Class
0.906     0.052     0.707       0.979       YES
0.948     0.094     0.986       0.979       NO

=== Confusion Matrix ===

   a     b   <-- classified as
  29     3 |   a = YES
  12   218 |   b = NO
```

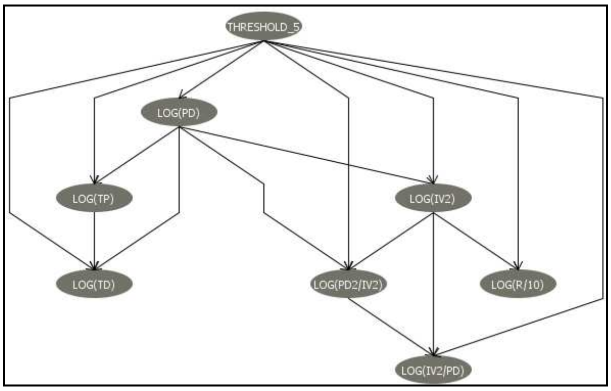


Fig. 13 Bayes Network with P=3

```
P=3    A=3.27

=== Evaluation on VALIDATION set ===

Correctly Classified Instances        1413            96.0571 %
Incorrectly Classified Instances      58              3.9429 %
Kappa statistic                        0.7027
Total Cost                            460.8
Mean absolute error                   0.0429
Root mean squared error                0.1659

Relative absolute error               29.838  %
Root relative squared error           62.0562 %
Total Number of Instances             1471

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   ROC Area   Class
0.667     0.015     0.792       0.983        YES
0.985     0.333     0.972       0.983        NO

=== Confusion Matrix ===

   a       b   <-- classified as
  76      38 |    a = YES
  20  1337 |    b = NO
--------------------------------------------------------------------------------
=== Evaluation on TEST set ===

Correctly Classified Instances        251             95.8015 %
Incorrectly Classified Instances      11              4.1985 %
Kappa statistic                        0.8211
Total Cost                            32.2
Mean absolute error                   0.0618
Root mean squared error               0.1947
Relative absolute error               34.0811 %
Root relative squared error           58.9282 %
Total Number of Instances             262

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   ROC Area   Class
0.938     0.039     0.769       0.984       SI
0.961     0.063     0.991       0.984       NO

=== Confusion Matrix ===

 a     b    <-- classified as
 30   2    |   a = SI
 9    221 |   b = NO
```

E. Models Comparison

The performances of the built models were calculated and compared.

In Table XIII the results of the three previous models are summarized.

In the next Fig. 14 the ROC curves of the best three models obtained are compared.

TABLE XIII
SOME COMPARATIVE METRICS

Nr	Parameters	Total Cost	AUC on TEST	Confusion Matrix on TEST set	
1	P=maxNrOfParents=1 A=alpha=8.4	64.2	0.95	30	2
				41	189
2	P=maxNrOfParents=2 A=alpha=11.3	46.8	0.979	29	3
				12	218
3	P=maxNrOfParents=3 A=alpha=3.27	32.2	0.984	30	2
				9	221

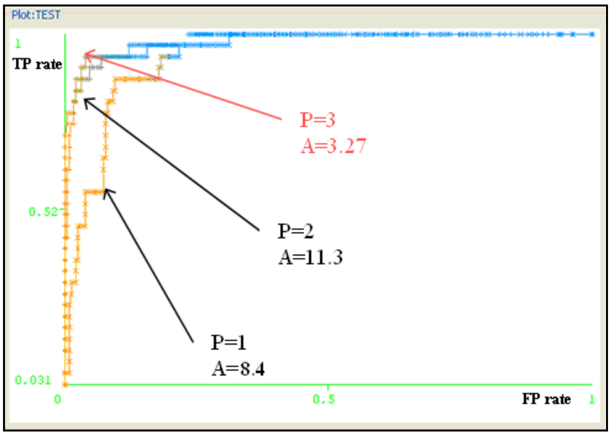


Fig. 14 Bayes ROC curves

As it is well known, the closer the curve is to the upper left corner, the better the classifier performs because the True Positive rate dominates over the False Positive rate.

In this case the classifier number 3, called RED classifier, with P=3 and A=3.27 offers the best results. Its corresponding AUC on the TEST set was higher and his total cost was lower.

F. Models Test on Irpinia Earthquake

On November 23, 1980, a powerful earthquake devastated the Irpinia area. Irpinia is a region of the Apennine Mountains around Avellino, a town in Campania, southern Italy about 40 km east of city of Naples. Measuring 6.9 on the Richter Scale, the quake, originated beneath the village of Conza, killed 2.914 people, injured more than 10.000 and left 300.000 homeless. This event produced vast damaging and strong amplitude shaking on a wide area. This event marked the beginning of quantitative seismic hazard assessment in southern Italy [6].

It is well known that there are no real seismograms for this great energy event because ISNet (par. III E) wasn't active in 1980. In order to test the obtained data mining models, synthetic seismic traces were used, that simulated the waveforms recorded by the ISNet stations. These synthetic seismograms are enclosed in 75 SAC files corresponding to 75 / 3 = 25 records (25 stations). The records were transformed accordingly to activities described in the Data Understanding and Preparation phases, and the seismic parameters were calculated from the first recording of the earthquake that exceeds the time-check seconds and they were set as input to the RED classifier which correctly classified the Irpinia event as dangerous. In other words, the record of the first synthetic registration of Irpinia earthquake was well classified in TRESHOLD_5 class as “YES”. In the following table XIV the 7 seismic parameters of the first registration are shown. The R distance is about 25 km. And the first time-check passed signal was registered after 4.3 seconds from the origin of the earthquake.

TABLE XIV
IRPINIA EARTHQUAKE FIRST RECORD

FIRST REGISTRATION OF IRPINIA SISMA	
LOG(PD)	-1.33458
LOG(TP)	-0.41132
LOG(TD)	0.166463
LOG(IV2)	-2.3263
LOG(PD2/IV2)	-0.34286
LOG(IV2/PD)	-0.99172
LOG(R/10)	0.409054

VI. DEPLOYMENT

The RED classifier was integrated within an EW system able to connect to a seismic monitoring sensor network using the most widespread seismic data format via TCP/IP protocol, to receive the data and process them in order to extract the physical indicators and evaluate the level of dangerousness of the running event just basing on the first registration of the earthquake. Flowing the synthetic data of the Irpinia 1980 earthquake into such EW system, very good results were obtained: the overall time interval from catching the event to the warning was 6.1 seconds and the probability of false alarm was less than 3%.

The EW system was enriched with advanced functionalities for the multidimensional analysis of historical seismic data, based upon data warehousing technologies.

The logical architectural view of the cited EW system is depicted in the following Fig. 15.

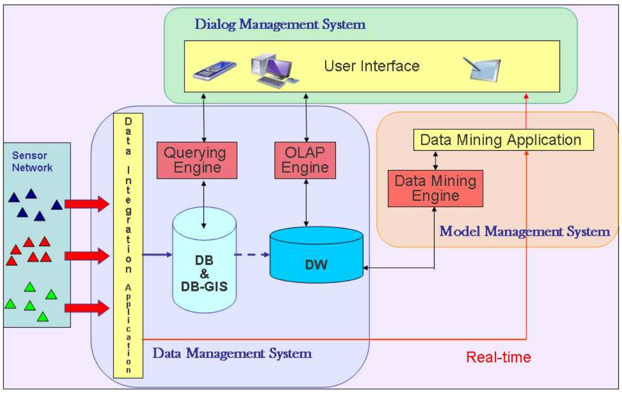


Fig. 15 Logical Architectural View of Seismic EW System

VII. CONCLUSION

In our knowledge, the application of Data Mining techniques for seismic early-warning purposes is not yet fully explored. Many future developments can be addressed starting from the creation of models based on more information coming from the reduction of the time interval allowed for passing the initial check, ending to the formulation of the estimation of the magnitude as a multiclass classification problem. In addition, other approaches can be set up to carry on the model phase: for example Neural Networks challenging the reduction of false alarms. Finally, it is worth to underline that one of the advantages assured by using Data Mining analysis methods was the availability of well-known missed and false alarms probability.

REFERENCES

[1] Bouckaert R.R., *Bayesian Network Classifiers in Weka*, The University of Waikato, September 1, 2004.

[2] Bouckaert R., Frank E., Hall M., Kirkby R., Reutemann P., Seewald A., Scuse D., *WEKA Manual for Version 3-6-2*, The University of Waikato, January 11, 2010.

[3] Elkan C., *The Foundation of Cost-Sensitive Learning*, Proceedings of the Seventeeth International Joint Conference on Artificial Intelligence, 2001.

[4] Gasparini P., Manfredi G., Zschau (Eds.), *Earthquake Early Warning Systems*, Springer, 2007.

[5] Han J., Kamber M., *Data Mining. Concepts and Techniques*, Morgan Kaufmann Publishers, 2001.

[6] Lancieri M., Zollo A., *Simulated shaking maps for the 1980 Irpinia earthquake, Ms 6.9: Insights on the observed damage distribution*, in Soil Dynamics and Earthquake Engineering 29, 1208-1219, 2009.

[7] Marketos G., Theodoridis Y., Kalogeras I.S., *Seismological Data Warehousing and Mining: a survey*, International Journal of Data Warehousing & Mining, 4(1), 1-16, 2008.

[8] Tan P-N, Steinbach M., Kumar V., *Introduction to Data Mining*, Pearson Addison Wesley, 2006.

[9] Witten H.I., Frank E., *Data Mining: Practical Machine Learning Tools and Techniques*, Elseiver, 2005.

[10] Zollo A., Iannaccone G., Convertito V., Elia L., Iervolino I., Lancieri M., Lomax A., Martino C., Satriano C., Weber E., Gasparini P., *Earthquake Early Warning System in Southern Italy*, in Encyclopedia of Complexity and System Science, Springer, 2395-2421, 2009.

[11] http://www.crisp-dm.org/

[12] http://www.cs.waikato.ac.nz/ml/Weka/

[13] http://www.iris.edu/software/sac/manual.html

[14] http://www.kik.bosai.go.jp/

[15] http://www.rissclab.unina.it/