Computational seismic algorithmic comparison for earthquake prediction

Sajjad Mohsin, and Faisal Azam.

faisal_azam@yahoo.com

***Abstract*—**Seismic data is generated in nature by the changes or movement of the earth crust. This data has evolutionary patterns. Since this data is based on time, a model can be formed to predict the future pattern. In this work we have focused on different statistical learning models to identify the potential seismic changes in the geography related to Pakistan. We used both deterministic and un-deterministic optimized algorithms to determine the future values. The results of different applied techniques show the possibility of future earthquakes in Pakistan region. This work also elaborates the comparative performance of statistical techniques for earthquake prediction. For this purpose, M8 and MSc algorithms have also been considered for critical overview.

***Keywords*—**Algorithms, Data Mining, Earth crust, OLAP, Time Series

I. Introduction

EARTHQUAKE data can be gathered from different World Wide Web sources and this data is heterogeneous in nature [1, 2]. This heterogeneity is because of measurement differences, recording sink and source distances, designs used for measurement and lot of other factors. This issue can be solved by merging different sources data into a single repository containing the metadata information about different seismic sources [3]. As this data has patterns related to the changes or movement of the earth crust, a time series analysis can be performed upon it. Difference between two consecutive occurrences of earthquakes (λ) can be normally distributed for some data sets and it can also have unspecified values. So we have chosen both normal and non-normal distribution algorithms in our experiment.

The analysis of seismic data mainly depend upon the nature of data being gathered, features selected for analysis, types of earth crust, change pattern in the past, nature of earthquakes after the change in patterns, and other associated factors that are developed or aborted on such changes. Data for analysis can be gathered from different seismic activities monitoring stations around the world [4, 5]. The main purpose of this work is to perform comparative analysis of the functioning of different statistical techniques on the specified earthquake data set and predict the probability of future earthquake magnitude and time. Some data quality issues regarding the analysis of earthquake data from different seismic sources have also been highlighted.

II. Earthquake Analysis

A. Data Sets

Earthquake prediction is possible using large scale data from any recording station. Methods used for prediction purpose are based upon the generic type and nature of seismic data. A typical data attributes contain ***unique identity of each event***, ***information about originating point***, ***latitude***, ***longitude***, ***depth of earthquake***, ***measurement unit***, and some other information as shown in figure 1. Seismic data recordings at different stations may vary in nature and measurement type [2]. For example if a station is close to the seismic zone then its computation is in *Local Magnitude (ML),* if a station is 2000 Km away from the recording station, then its computation is in *MB*. Few stations placed records in *MS,* that is the surface wave computations traveling along the of the earth surface. Moment Magnitude *MW,* is directly related with the physics of the earthquake sources.

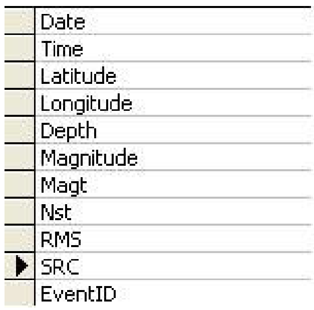

Figure 1: Seismic data with multiple features

Earthquake data can be mined using supervised learning techniques and results can be predicted [6]. This problem is not as simple as it seems to be. Mainly earthquake is originated from a single source is measured by different stations around the globe. When we are studying change patterns in earthquake with respect to change patterns in the earth crust, there may be different recoding stations and problem can be much bigger and complex. There exist many techniques to forecast future earthquakes in a specified region. In this regard, a famous M8 and MSc algorithms have been overviewed hereunder.

B. The M8 Algorithm

M8 algorithm is used to predict the earthquakes of 8 or above 8 magnitudes worldwide [7]. This algorithm uses specific diameter range in earthquake region. It removes the aftershock affects of the earthquake from the data set. It considers origin time, magnitude, depth, and number of aftershocks after some days. This sequence of earthquake data is then normalized by removing the lower earthquakes.

After minimizing the record sets by normalizing the sequence or removing the lower intensity quakes, we find the running averages queries that are the sliding window with specific magnitude range. These different averages show the different intensity of earthquake flows. We can now identify deviation or trend of earthquake over different period of time using these different moving averages.

In these moving averages following things are included,

a) Number of main shocks,
b) Deviation values for a long term trend
c) Cumulative number of shocks
d) Linear concentration is calculated as the ratio of average diameter of source to the average distance between the shocks.
e) This is a special scenario case in which earthquake sequence is considered in a specific time window and of specific magnitude.
f) The average number of earthquake for calculation purpose is considered to be greater than 10 and less then 20. Means only quakes in moving average function with count greater then 10 and less than 20 are considered.
g) Higher resultant values of earthquake magnitudes are then identified.

C. MSc Algorithm

MSc or *The Mendocino Scenario* algorithm (Kossobokov et al., 1990) was designed to find the seismic prediction of the earthquakes with 7.2 or above magnitude [7]. MSc working algorithms has following main steps,

a) The larger area or territory U is divided into small squares of s*s size in a two dimensional coordinate space.
b) In each such square, number of earthquakes is calculated using short time sliding window function. We sequence all such time based windows using an array *k*.
c) Using this technique we actually divided global time space of earthquake records into two dimensional small boxes containing the value of k. These boxes are of discrete size.
d) *Quiet* boxes in this space are identified, they are ruled out using a specified formula fore example our required boxes of interest have Q percentile of earthquakes. This will shed the load of such data which is not required.
e) We can now form the cluster of earthquakes with specified percentile and others of quiet boxes and then project such areas for statistical analysis.

D. Statistical Significance

Structure of the earth layers and parameters in the terms of seismic waves can be determined using statistical approaches of data mining [8]. In general four types of analysis can be performed for earthquake prediction.

a) Earthquake data sets gathered in time points
b) Earthquake data sets gathered in time intervals
c) Earth curst layers gathered data in time points
d) Earth crust layer data sets gathered in time intervals

Time points data sets means recording of seismic data in specific time points. In other words we are taking different snap shots of data with respect to time. In the time interval case we are intended to capture the evolution of seismic data sets over the time. This scenario can be illustrated from the case that we are interested in recording the duration of an earthquake and how certain parameters of the phenomenon vary throughout the time interval of its duration.

Layers mean the dispersion of different geographical regions represented in the form of data sets. The combination of data about layer and time point results in the snapshot of the data sets. In layers data set with respect to time intervals, that we are willing to model and analyze the seismic changes in different layers in different period of time in order to analyze and predict the bigger earthquake. If we achieve this target we can build a future knowledge base system to predict the seismic wave's propagation with respect to the quantum physics. On achieving this target future layers can be forecasted.

III. DATA QUALITY AND PREPROCESSING

When the large amount of data is being used for earthquake data analysis, we can not directly apply algorithms to analyze and data mine as we have to handle the data quality issues. While studying an open database for earthquake data, we came up with the several problems of data quality matters, that included,

i) Handling of missing or null values in seismic data sets
ii) Handling of incorrect values like negative values or misleading calculations.
iii) Time format correction handling problem that is conversion of different time formats into a single standard format for computations. We used military time conversion function to convert different time format into a unique standard. Earthquake data is then sorted based upon the time of its occurrence.
iv) *MB, MS, MW, ML* data handling. We used records from unique online source.

v) Duplicate entries handling. As same recording may be repeated in the data sets. Also keeping in view the after shock effects of earthquake which may be less in magnitude.

vi) Features selection to remove unnecessary data fields for analysis. We used *latitude*, *longitude*, *time*, *depth* and *magnitude* in our work.

Beside quality matters, several preprocessing steps are involved to quantify and extract useful knowledge from the data sets. Here is presented the summary of different techniques adopted for data preprocessing phase.

i) Delta time ΔT is calculated. This is the measured difference of two successive earthquake events. Such that

$$\Delta T = T(N) - T(N-1) \quad (1)$$

Where N is the current record number, the resultant ΔT is shown as lambda (λ). This λ has been used in different algorithms for analysis. We need to find the cases in which λ is roughly constant over some records. We can also apply moving average function with sliding window effects.

ii) Modern OLAP [9] tools use accumulative function to sum up the series of data. We added time instances and developed a customized accumulative function to sum up time. Idea is to sum up the series as follows,

$$t1, t2, t3, \ldots\ldots, tn-1, tn, tn+1 \quad (2)$$

Where t is the time instance, this series will become,

$$t1,\ t2 + t1,\ t3 + t2 + t1, \ldots\ldots.., \quad (3)$$

This is used to data mine the increase in the time trends of earthquake instances.

iii) It is highly inefficient to apply data analysis techniques on all the data set at once. It cost high computation requirements and many other associated problems. Based upon the requirements and the specified usage of different algorithms, features were selected from the existent database. Figure 2 shows different features and their values

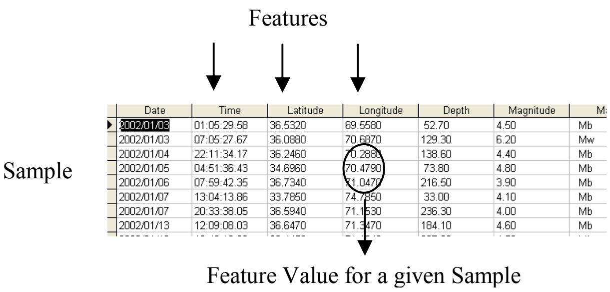

Figure 2: Earthquake data set with features

IV. Algorithms tested for earthquake data sets

In the experimentation task, we used following algorithms to find the probability and hidden patterns using statistical learning techniques. They are,

a. Binomial distribution
b. Erlang distribution
c. Exponential smoothing
d. Poisson distribution
e. Same Birthday Paradox
f. Linear Regression

The brief working and result output these algorithms based upon the data sets selected specifically for Pakistan region is described. Pakistan is situated between latitude 24 and 37 degrees North and longitude 62 and 75 degrees East [10]. This work has considered seismic records for Pakistan, from the Arabian Sea to the Himalayas and also the Kahsmir valley. This work also predicts the possibility of different magnitudes of earthquakes after specified interval of time.

A. Binomial distribution

We have a data set consisting of time and magnitude of earthquake. We are interested in finding the probability of the occurrence of a specific level earthquake in the existent data set. Our specified magnitude can occur at any time instance t randomly distributed across the data. For this purpose we used binomial distribution. The binomial distribution [11, 12, 13, 14] each successive experiment gives probability p and these experiments are independent of each other. We apply binomial on earthquake data records.

For example, the probability of earthquake above 7 is found to be 5% in the existent data set contained 26481 earthquake records. We considered what will be the probability of above 7 or above earthquakes in 26481 earthquakes record sets on random basis or how likely is it that we get 30 or more above 7.0 earthquakes?

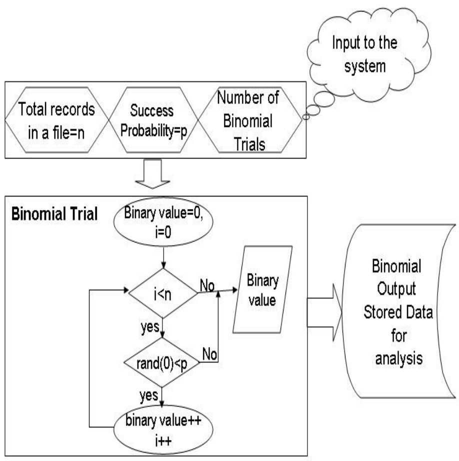

Figure 3: Work flow of binomial trial experiment

The number of earthquakes above 7.0 we pick is a random variable X which follows a binomial distribution with $n = 26481$ and $p = 0.05$. We are interested in the probability $\Pr[X \geq 30]$.

In general, if the random variable X, which is the probability of above 7 earthquakes follows the binomial distribution with parameters n, which is 26481 number of records in this case, and p which is 7%, we write X ~ B(n, p). The probability of getting exactly k successes is given by the probability mass function,

$$f(k;n;p) = (n \ldots k)\, pk(1-p)n-k \qquad (4)$$

For k=0,1,2,…..,n and where

$$(n \ldots k) = n!/k!*(n-k)! \qquad (5)$$

is the binomial coefficient. We want k successes (k) and n − k failures. However, the k successes can occur anywhere among the 26481 records, and there are C(n, k) different ways of distributing k successes in a sequence of n trials [11]. This means that k successes or the magnitude of earthquake above 7 can occur at the first time or it may occur at any other time during the earthquake sequence. We can not determine the exact location or time sequence.

We gave sample delta time of 26481 records to the binomial trial algorithm. We found that the probability of above 7 earthquakes and aftershocks to occur is 7.8%. This probability is very near to the real time occurrences.

B. Erlang Distribution

Suppose that for some interval of time, earthquakes happen at some average rate. For example after a bigger earthquake usually after shocks happens at some continuous rate. Erlang distribution [11] can be used to model the time between two independent earthquake events. Typical working flow chart of Erlang distribution is shown in the figure number 4.

We found a typical scenario in our accumulated time data set. It has been observed that average rate for earthquake record set for a typical time period was 12 days. On using Erlang distribution upon the 4445 experimented records upon which average rate were found, we got the result that minimum after 360.376749 (about 1 year) days their exists a Erlang probability of earthquake and aftershocks.

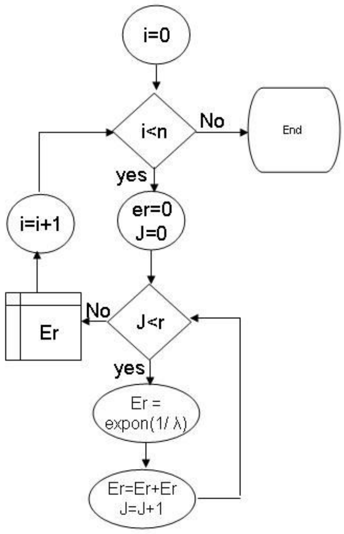

Figure 4: Work flow of Erlang distribution in our experimentation

C. Exponential smoothing

Exponential smoothing [11] we used to predict the next value in the time series of earthquake records. If we measure the probability of the time of earthquake by simple moving average function, the weights assign to each observation will be the same. This means that we are least interested in the earthquake records after the latest physical changes of the earth crust. In order to keep in view the latest changes, we need to assign more weight to the latest time series values. Exponential smoothing fulfills this requirement by assigning more weight to the latest values and less weight to the old values.

Let us suppose that we want to predict Xn+1, which is a weighted sum of the past observations, then

$$Xn+1 = c0X_n + c1X_{n-1} + \ldots$$

The set of weight given to the past observations is decreased by the constant ratio. C is the constant ratio in the above diagram. Following is presented the workflow of exponential smoothing algorithm used in our experiment.

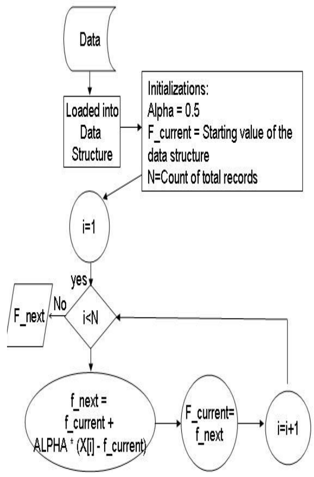

Figure 5: Work flow of exponential smoothing algorithm in our experimentation

We gave accumulative time series data as an input to the exponential smoothing. As exponential smoothing gives more weight to the recent moving averages, so we can get good forecast with respect to the earthquake data. We got next accumulative exponential smooth value to be 3.7 (approx.) years from the selected samples of our record set. This show the next accumulative time stamp value in days of the earthquake. We can then model to see wither earthquake occurred after the 3.7 years (λ) of the past value or not. Based upon the accuracy of our results we can change the noise distribution function or alpha.

D. Poisson Distribution

Let us suppose that the earthquake events are occurring with the fixed rate independent of the time since the last event, we can use Poisson distribution [11]. Suppose that an earthquake occur λ times over a specified time interval, the probability of exactly x occurrence of earthquake will be,

$$F(x, \lambda) = (\lambda^{x} e^{-\lambda}) / x! \quad (6)$$

λ is a positive real number, equal to the expected number of occurrences that occur during the given interval. For instance, if the events occur on average every 4 months, and we are interested in the number of events occurring in a 10 months interval, we will use as model a Poisson distribution with Lambda = 10/4 = 2.5.

We got the Poisson distribution probability of 30% for the earthquakes. This means that if the lambda is 2.5 for below 4 earthquakes, after application of Poisson distribution function we got that there is 30% probability of earthquake and aftershocks within our specified interval of time.

E. Same birthday paradox

For example we have the set of 26481 records, using same birthday paradox [15]; we can find the probability of two or more earthquakes of the same magnitude. Later we can model these same magnitudes to find the highest threat areas.

In the result we got there is 78% probability of two earthquakes with the same magnitude.

F. Linear Regression

We are interested in the magnitude of earthquake with respect to the depth of its occurrence. We can perform linear regression for this typical scenario. A regression is actually a running series of means of the expected value of Y for each value of X and is calculated from the following equation [11].

We tried regression over 13241 records. This thing is interested to note that the higher the depth of earthquake lower will be the intensity of its occurrence. So our linear regression model can predict the next value for the depth with respect to the magnitude of earthquake

V. Discussion on Results

M8 and MSc algorithms are designed to forecast earthquake in specific source region. If the source region is distributed among different earth crust layers, we need a bigger scenario to handle such issues. There exists number of ways to analyze and predict earthquakes with respect to time points and time intervals. Our chosen methodology is different from M8 and MSc algorithms. In spite of predicting earthquake using a specified algorithm at once we used step by step approach. We analyzed the results of different algorithms. Based upon this analysis we can determine our best prediction goals and further optimize our algorithms.

M8 algorithm considers the earthquake records of high magnitude. For this purpose it draws the diameter from the source to measure the wave's propagation recording of earthquake records. It then predicts the earthquake for next five years by generating the *Alarm.* This thing is interested to note that it takes the moving averages over the sliding window function to match the patterns of different time period earthquakes. We used exponential smoothing, which gives more weight to the current quakes and less weight to the past quake. We can implement a system to judge the performance of both the systems over the time.

When the result of exponential smoothing does not match with the real time value we simply need to adjust the value of alpha or noise distribution function. With accurate measure of alpha variance, good prediction results could be obtained. This process is done in iterative way so that we get our optimized target.

M8 algorithm counts the number of aftershocks and treats it as a whole. Means it does not steer into the details of after shocks and its results. Although the proof of working of M8 algorithm is much better yet this work considers that if we keep in view the specific structure or changes in the earth crust after certain sliding window affect, we might better suggest a system which is called a *knowledge base system* for earthquake in future. This system will contain the probability results and knowledge about the change in the earth crust with respect to this probability. Using such scenario we better model real time future changes. Moreover when we desire to find the earthquakes with respect to the change in earth crust, we need a bigger paradigm to study. Moving averages of different stations can be converted into a cube to analyze changing regional patterns.

MSc algorithm seems to generate best result as it divide N task into n1, n2, n3,……,k number of instances in a two dimensional vector space. Yet in our view, when there is a bigger earthquake, many smaller after affects seems to occur in specified period of time. These short coming affects are actually due to larger quake. So if we model both separately, we might get good probability for larger quakes, but smaller changes might be missing. Also earth crust formation might be changing with the lot of smaller changes over the time. These scenarios seem to be best for current real time prediction, when we will move onto the study of earth crust changes and movement with respect to time intervals; we need a bigger scenario which should monitor each and every thing.

Here is presented the brief working of different statistical technique.

Algorithm	Working Type	Result output	Comments
Binomial distribution	• Discrete • X ~ B(n, p) *where* • X is the result, n is number of records and p is probability.	Number of times a specified quake can happen in n number of records.	• All the earthquake data is provided at once in a file. • Rand() function is used and performance highly depends upon its output and number of Binomial trials

Erlang distribution	• Continuous • A uniform λ is found in the data set for some specified record set using sliding window.	Number of days after which earthquake can occur after provided the specified intervals data file.	• Best for studying after shock affects as they often happen to be with some uniform average rate.
Exponential smoothing	• Works by giving more weight to current data and less weight to the past data. • $Xn+1 = c0X_n + c1X_{n-1}+$ • N+1 value of X is predicted.	Gives the next time sequence in which a quake can happen	• A good technique for monitoring and predicting real time changes as past changes are given less weight and latest are more so working might depend upon the use of weight function C.
Poisson distribution	• Discrete • Uses function $f(x, \lambda)= (\lambda^x e^{-\lambda})/x!$ • λ is the time interval provided, x is the probability of number of earthquakes	Gives the probability percentage of earthquakes on giving λ value and record set	• A good algorithm if we are known with the uniform λ distribution in our records.
Same birthday paradox	• A simple probability function	Gives the probability of specified magnitude earthquake in record set.	• It is used to find same occurrence of specified magnitude earthquakes
Linear regression	• Formal statistical technique	Gives the probability value of depth or magnitude	• Best if linear relationship exists between variables
Auto-Correlation	• Identify non-random data walks in the existent data set	Gives the coo-relation factor	• Good for earthquake data set as data have different non-random data .
M8 Algorithm	• Uses moving average function on the data set and match the pattern of different results.	Gives the probability of earthquake in specified source with defined range	• Used for predicting above 8 magnitude earthquakes.
MSc Algorithm	• Divide N global problem in n1, n2,…,k smaller problems and then cluster together same smaller problems.	Gives the probability of earthquake in specified source with defined range.	• Used to predict above 7 earthquakes.

Table1: Algorithms for earthquake data set analysis

VI. Conclusion

Statistical learning techniques are utilized in variety of ways to predict the rate of earthquake probability. This work used seismic data sets with respect to its occurrence, in time intervals and time points. Successful occurrences of earthquake records can be predicted using statistically significant algorithms and applying data mining methodologies to an existent database. Data quality is the prime issue in the seismic monitoring as it may lead to good quality results. This has been observed that for monitoring a change in the physical structures after the earthquake needs more adequate methodology. Techniques from evolutionary programming or dynamic programming can later be utilized in this regard. An intelligent algorithm is needed to be developed which could measure the earthquake keeping in view all possible patterns of data sets. In the data collection phase an open architecture is needed to normalize data. We can perform much stronger analysis as actual standardized recordings from all around the world could have been available for the development. Applying the general statistical techniques to earthquake data, we found that the probability exists for different record set patterns to be repeated in future.

REFERENCES

[1] N.R.C.(.P.O.D.P.I.) Seismology, Effective use of earthquake data, National Academies, 1983
[2] S. Stein and M. Wysession, An introduction to seismology, earthquakes, and earth structure, Wiley-Blackwell, 2003
[3] R. Kimball and M. Ross, The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling, Wiley, 2002
[4] I.R.I.F. Seismology, IRIS newsletter, IRIS Consortium, 2006.
[5] US geological survey hazards program website. earthquake.usgs.gov/ Retrieved 2010-2-10.
[6] G. Giacinto, R. Paolucci, and F. Roli, "Application of neural networks and statistical pattern recognition algorithms to earthquake risk evaluation," Pattern Recognition Letters, vol. 18, 1997, pp. 1353–1362.
[7] V.G. Kossobokov, L.L. Romashkova, V.I. Keilis-Borok, and J.H. Healy, "Testing earthquake prediction algorithms: statistically significant advance prediction of the largest earthquakes in the Circum-Pacific, 1992–1997," Phys. Earth Planet. Inter, vol. 111, 1999, pp. 187–196.
[8] D. Pfoser and N. Tryfona, "Requirements, definitions, and notations for spatiotemporal application environments," Proceedings of the 6th ACM international symposium on Advances in geographic information systems, Washington, D.C., United States: ACM, 1998, pp. 124-130
[9] S. Chaudhuri and U. Dayal, "An overview of data warehousing and OLAP technology," SIGMOD Rec., vol. 26, 1997, pp. 65-74
[10] Geological Survey of Pakistan, Geological Information, Government of Pakistan, http://www.gsp.gov.pk
[11] Y. Dodge, The Concise Encyclopedia of Statistics, Springer, 2008
[12] Y.Y. Kagan, "Statistical distributions of earthquake numbers: consequence of branching process," Geophysical Journal International, vol. 180, 2010, pp. 1313–1328.
[13] H. Ryu, J.K. Kim, and J.W. Baker, "A Probabilistic Method for the Magnitude Estimation of a Historical Damaging Earthquake Using Structural Fragility Functions," Bulletin of the Seismological Society of America, vol. 99, 2009, p. 520.
[14] D. Schorlemmer, J.D. Zechar, M.J. Werner, E.H. Field, D.D. Jackson, and T.H. Jordan, "First results of the Regional Earthquake Likelihood Models experiment," Pure and Applied Geophysics, 2009, pp. 1–18.
[15] M. Mitzenmacher and E. Upfal, Probability and computing, Cambridge University Press, 2005.

Dr. Sajjad Mohsin is an illustrious researcher and practitioner in the field of Computer Sciences for over two decades, with credentials as a scholar and researcher of remarkable intellect and commitment He acquired his PhD from MIT, Japan in 2005. During his academic research career, spanning well over a decade, he has published/presented over 30 research papers in prestigious international Journals/conferences and is the first Impact Factor faculty member of CS CIIT. He is the approved HEC Pakistan PhD Supervisor, supervising 22 scholars of MS and PhD, is a former member of Australian Computer Society and is the recipient of MONBUSHO Scholarship thrice, for Research, MS and PhD from Ministry of Education, Japan in 1998 and 2002 to 2005.

Initially, in Japan, his research interest was in the applications of Neural Networks for Pattern Recognition in Electrocardiogram (ECG) and also developed a novel methodology of using Genetic Algorithm for optimization of Fuzzy Model (FCRM), consequent parameters without fine tuning and was able to get substantial results. Lately he has switched to the field of Computational Modeling of Cognitive Process of the Brain and in this regard has done research in employing the connectionist model, to compute the cognitive processes in memory functioning which is comprised of information processing, storage, and recall and also how brain derives new information based on existing information. He is now interested in applying AI Techniques in Seismic Data.

Faisal Azam is a PhD student at COMSATS Institute of Information Technology, Islamabad. He has background of research and development spanning over 10 years. He planned and implemented various software and research projects during his carrier. His recent development work includes the Pakistan's first Ranking System of Universities by Higher Education Commission. After realizing the true developmental issues and problems regarding ranking system, he performed the comparison of his implemented ranking system with other systems. This comparative research work has been published in INIT conference at Cyprus. Seismic forecasting and prediction remains his core area of interest. Currently he is working on agent development framework and optimization techniques related to bio-inspired computing.