Fuzzy Pattern Classification of Strong Ground Motion Records

Arzhang Alimoradi, Shahram Pezeshk, Farzad Naeim, and Hichem Frigui

Corresponding (First) Author:
Arzhang Alimoradi, Ph.D.
Seismic Research Engineer
John A. Martin & Associates, Inc.
R&D Department
1212 S. Flower Street, 5th Floor
Los Angeles, CA 90015
U.S.A.
Voice: (213) 483-6490
FAX: (213) 483-3084
Email: arzhang@memphis.edu

1 Ph.D., Seismic Research Engineer (formerly graduate student at The University of Memphis), John A. Martin & Associates, Inc., Los Angeles, CA
2 Ph.D., P.E., Emison Professor of Civil Engineering, The University of Memphis, Memphis, TN
3 Ph.D., S.E., Esq., Vice President and General Counsel, John A. Martin & Associates, Inc., Los Angeles, CA
4 Ph.D., Assistant Professor of Electrical and Computer Engineering, The University of Memphis, Memphis, TN
Classification of earthquake strong ground motion (SGM) records is performed using fuzzy pattern recognition to exploit knowledge in the data that is utilized in a genetic algorithm (GA) search and scaling program. SGM records are historically treated as “fingerprints” of certain event magnitude and mechanism of faulting systems recorded at different distances on different soil types. Therefore, databases of SGM records of today present data of complex nature in high dimensions (many of the dimensions—or SGM parameters in time and frequency domain—are presently available from different archives). In this study, simple ground motion parameters were used but were combined and scaled nonlinearly such that the physical properties of the data could be preserved while reducing its dimensionality. The processed data was then analyzed using fuzzy c-means (FCM) clustering method to explore the possibility of meaningfully representing earthquake SGM data in lower dimensions through finding subsets of mathematically similar vectors in a benchmark database. This representation can be used in practical applications and has a direct influence on the processes of synthesizing ground motion records, identifying unknown ground motion parameters (e.g. soil type in this study), improving the quality of matching SGM records to design target spectra, and in rule generalization for response. The results showed that the stochastic behavior of earthquake ground motion records can be accurately simplified by having only a few of motion parameters. The very same parameters may also be utilized to derive unknown characteristics of the motion when the classification task on “training” records is performed carefully. The clusters are valid and stable in time and frequency domain and are meaningful even with respect to seismological features that were not included in the classification task.

**Keywords:** Strong Ground Motion, Fuzzy Pattern Recognition, Seismic Design, Unsupervised Classification, Ground Motion Scaling
1. Introduction

Seismic design is traditionally carried out (for most types of common structures) with equivalent lateral static loading or by modal spectrum analysis. Availability of high performance computing and recommendations of the new seismic guidelines (e.g. [FEMA-350, 2000]) has made nonlinear response history analysis of structures a practical method for design and evaluation of modern structures with complex topology and functionality under extreme loading scenarios. One important question that challenges design engineers in performing time history analysis is often the criteria for selecting SGM records as input to the mathematical model of the structure to estimate structural demands. Exponential growth in the size and number of the available earthquake ground motion records and databases around the globe and on the Internet have created a research opportunity for data mining and classification towards better estimation of hazard. Recently, Naeim et al. [2004] developed an optimal solution using Genetic Algorithms (GAs) to select and scale a set of near optimal and realistic ground motion records in a large database with functionality to match a given target design spectrum. This study is a continuation of Naeim et al. [2004] work. The procedure outlined in Naeim et al. [2004], in general, results in a set of time histories of earthquake records with different epicentral distances, site characteristics, and magnitudes. Therefore, it is prudent to extend the methodology to add more constraints (such as de-aggregation of hazard, site-soil characteristics, and regional tectonics) to the search and scaling procedure. This may result in a more accurate estimation of hazard provided that the nature of expected seismic input motion at the site is understood. Even in those cases, since there is no guarantee that all possible causitive faults in a region are known, one may want to include diverse records in the input design bin due to random nature of ground motion generally expected. There has long been a consensus amongst researchers and practitioners that better representation of hazard can be obtained by including representatives form different types of motion that may be experienced at a site. In either cases, whether to have a “conforming” set of motions with site parameters and appropriate scaling factors to match the site’s design spectrum, or to have a “non-uniform” set of motions, the first question to answer is whether or not there are patterns of similar records in the database so the task of search could be performed more efficiently in one cluster of similar data or on any subset of few clusters. This can be addressed in the realm of pattern recognition, which is the technique of identification and classification of similar data in a set. Pattern classification was originally developed and successfully applied in the areas of system sciences, speech and image recognition. It is one of the most appropriate analytical procedures to carry on this task with.

In this paper, we studied different pattern recognition techniques (supervised and unsupervised¹) and applied them to a benchmark database to find meaningful clusters of

¹ Supervised methods of pattern recognition use a training sample to identify "test data" into groups through learning the statistical distribution of a number of distinct patterns. On the other hand, unsupervised methods
similar data and to identify the unknown characteristics of ground motion records important for design. This has three major advantages:

- Searching in small clusters can save computational time in large databases,
- Having the database classified enables the user to have more choices when it comes to selection of data characteristics to be included in the design bin, and
- By studying statistics of complex data, inference can be made in the form of rule generalization to estimate unknown data properties based on the available information.

Soft computing, originated mainly by the pioneering work of Lotfi Zadeh in 60’s to model the human mind’s ability to “exploit the tolerance for imprecision, uncertainty and partial truth to achieve tractability, robustness and low solution cost,” has now received considerable application in structural and earthquake engineering practice [Foley et al., 2003], [Alimoradi et al., 2000], [Pezeshk et al., 1999], [Adeli, 1999], [Ghaboussi et al., 1998], [Zadeh, 1965] and [Zadeh, 1974]. This is due to the fact that the nature of seismic design, by itself, is involved with processing of uncertain data and uncertain processes. The dimensions of today’s analytical models, especially in the area of nonlinear response estimation of complex dynamic systems, make it often difficult for the conventional “hard-computing” methodologies to efficiently meet all the real-world constraints necessitated by large-scale problems. Many of the principal constituents of today’s soft computing technology, such as fuzzy logic, neural network theory, and evolutionary computation are studied in this research to balance between computational cost and analytical accuracy of the ground motion selection and representation problem.

A brief description of the data dimensions, properties and feature selection is followed by data classification techniques used, cluster validation, and a discussion of the results.

2. Classification of Design Input Ground Motion Records

Bonnin et al. [1991] used SPARS (a syntactic pattern recognition scheme) to algorithmically distinguish between Californian, Japanese and Italian ground motion waveforms (from near field main shocks). Due to the lack of ground motion data in the Mediterranean region, it was hoped that by doing so, similar data from other parts of the world could be used in seismic design of structures in the Mediterranean region. Although the features used in the Bonnin’s study (logarithms of energy and zero-crossing rate) were possibly adequate but the high error rate obtained (in the range of 15% to 25%)

classify data into clusters based on some statistical or distance measures such as Mahalanobis or minimum variance rule without needing a training sample.
could be attributed to the relatively small size of their training set (62 three-component records).

Celebi et al. [1993] identified the parameters affecting ground motion amplification during the Loma Prieta 1989 earthquake (M = 7.1) using supervised pattern recognition. An area of study was chosen in the San Francisco peninsula between the Santa Cruz mountains and the Santa Clara basin with San Andreas fault running through it. This area was divided into 133 squares with known topographical and geological properties. Recorded ground motion data in the region, and azimuth and distance of each element in the test area to the rupture zone of the Loma Prieta and San Andreas fault were used for training of CORA-3 (a classifier developed by the academy of sciences of Russia). By studying classification results, Celebi et al. [1993] identified the most significant parameters in amplification of earthquake strong motion records as elevation, slope, distance to the San Andreas fault, predominant topographic landform, and predominant soil type. Although the training and the testing sets in their study were relatively small but careful selection of features and training made the procedure successful.

In an effort to characterize the motion in time and frequency domain, Suzuki [1984] used patterns of Running Power Spectrum Density (RPSD) to describe 30 earthquake records for classification and simulation. Classification was performed using the Complex Threshold Method based on the number of dominant peaks, peak shape, and spatial relation between the peaks. When carefully used, RPSD was shown to be a promising tool in representing nonstationary behavior in time and frequency domains.

2.1. Pattern Recognition (PR) Methods

Data classification methods are generally divided into supervised, unsupervised, and reinforcement learning (learning with a critic). Supervised methods “incorporate information from learning samples in the design of a classifier” [Duda et al., 1999] by understanding statistical distribution of a number of distinct patterns. Having class-conditional densities established, new entries can be associated to patterns upon checking their features by the classifier. Bayesian decision theory, in most cases, is used to incorporate the priori knowledge in the decision-making process. In learning class probability density functions, if the form of the underlying function is known, parametric methods such as Maximum Likelihood Estimation (MLE) may be used to derive the parameters of density function. On the other hand, in cases where the form of class-conditional density is multi-modal or unknown, non-parametric methods such as k-Nearest-Neighbor (KNN) are used to derive the class-density functions. Whether parametric or nonparametric, performance of all supervised PR methods depends on the size and quality of learning set; whereas in unsupervised PR, data is clustered as is presented to the classifier based on some statistical or distance measures such as Mahalanobis or minimum variance rule. Reinforcement learning uses a critic to provide
binary feedback when learning data is presented to the classifier. In this study, we used an unsupervised method, Fuzzy C-Means (FCM), for general classification of the database but to examine the distribution of some additional features in the resulting classes a supervised nonparametric method, (KNN), was used for validation purposes.

2.1.1. Data Properties

There has been a sustained effort in characterizing ground motion in time and frequency domains in the past few decades. Many different parameters of SGMs are proposed and studied [Kramer, 1996]. The SGM characteristics, or data features, can be taken any parameter or combination of parameters in time or frequency that can describe a particular behavior of data. An early attempt in data classification of a large database of SGM records for design was made by Naeim and Anderson [1996] where 1470 horizontal and 527 vertical components of mainly western U.S. events from 1933 to 1994 were processed in time and frequency domain to determine and rank their damage potential. Besides those commonly used in identifying an event, data parameters such as peak acceleration, peak velocity, peak displacement, peak incremental velocity, peak incremental displacement, and bracketed duration were reported from time domain processing for each record. From frequency domain, quantities such as elastic response spectra, effective peak acceleration, effective peak velocity, inelastic response spectra, elastic input energy spectra, and hysteretic energy spectra were computed. In this study, we used the database of the horizontal ground motion components developed by Naeim and Anderson [1996].

2.1.2. Hypothesis

Although the present state of knowledge treats the process of earthquake generation (and properties of signals temporally and globally) as non-deterministic, non-stationary random but shall enough accurate data parameters be available to represent all the natural processes involved (such as tectonics of the region and focal mechanism, background seismology, geological and topographical features of different wave propagation media, and local geotechnical site effects) then it would be possible to physically classify the nature of motion in few distinct types. Some of the practical benefits of such classification are mentioned at the end of this paper but the main advantage may be a new way of utilizing earthquake ground motion records in synthetic earthquake motion generation.

The criteria used in this study for feature selection and class description was based on some simple observations on the past recorded ground motions and the fact that SGM records are in general:
• Narrowed in time domain (as a spike) or are pronounced intensively during a long duration of shaking, and/or
• Wide-band or narrow-band in frequency domain, and/or
• Have large pulses of displacement accompanied by rapid forward/backward velocity pulses or are of a far-field type.

It is assumed that these three criteria, in essence, summarize many of the aforementioned characteristics of earthquake motion generation and propagation process. For example, contribution of low frequencies and wide-band nature of the response spectra of motions recorded on top of soft soil deposits is now a well-recognized fact. Besides generating insight into earthquake ground motion processes, such classification makes it possible for design engineers to have more control on the type of data they use out of a database.

The following parameters are available for the benchmark problem and are used in this study to describe classes:

• Peak Ground Acceleration (PGA),
• 5%g bracketed duration, ([D]),
• Effective Peak Acceleration (EPA) and Effective Peak Velocity (EPV) - (average response spectral ordinates in selected period bands; i.e., 0.1-0.5 second for EPA and about 1.0 second for EPV [ATC-3-06, 1978], and
• Maximum Incremental Velocity (IV) and maximum Incremental Displacement (ID) – sudden changes in velocity and displacement due to existence of near-source pulses [Anderson and Bertero, 1987].

The general behavior of these parameters over the whole database is shown in Figure 1. Although the behavior of a single parameter is erratic and hard to classify into any groups but when plotted two-by-two, earthquake ground motion parameters are more or less statistically correlated and grow monotonically together with higher dispersion associated with higher intensities (Figure 2). This is because most ground motion characteristics are defined, in a way or another, to measure the severity of an event. Therefore, good localization of similar data is hard to achieve when only simple parameters are used. In these cases, extraction of new variables in the feature space is necessary to physically define the similarity measures.

2.1.3. Feature Extraction

A schematic description of feature extraction and classification procedure is shown in Figure 3. The hypothesis is that earthquake SGM records in a large database naturally come into clusters of similar forms, when the data description is performed utilizing enough independent and appropriate features. Because of the problem with simple data characteristics, ground motion parameters were nonlinearly combined to create a set of
new features that can meaningfully describe higher dimensions in a reliable physical sense with significant design implications. The form of nonlinear combination is derived based on the quality of classification, distribution of new features, experience, and the error rate (see Figure 8.) Historically, the probability distribution of SGM parameters is assumed to be of a Poisson’s type (although this is being seriously questioned recently – see [Jordan, 2004]), which is the case in Figure 5 representing data of 61 years, with variable $x$ substituted by $-x$. This is because of the scheme of scaling adopted and will be explained later on in this paper.

For the preliminary analyses, it was aimed to find out whether clusters with the following design implications are identifiable in the database:

- Long duration records, for systems with deteriorating properties during vibration,
- Wide frequency band records for nonlinear systems,
- Near-fault ground motion records, when the burst of input energy is important, and
- Moderate records.

The following features were considered based on their strong relationship with the classes, the form of their probability density function, the quality of classification and by previous experience with the data, which is essential in learning the priori information in pattern recognition:

Scaled $\left\{ \frac{PGA}{D} \right\}$

Scaled $\left\{ (EPA + EPV)^{2.1} \right\}$

Scaled $\left\{ (IV \cdot ID)^{1.5} \right\}$

Two common scaling schemes were tried on the database; scaling based on the length of the features (0 to 1) and scaling for mean and standard deviation in logarithmic scale as follow:

Scaled Feature $\psi_i = \log \left| \frac{\psi_i - \text{mean}(\Psi)}{\sigma(\Psi)} \right|$  \hspace{1cm} (2)

in which, $\psi_i$ is the $i^{th}$ component of feature set $\Psi$, and $\sigma$ is the standard deviation of the feature set. The results of scaling based on length of features were presented in Alimoradi et al. [2004]. The second scaling scheme (Equation 2), presented in this paper, yields better distribution of new features over the database as shown in Figure 4 and therefore is used for further analyses. When plotted in feature space in Figure 6, and two-
by-two in Figure 7, new scale invariant data tend to disclose patterns of similar data. Also lower dispersion is achieved through appropriate nonlinear combination of simple ground motion parameters. Raw data and scale invariant feature statistics are presented in Tables 1 and 2, respectively.

2.1.4. Fuzzy C-Means Method

Fuzzy c-means (FCM) can be taken as a modification of hard c-means classification. These are methods for unsupervised minimum variance partitioning of data, in which no training (labeled data) is required. During classification, the algorithm initializes a number of clusters as well as cluster centers and then assigns every data point to a cluster to which the data point has the closest distance. This consists of one step of an iterative procedure that converges to a solution when there is a least amount of dispersion due to the location of cluster centers (minimization of objective function of Equation 3). At each step, new cluster centers are updated by finding the mean value of the data points associated to the cluster at the previous step [Duda et al., 1999].

The objective function to be minimized is a measure of intraclass dispersion summed over all classes, as follows [Duda et al., 1999]:

\[
L = \min \sum_{i=1}^{c} \sum_{j=1}^{b} \left[ \hat{P}(\omega_j | \psi_j, \hat{\theta}) \right]^{\frac{1}{b}} \left\| \psi_j - \mu_i \right\|^2
\]  

(3)

in which \( \psi_j \) is feature \( \psi \) of point \( j \), \( \mu_i \) is the center of cluster \( i \), \( \hat{P}(\omega_j | \psi_j, \hat{\theta}) \) is the “fuzzy” cluster membership of point \( j \) in cluster \( i \), \( b \) is the “blending” parameter usually taken as 2.0, \( n \) is the total number of data points, and \( c \) is the number of clusters.

\[
P(\omega_i | \psi_j) = \frac{(1/d_\theta)^{\frac{1}{b(i-1)}}}{\sum_{j=1}^{n} (1/d_\theta)^{\frac{1}{b(i-1)}}}, \quad d_\theta = \left\| \psi_j - \mu_i \right\|^2
\]

(4)

\[
\mu_j = \frac{\sum_{j=1}^{n} \left[ P(\omega_i | \psi_j) \right]^{\frac{1}{b}} \psi_j}{\sum_{j=1}^{n} \left[ P(\omega_i | \psi_j) \right]^{\frac{1}{b}}}
\]

(5)

\[
\sum_{j=1}^{n} \hat{P}(\omega_j | \psi_j) = 1, \quad j = 1, \ldots, n.
\]

(6)
Fuzzy c-means differ from hard c-means in that every point belongs to every cluster to some degree, based on its fuzzy membership (Equation 4), whereas in hard c-mean a point can only belong to one cluster.

Fuzzy c-means clustering is applied to the data plotted in Figure 6 in a Matlab environment [The MathWorks, 2001], with \( c = 2, 3, \ldots, 8 \) assumed for each clustering case. The best clustering strategy (and number of clusters) is obtained by verifying, for each case, the optimization trajectories (plotted on top of Figure 8) and by checking the resulting clusters as shown in Figures 9 to 12. Six distinct clusters are detected in the data with clusters 1 and 6 presented as dense data around the center and the rest of clusters with higher dispersion surrounding the core. Cluster centers are presented using thick black symbols with the record number of the centers in the database printed besides them. Ten closest data points to each cluster center (year, event name, station, and azimuth) are shown in Table 3. Further investigations in time and frequency domain show that FCM clustering has successfully identified similar data within the database. Time histories of acceleration, velocity, and displacement of sample data in different clusters along with the acceleration response spectra are plotted in Figure 13 for verification. Some physical interpretation and possible practical applications of such clustering strategy are discussed in the next sections.

2.2. Input Motion Decomposition and Synthetic Ground Motion Generation

Synthesizing earthquake ground motion records has made a significant contribution to understanding of processes of earthquake ground motion generation and propagation and to seismic design in areas of infrequent seismicity such as the New Madrid seismic zone [Pezeshk et al., 1998]. A classical method in generating artificial earthquake records utilizes random vibration theories by using a “seed” spectrum that represents the hazard and the seismicity of the site to create time series that match the seed spectrum. A common problem that usually arises from using this method is that single record matching and scaling can be misleading in realistic estimation of energy demand expected at a site [Naeim, 1994] and [Naeim, 1995]. What is empirically observed in this study is that the expected type of motion at a site, in general, falls into a set of finite distinct spectral shapes, hereby called Basis Spectra – to make an analogy with basis vectors in linear algebra. When careful classification of ground motion is performed in a large database, spectral shapes naturally come into groups of similar forms such as those in Figure 14 (in linear scale). The square-root-of-the-sum-of-the-square (SRSS) average of ten closest points to each cluster center is plotted on top of each other in Figure 15 (in log scale). When scaled based on their length in Figure 16 (taken as vectors, spectra divided by their Euclidean norm), these vectors show distinct spectral shapes of basis spectra that may be used in generation of earthquake ground motion records in order to create a set of compatible time series for a site, instead of representing hazard by only one target spectrum. In mathematical form:
in which $E[FM_{class}]$ is the expected value of the interclass fuzzy membership. This way of presenting ground motion can be referred to as decomposition of earthquake ground motion records and is similar to the method of Principal Component Analysis widely used in face, handwriting and fingerprint recognition.

Another interesting observation is that by such dividing data into distinct groups, knowledge can be extracted with regards to unknown input parameters of the process. For example, the prevailing spectral form of class two (C2) records in Figure 15 is a wide-band low-frequency that is usually associated with records on top of soft-soil deposits. To examine this, we tested ten closest data points to the cluster center of C2 (ten highest fuzzy membership in the group) in the Cosmos database (The Consortium of Organizations for Strong-Motion Observation Systems) for soil-site type. Out of ten, nine records for which site information was available, all turned out recorded on quaternary soft deposits with shear wave velocity of less than 333 m/s. Other interpretations such as the pulse of near-fault at around 1.0 sec spectral ordinate, needs further investigation. It is worth mentioning though, that no specific information about the site-soil type was available nor was used in the classification procedure. The conclusion was rather drawn from the general behavior of grouped records that indicates there are distinct types of motion in the database associated with distinct physical conditions. The following fuzzy rule (Equation 8) was hence generated and tested using 30 data points and their dispersion to identify soft-soil stations by having simple characteristics of a ground motion recorded at a site:
\[
\text{if: } -1.26339 \leq \log_{10} \left( \frac{(IV \cdot ID)^{1.15} - 1.2222 \times 10^3}{5.0675 \times 10^3} \right) \leq -0.98336
\]

and

\[
\text{if: } -0.78850 \leq \log_{10} \left( \frac{(PGA^{1.1}\sqrt{D}) - 4.0415945 \times 10^6}{1.83041808 \times 10^7} \right) \leq -0.65387
\]

and

\[
\text{if: } -0.67624 \leq \log_{10} \left( \frac{(EPA + EPV)^{2.1} - 6.93225 \times 10^4}{2.007923 \times 10^5} \right) \leq -0.46187
\]

then:

Ground motion is recorded on quaternary deposits with \( S_{\text{wave}} \leq 333 \text{ m/s} \) with more than 89.1% certainty.

2.3. Implementation in GAGMS

GAGMS is software system with graphical user interfaces that can be used to find sets of input motion records that in combination match any target design spectrum over a range of frequencies specified by the user (Figure 17). A genetic algorithm procedure minimizes the SRSS of the deviation of the solution set from the target. As a result, the final solution set may contain records with different characteristics that although would satisfy the code requirements of matching a design spectrum, may not be directly applicable to a site’s conditions. Therefore, data classification is to be used to enhance the features of GAGMS by including constraints on site-soil types, ground motion characteristics, and the aggregation of hazard.

A supervised nonparametric method of pattern classification (k-Nearest-Neighbor-Rule) was applied to subsets of local data at each cluster to check the probability density of magnitude, epicentral distance, and depth in each class. Interestingly, different patterns of such seismological parameters were detected, as can be seen in Figures 18 to 20, although again no information about seismological parameters of records were used in the classification task. Should records of different epicentral distance, depth and magnitude be detectable in a database, then it is computationally possible to represent the aggregation of hazard into GAGMS with relative ease. Examples are:

- GAGMS can perform search in the records of a specific class, when the target hazard spectrum of the site lies close to the SRSS spectra of that specific class.
- If the target falls between few classes, then GAGMS includes only those classes of motion in the search.
2.4. Fuzzy Rule Generalization

As mentioned before, once the mean value and standard deviation of features are computed in each cluster and verified on local and global data, one can generate fuzzy rules for structural response or earthquake event characteristics by having a few simple parameters such as PGA, duration, or in general, any features used in the classification. Examples are given below:

For Response:

If: 
\[-1.5 \leq \text{scaled} \left( \frac{\text{PGA}}{\sqrt{D}} \right) \leq +0.5, \quad \text{and} \quad -1.5 \leq \text{scaled} \left( \frac{\text{EPA} + \text{EPV}}{2} \right) \leq +0.5, \quad \text{and} \]
\[\text{scaled} \left( (IV \cdot ID)^{1/11} \right) \leq -0.75 \]

Then: Data belongs to cluster C2.

If: 
\[-2.0 \leq \text{scaled} \left( \frac{\text{PGA}}{\sqrt{D}} \right) \leq -0.5, \quad \text{and} \quad \text{scaled} \left( \frac{\text{EPA} + \text{EPV}}{2} \right) \leq -1.0, \quad \text{and} \]
\[-1.5 \leq \text{scaled} \left( (IV \cdot ID)^{1/11} \right) \leq +0.5 \]

Then: Data belongs to cluster C3.

For Events:

If: 
\[5\text{km} \leq \text{Depth} \leq 30\text{km}, \quad \text{and} \quad 40\text{km} \leq \text{Distance} \leq 80\text{km}, \quad \text{and} \]
\[6.2 \leq \text{Magnitude} \leq 7.5 \]

Then: Data belongs to cluster C2.

3. Discussion of the Results and Concluding Remarks

Pattern recognition was applied to a database of recorded earthquake ground motions in order to localize clusters of data with significant similar characteristics. A feature extraction experiment was performed to find parameters that can physically distinguish SGM data based on sharpness in time and frequency domain, existence of large input
velocity pulses accompanied by large displacement pulses and the bandwidth in spectral ordinates. Fuzzy c-means was shown as a promising analytical tool in classification of design ground motion records. A fuzzy system of inequality equations was derived to identify soft-soil records in a database by only having simple ground motion characteristics of a record using inverse engineering. The results obtained can be utilized for increased functionality and performance of a GA-based software system that finds optimum sets of input motion records for a given target design spectrum.

Further insight can be obtained by applying methodologies presented in this benchmark study on a larger database. More accurate basis spectra and fuzzy rules can only be achieved through testing various feature selection schemes in a large database, depending on the degree of complexity of the process. It was also noted that the task of feature extraction could be extended to include some other SGM parameters such as depth, magnitude and distance. The optimum number of clusters and features to choose is the subject of further studies.

Acknowledgement

Many constructive comments and suggestions received by the first author during discussions with Dr. Charles A. Langston, Professor of Seismology and Earth Structure at The University of Memphis is appreciated.

References


PEER Strong Motion Database, <http://peer.berkeley.edu/smcat/>


List of Tables:

Table 1. Raw features statistics
Table 2. New feature statistics
Table 3. Ten Closest Points to Each Cluster Center

Table 1. Raw features statistics

<table>
<thead>
<tr>
<th>Feature</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duration, Sec.</td>
<td>8.88</td>
<td>8.82</td>
<td>0.00</td>
<td>76.00</td>
</tr>
<tr>
<td>PGA, gals</td>
<td>171.03</td>
<td>156.55</td>
<td>18.00</td>
<td>1745.00</td>
</tr>
<tr>
<td>IV, Cm/Sec</td>
<td>23.18</td>
<td>23.25</td>
<td>1.57</td>
<td>189.09</td>
</tr>
<tr>
<td>ID, Cm</td>
<td>8.43</td>
<td>11.46</td>
<td>0.08</td>
<td>121.62</td>
</tr>
<tr>
<td>EPA, gals</td>
<td>134.80</td>
<td>124.34</td>
<td>0.00</td>
<td>1336.00</td>
</tr>
<tr>
<td>EPV, Cm/Sec</td>
<td>9.98</td>
<td>10.51</td>
<td>0.00</td>
<td>97.93</td>
</tr>
</tbody>
</table>
Table 2. New feature statistics

<table>
<thead>
<tr>
<th>Feature</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature 1</td>
<td>1.222E+03</td>
<td>5.067E+03</td>
<td>0.000E+00</td>
<td>1.037E+05</td>
</tr>
<tr>
<td>Feature 2</td>
<td>4.042E+06</td>
<td>1.830E+07</td>
<td>-5.223E+04</td>
<td>4.106E+08</td>
</tr>
<tr>
<td>Feature 3</td>
<td>6.932E+04</td>
<td>2.008E+05</td>
<td>0.000E+00</td>
<td>4.081E+06</td>
</tr>
</tbody>
</table>
Table 3. Ten Closest Points to Each Cluster
Center-Values in Parentheses are the Azimuth of Records

<table>
<thead>
<tr>
<th>Cluster 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1  1983 Coalinga, CA, Parkfield, Gold Hill 2 E (90)</td>
</tr>
<tr>
<td>2  1985 Nahanni, NWT, Canada, Site 3 (Battlement Creek) (270)</td>
</tr>
<tr>
<td>3  1994 Northridge, Jan. 17, Leona Valley #3 (0)</td>
</tr>
<tr>
<td>4  1983 Coalinga, CA, Parkfield, Gold Hill 2 E (0)</td>
</tr>
<tr>
<td>5  1994 Northridge, Jan. 17, Leona Valley #3 (90)</td>
</tr>
<tr>
<td>6  1985 Guerrero, Mexico, Atoyac (180)</td>
</tr>
<tr>
<td>7  1979 Coyote Lake, CA, San Juan Bautista FF (303)</td>
</tr>
<tr>
<td>8  1994 Northridge, Jan. 17, Leona Valley #4 (0)</td>
</tr>
<tr>
<td>9  1983 Coalinga, CA, Parkfield, Gold Hill 4 W (90)</td>
</tr>
<tr>
<td>10 1994 Northridge, Jan. 17, Newport Beach - Newport Blvd &amp; Coast HWY (180)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cluster 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1  1971 San Fernando, 2500 Wilshire Blvd, bsmt (299)</td>
</tr>
<tr>
<td>2  1992 Landers, June 28, Indio - Coachella Canal (90)</td>
</tr>
<tr>
<td>3  1983 Coalinga, CA, Parkfield Fault Zone 2 (0)</td>
</tr>
<tr>
<td>4  1983 Coalinga, CA, Parkfield Fault Zone 2 (90)</td>
</tr>
<tr>
<td>5  1975 Island of Hawaii, Hilo, Univ of Hawaii (74)</td>
</tr>
<tr>
<td>6  1983 Coalinga, CA, Parkfield Fault Zone 7 (0)</td>
</tr>
<tr>
<td>7  1971 San Fernando, 435 N Oakhurst Beverly Hills (270)</td>
</tr>
<tr>
<td>8  1983 Coalinga, CA, Parkfield Fault Zone 10 (0)</td>
</tr>
<tr>
<td>9  1971 San Fernando, CMD Bldg, Vernon (187)</td>
</tr>
<tr>
<td>10 1979 Southern Alaska, Icy Bay (90)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cluster 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1  1971 San Fernando, CIT Millikan Lib, basement (0)</td>
</tr>
<tr>
<td>2  1979 Coyote Lake, CA, San Martin (250)</td>
</tr>
<tr>
<td>3  1984 Morgan Hill, CA, San Jose - IBM Building 28, F (315)</td>
</tr>
<tr>
<td>4  1994 Northridge, Jan. 17, 12001 Chalon RD., Los Angeles, CA (70)</td>
</tr>
<tr>
<td>5  1980 Mammoth Lakes, Long Valley Dam (Upper L Abut) (90)</td>
</tr>
<tr>
<td>6  1992 Petrolia, April 25, Shelter Cove - Airport (0)</td>
</tr>
<tr>
<td>7  1994 Northridge, Jan. 17, Los Angeles, Wadsworth V.A. Ho (325)</td>
</tr>
<tr>
<td>8  1934 Baja California, El Centro - Imp Val Irr Dist (270)</td>
</tr>
<tr>
<td>9  1994 Northridge, Jan. 17, Prado Dam (90)</td>
</tr>
<tr>
<td>10 1994 Northridge, Jan. 17, Leona Valley #6 (90)</td>
</tr>
</tbody>
</table>
### Cluster 4

<table>
<thead>
<tr>
<th></th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1983 Alaska, SIM Simeonof Island (340)</td>
</tr>
<tr>
<td>2</td>
<td>1994 Northridge, Jan. 17, Long Beach - City Hall Grounds (90)</td>
</tr>
<tr>
<td>3</td>
<td>1983 Mexico, Tuxla Gutierrez (0)</td>
</tr>
<tr>
<td>4</td>
<td>1994 Northridge, Jan. 17, Newport Beach - Irvine Ave. Fire Station (0)</td>
</tr>
<tr>
<td>5</td>
<td>1992 Landers, June 28, Silent Valley - Poppet Flat (90)</td>
</tr>
<tr>
<td>6</td>
<td>1992 Landers, June 28, Twentynine Palms (90)</td>
</tr>
<tr>
<td>7</td>
<td>1994 Northridge, Jan. 17, Long Beach - City Hall Grounds (360)</td>
</tr>
<tr>
<td>8</td>
<td>1994 Northridge, Jan. 17, Anaverde Valley - City Ranch (90)</td>
</tr>
<tr>
<td>9</td>
<td>1992 Landers, June 28, Silent Valley - Poppet Flat (0)</td>
</tr>
<tr>
<td>10</td>
<td>1986 North Palm Springs, CA, Colton Interchange - Vault (82)</td>
</tr>
</tbody>
</table>

### Cluster 5

<table>
<thead>
<tr>
<th></th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1952 Kern County, Hollywood Storage - PE Lot (180)</td>
</tr>
<tr>
<td>2</td>
<td>1966 Mexico, Infiernillo N-120 (338)</td>
</tr>
<tr>
<td>3</td>
<td>1974 Alaska, SAN Sand Point School (30)</td>
</tr>
<tr>
<td>4</td>
<td>1984 Morgan Hill, CA, Saratoga, WVC Gym, GF E Wall (0)</td>
</tr>
<tr>
<td>5</td>
<td>1986 North Palm Springs, CA, Colton Interchange - Vault (352)</td>
</tr>
<tr>
<td>6</td>
<td>1983 Mexico, Isla Palma, Mich. (0)</td>
</tr>
<tr>
<td>7</td>
<td>1967 Northern CA, Ferndale City Hall (314)</td>
</tr>
<tr>
<td>8</td>
<td>1980 Anza, Pinyon Flat (45)</td>
</tr>
<tr>
<td>9</td>
<td>1994 Northridge, Jan. 17, Rosamond - Airport (0)</td>
</tr>
<tr>
<td>10</td>
<td>1954 Wheeler Ridge, CA, Taft - Lincoln School (111)</td>
</tr>
</tbody>
</table>

### Cluster 6

<table>
<thead>
<tr>
<th></th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1983 Coalinga, CA, Parkfield Cholame 4 W (0)</td>
</tr>
<tr>
<td>2</td>
<td>1983 Alaska, VDZ Valdez City Hall (360)</td>
</tr>
<tr>
<td>3</td>
<td>1983 Coalinga, CA, Parkfield Cholame 6 W (0)</td>
</tr>
<tr>
<td>4</td>
<td>1987 Whittier, Oct. 1, Pasadena - Caltech Brown Athletic Center (270)</td>
</tr>
<tr>
<td>5</td>
<td>1984 Morgan Hill, CA, BM Almaden, F.F. (90)</td>
</tr>
<tr>
<td>6</td>
<td>1980 Mexico, Oaxaca Fac. Medicina (270)</td>
</tr>
<tr>
<td>7</td>
<td>1980 Trinidad, CA, Offshore, Rio Dell Overpass, E Ground (0)</td>
</tr>
<tr>
<td>8</td>
<td>1987 Whittier, Oct. 1, Downey - County Maint. Bldg. (270)</td>
</tr>
<tr>
<td>9</td>
<td>1994 Northridge, Jan. 17, 8510 Wonderland Ave., Los Angeles, CA (185)</td>
</tr>
<tr>
<td>10</td>
<td>1994 Northridge, Jan. 17, 2369 E. Vernon Ave., Los Angeles, CA (180)</td>
</tr>
</tbody>
</table>
List of Figures:

Figure 1. General Distribution of Data Parameters in the Database.
Figure 2. Correlation Between Raw Data Features.
Figure 3. Flowchart of the Process.
Figure 4. Distribution of the New Scale Invariant Date Features.
Figure 5. Histograms of the Scale Invariant Data Features.
Figure 6. Three Dimensional Distribution of the New Data Features.
Figure 7. Two Dimensional Correlation Between New Data Features.
Figure 8. FCM Optimization Trajectories for Classification for Different Number of Clusters.
Figure 9. Six Distinct Data Clusters Plotted With Cluster Centers in the Feature Space.
Figure 10. Plot of Data Clusters in Two Dimensions.
Figure 11. Plot of Data Clusters in Two Dimensions.
Figure 12. Plot of Data Clusters in Two Dimensions- Decision Borders in Black.
Figure 13. Samples of Data in Different Clusters- Records Closest to the Centers.
Figure 14. Sample of Response Spectra of Ten Closest Records to Each Cluster Center Plotted with the SRSS Spectra of Data in Each Group.
Figure 15. SRSS Spectra of the Ten Closest Records to the Centers in Each Cluster.
Figure 16. Normalized SRSS Spectra of Ten Closest Records to the Centers in Each Cluster.
Figure 17. GAGMS Main Input Page.
Figure 18. Histograms of Magnitude, The Whole Database vs. Thirty Closest Points to the Centers of Each Cluster.
Figure 19. Histograms of Epicentral Distance, The Whole Database vs. Thirty Closest Points to the Centers of Each Cluster.
Figure 20. Histograms of Focal Depth, The Whole Database vs. Thirty Closest Points to the Centers of Each Cluster.
Figure 1. General Distribution of Data Parameters in the Database.
Figure 2. Correlation Between Raw Data Features.
Figure 3. Flowchart of the Process.
Figure 4. Distribution of the New Scale Invariant Data Features.
Figure 5. Histograms of the Scale Invariant Data Features.
Figure 6. Three Dimensional Distribution of the New Data Features.
Figure 7. Two Dimensional Correlation Between New Data Features.
Figure 8. FCM Optimization Trajectories for Classification for Different Number of Clusters.
Figure 9. Six Distinct Data Clusters Plotted With Cluster Centers in the Feature Space.
Figure 10. 2D Plot of Data Clusters.
Figure 11. 2D Plot of Data Clusters.
Figure 12. 2D Plot of Data Clusters- Decision Borders in Black.
Figure 13. Samples of Data in Different Clusters- Records of Closest Distance to the Centers.
Figure 14. Sample of Response Spectra of Ten Closest Records to Each Cluster Center Plotted with the SRSS Spectra of Data in Each Group.
Figure 15. SRSS Spectra of the Ten Closest Records to the Centers in Each Cluster.
Figure 16. Normalized SRSS Spectra of Ten Closest Records to the Centers in Each Cluster.
Figure 17. GAGMS Main Input Page [Naeim et al., 2004].
Figure 18. Histogram of Depth for the Whole Database vs. Thirty Closest Points to Each Center.
Figure 19. Histogram of Distance for the Whole Database vs. Thirty Closest Points to Each Center.
Figure 20. Histogram of Magnitude for the Whole Database vs. Thirty Closest Points to Each Center.