An Advanced Pseudo-Random Data Generator that improves data representations and reduces errors in pattern recognition in a Numeric Knowledge Modeling System

Errol Davis
Director of Research and Development
Sound Linked Data Inc.

Erik Arisholm
Lead Engineer
Sound Linked Data Inc.

Artificially generated data proved to be more cost effective, easier to collect, result in reduced development time and produced better training data for Numeric Knowledge Modules than real life data. Having a domain experts knowledge of what data samples or variables best represent a specific domain of knowledge and an Advanced Pseudo-Random Data Generator capable of producing the samples, was the most accurate way to generate data sets for numeric knowledge modeling. Data representation or data sets are defined for this article as set of test results or other data necessary to represent all the variables that are needed to develop the Numeric Knowledge Modules that will ultimately classify accurately patterns within a specific market or domain. The term Pseudo-Random was used as the sample data was restricted to domain specific categories prior to being analyzed by the numeric knowledge module generator. Improving the data sets and not adjusting the mathematical models has proven in one domain, hearing, to have improved accuracy, reduced development time and increased acceptance of this artificially intelligent tool. The Advanced Pseudo-Random Data Generator reduced the domain experts learning curve related to data acquisition, improved statistical accuracy, simplified data collection and improved acceptance of numerical knowledge modules. This method of data representation for producing Numeric Knowledge Modules for pattern identification can likely be applied to other knowledge domains. There has been a recent trend in Artificial Intelligence to combine Expert Systems and Numeric Knowledge Modeling as one method to deal with fuzzy logic (gray areas) or pattern overlaps (statistically invalid results). The combining of two AI approaches could be the result of domain experts inability to communicate the complex problems that exist within his domain. In many cases, the Numeric Knowledge Modeling tools tend to be too complicated for the domain expert. It is possible that improved data sets from flexible random generators could prove to be a more efficient method to deal with complex specific domain knowledge acquisition.

The graphical representation of data and flexibility in instrument configuration afforded by the Advanced Pseudo-Random Data Generator had a significant impact on the reduction in development time and improved accuracy of the Numeric Knowledge Modules over a typical real life data set. Domain Experts seemed to better understand and accept numerical knowledge modules when they were part of the data generation process.

With the advent of object oriented instrumentation the random generator enabled the development of a more flexible tool capable of providing domain specific random samples of almost any type of data. Input data could be limited by graphical representations that were familiar to the domain expert. A Random Generator developed in Sound Linked Data labs generated representative raw data samples of hearing test patterns. The domain expert established which data variables were to be represented in the data set and the domain specific Pseudo-Random Data Generator (e.g. in this case an Audiogram Generator) was configured for a specific audiometric or hearing classification. Classifications were developed for a variety of hearing patterns and the resultant numerical knowledge module could be tested within minutes rather than hours or days as with the traditional techniques.

The Random Data Generator for audiogram classification and hearing aid recommendations allowed the audiologist (domain expert) to construct which patterns best represented the site of a hearing problem in the auditory system and what degrees of hearing loss needed to be represented for the appropriate hearing aid. The generator configured for this project had 12 frequency ranges in octave and half octave intervals from 125 Hz to 12,000 Hz. The loudness range in decibels was -10 dB HL to +110 dB HL.. The hearing patterns vs. hearing level were set by the domain expert (audiologist in this case) and saved in one of 5 diagnostic categories in combination with 5 hearing ranges. The hearing aid training data generator had additional audiological parameters required to accurately build the numerical knowledge models for hearing aid recommendations. These combinations were all graphically represented by the random audiogram generator and fully programmable. The number of samples, the upper and lower limits of categories and patterns were also set by the domain expert. Several other parameters relating to hearing loss classification were available but are beyond the scope of this article. All parameters settings were stored in the training directory for future reference. The Random Generator automatically formatted the training sets for the numerical modeling tool. The training sets were then imported directly into an advanced numerical modeling system.

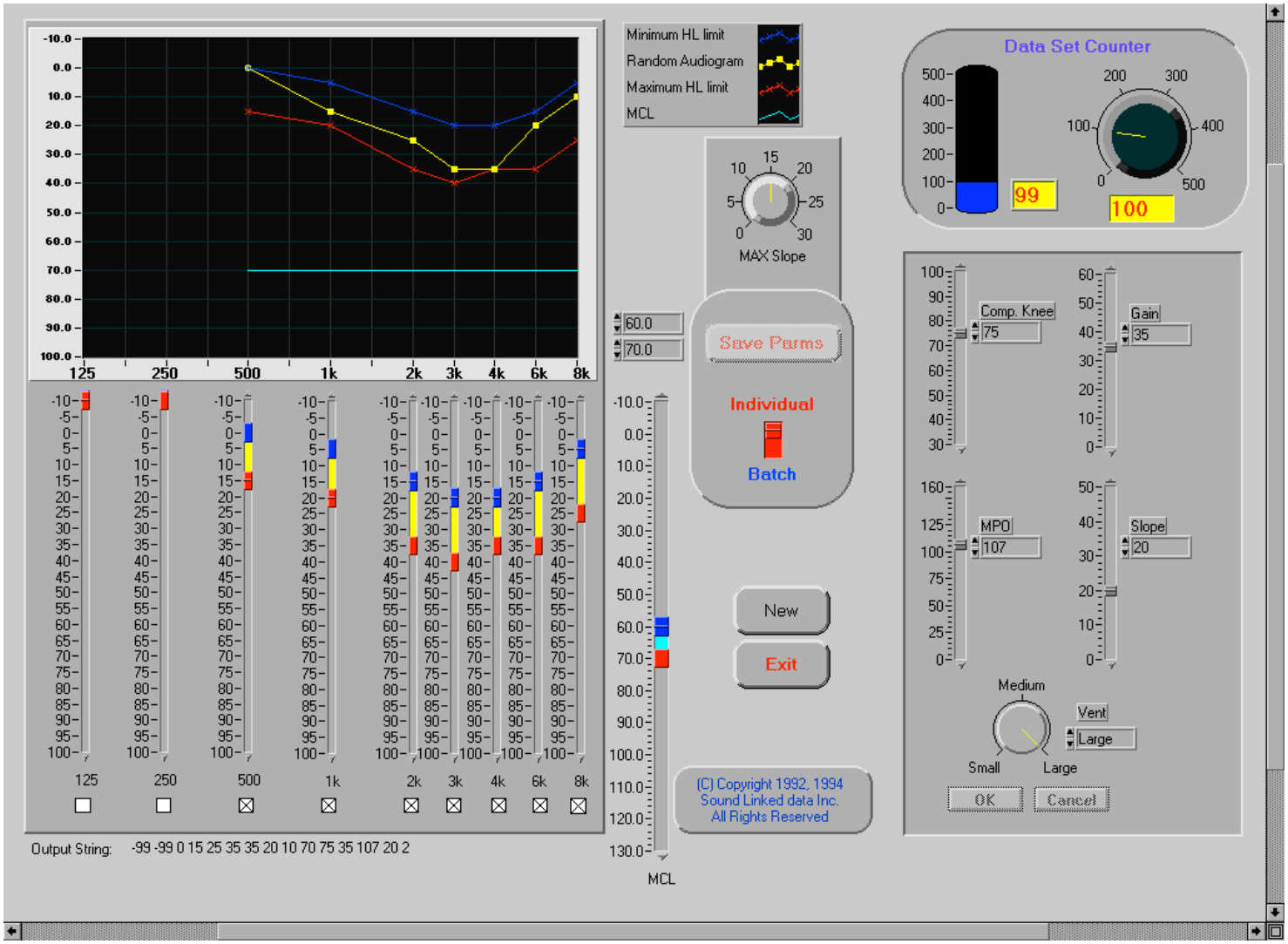

Figure 1. Random Training Data Generator for Hearing Aid Recommendation: Input data models are adjusted with the vertical slider controls to the left. Random audiograms are shown on the chart. The domain expert adjusts output values for each new generated random data record. The training data record set is saved in a format compatible with the advanced learning tool.

	f500	f1k	f2k	f3k	f4k	f6k	f8k	MCL	CK	GAIN	MPO	SLOPE
1	0	5.000000	20.000000	30.000000	25.000000	20.000000	25.000000	65.000000	70.000000	20.000000	102.000000	15.000000
2	10.000000	10.000000	20.000000	30.000000	30.000000	15.000000	25.000000	58.000000	70.000000	20.000000	102.000000	15.000000
3	15.000000	5.000000	15.000000	30.000000	25.000000	30.000000	15.000000	60.000000	70.000000	20.000000	102.000000	15.000000
4	15.000000	10.000000	20.000000	30.000000	25.000000	15.000000	20.000000	58.000000	70.000000	20.000000	102.000000	15.000000
5	10.000000	15.000000	35.000000	25.000000	30.000000	35.000000	25.000000	58.000000	70.000000	20.000000	102.000000	15.000000
6	0	15.000000	35.000000	35.000000	20.000000	20.000000	5.000000	62.000000	70.000000	20.000000	102.000000	15.000000
7	15.000000	15.000000	35.000000	40.000000	25.000000	35.000000	20.000000	60.000000	70.000000	20.000000	102.000000	15.000000
8	10.000000	15.000000	30.000000	35.000000	25.000000	20.000000	15.000000	55.000000	70.000000	20.000000	102.000000	15.000000
9	0	15.000000	25.000000	25.000000	30.000000	30.000000	15.000000	55.000000	70.000000	20.000000	102.000000	15.000000
10	10.000000	15.000000	35.000000	40.000000	30.000000	35.000000	25.000000	55.000000	70.000000	20.000000	102.000000	15.000000
11	0	5.000000	15.000000	30.000000	30.000000	20.000000	20.000000	65.000000	70.000000	20.000000	102.000000	15.000000
12	10.000000	10.000000	25.000000	30.000000	30.000000	25.000000	10.000000	60.000000	70.000000	20.000000	102.000000	15.000000
13	10.000000	15.000000	20.000000	40.000000	35.000000	30.000000	20.000000	55.000000	70.000000	20.000000	102.000000	15.000000
14	0	5.000000	20.000000	40.000000	30.000000	20.000000	5.000000	60.000000	70.000000	20.000000	102.000000	15.000000
15	5.000000	10.000000	20.000000	40.000000	35.000000	20.000000	20.000000	55.000000	70.000000	20.000000	102.000000	15.000000
16	0	15.000000	20.000000	20.000000	35.000000	35.000000	20.000000	62.000000	70.000000	20.000000	102.000000	15.000000

Figure 2. Example Randomly Generated Training Data Sets.

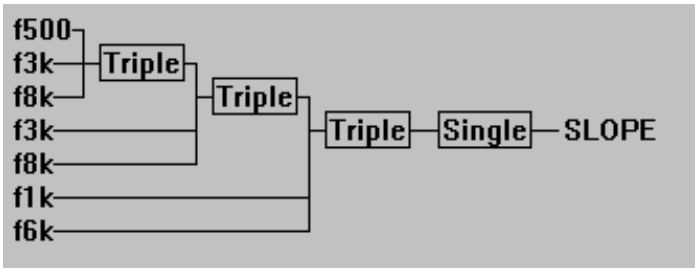

Figure 3. Hearing Aid Recommendation numeric knowledge module:
The function SLOPE uses patient audiogram input data (f500, f3k, f8k, f3k, f8k, f1k, f6k) to determine the acoustical frequency roll-off for the recommended hearing aid.

The modeling tool generated Numerical Knowledge Modules (fig.3) and because of the flexibility of the modeling tool the domain expert was able in real time (on the fly) to test the effectiveness and adjust training sets (fig. 2) where necessary. The learning tool then generated numerical knowledge modules in the form of “C” code that were callable by the hearing instrument through analysis libraries (DLL’s in this case) to classify hearing test results. The modules had the representative characteristics that reduced over and under fitting, increased confidence coefficients and reduced pattern recognition overlap that has been a common complaint of many advanced learning tools. Problems in over and under fit, pattern overlap and a high number of unclassifiable patterns were encountered in a pre-release version of the hearing instrument when real life hearing test results were used to develop the representative data sets.

Since the use of the Random Audiogram Generator (RAG) to provide the representative data samples for the advanced learning tool, the accuracy of the hearing instrument

system has improved and development time reduced. Artificially generated data improved numerical knowledge models in four areas: Improved confidence coefficients, reduced computing time, reduced category overlap and most importantly audiological or domain expert confidence in the resultant classifications. The result has been improved acceptance of the hearing instrument by end users based on increased accuracy and its ability to learn new patterns. The hearing instrument itself now has a built in *Learn* feature which allows the domain expert to build personalized updates of the Numerical Knowledge Modules. The adaptive *Learn* feature can be enabled by embedding the numeric modeling tool into the hearing instrument. This has further increased acceptance of the hearing instrument, as any knowledgeable domain expert can build his own interpretive style into the hearing instrument. The *Learn* function in the hearing instrument reduces the cost associated with the collection of training data and re-synthesis of the personalized Numeric Knowledge Modules.

Another influencing factor for acceptance of hearing instruments with embedded Numeric Knowledge Modules has been reduction of the domain expert involvement in many of the initial aspects of an individual hearing assessment. The cost savings occur because the hearing instrument can function as an intelligent audiometric technician. Numerical knowledge modules in the hearing market have been shown to reduce domain expert involvement up to 70%. This fact is especially important in market segments where cost containment decisions can have a negative influence on individual health care. In addition to generators for hearing analysis and hearing aid recommendations, the SLD group has plans to experiment with the Random Generator as a means to improve the understanding of speech signals in background noise and apply a similar approach to EKG, Vision and pulmonary function. This will mean the addition of a cost effective but advanced real time digital signal processing system.

The cost of data collection for real life data sets often precludes a complete representative sample that the modern learning systems require to develop accurate mathematical knowledge modules. Even when data collection cost are no object, it is often difficult to get the domain expert to understand the extreme importance of complete and accurate representative data. Domain experts often cannot understand the negative effect of non-representative raw data and the resultant inability of mathematical knowledge modules to identify a pattern accurately. Often the numeric knowledge modeling software or analyst is blamed rather than imperfect data collection. Over or under fitting is the ultimate result and the resultant low confidence coefficients produce pattern identification overlaps. These gray or fuzzy areas are the downfall of an modeling system. Mathematical knowledge acquisition professionals may not consider input data filtering as an important method for improving confidence coefficients. This usually is because their area of expertise in not related to the domain the data represents and

therefore are reluctant to pass judgment on its completeness. They depend usually too much on the domain experts understanding of his representative sample. The analysts are often fooled by a misguided confidence in their own ability to adjust modeling criteria in order to obtain acceptable pattern recognition. This constant tweaking often produces results that show high confidence in patterns the system recognizes but fuzzy areas grow to such a size that experts are required to maintain a constant vigil over system results. This defeats the whole purpose of numeric modeling as tool to reduce costs and improve accuracy. A significant confidence gap often develops as the end user often judges system results on a simple "*does it works or it does not*". The baby is often thrown out with the bath water because confidence coefficients are yielding significant pattern overlap. What progress had been made in defining or understanding a domain specific problem is dismissed because of too many pattern overlaps and budget over runs. Many times a process or adjusting the outputs of the modeling tool to compensate for overlaps are attempted to salvage the project. Adjusting the modeling parameters is often considered a form of voodoo by many domain experts because of inexperience in such matters. The domain expert begins to look immediately for pattern recognition flaws when adjustments are made. A competition develops to prove the machine wrong because the domain expert feels he no longer is in control. At SLD we maintain this has caused a confidence gap in the development process of utilizing advanced modeling to recognize patterns in many market segments.

Many examples exist where a single person or small group within an organization develops the understanding of acquiring complete data sets only to be told that it is too expensive and time consuming to reformat all the necessary input data to even bring a project to proof of concept. Many larger organizations have a ludite reaction to advanced modeling tools. Another complicating factor is that many organizations are not inclined to share the necessary data and in-house expertise with outside consulting firms, that are capable of developing accurate and useful mathematical knowledge modules. This is especially true in medicine where many fuzzy or gray areas exist and many experts feel their are more artist than scientists.

An Advanced Random Data Generator has been used successfully to generate representative, as well as accurate artificial data sets. Over fitting and under fitting were eliminated by improving the accuracy of input data rather than adjusting the modeling parameters. Real life data suffered more from over fitting and under fitting problems than did artificial data. Many neural-net systems and related numeric modeling systems rely on adjustment of mathematical models to the raw data rather than adjusting the raw data to the mathematical model. The learning curve for capturing domain expert information with a Random Data Generator becomes a function of generating new representative data sets when errors are encountered. Representative data sets are something a

domain expert can see graphically and understand quickly. Attempting to make the domain expert understand the traditional process of tweaking modeling parameters to deal with over or under fit has been less than successful. By giving the domain expert the tools to generate representative data sets they become part of the development process and the modeling tool remains just that - a tool. The result should be greater acceptance of advanced artificial intelligence systems in certain markets segments.

Errol Davis
103-523 The Queensway, Toronto, M8Y1J7
errol@microlab.ca (416-251-7508) ©