Artificial Intelligence and Social Computing, Vol. 28, 2022, 187–193
https://doi.org/10.54941/ahfe1001462

Proposal for the Generation of Profiles using a Synthetic Database

Andres Viscaino-Quito[1] and Luis Serpa-Andrade[2]

[1]Research Group on Artificial Intelligence and Assistive Technologies–GIIATa, Universidad Politécnica Salesiana, Cuenca, Ecuador
[2]Research Group on Applied Embedded Hardware–GIHEA, Universidad Politécnica Salesiana Cuenca, Ecuador

ABSTRACT

The lack of data to perform various models that feed an artificial intelligence with which you can get or discover various patterns of behavior in a set of data. Therefore, due to this lack of data, the systems are not well nourished with data large enough to fulfill its learning function, presenting a synthetic database which is parameterized with restrictions on the characteristics of graphomotor and language elements, which develops a set of combinations that will be the model for the AI. As effect to all this gave a commensurable amount of 777,600 combinations at the moment of applying the first filter with the respective restrictions, when taking the valid combinations that are 77304 a second filter is applied with the remaining restrictions that gave 57,672 valid combinations for the generation of the synthetic database that will feed the AI. It is concluded that the generation of synthetic data helps to create, according to its importance, more or less similar to real data and in this way ensures a quantity and no dependence on real or original data.

Keywords: Exemplary paper, Human systems integration, Systems engineering, Systems modeling language

INTRODUCTION

The advance of Information and Communication Technologies (ICT) has greatly increased the value of information from different areas of development as a source of actionable knowledge. But it is not only the volume of information that makes them interesting. It is data as such and their behavior that transforms them into complementary elements in the search for knowledge, characterized by the high diversity they present. The large number of applications of information generates different solutions related to the area of application, the particularity of the problem and the value of the results (Berrezueta-Guzmán et al., 2016; Méndez & Rubier, 2018). The results will always be an imperative factor in determining the qualitative value generated by a study or research, however, on many occasions, being able to test the different methods that comprise the different levels of development generates certain limitations arising from the acquisition of information, as well as the interpretation of the same. And despite having the necessary information, they do not always fit the needs of the problem in general. The use of synthetic data is an alternative to address this problem, since they are intended to emulate the circumstantial behavior of the study approach, generating a level of realism and detail that does not provide any type of risk (Berrezueta & Abad, 2022; León et al., 2019; Morales et al., 2020).

In accordance with the above, this article proposes the generation of synthetic profiles to define activities that evaluate the level of graphomotor and language skills established as an objective function to measure the level of skills in children with gross motor disabilities, writing problems, etc. This process involves an initial evaluation phase where the system determines the activity or activities to be prioritized according to a set of interrelated constraints that will serve to achieve the proposed objective, so it is important to have a considerable amount of information that allows us to study and analyze the different probabilities of suggestion provided by the system.

RELATED WORK

The use of synthetic data has been employed for many fields in various areas of development. For example, (Torres-Vásquez et al., 2019) employs a data balancing methodology that allows optimizing the different predictive models used to determine the SGB. These models were developed by means of simple and combined classifiers in relation to the initial input values (unbalanced data) with the help of SMOTE (Synthetic Minority Oversampling Technique), and then their effectiveness was evaluated in percentage levels and an operating characteristic curve specified by the ROC receiver that summarizes the performance obtained per model.

Similarly (Ilasaca Cahuata et al., 2018) conducted a study to generate synthetic indicators and set a ranking of sustainable development in the regions of Peru, which consists of a regional and local environmental information system from which the information required for the development process is obtained. These indicators are grouped into eight indicators divided into three dimensions (economic, social, environmental) that are recorded in a data matrix as the basis for a correlational matrix, which is generated through statistical software (SPSS v. 23) for the group of indicators considered. During this process a factor analysis is applied - principal component method for the interpretation of factors, factor rotation, and applying the respective methodology for the construction of the indicators that involves both the imputation of missing data, normalization, weighting and assignment of weights, aggregation of values and evaluation of the results obtained.

Equally (León et al., 2019) presents a methodology for the generation of synthetic signals of pseudo-realistic character centered on functions of statistical distributions of random samples to schematize a signal with a set of samples within a specified range. It also employs a set of probability distributions governed to certain parameters that give rise to a particular probability distribution, Gamma and Gaussian, which provides a coherent, level-defined signal. This model is presented in a web tool called “SysGpr”, where signals can be configured and produced. The validation of the results is performed by means of three metrics (learning algorithms, autocorrelations and cross-correlations, mean opinion score).

On the other hand, (Vallez et al., 2019) proposes the generation of a scenario that shows the corridor of a high school, built with the help of a Unreal graphics engine, where a surveillance system is projected that focuses clearly on the detection of dangerous objects, mainly in the detection of guns. This system makes use of YOLO, Tiny-YOLO and VGG-SSD for object detection, which are trained with both real and synthetic images. The synthetic images are obtained from the videos generated in the projection of the scene, extracting the relevant information in relation to the position coordinates where dangerous objects are located. Once all this process is done, the information is saved in an XML file and validated for each of the models.

METHODOLOGY

A proposal is made for the generation of synthetic profiles, so initially we started with the creation of a synthetic database, whose information has been increasing during the development of the proposed objective. For the creation of the data, the structure to be used was initially defined, and with the help of an expert in the area, a list of characteristics of the focus of study and analysis (children with disabilities) was determined to measure the level of graphomotor skills and language.

The characteristics in the graphomotor part are the following:

- Body control
  - Maintains full control of his body (A1)
  - Maintains control of his body with difficulty in balance and likes gross motor activities (A2).
  - Maintains control of body with difficulty in balance and dislikes gross motor activities (A3)
  - Maintains control of body, but has difficulty with gross motor activities (A4)
  - Does not maintain control of his body, needs full support and supervision (A5)
- Hand-eye coordination
  - Good hand-eye coordination (B1)
  - Difficulty in maintaining hand-eye coordination (B2)
- Attention
  - Maintains attention during an assigned task (C1)
  - Maintains attention during an assigned task even though delayed, but completes it (C2)
  - Most of the time maintains attention during an assigned task (C3)
  - Has difficulty sustaining attention on an assigned task, is easily distracted (C4)
  - Has a hard time maintaining attention on an assigned task, wants to finish quickly (C5)

  - Holds attention for short periods of time on objects with lights and sounds (C6)

- Laterality
  - Definite (D1)
  - Switches hands constantly when performing an activity (D2)
  - Changes hands infrequently, but uses one hand more often (D3)
- Dissociation of hands
  - Constantly uses the support hand (E1)
  - Occasionally uses the support hand (E2)
  - Does not use the support hand (E3)
- Disposition for a task at a table
  - Remains seated during a task (F1)
  - Has difficulty remaining seated during a task (F2)
  - Does not remain seated (F3)
- Pencil manipulation
  - Is attracted to pencils and manipulates pencils (G1)
  - Not attracted to pencils and does not manipulate them (G2)
  - Is not attracted to pencils and does not manipulate pencils unless they are necessary for the task assigned (G3)
- Sensitivity to touch
  - Tolerates activities that involve getting hands dirty and wet (H1)
  - Tolerates activities that involve getting hands dirty or wet (H2) o Does not tolerate activities that involve getting hands dirty or wet (H2)
  - Does not tolerate activities that involve getting hands dirty or wet (H3).

The characteristics on the language side are:

- Comprehensible language
  - Has good comprehension (I1)
  - Comprehension is good, but needs visual aids and sign language (I2)
  - Comprehension is regular, the instructions need to be repeated a few times (I3)
  - Comprehension is regular despite verbal repetition, needs visual aids (I4)
- Expressive language
  - Has good expressive language (J1)
  - His language is clear, rigid, repetitive, and he finds it difficult to participate in conversations (J2)
  - Expresses self through oral language with some distortion in voice intonation (J3)
  - Expresses self through guttural sounds, shouting and crying (J4)

Table 1. Graphomotor constraints.

A1	B2, C6, E1, E2, F3, G3, I2, J6, G1
A2	C6, E3, F3, G3
A3	C6, E3, F3, G3
A4	B1, E3, F3, B2, C1, E1, G4
A5	B1, C1-C5, E3, F1, E1, G4, D1
B2	C1, G3
C2	I5, J4

Table 2. Language constraints.

I1	J4
I2	J5
I3	J6
I4	J7
I5	J1, J2, J3, J5-J8

- Expressed by spoken language, oral with monosyllabic and bisyllabic words (J5)
- Expresses self through Alternative and Augmented Communication Systems (pictograms) (J6)
- Expresses him/herself through sign language and Alternative and Augmentative Communication Systems (J7)

Development

Python was used to generate the possible combinations with all the above mentioned characteristics, with a total of 777600 combinations, however, there were problems with them since there are characteristics that do not match physically and logically, for example in the graphomotor part the values of "Maintains full control of his body" and "Difficulty in maintaining hand-eye coordination", is wrong, so a set of restrictions for the combinations was proposed, as we can see in the following table.

The constraints in Table 1 help us to have a better control of the combinations. In the first restrictions, 77304 valid combinations were obtained, however, some combinations were still incoherent, so Table 2 of restrictions was applied, which generated 57672 valid combinations, one of the valid combinations is shown below:

With the valid combinations we have a database to be used in any learning process such as profiling through synthetic data.

Proposed Synthetic Profiles

The basis of this article is the creation of synthetic profiles to develop a case-based expert system. Thanks to the synthetic database generated, several cases of children with different characteristics can be seen in the execution

Table 3. Valid Combinations.

	Maintains body control with difficulty in balance and dislikes gross motor activities.
B2	Difficulty in maintaining hand-eye coordination
C4	Difficulty maintaining attention on an assigned task, easily distracted
D2	Constantly switches hands when performing an activity
E1	Constantly uses support hand
F1	Remains seated during a task
G2	Has little attraction to pencils and manipulates them little
H2	He has little tolerance for activities that involve getting his hands dirty or wet.
I4	His comprehension is regular in spite of repeating verbally, he needs visual aids.
J8	He expresses himself through sign language and Alternative and Augmentative Communication Systems.

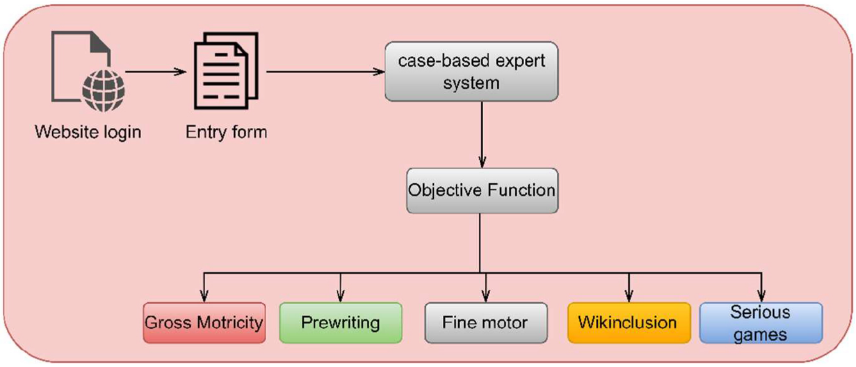

Figure 1: Diagram of operation using synthetic profiles.

of the model, presenting a set of models of activities defined by each profile, and even optimization criteria for the same system.

The system reviews each profile and its characteristics to determine the activities for the profile to be applied. These activities are divided into areas or modules such as:

- Gross Motor Skills
- Pre-Writing
- Writing
- Serious Games
- Wikinclusion

In the diagram of figure 1 we can see how it would work when applying the synthetic profiles, first the person in charge of the system will enter the page, at the moment of filling the characteristics form the data will enter the expert system and the output data will enter the function that will be in charge of giving the number of repetitions by activities in each one of the modules to which it is related.

CONCLUSIONS

One of the alternatives to facilitate the learning process of children with disabilities is to develop activities that generate interaction such as serious games. These help the child to learn or acquire skills, although in some cases the implementation and development of activities incorporates a level of complexity due to the fact that each child has a different level of learning and development than the others.

In the present research, a synthetic database was developed that incorporates all possible combinations, using constraints for its generation, and subsequently the creation of synthetic profiles for the training of a case-based expert system, and an objective function capable of assigning specific activities to each child and for each disability.

REFERENCES

Berrezueta, S., & Abad, K. (2022). Doctoral Symposium on Information and Communication Technologies - DSICT. Springer, 846.

Berrezueta-Guzmán, J., Serpa-Andrade, L., Robles-Bykbaev, V., & Pinos-Velez, E. (2016). Digital trainer for the development of the fine motor ability in children with cerebral palsy. MATEC Web of Conferences,

Ilasaca Cahuata, E., Tudela Mamani, J. W., Zamalloa Cuba, W., Roque, B., & Fernandez, E. (2018). Generación de indicadores sintéticos de desarrollo sostenible-Perú 2015: Generation of sustaintable development syntethic indicator-Peru. Revista de Investigaciones Altoandinas, 20(2), 251-260.

León, F., Rodríguez-Lozano, F. J., Cubero-Fernández, A., Palomares, J. M., & Olivares, J. (2019). SysGpr: Sistema de generación de señales sintéticas pseudo-realistas. Revista Iberoamericana de Automática e Informática industrial, 16(3), 369–379.

Méndez, N. P., & Rubier, J. P. (2018). Ciencia de datos: una revisión del estado del arte. UCE Ciencia. Revista de postgrado, 6(3).

Morales, G. R., Salgado, J. P., Pérez-Gosende, P., Cordero, M. O., & Berrezueta, S. (2020). Information and Communication Technologies: 8th Conference, TICEC 2020, Guayaquil, Ecuador, November 25–27, 2020, Proceedings (Vol. 1307). Springer Nature.

Torres-Vásquez, M., Hernández-Torruco, J., Hernández-Ocana, B., & Chávez-Bosquez, O. (2019). Balanceo de datos del Síndrome de Guillain-Barré utilizando SMOTE para la clasificación de subtipos. Res. Comput. Sci., 148(7), 113–125.

Vallez, N., Velasco Mata, A., Cotorro, J. J., & Deniz, Ó. (2019). Â¿Es posible entrenar modelos de aprendizaje profundo con datos sintéticos?