Modeling of Small Data with Unsupervised Generative Ensemble Learning

Serge Dolgikh^a

^a National Aviation University, 1 Lubomyra Huzara Ave, Kyiv, 03058, Ukraine

Abstract

Modeling and analysis of small data raises essential problems and challenges stemming from insufficient sampling of unknown observable distributions, complicating and exacerbating confident analysis and often reducing statistical confidence of the conclusions. In this work, an original approach to analysis of small data is proposed that is based on an ensemble of generative neural network models with the intent of identifying stable clusters of data in informative generative representations. We demonstrate how characteristic structure of stable clusters in generative representations of a dataset of images of basic geometric shapes can be determined from representations produced by a generative ensemble. The method can be used to identify characteristics structure, perform correlation analysis and augment data of different types and under some conditions that were discussed, improve the performance of supervised classification in cases with a deficit of training data.

Keywords 1

Unsupervised learning, ensemble learning, clustering, statistical analysis, small data

1. Introduction

Modeling and analysis of small data raises essential problems and challenges stemming from insufficient sampling of unknown observable distributions, complicating and exacerbating confident analysis and often producing lower statistical confidence of the conclusions. Nevertheless, early analysis of structure and trends in emerging data can be essential in situations and events of novel or rare nature / condition where large volumes of confident decisions may not be available for any reason [1].

Among the admitted challenges in practical applications of methods of machine intelligence in the analysis of small data are those of stability of learning and produced results. It can be observed for example, as a strong dependency of the learning success on the choice of training parameters, selection, temporal ordering of batches and other training factors. Issues that have been noted [2,3] include reproducibility of the results, overfitting, inability to generalize and others. Issuing from these challenges, results produced by models of similar architecture with the same datasets can be inconsistent and volatile, and the ability to generalize characteristic patterns, more limited than in conventional applications. Not in the least, reproducibility of the results that is essential in establishing confidence in the methods can be less certain, significantly complicating comparison of methods and models.

Numerous efforts attempted to examine the problem of stability of small data learning and a number of promising approaches and directions described, including: cross-validation; ensemble methods [4]; Radial-Basis Function (RBF) networks [5,6] and other methods [7,8]. However, whereas some of the methods shown success in a number of specific applications, generality and applicability to different types of analyzed data and problems, to the best of our knowledge could not be established due to specialized structure, architecture and essential assumptions about the distributions.

IDDM-2022: 5th International Conference on Informatics & Data-Driven Medicine, November 18–20, 2022, Lyon, France EMAIL: sdolgikh@nau.edu.ua (A. 1) ORCID: 0000-0001-5929-8954



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

In parallel to these developments, methods of unsupervised generative learning [9,10] demonstrated effective ability to achieve significant simplification of complex data in the process of unsupervised generative learning via reduction of redundancy in the observable parameters and identification (extraction) of informative features. In a growing number of instances, these methods were instrumental in the analysis of patterns in complex real-world data [11,12] including data strongly constrained by the size of the sample [13]. Critically for success in the identified problem area, application of such methods is not limited by availability of prior data including labeled datasets, and in many cases can be successful with smaller samples than conventional methods of supervised learning. These traits set methods of unsupervised generative learning as good candidates for analysis of data constrained by both size and availability of confident prior knowledge, without precluding aggregation of confidently known data for subsequent analysis with conventional methods.

To address these challenges as outlined above, we propose the ensemble approach [14] based on a collective of unsupervised generative neural networks to identify stable patterns and structure in the observable (training) data, that simultaneously addresses both problems of the deficit of known labels, and stability of learning. Stable structures in the informative low-dimensional representations of observable data produced by generative models can be identified via a process that was developed and used for several purposes, including correlation analysis by factors of interest, augmentation of small data by producing newly generated data points from identified characteristic latent structure. In contrast to some of the methods mentioned earlier, this approach does not have strong dependencies on specific assumptions about observable distributions and can be used in a generic manner with data of different types and origin.

2. Methods

An ensemble of generative neural network models with an architecture of a deep convolutional autoencoder with a strongly compressed representation layer was used to produce two or three-dimensional representations of small datasets of images of basic geometrical shapes.

The advantage of the proposed architecture stems from previous applications in producing informative low-dimensional representations of complex data as well as universal approximation capacity of neural network models [15] making them a useful and versatile tool in modeling diverse distributions of complex data.

Following successful generative training of the models in the ensemble, embedded distributions of the evaluated datasets were produced in the spaces of latent coordinates. We then attempted to identify stable populations of recognizable latent structures, such as visually identifiable clusters. The resulting structure of identified stable clusters had to be invariant with respect to individual training model and therefore represent innate characteristics of the input data that was possible to verify due to known composition of the dataset.

2.1. Generative Neural Network Architecture

A deep convolutional autoencoder neural [16,17] had the input layer of dimension p with 2-3 convolutional layers common in the practice of learning visual data. Models had convolutional layers for acquisition of visual features, one deep layer and a central encoding layer of size d, creating twoor three-dimensional (i.e. d = 2, 3) latent representations of the input data defined by activation values of the latent neurons in the encoding layer.

Overall, generative autoencoder models in this work had 48,000 - 96,000 trainable parameters depending on configuration of layers. The decoding or generating stage was fully symmetrical to the encoder. The models were implemented in Tensorflow / Keras [18] with data processing, plotting and visualization Python packages used in the analysis of the results.

An architecture diagram of the model is shown in Figure 1.



Figure 1: Deep convolutional autoencoder architecture with a strong dimensionality reduction.

Training of the models proceeded in an unsupervised process with minimization of generative error (i.e., the distance of the generated output on the input training batches) with MSE (mean squared error) and CCE (categorical cross-entropy) cost. Unsupervised training in this process over 10 - 25epochs produced a strong reduction in the value of the cost function with majority in the ensemble of learning models.

A success of generative learning was measured by two criteria: 1) drop in the value of the cost function on the validation dataset; 2) the ability of trained models to generate a randomly selected subset of training data (Figure 2). Up to 80 - 90% of training models trained successfully based on these criteria.



Figure 2: Generative training with geometrical shapes dataset: evaluation of training success.

2.2. Data

Table 1

For verification of the approach, we used a dataset of greyscale images of basic geometrical shapes including circles, triangles and empty backgrounds with variation in size and contrast. The composition of the dataset described in detail in [19]. Two small datasets of different sizes were used: G-150, 50 samples per shape, with overall size 150; and G-300, 100 samples per shape. Images of different geometrical shapes represented different characteristic patterns in small datasets of observable data.

The dataset of geometrical shapes is described in Table 1.

Main characteristics, Geometric shapes dataset								
Dataset	Size	Input size	Composition	Variation parameters				
G-150	150	32 x 32	3 shapes: circle, triangle, greyscale background	size, contrast				
G-300	300	32 x 32	3 shapes	size, contrast				

2.3. Unsupervised Ensemble Learning

We decided to approach the question of stability of learning with small data with a set (an ensemble) of generative neural network models that do not require prior knowledge, including in the form of labeled data, for successful training. An ensemble of trained generative models of a size nthus produced an array of two-dimensional representations of the input data as shown in Figure 3.



Figure 3: 2D latent distributions, G-150 dataset, three independently trained models.

As a result of this phase of unsupervised generative ensemble learning was produced a set of pairs: $R = \{$ (trained model, map(input data point, 2D latent position) $\times n \}$. Association of a unique id to a latent position of an input data point (Figure 2, e1) relative to that of the other points in the set allowed to identify stable clusters K in the input data by entirely unsupervised process as follows (pseudocode, where D is input dataset, E: encoding phase of the generative model, where e(x) = E(x) encoded (latent) image of observable data point, m: the number of identified clusters):

for
$$x(k)$$
 in D:
if $e_k = E(x(k))$, e_{k-1} in $K_1 : K(x) = K_1$ (known cluster)
else if $conf(e_k, L) > \gamma : K(x) = L$; $m := m+1$ (a new cluster)
else: $K(x) = A$ (not in cluster)

where A: an arbitrary id for elements with uncertain cluster; γ : confidence of identification of the latent position e_k belonging to cluster L. The process can be described as follows: if the next element is in the same cluster as an earlier one (i.e., with a lower sequential id in the dataset), the cluster of that element is assigned. If cluster of the element is uncertain, an arbitrary constant number is assigned; finally, if neither of the conditions is satisfied, a new cluster is assigned and the process is repeated until the dataset is exhausted. This process is deterministic, and as can be seen does not depend on selection of ordering sequence, as long as the same sequence is maintained throughout the process.

There result is a matrix K(D) of (*data point*, *cluster id*) pairs of a dimension ($M \times n$) where M is the size of the dataset and n, the of the ensemble of generative models, where points in the same cluster (including uncertain cluster association) have the same cluster id.

In the final step one can obtained the set of stable clusters $K_s(D)$ identified by the ensemble as a subset of K(D) satisfying certain confidence criteria c_s :

$$x \in K_s(D): card(K(x)) \ge c_s \tag{1}$$

For example, if correlation of K(x) in the matrix K(D) was found to be 0.9 (i.e., 9 out of 10 models in the ensemble produced the same cluster id for a given element) and the size of the ensemble, 20 then the 95% confidence interval for the correlation coefficient of the element and the cluster would be [0.76, 0.96] indicating a strong and confident association of an element to a cluster.

The resulting subset of stable clusters $K_s(D)$ identified in the described process can be used in an analysis of composition of the input data and a number of other applications as discussed in the subsequent sections. It may be worth reiterating that at no point in the analysis any true known samples of classes in the input dataset were used.

3. Results

3.1. Cluster Structure

The results of evaluation of the cluster structure with the datasets of images in a process described in the preceding section are presented in Table 2 and Table 3. Identification of clusters was performed by visual analysis (examples in **Figure 3**) that demonstrate both stability and accuracy of the identified cluster structure with the data in the study. In the future, unsupervised clustering methods such as DbScan [20], Meanshift and others can be applied to identify stable cluster structure in an automated unsupervised process.

Table 2

Cluster composition, G-150 and G-300 datasets, $n^{(1)} = 20$

Dataset	Number,	Clustered fraction ⁽²⁾	Clustered fraction	Visible cluster	
	clusters	at conf = 0.8	at conf = 0.95	separation	
G-150	3	1.0	0.92	high	
G-300	3	1.0	0.96	high, very high	

⁽¹⁾ Ensemble size

⁽²⁾ Fraction of the dataset in identified stable clusters, at confidence level

Table 3

Cluster confusion matrix, G-150 dataset

Cluster, type	Cluster 1	Cluster 2	Cluster 3
0 (circle)	1.0	0.	0.
1 (triangle)	0.25	0.75	0.
2 (background)	0.	0.15	0.85

With a large unsupervised dataset, from G-150 to G-300 significantly improved accuracy of the cluster to type association was observed, rising to the level of 95 - 100%. Stability of the latent structure of clusters is a key observation and a necessary requirement for successful generation of new data, confirming that clusters identified by the unsupervised generative ensemble method indeed described stable characteristic patterns in the observable data.

3.2. Generation and Prototypes

The architecture of generative models provides a direct way to propagate positions in the latent space of generative models to the space of input (observable) parameters. The mapping can be obtained by taking a latent position with coordinates $p = (l_1, l_2)$ as input to the generator component of the model (Figure 1) to produce an observable position X_{obs} :

$$X_{obs} = G(p) \tag{2}$$

where G: $R \rightarrow O$, is the generator component of the model, operating from the latent space R into the space of inputs, O.

Based on (2) generative ability of successfully trained generative models can be used to create new data points with "similarity" to identified characteristic patterns by selecting positions in the latent regions of identified stable clusters as illustrated in **Figure 4**.



Figure 4: Cluster-based data generation. Green, red dots: stable latent clusters; cross: newly generated data points; blue: other data points.

The effectiveness of the proposed method of ensemble cluster-based data augmentation can be supported by the arguments:

Consider a small data set *S* of size *N* with *p* input parameters. With a conventional method of approximation, e.g. Gaussian, the error of the mean in each of the parameters of the data can be estimated as: p_{mean} / \sqrt{N} [21]. Where N is small the dispersion of observable parameters can be sufficiently large. Next, if a successful generative representation of a lower dimensionality *d* with a good cluster structure existed, the data can be approximated with a good accuracy by a quasi-multi-modal distribution with the number of modes N_c , where N_c : the number of stable latent clusters, and the dispersion, $d_{mean} / \sqrt{n_{clus}}$ (the size of a latent cluster). Where *d* is small: $d \ll p$ (i.e., a strong reduction of dimension of observable data) and the number of samples in the principal clusters, sufficiently large one obtains a statistical problem of significantly lower complexity and dispersion. As an example, in this study, the reduction of dimensionality from 4,096 input parameters (grayscale images with resolution 64×64) to 2, i.e. by a factor of ~2,000 was achieved.

4. Applications

4.1. Data and Factor Analysis

Decomposition of unsupervised datasets into a structure of stable latent clusters, where successful, can be helpful in the analysis of the distribution of data and associated factors of interest. As discussed earlier, the proposed approach offers a general, independent of specific types of data capability to decompose observable data into a structure of more homogenous regions, or clusters. Even without known samples of classes of interest, distributions of clusters, both latent and observable can be analyzed in detail, including observation of characteristic representatives of clusters, prototypes.

Generative ability of models in the study combined with cluster decomposition of informative representations allows non-trivial analysis of composition of input data by generating typical observable instances of stable clusters, or prototypes of characteristic natural classes of data. With stable clusters identified with proposed methods, one can generate specific latent positions associated with clusters for example, as a mean of cluster member positions and propagate them to the observable space with generative transformation (2).

$$K:t(K) = (x_k \in K); P(K) = G(t(K))$$
 (3)

where K: stable latent cluster, P(K): observable prototype. Examples of observable prototypes of clusters in the dataset G-150 are shown in **Figure 5**.



Figure 5: Cluster prototypes, G-150 dataset.

Importantly, cluster decomposition can provide insights into essential associations of the data to the factors of interest, again without any previously known context or associations. For an example consider a hypothetical dataset where input datapoints $\{x\}$ in the dataset G-150 represent patient data and are associated with certain factor of interest f(x), such as a reaction to an infection. A direct correlation analysis in the observable space can be challenging due to large number of observable parameters (close to 5,000). On the other hand, cluster decomposition as discussed earlier can provide a mapping of input data to its stable cluster: $x \rightarrow K(x)$ resulting in a single dimension correlation problem (f(x), K(x)), a massive reduction of complexity.

4.2. Data Generation and Augmentation

Based on the discussion in Section 3.2 cluster decomposition can be used to generate new data points and augment small datasets, again without any limitations on the prior knowledge of the distribution of the input data. Once a structure of stable latent clusters has been identified, it is

straightforward to determine their distribution regions and produce latent candidate positions for augmentation of the original data. Generative transformation (2) can then be applied to obtain the related data points in the space of observable parameters.

To summarize the results and discussion on data generation, ensemble-based generative augmentation of small data can be successful under these conditions:

- The latent dimensionality is sufficiently small: $d \ll p$.
- The models demonstrate good learning success and consistent, stable, cluster structure.

• The number of stable clusters is small compared to the size the dataset (the number of samples), and the population of main clusters is sufficiently large: $N/n_{clus} \propto 1$.

If the conditions are satisfied, the original data can be described by a multi-modal distribution of stable latent clusters that can be identified with density clustering or another method as demonstrated earlier; further, augmentation of data can be performed in an unsupervised process based on the identified distribution in the latent space and will provide stable results invariant with respect to the selection of a specific instance of generative model.

4.3. Classification

The method of augmentation based on unsupervised cluster structure produced in generative learning can be employed to improve the success of classification with models of supervised learning trained with small datasets, as in the conventional practice of supervised learning, the size and representative quality of training data can have strong influence on the accuracy of classification [22].

The success of the method essentially depends on the presence of a correlation between stable latent structure of clusters and the factor of interest for classification that can be used as a label in supervised learning. If such a correlation can be established between the data points in identified stable clusters and the factor of interest as discussed in Section 4.1, an augmentation process outlined in Sections 3.2, 4.2 can be applied, with class labels assigned to generated data points based on the established association between latent clusters and known classes. Such a process of augmentation can produce an improvement in classification accuracy due to larger and more representative dataset in supervised learning.

For example, a clustering analysis performed with the ECE dataset [13] demonstrated a strong correlation of the identified cluster structure with the classification factor of interest, an epidemiological outcome. In that and similar cases where correlation of identified cluster structure with an observable factor of interest can be established, augmentation of data with the method described in this work can produce substantial improvements in classification.

5. Conclusions

The method of identification of unsupervised cluster structure in generative representations of small datasets with an ensemble of unsupervised generative models has been described and verified with small datasets of visual data of basic geometrical shapes. It was shown that a stable structure of clusters can be identified in the latent representations of successful generative models with high confidence in the interval 90–100% with the dataset used in the study. The structure of stable clusters representing characteristic types in the input data can be used to augment small datasets by generating new data points with several potential applications, including enhancement of labeled datasets with an objective to improve the success of classification in supervised learning.

It needs to be noted that the method described in this work may not have universal applicability to all datasets and its effectiveness is defined not only by the observable parameters, but also by the composition and characteristics of the dataset such as: the size, the number and population of principal clusters in the latent representation of the data and their correlation with the factors of interest. Where the conditions described in Section 3.2 are met, augmentation can produce additional data points associated with principal clusters and improve the performance of conventional classification methods trained with augmented datasets.

In novel, rare, non-standard cases, events and environments, large volumes of known data may be needed for confident analysis with conventional methods of machine learning. Availability of training data may present strong challenges in such cases. Methods of unsupervised generative learning, including the ensemble approach presented in this work, can be successful in identification of characteristic structure and patterns even with smaller sets of observable data without requirements of massive prior knowledge offering a practical direction toward improving the confidence of the analysis.

6. References

- [1] Hekler E.B., Klasnja, P., Chevance G. et al.: Why we need a small data paradigm, BMC Medicine, 17 (1) 133 (2019).
- [2] Wasserman P.D.: Neural computing: theory and practice. Van Nostrand-Reinhold, New York (1989).
- [3] LeBaron B., Weigend A.S.: A bootstrap evaluation of the effect of data splitting on financial time series. IEEE Trans. Neural Networks 9 213–220 (1998).
- [4] Cunningham P., J. Carney, S. Jacob S.: Stability problems with artificial neural networks and the ensemble solution. Artificial Intelligence in Medicine, 20 (3) 217–255 (2000).
- [5] Karar M.E., Robust RBF neural network-based backstepping controller for implantable cardiac pacemakers, Int. J. Adap. Cont. Sign. Proc 32 1040–1051 (2018).
- [6] Izonin, I., Tkachenko R., Dronuyk I. et al.: Predictive modeling based on small data in clinical medicine: RBF-based additive input-doubling method. Math Biosc. Eng, 18 (3) 2599–2613 (2021).
- [7] Forman G., Cohen I.: Learning from little: comparison of classifiers given little training. In: Proceedings of PKDD, 19 161–172 (2004).
- [8] Geris L.: Computational modeling in tissue engineering. Springer-Verlag, Berlin (2013).
- [9] Bengio, Y.: Learning deep architectures for AI. Foundations and Trends in Machine Learning 2(1), 1–127 (2009).
- [10] Coates, A., Lee, H., Ng, A.Y.: An analysis of single-layer networks in unsupervised feature learning. In: Proceedings of 14th International Conference on Artificial Intelligence and Statistics 15, 215–223 (2011).
- [11] Gondara, L.: Medical image denoising using convolutional denoising autoencoders. In: 16th IEEE International Conference on Data Mining Workshops (ICDMW), Barcelona, Spain, 241– 246 (2016).
- [12] Shi, J., Xu, J., Yao, Y., and Xu, B.: Concept learning through deep reinforcement learning with memory augmented neural networks. Neural Networks 110, 47–54 (2019).
- [13] Dolgikh, S.: Unsupervised clustering in epidemiological factor analysis. The Open Bioinformatics Journal 14(1), 63–72, 2021.
- [14] Opitz, D., Maclin, R.: Popular ensemble methods: An empirical study. Journal of Artificial Intelligence Research, 11 169–198, 1999.
- [15] Hornik K., Stinchcombe M., White H.: Multilayer feedforward neural networks are universal approximators. Neural Networks 2(5), 359–366 (1989).
- [16] Le, Q.V.: A tutorial on deep learning: autoencoders, convolutional neural networks and recurrent neural networks. Stanford University, 2015.
- [17] Dolgikh, S.: Low-dimensional representations in generative self-learning models. In: Proc. 20th International Conference Information Technologies – Applications and Theory (ITAT-2020), Slovakia, CEUR-WS.org 2718, 239–245 (2020).
- [18] Keras: Python deep learning library. https://keras.io/, last accessed: 2021/08/21.
- [19] Dolgikh, S.: Topology of conceptual representations in unsupervised generative models. In: Proc. 26th International Conference on Information Society and University Studies (IVUS 2021) Kaunas, Lithuania, CEUR-WS.org 2915, 150–157 (2021).
- [20] Ester, M., Kriegel, H-P., Sander, J., et al.: A density-based algorithm for discovering clusters in large spatial databases with noise. Proc. Second International Conference on Knowledge Discovery and Data Mining (KDD-96) 226–231 (1996).

- [21] Wendland H.: Scattered data approximation. Cambridge University Press 9 (2005).[22] Richards J.A.: Supervised classification techniques. In: Remote Sensing Digital Image Analysis. Springer, Berlin, Heidelberg 247–318 (2013).