Machine Learning

CS-433

Project 2 - Machine Learning for Science

3D Spatiotemporal clustering of mixed-type medical data in Tanzania

Students
Paloma Cito
Lorena Egger
Jules Triomphe

Professors
Martin Jaggi
Nicolas Flammarion

Supervisors
Mary-Anne Hartley
Kuan Tung

Abstract

This project aims to detect patterns in tabulated clinical and satellite data by using unsupervised machine learning methods, such as Variational Autoencoders. By tuning the hyperparameters and choosing the right distance function in our final algorithm (MMD), we obtained a 30% accuracy increase for one of the features. The model was shown to be useful to identify clusters for existing infectious diseases.

December 17, 2020

1 Background

COVID-19 has shown that emerging infectious diseases are a major global health threat. Prompt identification and intervention is critical to mitigating their effect. However, traditional disease surveillance can be costly [SSM04] and some countries do not have the resources to efficiently detect outbreaks at their earliest stages [MMR13].

With climate change and urbanisation, new and unexpected infectious diseases are emerging more frequently. And inappropriate antibiotic prescription is driving drug resistance amongst existing infections. Some of the highest risk regions for these new and newly resistant diseases are in low resource settings like Tanzania [MCM16]. Electronic clinical decision algorithms (eCDAs) like the electronic point-of-care test (ePOCT) have been developed to assist in efficient diagnostic assessments and by guiding antibiotic use [KD18]. The systematic clinical information collected in such tools could be used for disease surveillance where spatio-temporal clusters may identify sub-populations with possible infectious spread. Thus, these clusters can be used to better target high-risk sentinel populations for the expensive testing required for pathogen screening.

Unsupervised machine learning models are designed to detect and cluster unexpected patterns in data with no pre-existing labels. One architecture of unsupervised learning that is particularly suited to this task in high dimensional data is the autoencoder (AE): an unsupervised neural network that reduces dimensionality for processing by encoding the features and then reconstructing them back into a representation that is as close to the original input as possible. Hierarchical variational AE (H-VAE) [Sim+19] allows the implementation of multiple AEs for various data types in mixed-type data, and we will explore their use in this project.

Additionally, we also seek to expand the spatial representation of disease surveillance by incorporating environmental satellite data (elevation) to provide further information of likely clusters of infectious vectors.

Thus, this project aims to create a framework for 3D spatio-temporal clustering of mixed type medical data.

2 Aim and objectives

In cooperation with the IGH Global Health research group, this project aims to detect patterns in tabulated clinical and satellite data by using unsupervised machine learning methods. The data used is made up of 2 data sets: a labeled ePOCT data set and a satellite elevation data set. The ePOCT data set is from a randomized, controlled non-inferiority study among 3192 children aged from 2 to 59 months presenting acute febrile illness in 9 outpatient clinics in Dar es Salaam, Tanzania from December 2014 to February 2016 [Kei+17]. The satellite data set was previously extracted using ArcGIS.

The objectives are to:

- Clean the dataset and correct misspelling errors
- Incorporate geographical elevation data in the dataset
- Build an autoencoder to use to cluster data
- Evaluate the effectiveness of using a H-VAE algorithm on several classification tasks

3 Methods used

3.1 Clean the dataset and correct misspelling errors

As mentioned in the work of Zeineb Sahnoun [Sah20], the misspelling of the geographic ward feature resulted in significant missingness. To address this issue, we implemented the SymSpellPy python library for a Symmetric Delete spelling correction algorithm, which objectively corrects possible misspellings compared to a reference list of correctly spelled words. This reference file of correct ward names was derived from the GeoJSON file of Tanzania's spatial data. Ward names were first corrected using this library, then latitude and longitude values were filled.

3.2 Incorporate elevation data in the dataset

Elevation data from ArcGIS is used between the latitude and longitude coordinates $(-8.0001388888699, 37.999861111111)$ and $(-5.999861111110973, 40.00013888886993)$. From the cleaned ePOCT dataset, elevation data is computed for all observations where latitude and longitude coordinates exist. After checking that the coordinates are within the frame defined by the satellite data, the corresponding elevation values are obtained by fetching the value in row *Latitude* and column *Longitude*. NaN values are assigned to the longitude and latitude row if these values don't fall in the bounds. The extraction of elevation data was made with our supervisor Kuan Tung during his semester project.

3.3 Build an autoencoder (AE) to cluster data

Due to the heterogeneity of our features, an autoencoder was needed to first encode our data in an homogeneous manner, which then allowed us to train the model efficiently in the latent space.

However, one of the limitations of this implementation is that once we encoded the features back to the original space after the training, we could not find a loss function that could optimally be used with our different types of data. To address this problem, we based our work on Kuan Tung's idea of using a hierarchical VAE (H-VAE) [Sim+19]. This implementation trains three autoencoders, one for the categorical features, one for the continuous features and the last one that takes as input the results of the two previous encoders, solving the type issue.

3.4 Evaluate the effectiveness of using a H-VAE algorithm on several classification tasks

To evaluate whether the H-VAE is useful in identifying clusters, it must be compared with a baseline. As no classic classifier exists which can classify mixed-data efficiently, another one must be chosen. The Gaussian Naive Bayes (GaussianNB) classifier is used as it does not require much training data and is comparatively more robust to the curse of dimensionality than other classifiers. Moreover, the embedded layer of the H-VAE is made up of continuous values. These might not be distributed following a Gaussian distribution, but they can be considered to be conditionally independent from one another and so the Gaussian Naive Bayes classifier is acceptable.

On one side, the ROC AUC is computed for the GaussianNB trained using the preprocessed data (including the elevation feature) and five different classes, namely *malaria* (yer or no), *fws* (fever without clinical source, yes or no, i.e. a fever without any kind of other clinical indication for its origin), *anemia* (yes or no), *malaria_hl* (malaria high or low, when *malaria* is yes) and *fws_bv* (fever without clinical source from bacterial or viral origin). Each class was extracted from the original ePOCT dataset as these features were removed in the preprocessed dataset. On the other hand, the classifier is trained using the latent H-VAE space obtained from training the model with all features, and all five classes. For both scenarios, the ROC AUC is computed using three-fold stratified cross-validation to preserve class imbalance and increase the accuracy of the estimated generalization error.

4 Results

4.1 Misspelling correction

To minimize the risk of introducing errors, corrections were only made for misspelled wards with a single suggested correction and an optimal distance between words was explored. The

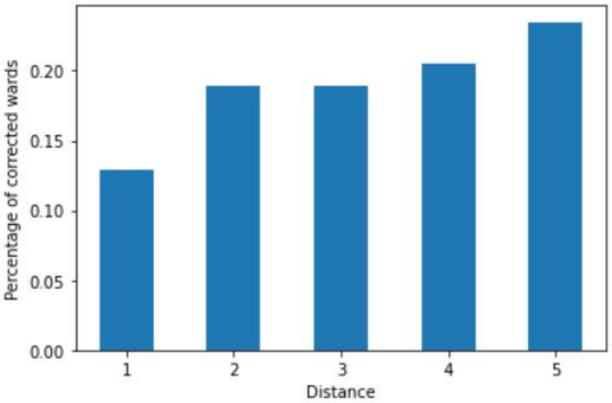

Figure 1: Percentage of misspelled ward names according to their spelling distance from the true ward name

distance between words (number of positions at which the misspelled word and the dictionary word differ) was plotted against the percentage of corrected ward names.

From this, we can deduce that the distance parameters of above 2 seem to offer little benefit, with an increasing risk of misallocation. With a final distance of 2 as parameter in the SymSpell functions, 133 misspelling errors were detected and 25 ward names corrected (meaning that 25/133 had only one suggestion). Finally, after the last pass of the filling missing latitude and longitude method, 29 new missing values are filled. The additional 4 values can be explained by the fact that the filling code also filled correctly spelled wards which did not have coordinates.

4.2 Evaluate the effectiveness of using a H-VAE algorithm

The results are shown in **Table 1**. We noticed that H-VAE using the classic KL distance was not performing very well. We then came across an interesting article [Zha] that explained why the MMD (maximum mean discrepancy) divergence was a better choice for VAEs. One such reason is that it compares the moments of the distributions, and doesn't assume these are Normal distributions. After changing that parameter in our implementation, we indeed saw a significant performance increase for groups of classes with sufficient numbers of observations. The most impressive result is for *fws*, with a 30% performance increase. Indeed, this heterogenous group of patients may have a variety of hidden clinical causes for their fevers and thus non-normal distributions.

Group (n. obs.)	No H-VAE	H-VAE with KL	H-VAE with MMD
malaria (2,920)	0.56	0.51	**0.76**
fws (2,920)	0.58	0.53	**0.83**
anemia (2,920)	0.68	0.70	**0.83**
malaria_hl (279)	0.52	0.58	**0.60**
fws_bv (769)	0.49	0.47	**0.51**

Table 1: Average model ROC AUC using a Gaussian Naive Bayes classifier

The results show that the classification is much better when using the embedded prediction than when using the preprocessed data directly except for *fws_bv*. For *malaria_hl*, the results are only somewhat better by using the H-VAE, but this is probably due to the very low number of training data points (279) and the high complexity of this task as there are no features explaining this result. The results for *fws_bv* can also be explained this way, having only 769 data points.

This validates the effectiveness of the HVAE model to improve classification and hence the potential effectiveness for new infectious disease cluster identification. The final model is therefore built using an unsupervised clustering algorithm. The Gaussian Mixture Model (GMM) is chosen as it predicts outcomes based on the covariance of the data, which we expect to be present in the case of infectious diseases.

4.3 Results with additional elevation feature

When running the model (with the GMM) on the preprocessed data with 5 clusters (chosen to be consistent with [Sah20]), we notice that one stands out (cluster 3 on **Figure 2**): its mean for elevation is almost an order of magnitude greater than the other clusters and is significantly different from the mean of the other clusters' distributions at the 90% confidence level. In fact, the cluster contains the highest percentage of malaria cases among all clusters.

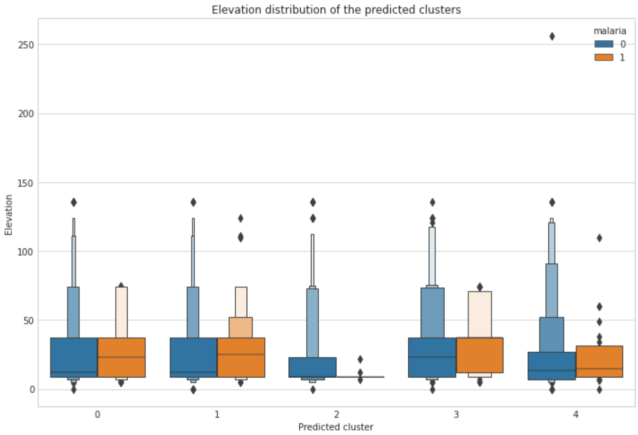

Figure 2: Elevation clusters obtained with the GMM

Moreover, two other (2 and 4) clusters' means for malaria are significantly different from the means of all other clusters. This is due to the fact that they have the lowest percentage of malaria cases among all clusters, by an order of magnitude from clusters 0, 1 and 3. This shows that the model has successfully clustered observations and that the 3rd spatial dimension of elevation adds meaningful insights for surveillance activities, especially for vector-borne disease such as malaria (where the habitat of the mosquito has circumcribed elevations).

5 Discussion

Although the model has been proven to be efficient, the process may still be optimized. In the preprocessing step, the latitude and longitude completion may be run only once, after the ward names have been corrected to reduce preprocessing time. If the ward names were input from a drop-down list during data collection as suggested by Zeineb Sahnoun [Sah20], this would also preserve more data points.

Having more data points would improve the training of the model. Indeed, the H-VAE is a neural network and by nature, these models require a lot of data to be trained. As there are only 2,920 useful data points, the model may perform poorly on new data as the new data's variance will differ from the training dataset's and so the latent space's representation may not be optimal for the new data. What's more, the training data is most likely unrepresentative of the Tanzanian population. Even in the training data itself, the dates range from December 10th, 2014 to February10th, 2016. The months of December, January and February are therefor over-represented in the dataset, giving more weight to seasonal diseases occurring in these months than to others. Nevertheless, the data was not truncated in order to improve training.

The quality of the prediction was checked for common infectious diseases in Tanzania with known and checked-against symptoms, but not against new infectious diseases. Therefore, the model has not been shown to improve the detection of emerging infectious diseases. If it was given enough historical data however, it may detect new clusters based on past trends. As a result, reemerging diseases could be detected.

Finally, the model is non-deterministic, so training it twice would yield different results. Hence, the whole model needs to be saved in order to be able to replicate them.

6 Future Work

This work can be improved by enhancing the quality of collected data through the use of a drop-down menu for ward names.

The same hyperparameters were used to train each of the three VAEs used in the H-VAE, which may not be optimal as they each work on different data. Therefore, the model would benefit from hyperparameter tuning for each VAE independently.

To determine the quality of the algorithm's detection capabilities for new infectious diseases, it would need to be evaluated on a much larger dataset containing years of data to see whether it can cluster new infectious diseases at the time when they appear. Such a dataset may also be used to evaluate the quality of detection for multiple reemerging infectious diseases to compare against current results.

References

[SSM04] Michael A. Stoto, Matthias Schonlau, and Louis T. Mariano. "Syndromic Surveillance: Is it Worth the Effort?" en. In: *CHANCE* 17.1 (Jan. 2004), pp. 19–24. ISSN: 0933-2480, 1867-2280. DOI: 10.1080/09332480.2004.10554882. URL: http://www.tandfonline.com/doi/full/10.1080/09332480.2004.10554882 (visited on 12/10/2020).

[MMR13] Stephen E Mshana, Mecky Matee, and Mark Rweyemamu. "Antimicrobial resistance in human and animal pathogens in Zambia, Democratic Republic of Congo, Mozambique and Tanzania: an urgent need of a sustainable surveillance system". In: *Annals of clinical microbiology and antimicrobials* 12.1 (2013), p. 28.

[MCM16] Nyambura Moremi, Heike Claus, and Stephen E Mshana. "Antimicrobial resistance pattern: a report of microbiological cultures at a tertiary hospital in Tanzania". In: *BMC infectious diseases* 16.1 (2016), p. 756.

[Kei+17] Kristina Keitel et al. "A novel electronic algorithm using host biomarker point-of-care tests for the management of febrile illnesses in Tanzanian children (e-POCT): A randomized, controlled non-inferiority trial". In: *PLOS Medicine* 14.10 (Oct. 2017), pp. 1–29. DOI: 10.1371/journal.pmed.1002411. URL: https://doi.org/10.1371/journal.pmed.1002411.

[KD18] Kristina Keitel and Valérie D'Acremont. "Electronic clinical decision algorithms for the integrated primary care management of febrile children in low-resource settings: review of existing tools". In: *Clinical Microbiology and Infection* 24.8 (2018), pp. 845–855. ISSN: 1198-743X. DOI: https://doi.org/10.1016/j.cmi.2018.04.014. URL: http://www.sciencedirect.com/science/article/pii/S1198743X18303525.

[Sim+19] Nikola Simidjievski et al. "Variational Autoencoders for Cancer Data Integration: Design Principles and Computational Practice". English. In: *Frontiers in Genetics* 10 (2019). ISSN: 1664-8021. DOI: 10.3389/fgene.2019.01205. URL: https://www.frontiersin.org/articles/10.3389/fgene.2019.01205/full (visited on 12/15/2020).

[Sah20] Zeineb Sahnoun. "Detection and Visualization of Patterns in Medical Data". en. In: (June 2020), p. 26.

[Zha] Shengjia Zhao. *A Tutorial on Information Maximizing Variational Autoencoders (InfoVAE)*. URL: https://ermongroup.github.io/blog/a-tutorial-on-mmd-variational-autoencoders/ (visited on 12/15/2020).