SEMI-SUPERVISED CLUSTERING FOR OIL PROSPECTIVITY

Anonymous authors
Paper under double-blind review

ABSTRACT

We create a new semi-supervised clustering method, that can be used in a wide range of problems. The clustering is based on a graph Laplacian similarity matrix and has a closed form solution, which enables us to cluster very large unstructured datasets rapidly ($> 10^7$ samples). In this work, we apply our method to oil prospectivity, based on seismic surveys and intersecting boreholes. Our clustering method can be extended to a pseudo-labelling deep learning scheme, and furthermore have the capability to utilize an octree mesh for larger datasets or faster clusterings.

1 INTRODUCTION

Semi-supervised learning (SSL) has been rapidly improving in the last few year, as evidenced by the Cifar 10 benchmark test (Krizhevsky et al., 2009), which in the last 5 years have had at least 15 different SSL state-of-the-art methods that have taken the accuracy from 79.6 % (Rasmus et al., 2015) to an impressive 97.3 % (Xie et al., 2019). These rapid improvements, combined with its broad domain of applicability, makes SSL particularly attractive for many problems in science.

While the usage of machine learning is growing rapidly, there are several issues holding it back, especially in science; one thing is its notorious reputation for being a black box, from which little insight is gained (Castelvecchi, 2016). Secondly, machine learning is often associated with a heavy computational burden, which prohibits a lot of people from using it (Karras et al., 2019).

In the following, we approach semi-supervised learning from a minimalistic point of view. We create an objective function, which we minimize in order to develop a graph Laplacian based semi-supervised clustering method (Von Luxburg, 2007). In this way, we are capable of creating a clustering algorithm with a closed form expression that requires very limited computational power.

Our objective function can be thought of as an extension/improvement of the objective developed in Zhou et al. (2004). The key difference between our objective function and theirs, is that ours includes a constraint. Without this constraint, the clustering can predict the probable order of classes (i.e. which classes are more/less likely than others), but not their actual probabilities. Furthermore, while the constraint does add extra steps into the calculations, the overall result is a boost in computational time, due to the fact that the constraint can replace one of the linear systems that would normally need to be solved. This is particularly relevant in our two class example, where the clustering time is nearly halved. Finally, the constraint enables us to handle problems where the known labels contain a mix of different classes, as shown in our example.

Iscen et al. (2019) used the objective given in Zhou et al. (2004) and combined it with a deep neural network in a pseudo-labelling approach to perform image classification. Their approach achieved state-of-the-art results, and shows how a clustering approach can be combined with a deep neural network. While their results are impressive, their approach does have one issue. Since they use the objective function developed in Zhou et al. (2004), their clustering return pseudo-probabilities, which should not be used as entropy weights. By including our aforementioned constraint, this could easily be remedied.

Instead of combining our clustering method with a deep neural network and doing the standard benchmarks, we chose to highlight the broad applicability of our method. Thus, we show how our method can be applied to oil prospectivity. Predicting oil is a highly sought target in petrophysics,

and is traditionally regarded as being a notoriously hard problem, due to the convoluted relationship between the parameters typically derived from a seismic survey and oil. Yu et al. (2008) predicts oil reservoirs using a fusing of genetic algorithms, simulated annealing and error back propagating neural networks. Jian & Fanhua (2009) predicts oil bearing sand using seismic inversion combined with 3D geologic modelling and petrophysical relations. Powers et al. (2018) utilize a Markov chain Monte Carlo method and combine it with a naive Bayesian classifier, which enables them to bin the well production and produce an oil prediction map.

A standard inversion of a seismic survey gives a set of parameters discretized in a 3D volume. In the framework of machine learning, each point in this 3D volume can be thought of as an unlabelled data point, with the parameters generating the features of each data point. Boreholes intersecting the data volume can be utilized to generate labelled data. Borehole data are prohibitively expensive to get in comparison to seismic data, so typically the amount of labelled data will be severely limited in comparison to the amount of unlabelled data, which is exactly the domain of semi-supervised learning.

METHOD

Assume that we are given a dataset $\mathbf{X} \in R^{n \times l}$, where n is the number of data points, and l is the number of features. Hence, each row of the matrix represents a data point. Let $\mathbf{X}_k$, be a known labelled subset of $\mathbf{X}$, with prior associated label probabilities $\mathbf{U}_k^{obs} \in R^{n_k \times n_c}$, where n_c is the number of classes in the data, and n_k is the number of known labelled points. The goal in semi-supervised learning is to find a probability matrix $\mathbf{U} \in R^{n \times n_c}$, which matches $\mathbf{U}_k^{obs}$ on the known subset and gives a reasonable estimate of the probabilities on all the remaining data points.

CREATING AN OBJECTIVE FUNCTION

Our objective function is given as:

$$\begin{aligned} \phi(\mathbf{Y}) = \frac{\alpha}{2n} \operatorname{Tr}\left[\mathbf{Y}^\top \mathbf{L} \mathbf{Y}\right] + \frac{1}{2}||\mathbf{Y} - \mathbf{Y}^{obs}||_W^2, \\ \text{st. } \mathbf{Y}\mathbf{e} = 0, \end{aligned} \tag{1}$$

with α being a scalar hyper-parameter, determining the relative strength between the two terms. $\mathbf{L}$ is the symmetric-normalized graph Laplacian (Von Luxburg, 2007), $\mathbf{e}$ is a unity vector, and $\mathbf{Y}$ is a pseudo-probability matrix, connected with the normal probability $\mathbf{U}$ through a softmax normalization:

$$\mathbf{U} = \exp(\mathbf{Y}) \oslash \left(\exp(\mathbf{Y})\mathbf{e}\mathbf{e}^\top\right). \tag{2}$$

The idea is that a minimization of equation 1 will lead to reasonable label probabilities, since the second term penalizes known data points that do not match their label pseudo-probabilities, while the first term connects all points with their most similar points and encourage them to have similar label probabilities. The constraint ensures that each probability corresponds to exactly one pseudo-probability, which enables us to find $\mathbf{Y}^{obs}$ based on $\mathbf{U}^{obs}$:

$$\mathbf{Y}^{obs} = \log(\mathbf{U}^{obs}) - \frac{\log(\mathbf{U}^{obs})\mathbf{e}}{\mathbf{e}^\top \mathbf{e}} \otimes \mathbf{e}^\top. \tag{3}$$

By taking the derivative of the objective function given in equation 1, a closed form minimization of $\mathbf{Y}$ can be found:

$$\mathbf{Y} = \left(n^{-1}\alpha\mathbf{L} + \mathbf{W}\right)^{-1} \mathbf{W}\mathbf{Y}^{obs}\mathbf{C}, \tag{4}$$

where $\mathbf{C} = \mathbf{I} - \frac{\mathbf{e}\mathbf{e}^\top}{\mathbf{e}^\top\mathbf{e}}$ is a centering matrix, and $\mathbf{W}$ is the diagonal normalized weight matrix, which fulfils the normalization, $\operatorname{Tr}(\mathbf{W}) = 1$, and has zeros on the diagonal for all unknown points.

Equation 4 can for some problems be ill-conditioned. In those cases a small identity matrix can be added to stabilize the solution. Note that the system is decoupled over the classes, so each class can be solved independently. Furthermore, since the solution by construction will fulfil the constraint $\mathbf{Y}\mathbf{e} = 0$, we only have to solve equation 4 for $n_c - 1$ classes. So the solution strategy is to solve equation 4 for $n_c - 1$ classes, use the constraint $\mathbf{Y}\mathbf{e} = 0$ to get the last class pseudo-probability, and then convert the pseudo-probabilities to probabilities using the softmax normalization given in equation 2.

EXAMPLE - SEAM LIFE OF FIELD

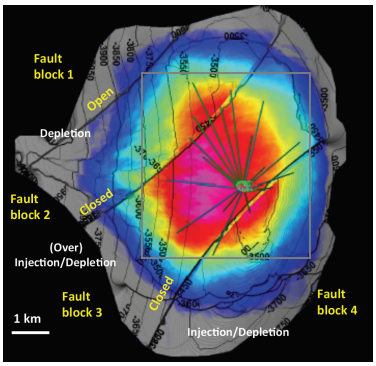
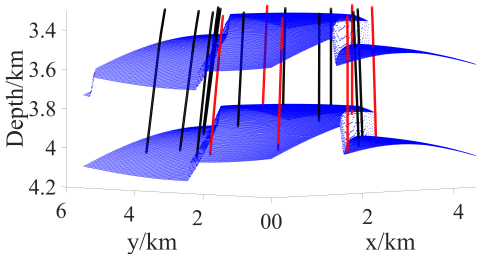
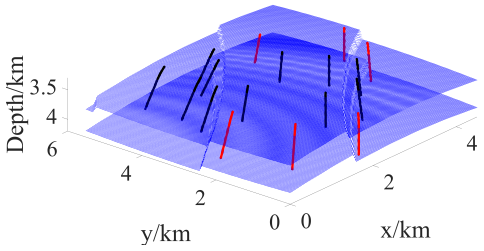


Figure 1: Overview of the SEAM Life of Field dataset, taken and modified from Oristaglio (2016). The model shows the fault lines, as well as the 17 boreholes that goes through the reservoir. The superimposed color scale shows vertically summed oil volume, with purple being high production and dark blue being low. The gray square indicates our modelling area.

We apply our clustering approach to the synthetic SEAM Life of Field dataset (Oristaglio, 2016; Oppert et al., 2017), which is a highly realistic synthetic oil reservoir, containing: a gas cap, an oil leg, a brine section below, and three fault lines - as shown in Figure 1. Our data covers the central region of the oil zone, and is shown in Figure 2. The data consist of 181 x 221 x 157 data points (x,y,z). In each data point we have five inverted parameters, provided from an amplitude versus angle inversion: density, porosity, lithology, VpVs, and acoustic impedance. Furthermore, we have 17 boreholes going through our dataset - 11 production wells, and 6 injection wells, each with a well-log.



(a) 3D overview



(b) 3D overview

Figure 2: Overview of our SEAM dataset. The two blue planes are the top and bottom layer of our data, while the black lines are production wells and the red lines are injection wells.

DATA PREPARATION

Each parameter is standardized by setting the mean to zero and the variance to one.

As features, we utilize the value and x/y/z-neighbours mean and difference of each of the 5 parameters given to us. This means that each parameter is represented by 7 features that gives information about its value as well as gradient. Finally, for added data-locality we chose to include the standardized x,y,z coordinates as features. In total this leads to each data point having 38 features.

The graph Laplacian is based on the approximate nearest neighbours approach known as Hierarchical Navigable Small World graphs (Malkov & Yashunin, 2018), with an ℓ_2 metric and 40 neighbours.

For this problem, we assign each data point two classes, oil and no-oil, and assume that the oil probability is linearly correlated with oil volume. As labelled data we utilize the data points closest to each borehole in each layer, and assign oil volume based on the well-log information.

The labels we assign to the points belonging to a well sums up to the oil produced by that well, with the amount of oil in each point being inversely proportional to the fractional shale content of the point, as provided by the well-log. This means that if the well logs lithology says there is 100 % shale in a point, its label will have 0 oil probability. For the injection wells we only label points containing 100 % shale, these gets labelled as 0 oil data points. On top of the labelled points we get from the well-logs, we also apply boundary constraints. Applying boundary constraints is not strictly necessary, but helps stabilize the solution. In this particular example, we have production wells relatively close to the edge of our simulation area, and thus we need to set very loose boundary constraints. We set the boundary points on all 4 sides to 0 with a weight of 10^{-4} (all other labelled points have a weight of 1).

In total this leads us to have 6,280,157 data points, 2068 labelled data points from boreholes, and 126,228 weakly labelled boundary points.

RESULTS

Following Iscen et al. (2019), we set $\alpha = 100$, which gives a reasonable balance between matching the labelled data points and fulfilling the regularization constraints. From the clustering we get an oil volume prediction in each data point, which shows a clear central oil deposit, bounded by the fault lines. Figure 3 features two slices through the oil deposit and shows that the clustering looks reasonable and consistent on an individual data point level.

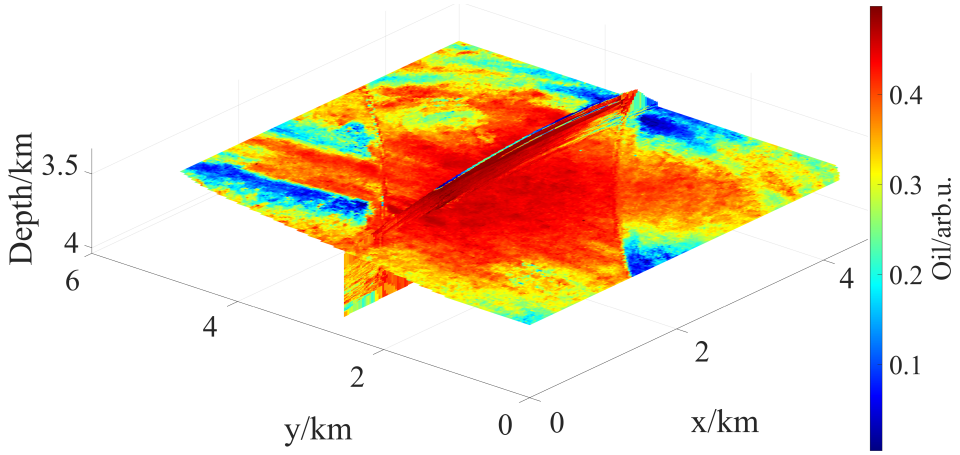


Figure 3: 3D slices of the clustering. The oil deposit is clearly visible in the central fault block. Furthermore, the clustering is showing the foldings in the earth.

From this clustering we can estimate the total oil volume, by vertically summing the oil in each layer. Figure 4 shows the total oil, scaled to match the boreholes. The scaling is made in order for us to estimate the error of our clustering and depends on how we define the oil produced by a well. We tested a few simple approaches and ended up using the following: Each well is approximated as a single vertical drilling, with its horizontal location given as the mean horizontal location of all points of the well intersecting the modelling area. Around each well we define all points $\mathbf{x}$ within 100 m as contributing to the well. The production of the well, P, is then given as:

$$P = \frac{1}{2}\min(\mathbf{p}) + \frac{1}{2n_x}\mathbf{p}^\top \mathbf{w}, \tag{5}$$

where $\mathbf{p}$ is the production belonging to the selected points $\mathbf{x}$, n_x is the number of point in $\mathbf{x}$, and $\mathbf{w}$ is the weight of each point, which is related to the points ℓ_2 distance, $\mathbf{d}$, from the well: $\mathbf{w} = \frac{1}{1+\mathbf{d}}$. Defining the oil production in this way, leads to a simple yet robust model.

Based on this, we can estimate the error of our clustering. We find an average production well error of 21.4 %. The oil volume shown in Figure 4 shows similar trends to the oil volume shown in Figure 1.

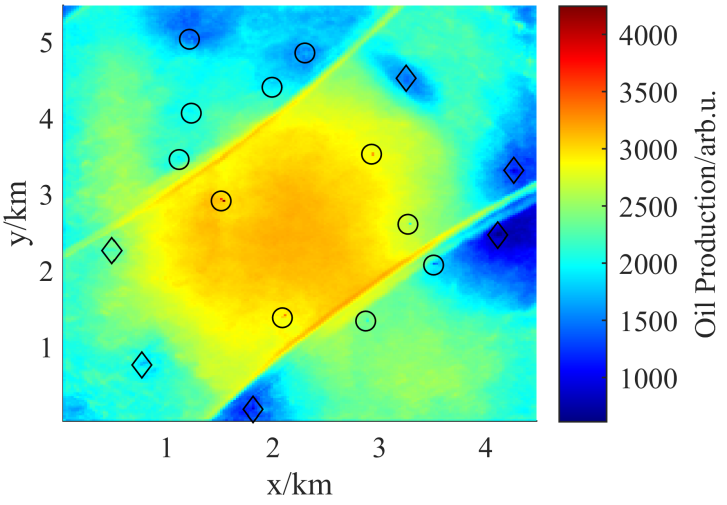


Figure 4: Oil prediction based on our semi-supervised clustering approach. The black circles designate production wells, while the diamonds designates injection wells.

To further test our method, we have done cross-validation on the above dataset, where we leave out 4 of the 11 production wells and test the predictive power across all possible 330 combination this result in, which increased the error to 36.15%.

PERFORMANCE

The approach presented here offers a simple method for predicting oil production that can be done on even modest computers. Calculating the graph Laplacian is the heaviest computational task, which takes ~ 5 hours for ~ 6 million data points with 40 neighbours. When the graph Laplacian has been computed, a clustering can be done in ~ 15 minutes on a single core on an Intel Xeon Processor E5-2670.

DISCUSSION & FUTURE WORK

Our oil clustering results are reasonable, but could likely be improved with more complex/realistic assumptions. Currently, we ignore any lateral change in the boreholes intersection path through the data volume, when we scale the oil volume to match the boreholes, even though all boreholes in this example have some lateral change. The dataset has previously been explored with Bayesian machine learning techniques as presented in Powers et al. (2018), which used the same amplitude versus angle inversion properties as this work. Their method required serious computational power, and had trouble predicting the sharp contrasts encountered around the fault lines, which our method seems to handle better.

We have also applied the method successfully to field data, where we had even larger datasets consisting of roughly 9.6M samples. In that case we utilized octree meshing (Haber & Heldmann, 2007; Horesh & Haber, 2011), to increase performance.

CONCLUSIONS

We have developed a simple semi-supervised clustering approach, which can be combined with deep neural networks to perform generative image classification, as done in Iscen et al. (2019). In this work however, we use the clustering approach to successfully predict oil. We hope our method and example can help illuminate a field which is often regarded as a black box, and show one way to recast problems into the framework of semi-supervised learning, by using simple assumptions.

ACKNOWLEDGEMENTS

Will be added after double blind review.

REFERENCES

Davide Castelvecchi. Can we open the black box of ai? *Nature News*, 538(7623):20, 2016.

Eldad Haber and Stefan Heldmann. An octree multigrid method for quasi-static Maxwell's equations with highly discontinuous coefficients. *Journal of Computational Physics*, 223(2):783–796, 2007. ISSN 00219991. doi: 10.1016/j.jcp.2006.10.012.

Lior Horesh and Eldad Haber. A Second Order Discretization of Maxwell's Equations in the Quasi-Static Regime on OcTree Grids. *SIAM Journal on Scientific Computing*, 33(5):2805–2822, 2011.

Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondrej Chum. Label propagation for deep semi-supervised learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5070–5079, 2019.

Wu Jian and Li Fanhua. Prediction of oil-bearing single sandbody by 3d geological modeling combined with seismic inversion. *Petroleum exploration and development*, 36(5):623–627, 2009.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. *arXiv preprint arXiv:1912.04958*, 2019.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

Yury A Malkov and Dmitry A Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

Shauna Oppert, Joseph Stefani, Daniel Eakin, Adam Halpert, Jorg V Herwanger, Andy Bottrill, Peter Popov, Lijian Tan, Vincent Artus, and Michael Oristaglio. Virtual time-lapse seismic monitoring using fully coupled flow and geomechanical simulations. *The Leading Edge*, 36(9):750–768, 2017.

Michael Oristaglio. Seam update: Integrated reservoir and geophysical modeling: Seam time lapse and seam life of field. *The Leading Edge*, 35(10):912–915, 2016.

Hayden Powers, Whitney Trainor-Guitton, and G Michael Hoversten. Classification of total oil production of wells in seam life of field from stochastic ava inversion attributes via machine learning. In *SEG Technical Program Expanded Abstracts 2018*, pp. 2131–2135. Society of Exploration Geophysicists, 2018.

Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. Semi-supervised learning with ladder networks. In *Advances in neural information processing systems*, pp. 3546–3554, 2015.

Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Unsupervised data augmentation for consistency training. 2019.

Shiwei Yu, Kejun Zhu, and Fengqin Diao. A dynamic all parameters adaptive bp neural networks model and its application on oil reservoir prediction. *Applied mathematics and computation*, 195 (1):66–75, 2008.

Dengyong Zhou, Olivier Bousquet, Thomas N Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In *Advances in neural information processing systems*, pp. 321–328, 2004.