Manifold Learning for Innovation Funding: Identification of Potential Funding Recipients

Vincent Grollemund[1,2(✉)], Gaétan Le Chat[2], Jean-François Pradat-Peyre[1,3], and François Delbot[1,3]

[1] Sorbonne Université, 75005 Paris, France
vincent.grollemund@lip6.fr
[2] FRS Consulting, 75009 Paris, France
[3] Nanterre Université, 92014 Nanterre, France

Abstract. finElink is a recommendation system that provides guidance to French innovative companies with regard to their financing strategy through public funding mechanisms. Analysis of financial data from former funding recipients partially feeds the recommendation system. Financial company data from a representative French population are reduced and projected onto a two-dimensional space with Uniform Manifold Approximation and Projection, a manifold learning algorithm. Former French funding recipients' data are projected onto the two-dimensional space. Their distribution is non-uniform, with data concentrating in one region of the projection space. This region is identified using Density-based Spatial Clustering of Applications with Noise. Applicant companies which are projected within this region are labeled potential funding recipients and will be suggested the most competitive funding mechanisms.

Keywords: Dimension reduction · Manifold learning · Clustering

1 Introduction

Given the diversity and quantity of unstructured information available on existing French funding mechanisms, innovative companies need guidance with regard to their financing strategy. finElink [4] is a recommendation system that meets this need. Developed by FRS Consulting, a French consulting firm specialized in public innovation funding, it was initially based on business knowledge of FRS Consulting associates. Analysis of financial data from former French funding recipients, using machine learning, helped identify applicant companies with a high potential and further enhance finElink's recommendation.

However, relevance of applicant companies cannot be solely assessed on former funding recipient data, as these data suffer from significant data sparsity,

Published by Springer Nature Switzerland AG 2020
I. Maglogiannis et al. (Eds.): AIAI 2020, IFIP AICT 583, pp. 119–127, 2020.
https://doi.org/10.1007/978-3-030-49161-1_11

bias and missing data constraints. Funding recipients data were obtained through cross-checking information from funding mechanism websites and the French national company registry where companies' financial statements are available. Financial information was frequently missing, specifically for newly created companies which are the targeted recipients of numerous funding mechanisms. Data collected had a significant number of missing features. Moreover, most funding mechanisms did not communicate on their recipients, especially for small funding mechanisms. As such, available funding recipient data were strongly biased towards well-known funding mechanisms. Supervised learning on data with these limitations would have easily led to overfitting. These limitations were avoided using unsupervised learning and another larger dataset of French companies obtained using a proprietary software. This other dataset was representative of all French companies and suffered from fewer missing data.

These representative company data were reduced and projected into a two-dimensional space with Uniform Manifold Approximation and Projection (UMAP), a manifold learning algorithm. Former funding recipient data were then projected into the new space. Funding recipient projections showed an uneven distribution pattern with funding recipients concentrating in one projection space area. This area was identified using a density-based clustering method: Density-based Spatial Clustering on Applications with Noise (DBSCAN) [3].

This study presents our approach to use this target population of funding recipients in order to isolate a sub-population of potential funding recipients within a large representative population. This approach is neighborhood-based and combines a manifold learning algorithm with a density-based clustering method. Section 2 will present the data used and the data processing steps. Section 3 will focus on data reduction results. The conclusion will be addressed in Sect. 4.

1.1 Previous Work

Dimension reduction intends to represent high-dimensional data in a low-dimensional space while preserving data structure. Linear dimension reduction algorithms, namely Principal Component Analysis (PCA), strive to preserve global input data structure but describe poorly the true geometry of nonlinear data [7]. In this study, former funding recipient and representative populations had similar spatial distributions in the low-dimensional space, hence PCA was unable to isolate the target population from the representative population. Nonlinear dimension reduction algorithms, also referred to as manifold learning algorithms, can describe a wider range of variable interactions [14]. They are usually divided into two categories based on whether they focus on local or global data structure preservation. Global nonlinear dimension reduction algorithms such as Kernel PCA [10] or Isometric feature Mapping (ISOMAP) [17] strive to preserve input data geometry at all scales: neighborhood and remoteness are preserved between the input and output spaces. Local nonlinear dimension reduction algorithms such as Locally Linear Embedding (LLE) [11] or t-Stochastic Neighbor Embedding (t-SNE) [8] focus primarily on local geometry preservation in

small neighborhoods of the manifold [14]. Recently developed, UMAP [9] falls into this last category. The local approach has two advantages over the global approach: first, lower computational complexity as computations involve sparse matrix manipulation, second, an enhanced ability to represent a wider range of manifolds, specifically when geometry is Euclidean at a local scale but is non-Euclidean at a global scale. t-SNE has proven to balance well local and global data structures on real life data giving t-SNE a competitive edge. This was not the case for LLE and its other nonlinear counterparts [8]. But t-SNE suffers from several drawbacks [13,18], which do not affect UMAP, such as:

- inability to scale computationally when working with widely used python libraries;
- non-convexity of its cost function, leading to potential initialization-based results;
- non-preservation of density and distances between the input and output spaces (neighborhood is nonetheless preserved).

Due to computational scaling constraints and the inability to use distances in the output space, UMAP was preferred to t-SNE. UMAP has already been successfully used in various medical contexts such as survival prognosis estimation of Amyotrophic Lateral Sclerosis (ALS) patients [6], gene co-expression analysis [5] and infection risk prediction of newly diagnosed B-cell chronic lymphocytic leukemia (B-CLL) patients [1].

Applying UMAP in the context of public innovation funding is original with regard to both testing this recent manifold learning algorithm and experimenting on novel data in a field where public data is sparse.

2 Methods

2.1 Data

The first dataset was our target population of French funding recipients which included 3,350 samples. The second dataset was our representative population of French companies which contained 152,899 instances, randomly sampled from the Amadeus database [2]. As such, companies sampled from that database were selected with less bias than funding recipient data. Features selected for this study were limited to turnover, net income, equity and headcount over a three-year period. Data were not processed as time-series. Feature selection was based upon finElink's use case: information asked to users needed to be easily available to improve user-friendliness. Age, business sector and location information was excluded as these features were not continuous. Age was discretized in years. Business sector and location were categorized with respectively NACE codes and region names. When categorical or discretized features are included in a UMAP projection, the algorithm primarily learns how to represent these different categories or bins without providing additional information on the data. As such, these UMAP projections were unable to isolate the target population from the representative population when these features were included.

2.2 Missing Data Analysis

Missing data were imputed using MissForest [15], a multiple imputation method based on a random forest model. Multiple imputation methods preserve input data distribution better than single imputation methods. MissForest has a good tolerance for high missing data rates and can handle Non Missing At Random (NMAR) schemes [16]. Multiple Imputation by Chained Equations (MICE) [12] is another multiple imputation method based on regression. Both MICE and MissForest deal with mixed data types (categorical and/or continuous). However, MissForest is non-parametric and, as such, can handle non-linearity and variable interactions in data, which MICE cannot. Initial missing data rates were 58% and 34% for respectively the target and representative populations. Given the high missing data rates, data imputation on the overall available population would have been inappropriate. Data with up to 7 missing features, on a total of 15 (age, business sector and location were included for missing data imputation) were selected. As such, initial feature distributions were not significantly altered after data imputation. Data were normalized prior to missing data imputation. After processing, the two datasets included 1,413 and 114,628 samples for respectively the target and representative populations.

2.3 Dimension Reduction

Data reduction was carried out using UMAP. The representative population was projected into a two-dimensional space. UMAP is neighborhood-based and works in two steps. First, a compressed embedding of the input space is built through topological analysis of the data structure using simplexes[1]. Second, a low-dimensional (in our case two-dimensional) data embedding is found through a cross-entropy[2] optimization process. UMAP preserves data neighborhoods, distances and density. The initial modeling step depends on whether the algorithm should focus on preserving the local or global input data structure. Data structure is estimated according to the size of the neighborhood investigated. The second compression step is mainly defined using two parameters which are the output dimension size and the minimum distance permitted between two points in the output space, i.e. how compact the output projection can be. In our study, UMAP parameters were set as follows:

- output dimensionality was set to 2, as adding an additional dimension did not provide more insight;

[1] In geometry, a simplex is defined as a set of points, where none is a barycentre of the remaining points. The convex hull of these points corresponds to the face of the simplex. In simpler terms, a n-simplex can be thought of as the generalization of a triangle in the n^{th} dimension.

[2] In machine learning, cross entropy is frequently used as a cost function to compare two probability distributions (p,q): p is optimized to approximate q the fixed target distribution.

- neighborhood size was set as high as possible given computational time (6,500) in order to obtain a global overview of data structures, funding recipients were not isolated when the focus was made on local data structure;
- no minimum distance in the output space was set to allow overlapping.

2.4 Clustering

The UMAP projection space was divided into a grid and density differences within that grid were examined using the ratio of funding recipient samples within each cell over the total cell samples. Centroid-based clustering methods are not relevant given the data distribution as they are unable to deal with noise. Density-based clustering methods, such as DBSCAN, manage noise through density analysis which meets our problem's constraints. In DBSCAN, cluster identification is carried out by assessing the neighborhood density of each sample, i.e. evaluating the number of neighbors within an ϵ radius of that sample. Provided the number of neighbors is above the user-defined threshold, that sample is said to be a cluster core point. If the sample does not have enough neighbors within an ϵ radius while having at least one core point as a neighbor, then that point is assigned to the core point's cluster. Otherwise, that point is labeled as noise. Projections from the target population were fed into DBSCAN to isolate the projection space area with a high density of target population samples. The remaining samples were labeled as noise. DBSCAN tuning led to the following setup:

- the ϵ distance was set to the first percentile of the target population distance distribution;
- the minimum number of points within a ϵ radius required to form a cluster was set to 20.

3 Results

3.1 Input Feature Distribution Analysis

As UMAP is a non-linear dimension reduction method, projection features cannot be analyzed to provide any interpretability. Output dimension analysis, as commonly performed for PCA, cannot be carried out. Nonetheless, analysis of input feature distribution in the UMAP projection space is an alternative as it gives a broad overview of variable importance with regard to the projection. Plot analysis can help identify strong correlations between projections and input features. This was the case for turnover and headcount variables for year N-1, presented in respectively Fig. 1a and Fig. 1c. These variables appeared to have an impact on the overall data projection pattern. Net income and equity variables did not show a high degree of correlation with the projection, as shown respectively in Fig. 1b and Fig. 1d, as feature distribution appeared to be random in some projection space areas. Results were plotted for year N-1, but the patterns were similar for the two other years (N-2 and N-3). Turnover and headcount

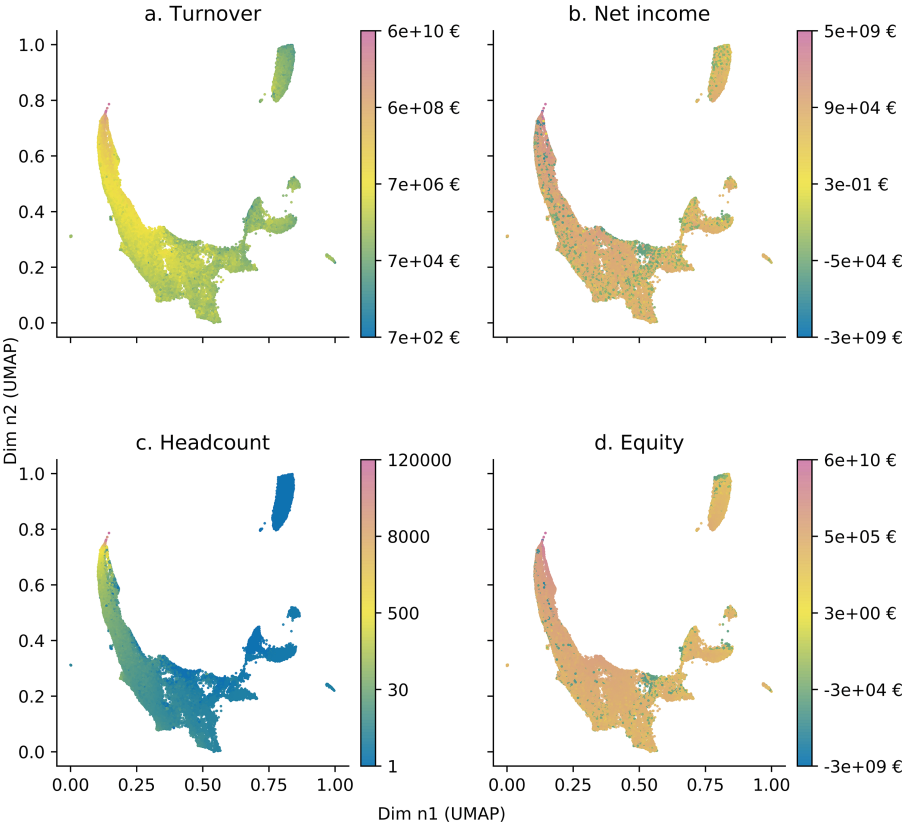

Fig. 1. Input feature distribution for samples from the representative population of French companies: turnover(a), net income (b), headcount (c) and equity (d) for the year N-1. Axes are dimensionless and come from UMAP dimension reduction (a, b, c, d).

appeared to be the variables that mattered the most distance-wise in the output space. Net income and equity, which showed a weaker or limited impact on the overall UMAP projection distribution, might have had a more local impact distance-wise in the output space.

3.2 Funding Recipient Distribution Analysis

Funding recipient samples were then projected onto the low-dimensional space. Distribution patterns for funding recipients were different from those observed for the representative population as shown in Fig. 2a. Funding recipients were prone to concentrate primarily in the left region of the projection in the shape of a curve. The curve went from the projection's upper left side to the lower center region, in the shape of a "backbone". Projection space division into a grid, presented in Fig. 1b, helped understand the projection space density distribution.

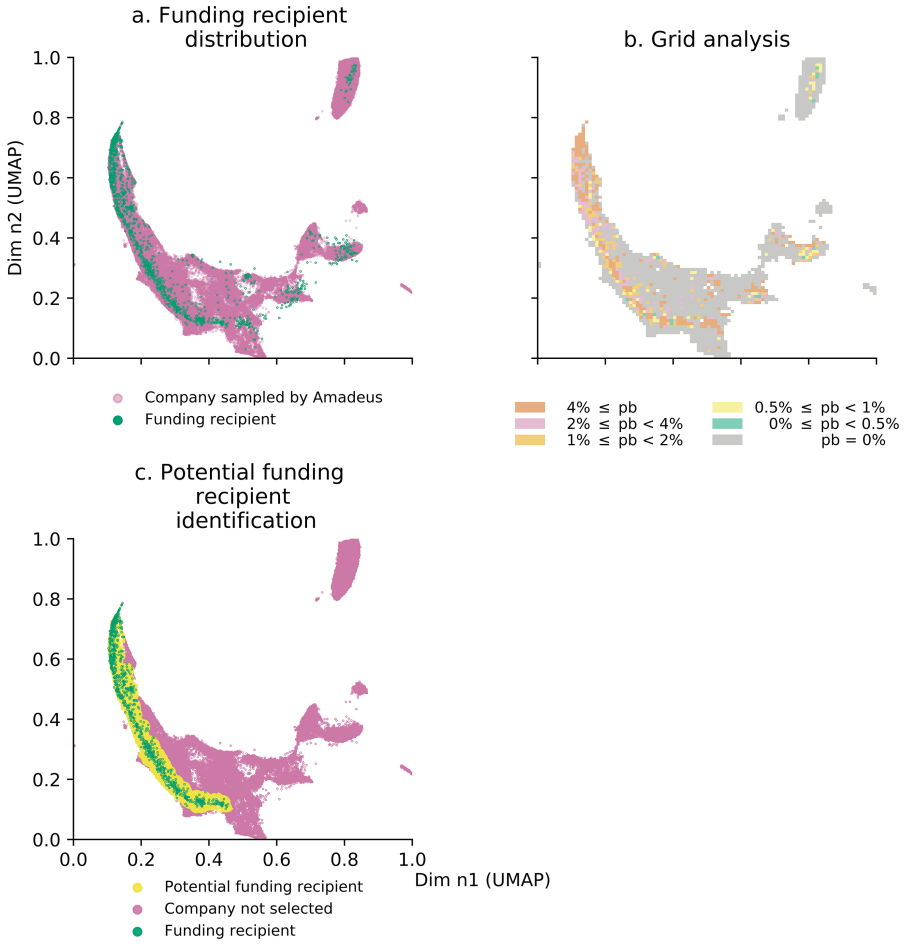

Fig. 2. Funding recipients were projected onto the two-dimensional plane with a non-uniform distribution pattern (a). The projection space was divided in a grid to analyze density of funding recipients within each cell(b). DBSCAN identified the main cluster of funding recipients. Companies close to the funding recipient cluster belonged to the potential funding recipient cluster (c). Axes are dimensionless and come from UMAP dimension reduction (a, b, c).

Density analysis confirmed the shape identification. Funding recipient samples were mainly located within the "backbone" shape as 74% of funding recipients belonged to it (i.e. 1,041 out of the 1,413 funding recipient samples). Funding recipient concentration within a specific projection space area confirmed that our similarity-based approach on financial features for potential funding recipient identification was relevant. DBSCAN was then applied on the target population and its main cluster was identified. The remaining funding recipients were labeled as noise. Company samples from the representative population that were

within an ϵ radius of cluster core points were labeled potential funding recipients as shown in Fig. 2c. Membership to the main cluster was fed into the recommendation system and potential funding recipients were suggested more competitive funding mechanisms.

4 Conclusion

Our study demonstrated that our approach successfully isolated a subset of companies which shared similarities with the target population of former funding recipients. Combining a novel non-linear dimension reduction method with a density-based clustering algorithm proved to be most instructive. Similarity was assessed using a limited set of financial features: turnover, net income, headcount and equity over a period of three fiscal years. Our approach can be summarized in three stages. First, representative company data were projected onto a low-dimensional space using the manifold learning algorithm UMAP. Second, former funding recipient data were projected onto that same low-dimensional space. Third, the cluster with the highest density of former funding recipients was identified using the density-based clustering algorithm DBSCAN. Companies close to that cluster, either from the representative company dataset or newly added from a finElink user, were separated from the rest. They were deemed to be more successful than their counterparts. Finelink suggestions were personalized to companies' financial information as companies with higher chances of success were proposed the most competitive funding mechanisms while the others were offered smaller funding mechanisms. Further recommendation system tuning includes analyzing funding recipients from mechanisms with similar characteristics in order to identify distribution patterns specific to these sub-groups. Additionally, this approach can be extended to other contexts for minority population identification within a larger population sample when facing strong data constraints. Notwithstanding significant data sparsity, bias and missing data constraints, we have demonstrated that combining non-linear dimension reduction with density-based clustering, important correlations can be unraveled.

References

1. Agius, R.: Machine learning can identify newly diagnosed patients with CLL at high risk of infection. Nat. Commun. **11**(1), 1–17 (2020)
2. Amadeus. https://www.bvdinfo.com/en-gb/our-products/data/international/amadeus. Accessed 25 Feb 2020
3. Ester, M., Kriegel, H., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: KDD, vol. 96, no. 34, pp. 226–231 (1996)
4. finElink. https://www.finelink.eu. Accessed 25 Feb 2020
5. Fornito, A., Arnatkevičiūtė, A., Fulcher, B.: Bridging the gap between connectome and transcriptome. Trends Cogn. Sci. **23**(1), 34–50 (2019)

6. Grollemund, V., Pradat, P.F., Delbot, F., Le Chat, G., Pradat-Peyre, J.F., Bede, P.: Manifold learning for ALS prognosis: development and validation of a prognosis model. Scientific Reports (submitted manuscript)
7. Lee, J., Verleysen, M.: Nonlinear Dimensionality Reduction. Springer, Heidelberg (2007). https://doi.org/10.1007/978-0-387-39351-3
8. Maaten, L., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**, 2579–2605 (2008)
9. McInnes, L., Healy, J., Melville, J.: UMAP: uniform manifold approximation and projection for dimension reduction. arXiv, preprint arXiv:1802.03426 (2018)
10. Mika, S., Schölkopf, B., Smola, A., Müller, K., Scholz, M., Rätsch, G.: Kernel PCA and de-noising in feature spaces. In: 11th International Conference on Neural Information Processing Systems, pp. 546–542. MIT Press, Denver (1998)
11. Roweis, S.: Nonlinear dimensionality reduction by locally linear embedding. Science **290**(5500), 2323–2326 (2000)
12. Schafer, J.: Analysis of Incomplete Multivariate Data. Chapman and Hall/CRC, London (1997)
13. Schubert, E., Gertz, M.: Intrinsic t-stochastic neighbor embedding for visualization and outlier detection. In: Beecks, C., Borutta, F., Kröger, P., Seidl, T. (eds.) SISAP 2017. LNCS, vol. 10609, pp. 188–203. Springer, Cham (2017)
14. Silva, V., Tenenbaum, J.: Global versus local methods in nonlinear dimensionality reduction. In: Advances in Neural Information Processing System, pp. 721–728 (2003)
15. Stekhoven, D.J., Bühlmann, P.: MissForest - non-parametric missing value imputation for mixed-type data. Bioinformatics **28**(1), 112–118 (2012)
16. Tang, F., Ishwaran, H.: Random forest missing data algorithms. Stat. Anal. Data Min.: ASA Data Sci. J. **10**(6), 363–377 (2017)
17. Tenenbaum, J.: A global geometric framework for nonlinear dimensionality reduction. Science **290**(5500), 2319–2323 (2000)
18. Wattenberg, M., Viégas, F., Johnson, I.: How to use t-SNE effectively. Distill **1**(10), e2 (2016)