Proceedings 1st International Workshop on Advanced Analytics and Learning on Temporal Data
AALTD 2015

Temporal Density Extrapolation

Georg Krempl
Knowledge Management & Discovery
Otto-von-Guericke University Magdeburg
Universitätsplatz 2, 39106 Magdeburg, Germany
georg.krempl@iti.cs.uni-magdeburg.de
https://kmd.cs.ovgu.de/res/driftmining

Abstract. Mining evolving datastreams raises the question how to extrapolate trends in the evolution of densities over time. While approaches for change diagnosis work well for interpolating spatio-temporal densities, they are not designed for extrapolation tasks. This work studies the temporal density extrapolation problem and sketches two approaches that addresses it. Both use a set of pseudo-points in combination with spatio-temporal kernel density estimation. The first, weight-extrapolating approach, uses regression on the weights of stationary-located pseudo-points. The second, location-extrapolating approach, extrapolates the trajectory of uniformly-weighted pseudo-points within the feature space.

Keywords: kernel density estimation, density extrapolation, density forecasting, spatio-temporal density, evolving datastreams, nonstationary environments, concept drift, drift mining

1 Introduction

Density estimation methods, like kernel density estimation [14, 13], allow to learn a model from instances observed at different positions in feature space, and to use this model to estimate the density at any position within this feature space. While the original work in [14, 13] is limited to spatial densities of a stationary distribution, the approach was extended for spatio-temporal density estimation of non-stationary distributions in [1, 2]. This so-called velocity density estimation allows to estimate and visualise trends in densities. However, these existing approaches are not directly applicable for predicting the density at *future time points*, for example to extrapolate trends in the evolution of densities and for building classifiers that work with delayed label information [10, 5]. Such *temporal density extrapolation* should predict the densities at spatio-temporal coordinates in the future, given a sample of (historic) instances observed at different positions in feature space and at different times in the past.

We propose and study two approaches to address this problem. Both use extrapolation of pseudo-points in combination with spatio-temporal kernel density estimation. The first approach extrapolates the weights of stationary located pseudo-points, while the second extrapolates the path of moving pseudo-points of fixed weight. Subsequently, these weight- or position-extrapolated pseudo-points are used in a spatio-temporal kernel density estimation.

In the following Section 2, we review the related work, before sketching the two approaches in Section 3 and concluding in Section 4.

2 Related Work

The task of estimating the probability density based on a sample of independently and identically distributed (iid) observations has been intensively studied. *Density estimation* methods, like kernel density estimation [14, 13], allow to learn a model from instances observed at different positions in feature space, and to use this model to estimate the density at any position within this feature space. The difficulties of the early kernel and near-neighbour density estimation techniques when extended to multivariate settings was addressed by approaches like *projection pursuit* density estimation, proposed in [4]. All these density estimation approaches, as well as related curve regression approaches, require an iid sample from a stationary distribution [11].

In the case of a nonstationary distribution, one might be interested in estimating the density at different points in time and space. In [1, 2], this problem of *spatio-temporal density estimation* is addressed by combining spatial kernel density estimation with a temporal weighting of instances. A framework for so-called *change diagnosis* in evolving datastreams is proposed, which estimates the rate of change at each region by using a user-specified temporal window to calculate forward and reverse time slice density estimates. This velocity density estimation technique is applicable for spatio-temporal density *interpolation*, for monitoring and visualising the change of densities in a (past) time window. However, it is not designed for extrapolating the density to (future) time points outside the window of observed historical data.

Related to change diagnosis is *change mining* [3], which aims to understand the changes in data distributions themselves. Within this paradigm, the idea of so-called *drift-mining* approaches [8, 9, 5] is to model the evolution of distributions in order to extrapolate them to future time points, thereby addressing problems of verification latency or label delay. The algorithm APT proposed in [8] uses matching between labelled old and unlabelled new instances to infer the labels of the later, thus indirectly estimating the class-conditional distributions of the new instances. Likewise, an expectation-maximisation approach is used in [9] to track components of a Gaussian mixture model over time. In [5], a mixture model is learned on old labelled data and compared to density estimates on current unlabelled data, thereby inferring changes such as of the class prior. However, these approaches are again not designed for directly extrapolating densities.

Density forecasting approaches [15, 16, 7], on the other hand, focus on the prediction of a single variable's (mostly unimodal) density at a particular future timepoint, based on an observed time series of this variable's values. In the simplest case, as discussed for example in [16], this is done by providing a confidence interval for a point estimate, obtained by assuming a particular distribution of this variable. More sophisticated approaches return (potentially multi-modal) density estimates by combining several predictions, which are obtained for example by different macroeconomic models, experts, or simulation outcomes, into a distribution estimation by kernel methods. Nevertheless, their multi-modal character originates from the different modes in the combined *unimodal* models. In addition, most works consider only a single variable. One exception is [7], where univariate forecasts of two explanatory variables are converted using conditional kernel density estimation into forecasts of the dependent variable.

In contrast to density forecasting above, we are concerned with temporal density extrapolation of a potentially *multi-modal density distribution*. Furthermore, instead of having a time series of single observations at any one time, our input data consists of multiple observations at any one time. This *temporal* density extrapolation is related to *spatial* density extrapolation [17, 6], which addresses the extrapolation of densities for feature values that have not been seen yet in historical instances. In [17], the authors suggest a Taylor series expansion about the point of interest to estimate the density, while in [6] a statistical test is provided to examine whether the data distribution is distinct from a uniform distribution at the extrapolation position. While modelling time as a feature is possible, there is an important difference in extrapolation between time and feature space: one expects the density to diminish towards unpopulated (and thus unseen) positions in feature space. However, there is no a priori reason to assume densities to decrease towards yet unseen moments in time. On the contrary, it is reasonable to assume that *at each point in time* (whether future, current, or past) the *density integrates to one* over the feature space.

3 Temporal Density Extrapolation

To address the problem of extrapolating the observed, potentially multi-modal density-distribution of instances to future time points, we propose an approach based on *pseudo-points*. These pseudo-points are used in the spatio-temporal kernel density estimation in lieu of the originally observed instances. The resulting kernel density estimation model can be interpreted as a mixture model, where each pseudo-point constitutes itself a component. The pseudo-points evolve over time, either by changing their weight (their component's mixing proportion), or by changing their position (their component's location). Therefore, the learning task is to fit a trend function to the evolution of each pseudo-point. We present each of the two variants in the next Subsections 3.1 and 3.2, before discussing their potential difficulties and limitations in Section 3.3.

3.1 Weight-Extrapolated, Stationary Pseudo-Points

Given a set of stationary pseudo-points, the first approach models their weights as functions of time. These functions are then fit on a window with historical data, such that the distribution therein is modelled with maximum likelihood.

The approach is illustrated for a one-dimensional feature space in Figure 1. At the first time point in the past ($time = 0$), a density estimate is calculated using historical data collected at that time (solid blue line). Then, a set of pseudo-points (here $1, 2, \cdots 4$) is generated, either by placing them equidistant on a grid or by drawing them at random. Next, the weights $(w_1, w_2, \cdots w_4)$ of all pseudo-points are calculated such that the divergence is minimised between the kernel density estimate over the weighted pseudo-points and the kernel-density estimate over the original data instances at that time point. The pseudo-point's weights are estimated in the same way for subsequent time points (e.g. $t = 1$), as soon as instances become available for them. This results for each pseudo-point in a time series of weight values, for which a polynomial trend function (red curves) is learned by regression. Finally, for a future time point (e.g. $time = 2$), the trend functions' values are predicted ($w_1, w_2, \cdots w_4$ in red at $time = 2$). Using these weighted pseudo-points in a kernel density estimate at $time = 2$, one obtains the extrapolated density (red dotted line), which is later evaluated against the observed density (solid blue-gray line).

3.2 Position-Extrapolated, Uniformly-Weighted Pseudo-Points

The second approach to address this problem is to use uniformly-weighted, but flexibly located pseudo-points. Thus, the pseudo-point's weights are uniform and constant, but their positions are functions of time, fitted such that the divergence on the available historical data is minimised.

In analogy to the previous figure, this approach is illustrated for a one-dimensional feature space in Figure 2. Given a set of historical instances and a specified number of pseudo-points, density estimates (solid blue lines) are made for historical time points ($time = 0$ and $time = 1$). Then, a mixture model with each pseudo-point as a single Gaussian component is formulated. Assuming polynomial trajectories (red solid lines) for the pseudo-points, the parameters of this model are the coefficients of the pseudo-points polynomial trajectories, which are learned using Expectation-Maximisation. Finally, for a future time point ($time = 2$), the pseudo-point's positions are predicted using the polynomial function, and the density (red dotted line) at this time point is estimated using kernel density estimation over the pseudo-points placed at their extrapolated positions.

3.3 Discussion

Both approaches above rely on a regression over time for extrapolating trends in the development of either weights or positions. In order to make this extrapolation more robust, we recommend using regularised trend functions that consider penalties for the models' complexities. The choice of the type of regression function depends on the type of drift, as for example polynomial functions require gradual drift, while trigonometric functions seem to be interesting candidates for modelling recurring context.

The weight-extrapolation in the first approach requires a normalisation, such that the extrapolated weights are all non-negative and sum up to one. An important question concerns the choice of the pseudo-point's location in this approach,

Fig. 1. Temporal Density Extrapolation Using Weight-Extrapolated Pseudopoints

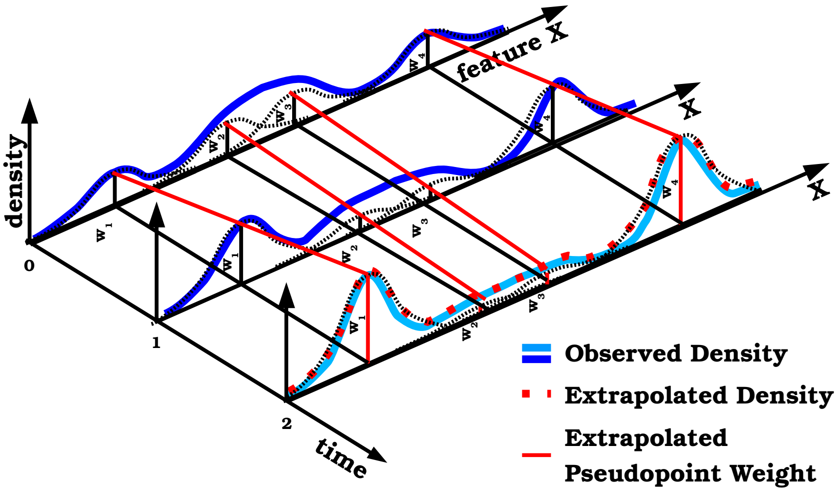
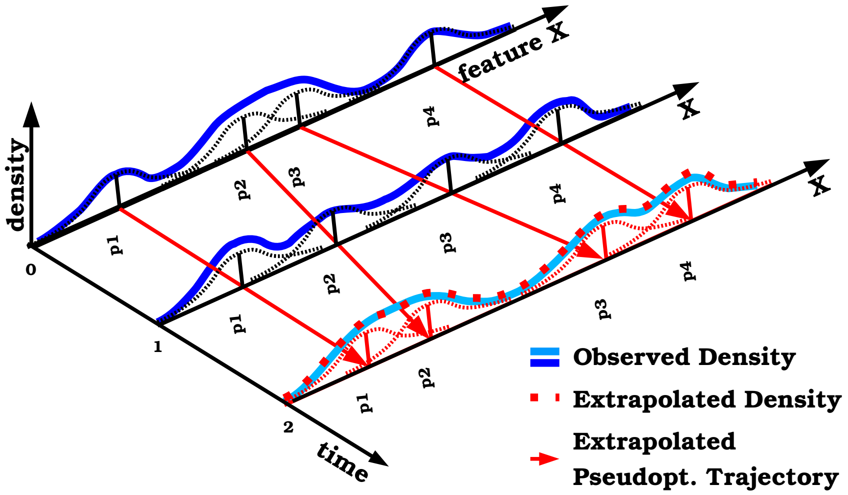


Fig. 2. Temporal Density Extrapolation Using Position-Extrapolated Pseudopoints



as it influences the precision of the extrapolated values: in regions with sparse pseudo-point populations, the model is less flexible than in densely populated ones. Therefore, this approach seems better suited for constricted (bounded) feature spaces. A simple equidistant placement of pseudo-points distributes the precision over the whole feature space. Alternatively, the pseudo-points might be placed at the coordinates of a subsample of the observed instances, thus concentrating the precision on areas with previously high density. However, if densities change largely over time, these areas might become less relevant.

In contrast, the second, position-extrapolating approach determines the positions of each pseudo-point automatically. It aims to adjust the future location of the pseudo-points such that they are densely placed in regions with a high expected density. However, in the case polynomial regression functions are used, a potential drawback is that their trajectories diverge in the long run. Thus, in contrast to the first approach, the second one seems to be better suited for infinite (unbounded) feature spaces.

Related to the choice of the pseudo-points' placements is the question of optimal bandwidth selection, which for kernel density estimation has already been reviewed in [18]. In short, we expect that with an increasing number of pseudo-points the optimal bandwidth decreases, while the extrapolation's precision increases. Furthermore, the number of pseudo-points is also an upper bound on the number of modes that both approaches are able to model.

4 Conclusion

In this paper, we have addressed the problem of *temporal density extrapolation*, where the objective is the prediction of a (potentially multi-modal) density distribution at *future time points*, given a sample of historical instances observed at different positions in feature space and at different times in the past. Two approaches based on pseudo-points were sketched: the first uses an extrapolation of time-varying weights of stationary located pseudo-points, while the second uses an extrapolation of the trajectory of the time-varying location of pseudo-points with uniform weights. Subsequently, these extrapolated pseudo-points are used in a kernel density estimation at future time points.

Having sketched the idea of the two temporal density extrapolation approaches, a more detailed specification and evaluation of these methods needs to be done in future work. Furthermore, the conjectures in the discussion above, in particular the usability of each approach for bounded and unbounded feature spaces, need to be verified. Finally, a known challenge for kernel-based approaches is the curse of dimensionality on multi-dimensional data. A naive approach is to combine multiple univariate temporal density extrapolations. However, an optimisation for multi-variate problems by using either projection pursuit [4] or copula [12] techniques seems worth investigating.

Acknowledgments. We thank Vera Hofer from Karl-Franzens-University Graz, Stephan Möhring, and Andy Koch for insightful discussions.

References

1. Aggarwal, C.C.: A framework for diagnosing changes in evolving data streams. In: Proceedings of the ACM SIGMOD Conference (2003)
2. Aggarwal, C.C.: On change diagnosis in evolving data streams. IEEE Transactions on Knowledge and Data Engineering 17(5), 587–600 (2005)
3. Böttcher, M., Höppner, F., Spiliopoulou, M.: On exploiting the power of time in data mining. ACM SIGKDD Explorations Newsletter 10(2), 3–11 (2008)
4. Friedman, J.H., Stuetzle, W., Schroeder, A.: Projection pursuit density estimation. Journal of the American Statistical Association 79(387) (1984)
5. Hofer, V., Krempl, G.: Drift mining in data: A framework for addressing drift in classification. Computational Statistics and Data Analysis 57(1), 377–391 (2013)
6. Hooker, G.: Diagnosing extrapolation: Tree-based density estimation. In: Knowledge Discovery in Databases (KDD) (2004)
7. Jeon, J., Taylor, J.W.: Using conditional kernel density estimation for wind power density forecasting. Journal of the American Statistical Association 107 (2012)
8. Krempl, G.: The algorithm APT to classify in concurrence of latency and drift. In: Gama, J., Bradley, E., Hollmén, J. (eds.) Advances in Intelligent Data Analysis X, Lecture Notes in Computer Science, vol. 7014, pp. 222–233. Springer (2011)
9. Krempl, G., Hofer, V.: Classification in presence of drift and latency. In: Spiliopoulou, M., Wang, H., Cook, D., Pei, J., Wang, W., Zaïane, O., Wu, X. (eds.) Proceedings of the 11th IEEE International Conference on Data Mining Workshops (ICDMW 2011). IEEE (2011)
10. Krempl, G., Zliobaitė, I., Brzeziński, D., Hüllermeier, E., Last, M., Lemaire, V., Noack, T., Shaker, A., Sievi, S., Spiliopoulou, M., Stefanowski, J.: Open challenges for data stream mining research. SIGKDD Explorations 16(1), 1–10 (2014), special Issue on Big Data
11. Nadaraya, E.A.: Nonparametric estimation of probability densities and regression curves. Kluwer (1989), originally published in Russian by Tbilisi University Press, Translated by S. Klotz
12. Nelsen, R.B.: An Introduction to Copulas. Springer (1999)
13. Parzen, E.: On estimation of a probability density function and mode. Annals of Mathematical Statistics 33, 1065–1076 (1962)
14. Rosenblatt, M.: Remarks on some non-parametric estimates of a density function. Annals of Mathematical Statistics 27(3), 832–837 (1956)
15. Skouras, K., Dawid, A.P.: On efficient probability forecasting systems. Biometrika 86(4), 765–784 (1999)
16. Tay, A.S., Wallis, K.F.: Density forecasting: A survey. Companion to Economic Forecasting pp. 45–68 (2002)
17. Terrell, G.R.: Tail probabilities by density extrapolation. In: Proceedings of the Annual Meeting of the American Statistical Association (2001)
18. Turlach, B.A.: Bandwidth selection in kernel density estimation: A review. Tech. Rep. 9307, Humboldt University, Statistic und Ökonometrie (1991)