Expert Systems with Applications 41 (2014) 7596–7607

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

An analysis of alignment and integral based kernels for machine learning from vessel trajectories

Gerben Klaas Dirk de Vries *, Maarten van Someren

Informatics Institute, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Available online 2 June 2014

Keywords:
Alignment measures
Integral over time similarity
Vessel trajectories
Trajectory kernels

ABSTRACT

In this paper we present an analysis of the application of the two most important types of similarity measures for moving object trajectories in machine learning from vessel movement data. These similarities are applied in the tasks of clustering, classification and outlier detection. The first similarity type are alignment measures, such as dynamic time warping and edit distance. The second type are based on the integral over time between two trajectories. Following earlier work we define these measures in the context of kernel methods, which provide state-of-the-art, robust algorithms for the tasks studied. Furthermore, we include the influence of applying piecewise linear segmentation as pre-processing to the vessel trajectories when computing alignment measures, since this has been shown to give a positive effect in computation time and performance.

In our experiments the alignment based measures show the best performance. Regular versions of edit distance give the best performance in clustering and classification, whereas the softmax variant of dynamic time warping works best in outlier detection. Moreover, piecewise linear segmentation has a positive effect on alignments, due to the fact that salient points in a trajectory, especially important in clustering and outlier detection, are highlighted by the segmentation and have a large influence in the alignments. Based on our experiments, integral over time based similarity measures are not well-suited for learning from vessel trajectories.

1. Introduction

In the maritime domain vessel movements are tracked by GPS sensors and radar. Storing these moving object trajectories and applying machine learning and data mining techniques to this data can assist operators and observers in this domain with analyzing vessel behavior. For instance, around large ports, grouping vessel trajectories into clusters of similar behavior can help operators to spot irregular movements. Or consider fishing vessels, which show very irregular movement patterns, and cargo ships, which have very regular behavior. Identifying to which class a vessel movement belongs can tell whether the observed behavior is unwanted and thus if further investigation is required.

One common machine learning approach for moving object trajectories is to use similarity measures between two trajectories (Nanni, Kuijpers, Körner, May, & Pedreschi, 2008). Using similarities is a natural fit for moving object trajectories, since two trajectories typically differ in number of samples, distances traveled and temporal length, which means that there is no natural attribute/feature vector representation. However, similarities can be more computationally intensive than an attribute/feature based approach.

In this paper we present an analysis of the application of the two main types of similarity measures for moving object trajectories in machine learning from vessel movement data. The first type of similarity are alignment measures, which seek an alignment between points in a trajectory and then compute a score for this alignment. The second technique is based on computing the integral over time of the distance between two trajectories. Both types of measures have advantages and disadvantages, and which type is better suited in the domain of vessel trajectories is the main question that is addressed in this paper. The measures are applied in three machine learning/data mining tasks: clustering, classification and outlier detection.

Previous research (de Vries & van Someren, 2010, 2012) has shown that piecewise linear segmentation has a positive effect on alignment measures used in machine learning on vessel trajectories. Furthermore, it significantly improves computation time. In our comparison we consider alignment measures with and without the use of piecewise linear segmentation. Compared to the work

* Corresponding author. Tel.: +31 20 525 7522.
E-mail addresses: g.k.d.devries@uva.nl (G.K.D. de Vries), m.w.vansomeren@uva.nl (M. van Someren).

http://dx.doi.org/10.1016/j.eswa.2014.05.025
0957-4174/© 2014 Elsevier Ltd. All rights reserved.

presented in de Vries and van Someren (2010), this paper includes additional tasks and the integral over time measures for comparison. In de Vries and van Someren (2012) the focus is on alignment measures as part of a machine learning framework for vessel trajectories, whereas this paper focuses on the comparison between different types of measures for machine learning with vessel trajectories and further explores the influence of piecewise linear segmentation.

Kernel methods (Schölkopf & Smola, 2001; Shawe-Taylor & Cristianini, 2004) provide state-of-the-art, robust algorithms for a range of machine learning problems. They only require a similarity function, i.e. the kernel, between objects, and therefore are a natural choice for moving object trajectories. We use kernel methods for the three machine learning tasks, following (de Vries & van Someren, 2010, 2012). Kernels are required to be positive semi-definite (PSD). However, in practice the regular non-PSD similarity measures often work well. Therefore, we use both regular and PSD variants for the similarity measures that are compared in this paper. To the best of our knowledge this is the first time that kernel variants of the integral over time similarities have been studied for moving object trajectories.

In summary, this paper provides the following contributions for machine learning from vessel trajectories in three different tasks.

- An extensive experimental comparison between the two most important types of similarity measures for moving object trajectories: alignment measures and integral over time similarities.
- An experimental investigation of the effect of applying piecewise linear segmentation when using alignment measures.
- All similarities and tasks are studied in the framework of kernel methods.

The rest of this paper is structured as follows. We present the necessary preliminaries in Section 2. Section 3 contains all the technical details on the different similarity measures and piecewise linear segmentation. Our experiments are described in Section 4. Finally, we end with some conclusions and suggestions for future work.

2. Preliminaries

The work presented in this paper fits into the field of spatio-temporal data mining (STDM). Spatio-temporal data mining aims at performing typical machine learning/data mining tasks on spatio-temporal data. For a good overview of the research done in this area, see Nanni et al. (2008).

The most prominent type of data studied in STDM is the moving object trajectory. These are trajectories of objects moving in a certain space, typically captured using things like GPS sensors. In a 2-dimensional space this means that they are sequences of $\langle x, y, t\rangle$ vectors, where $x,\ y$ is the position at time t. However, for most of this paper, a moving object trajectory T in a 2-dimensional space is represented by a sequence of vectors: $T = \langle x_1, y_1\rangle, \langle x_2, y_2\rangle, \ldots, \langle x_n, y_n\rangle$. The number of vectors of a trajectory is denoted as: $\mid T \mid$. Furthermore, let $T(i) = \langle x_i, y_i\rangle$ and $T(i,j) = \langle x_i, y_i\rangle, \ldots, \langle x_j, y_j\rangle$. In the following we refer to a vector $\langle x_i, y_i\rangle$ as a trajectory *point* or *element*. Note, that when the sample rate of a trajectory is fixed, then the time component is implicitly represented in the trajectory.

2.1. Similarity measures

For the data mining tasks with moving object trajectories that we study in this paper we use similarity measures. There are a number of different approaches to similarities between trajectories, which can be broadly divided into two classes. Note that we use the term 'similarity' generically for measures that are distances or similarities.

The first class is that of alignment based similarities, which treat a trajectory as a *sequence of points*. For two trajectories we try to find an optimal alignment between the points according to some scoring scheme. The alignment computation algorithm and the definition of the scoring scheme lead to different similarity measures. There exists Dynamic Time Warping (DTW) (Vlachos, 2004), and various forms of Edit Distance (ED), such as edit distance with real penalties (Chen & Ng, 2004), edit distance on real sequences (Chen & Özsu, 2005) and Longest Common SubSequence (LCSS) (Vlachos, Gunopoulos, & Kollios, 2002; Vlachos, Kollios, & Gunopulos, 2005). The DTW similarity measure has its origin in the time-series/speech processing literature, whereas edit distance measures were originally defined for strings of characters.

The similarity measures considered in Nanni (2002), Nanni and Pedreschi (2006), van Kreveld and Luo (2007), Frentzos, Gratsias, and Theodoridis (2007) and Buchin, Buchin, van Kreveld, and Luo (2009) interpret a trajectory as *a continuous function* instead of a sequence. Nanni (2002), Nanni and Pedreschi (2006), van Kreveld and Luo (2007), Frentzos et al. (2007) and Buchin et al. (2009) use a similarity measure that takes the integral over time for the distance function that gives the distance between two trajectories for each time point. The measure given in Buchin et al. (2009) generalizes previous versions of this measure with both a variable time-shift and a variable length of matching. Pelekis et al. (2007) describe a different method that calculates the surface between trajectories, instead of calculating the integral over time, projections to the *xy*- and *t*-plane are used. These similarity measures that treat trajectories as continuous functions are generally more precise than alignment measures, because they do not depend on the specific samples that are taken, but they are also typically slower to compute.

Similarity measures are designed to consider two trajectories similar if they (more or less) describe the same absolute spatio-temporal path. However, no two trajectories are exactly the same. Trajectories can only share part of their spatio-temporal paths, e.g. they might start or end at a different location, but are similar for certain tasks. Vessels might move at a different speed along different parts of the trajectory, but are still similar for some tasks if this difference is not too large. Furthermore, there are small noisy differences between trajectories, for instance caused by the sensors measuring them. The two classes of similarity measures deal with these trajectory differences in their own way. The integral based measures allow for time shifts and/or subtrajectories, which is especially suited for dealing with differences in overlap between trajectories. Alignment measures allow for repetition of elements (dynamic time warping) or gap penalties (edit distance), which deal both with overlap differences and (slight) differences and speed.

In this paper we study representative members of both classes of similarity measures in machine learning for vessel trajectories.

2.2. Data mining tasks

This paper covers three machine learning tasks applied to vessel trajectory data: clustering, classification and outlier detection. Clustering means grouping trajectories into clusters of similar behavior. Classification is the task of assigning a class to a trajectory, for instance the vessel type that generated the trajectory. In outlier detection the aim is to discover irregular movement patterns among a set of normal patterns.

Of these three tasks, most of the prior work in STDM concerns clustering. In Vlachos et al. (2002) longest common subsequence measures are combined with a density based algorithm. An integral over time similarity measure is used in Nanni (2002) with