Data Min Knowl Disc (2018) 32:1179–1199
https://doi.org/10.1007/s10618-018-0554-1

Analyzing concept drift and shift from sample data

Geoffrey I. Webb[1] · Loong Kuan Lee[1] · Bart Goethals[1,2] · François Petitjean[1]

Received: 11 September 2017 / Accepted: 7 February 2018 / Published online: 12 March 2018

Abstract Concept drift and shift are major issues that greatly affect the accuracy and reliability of many real-world applications of machine learning. We propose a new data mining task, *concept drift mapping*—the description and analysis of instances of concept drift or shift. We argue that concept drift mapping is an essential prerequisite for tackling concept drift and shift. We propose tools for this purpose, arguing for the importance of quantitative descriptions of drift and shift in marginal distributions. We present quantitative concept drift mapping techniques, along with methods for visualizing their results. We illustrate their effectiveness for real-world applications across energy-pricing, vegetation monitoring and airline scheduling.

Keywords Concept drift · Concept shift · Non-stationary distribution · Visualisation · Mapping

Responsible editor: Jesse Davis, Elisa Fromont, Derek Greene, and Bjorn Bringmann.

✉ Geoffrey I. Webb
geoff.webb@monash.edu

Loong Kuan Lee
lklee9@student.monash.edu

Bart Goethals
goethals@gmail.com

François Petitjean
francois.petitjean@monash.edu

[1] Faculty of Information Technology, Monash University, Clayton, VIC 3800, Australia

[2] Department of Mathematics and Computer Science, University of Antwerp, Antwerp, Belgium

1 Introduction

The world is dynamic, in constant flux. But machine learning usually creates static models from historical data. As the world changes, these models can grow increasingly unreliable. A distribution that changes is called *non-stationary* and a change in the distribution from which a model is learned is called *concept drift*.

There has been a very substantial effort investigating methods for *detecting* concept drift (Widmer and Kubat 1996; Kifer et al. 2004; Gama et al. 2004; Baena-Garcıa et al. 2006; Nishida and Yamauchi 2007; Dries and Rückert 2009; Gama and Rodrigues 2009; Žliobaite 2010; Hoens et al. 2012; Moreno-Torres et al. 2012; Bifet et al. 2013; Gama et al. 2014; Qahtan et al. 2015; Yu and Abraham 2017). This paper introduces a complementary capability. Whereas drift detection seeks to detect whether or not change is present, we instead seek to generate a detailed description of the nature and form of whatever drift there may be. We call such a description a *concept drift map*.

Concept shift is closely related to concept drift. This occurs when a model learned from data sampled from one distribution needs to be applied to data drawn from another. For example, a model learned in one region might be applied in another region, or a model learned from customer data might be applied to potential customers. For ease of exposition, this paper focuses only on the issue of analyzing concept drift, but the approaches and discussion generalize directly to the equally important issue of concept shift analysis.

Figure 1 shows some example raw data and a corresponding simple map from a data set describing the Australian electricity market, explained in detail in Sect. 6.1. This simple example concept drift map plots the drift in two key variables `nswprice` and `vicprice` both individually and jointly. The first is the price in the state of New South Wales. The second is the price in the state of Victoria. This map shows how `nswprice` determines the drift up until May 1997. At this point the Victorian price is deregulated and `vicprice` briefly dominates the drift before the market settles and each attribute contributes to the joint drift. The map identifies the relative contribution of each attribute to change within the system, revealing the relative rate of change in the underlying distributions much more clearly than direct examination of the original data. By way of contrast, a drift detection mechanism should have no difficulty in detecting that drift has occurred around the times corresponding to each spike in the magnitude of the drift in the joint distribution. However, a drift detection mechanism would not provide any insight into which attributes were responsible for this drift.

In this paper we present techniques for generating drift maps from data. In Sect. 2 we provide a formal definition of the problem and related terminology. In Sect. 3 we present methods for measuring total drift magnitude. In Sect. 4 we present methods for measuring marginal drift magnitudes. Section 5 describes graphical methods for communicating the detailed maps that our quantitative techniques produce. Section 6 evaluates the effectiveness of our techniques on three real-world datasets. Section 7 discusses related research. We present conclusions in Sect. 8.

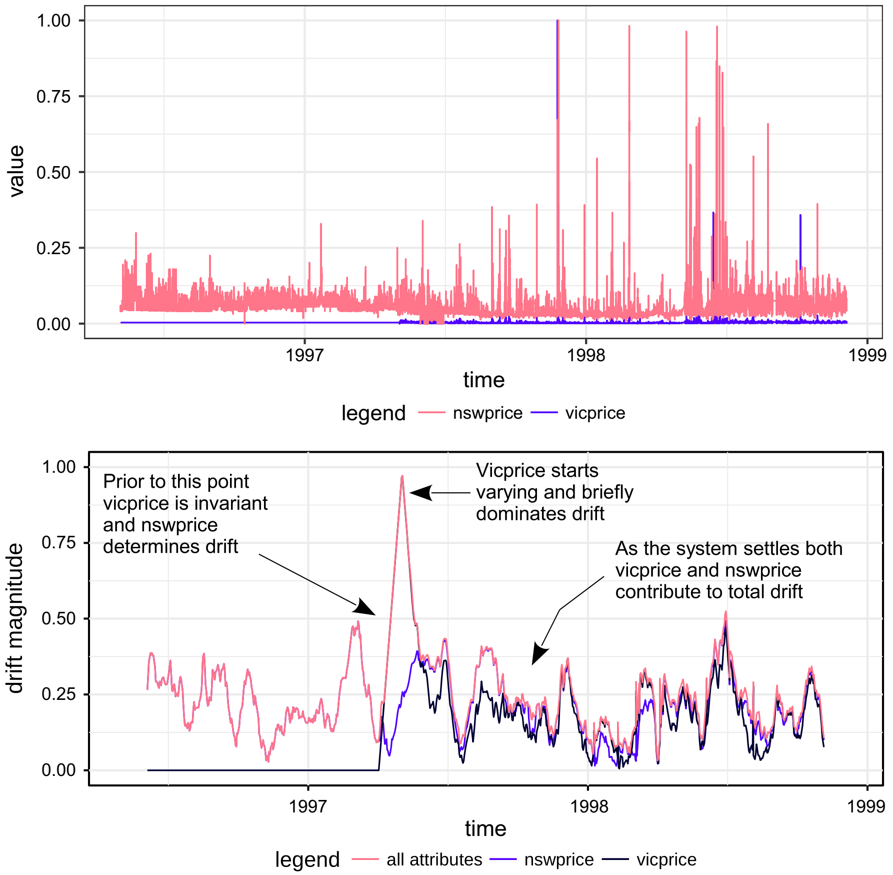

Fig. 1 Some example raw data and corresponding concept drift map plotting the changing drift magnitude of two variables both individually and jointly. These variables record the electricity price in each of the Australian states of Victoria and New South Wales. Prior to May 2nd 1997 there was no interstate electricity market and the Victorian price was invariant. The top graph plots the raw price data. The bottom plot shows the drift magnitude calculated at each day between the preceding 30 days and the following 30 days

2 Problem description

A data stream is a data set in which the objects have time stamps, which, depending on the granularity of the stamps, induces either a total or a partial order between observations. While our techniques generalize in a straightforward manner to situations in which there is no target attribute, in the current work we assume a classification learning context. In consequence, we can consider the process that generates the stream to be a joint distribution over random variables Y and $X = \{X_1, \ldots, X_n\}$, where $y \in \mathrm{dom}(Y)$ are the class labels and the $x_i \in \mathrm{dom}(X_i)$ are the attribute values. We provide a summary of the key symbols used in this paper in Table 1.

In order to reference the probability distribution at a particular time we add a time subscript, such as $P_t(X, Y)$, to denote a probability distribution at time t. It is often

Table 1 List of symbols used

Symbol(s)	Usage	Scope
t, u, v, w	Points of time	$\mathbb{R}_{>0}$
$X = \{X_1, \ldots, X_n\}$	A random variable over covariates	Stream dependent
x_i	A value of a covariate	$\mathrm{dom}(X_i)$
$\bar{x} = \langle x_1, \ldots, x_n \rangle$	A simultaneous assignment of values to all covariates	$\mathrm{dom}(X)$
Y	A random variable over class labels	Stream dependent
y	A class label	$\mathrm{dom}(Y)$
Z	A vector of random variables over either covariate values or class labels	Stream dependent
$\bar{z}_i$	A value in $\mathrm{dom}(Z_i)$	$\mathrm{dom}(Z_i)$
$P_t\,(X, Y)$	A probability distribution at time t	
$P_{[t,u]}\,(X, Y)$	A probability distribution over time period $[t, u]$	
$\sigma_{t,u}(Z)$	Total variation distance measure of drift between each of the probability distributions of vector of random variables Z from time period t to time period u	[0,1]
$\sigma_{t,u}^{X\|Y}$	Total variation distance measure of drift between each of the conditional covariate probability distributions from time period t to time period u	[0,1]
$\sigma_{t,u}^{Y\|X}$	Total variation distance measure of drift between each of the conditional class probability distributions from time period t to time period u	[0,1]

not practical to estimate the distribution in effect at a specific point in time and for this purpose we often deal with concepts and probability distributions over a time interval, $P_{[t,u]}(X, Y)$.

We follow recent practice and adopt Gama et al's (2014) definition of a concept.

$$\text{Concept} = P\,(X, Y)\,. \tag{1}$$

In the context of a data stream, we need to recognize that concepts may change over time. To this end we define the concept at a particular time t as

$$P_t\,(X, Y) \tag{2}$$

and at a specific time period $[t, u]$ as

$$P_{[t,u]}\,(X, Y)\,. \tag{3}$$

Concept drift occurs between times t and u when the distributions change,

$$P_t(X, Y) \neq P_u(X, Y) \tag{4}$$

and similarly between time periods $[t, u]$ and $[v, w]$,

$$P_{[t,u]}(X, Y) \neq P_{[v,w]}(X, Y). \tag{5}$$

We define the *concept drift mapping task* as taking as input a data stream and generating as output useful descriptions of the drift in the process that generates the data. Note that in the concept shift mapping task, the input is sample data from each of two or more related distributions and the output is useful descriptions of the differences between the distributions that generate the data.

3 Measuring total drift magnitude

Webb et al. (2016) proposed four quantitative measures of concept drift including the key measure *drift magnitude* which measures the distance between two concepts $P_t(X, Y)$ and $P_u(X, Y)$. Any measure of distance between distributions could be employed. Webb et al. (2016) use Hellinger distance (Hellinger 1909; Hoens et al. 2011) for this purpose. In the current work we employ total variation distance (Levin et al. 2008):

$$\sigma_{t,u}(Z) = \frac{1}{2} \sum_{\bar{z} \in \mathrm{dom}(Z)} |P_t(\bar{z}) - P_u(\bar{z})| \tag{6}$$

where Z represents a vector of random variables.

Of all the standard measures of distance between probability distributions we favour Hellinger distance and total variation distance because they are metrics and it is highly desirable that a measure of drift between two periods should be symmetric. In this paper we use total variation distance because it is slightly less complex to analyse than Hellinger distance and more efficient to compute. However, our approaches trivially generalize to any measure of distance between probability distributions such as Kullback–Leibler divergence or Wasserstein distance.

Note that our techniques are designed for discrete valued data. While there are techniques for computing total variation and Hellinger distance for continuous data drawn from specific distributions, such as a Gaussian, these require strong assumptions about the form of the distribution and hence are not applicable to numeric data drawn from arbitrary distributions. In consequence, we discretize all numeric attributes, using 5 bin equal frequency discretization of each attribute across all time periods. Should an appropriate method for calculating distances between arbitrary continuous probability distributions be developed, the approaches we describe herein can be applied directly, using them in place of discretization.

Webb et al. (2016) propose a number of quantitative measures for drift that provide gross summaries of the drift between two time points. These include using any measure of distance between probability distributions to measure drift magnitude. They demonstrate that these measures enable insights to be derived that are otherwise not possible, such as how different algorithms perform in the face of drift of varying magnitude. However, our subsequent uses of these measures in real world applications have revealed that it can be important to augment these overview measures with further finer grained analysis.

One limitation of a single gross measure of drift magnitude arises from both total variation distance and Hellinger distance being monotonic as the dimensionality of data increases. We provide a proof of this in Appendix A. As a result, in practice, in high dimensional data these measures are likely to be close to their maximum, 1.0, simply through accumulation of small differences across many dimensions. This reduces their capacity to distinguish between different types of drift.

Second, a single value measure of drift provides only a very gross description of a complex drift phenomenon. It fails to recognize or to describe the details of how drift differs across the subspaces defined on different attributes of the data. In the real world, drift is often not uniform, as we show in Sect. 6. For example, not all factors are subject to inflation and those that are may increase at varying rates. A change in technology may cause a sudden abrupt change in some attributes of the data but have no affect whatsoever on others. Some factors may drift in cycles with differing periodicity and other factors may be subject to drift that is not cyclical. In many real world applications it is likely to be useful to be able to understand which attributes and combinations of attributes are drifting in which manners at any particular time.

For these reasons we investigate the introduction of *concept drift maps*, methods for describing the drift affecting different subspaces of the data.

4 Measuring marginal drift magnitude

The key to describing drift in different attribute subspaces is to measure the drift in the marginal distributions defined over different combinations of attributes.

A problem that arises is how to estimate the required probability distributions from the available data. In order to manage the variance in the estimates it is important to derive them from sufficiently large data samples. This will usually preclude the possibility of deriving instantaneous estimates—estimates of the probability distribution at any single point in time. Rather it will often be necessary to derive estimates of the distribution over some time interval, such as the distribution for a given hour, day or week. However, this practical driver is not the only reason for considering drift between extended periods rather than drift between instantaneous points in time. As we show in Sect. 6, consideration of drift between periods of differing granularity can also be extremely revealing. In consequence, our techniques estimate the drift between two time intervals by first estimating the distributions in each interval and then calculating the magnitude of the drift between them.

In the current work we use maximum likelihood estimates.

It turns out to be useful to map not only the drift in the *joint distribution* $P(X, Y)$, but also the *covariate distribution* $P(X)$, the *class distribution* $P(Y)$, the *conditional class distribution* $P(Y \mid X)$ and the *conditional covariate distribution* $P(X \mid Y)$, as each reveals different facets of a potentially complex drift.

To give extreme examples, drift might occur because there is a change in the relative frequencies of the classes, $P(Y)$, a change in the relative frequencies of the covariates, $P(X)$, or a change in the relationship between the classes and covariates, sometimes called *pure concept drift*, $P(Y \mid X)$ and $P(X \mid Y)$. By analysing all of these simultaneously, drift maps allow the user to understand the extent to which each of these forms of drift is affecting their data and to act accordingly.

For joint, covariate and class drift Eq. 6 applies directly. However, for the two conditional drifts it is necessary to deal with multiple distributions, one for each value of the conditioning attributes. We address this by weighted averaging, as described in the next two subsections.

4.1 Conditional marginal covariate drift

For a given subset of the covariate attributes there will be a conditional probability distribution over the possible values of the covariate attributes for each specific class, y. The *conditional marginal covariate drift* is the weighted sum of the distances between each of these probability distributions from time period t to u, where the weights are the average probability of the class over the two time periods.

$$\sigma_{t,u}^{X|Y} = \sum_{y \in Y} \left[\frac{P_t(y) + P_u(y)}{2} \frac{1}{2} \sum_{\bar{x} \in X} |P_t(\bar{x} \mid y) - P_u(\bar{x} \mid y)| \right] \tag{7}$$

4.2 Conditional class drift

For each subset of the covariate attributes there will be a probability distribution over the class labels for each combination of values of those attributes, $\bar{x}$ at each time period. Therefore, the Conditional Class Drift can be calculated as the weighted sum of the distances between these probability distributions where the weights are the average probability over the two periods of the specific value for the covariate attribute subset.

$$\sigma_{t,u}^{Y|X} = \sum_{\bar{x} \in X} \left[\frac{P_t(\bar{x}) + P_u(\bar{x})}{2} \frac{1}{2} \sum_{y \in Y} |P_t(y \mid \bar{x}) - P_u(y \mid \bar{x})| \right] \tag{8}$$

5 Methods for communicating drift maps

Our primary technique measures marginal drift magnitudes between time periods. Sometimes it will be interesting to consider a single such comparison at a time. At other times it will be useful to consider how drift unfolds over an extended period of time. This can result in very large numbers of individual drift values. Here we present methods for succinctly communicating these large amounts of information.

For drift over the marginals between two time periods the key information that we want to convey is the relative magnitude of the drift in each combination of attributes. We find that heat maps provide a highly effective means of doing so, clearly highlighting the interactions between the variables. We provide examples in Figs. 11, 12, 13, and 14 below.

We use line plots to communicate the evolution of drift over extended periods of time. In doing so we use two periodicity parameters. The first parameter is how frequently should the drift be calculated. In the electricity and airlines domains discussed below, we calculate the drift daily. The second parameter is the period over which to determine the distributions to be compared. In the airlines domain we use two periods for this purpose, daily and weekly, and show that each reveals different insights.

6 Illustrative examples

We illustrate the proposed techniques by application to a number of real-world datasets.

6.1 Electricity

The first example is electricity pricing in South-East Australia, a multivariate time series dataset downloaded from the MOA dataset repository (MOA 2017) and described by Harries (1999). The covariates are `nswprice`, `nswdemand`, `vicprice`, `vicdemand`, and `transfer`, recording the price and demand in the states of New South Wales and Victoria and the amount of power transferred between the states. The class label identifies whether the transfer price is increased or decreased relative to a moving average of the last 24 h. Examples are generated for every 30 min period from 7 May 1996 to 5 December 1998. The values have been normalized to the interval [0,1].

Figures 2 and 3 present the covariate drift and conditional marginal covariate drift respectively. Each point corresponds to a day and presents the drift from the 30 day period prior to that day compared to next 30 days.

As can be seen, there is a sudden increase in covariate drift on 2nd of May 1997. This is the date at which the process of introducing a national electricity market (NEM) commenced. From this date a trial NEM allowed wholesale electricity sales between the states of New South Wales, Victoria, the Australian Capital Territory, and South

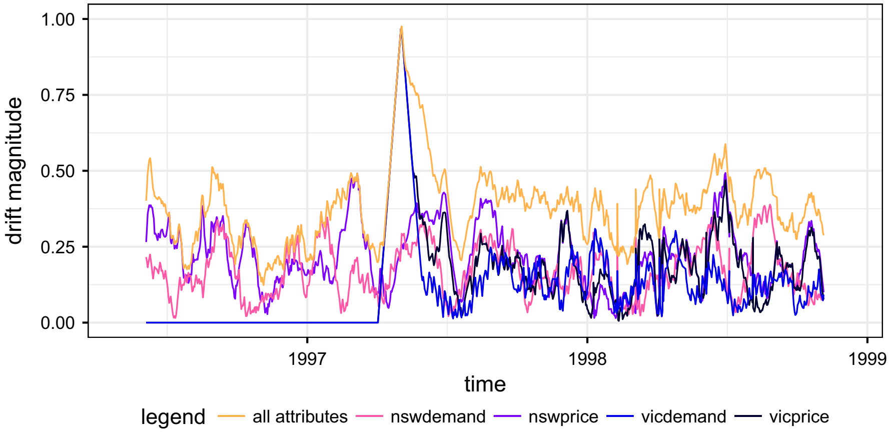

Fig. 2 Covariate drift for the electricity data. Values calculated daily for the drift between the 30 days prior to the current day and the 30 days thereafter

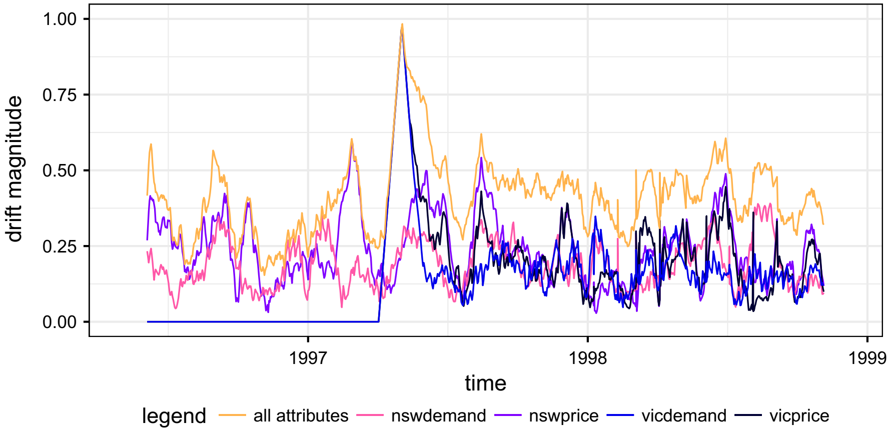

Fig. 3 Conditional marginal covariate drift for the electricity data. Values are calculated daily for the drift between the 30 days prior to the current day and the 30 days thereafter

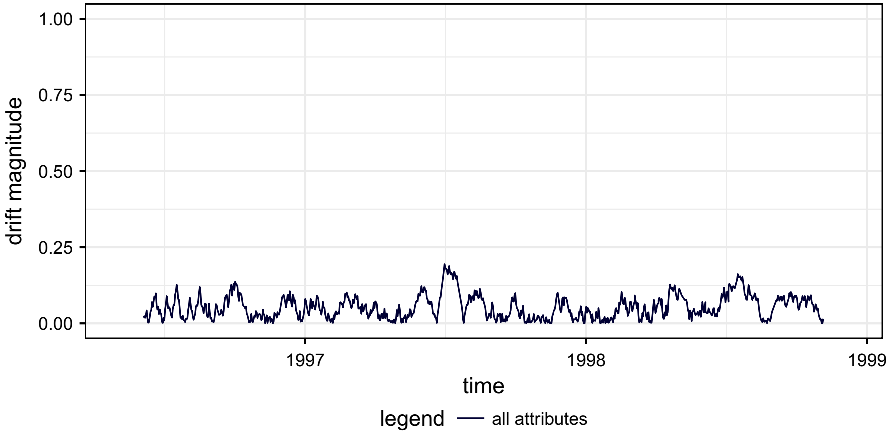

Fig. 4 Class drift for the electricity data. Values are calculated daily for the drift between the 30 days prior to the current day and the 30 days thereafter

Australia (Roarty 1998). `Vicprice`, `vicdemand`, and `transfer` have no drift prior to this date. Indeed these three variables are constant until the market is introduced. Past May 1997, drift in `nswprice` and `nswdemand` stays similar to before, but substantial variability is apparent in the drift within `vicprice`, `vicdemand`, and `transfer`. The conditional covariate drift closely follows the unconditional covariate drift indicating that there was little difference in drift of the covariates between classes. This illustrates how our proposed mapping of drift over both marginals as well as the distribution as a whole can provide additional useful information.

The class drift, depicted in Fig. 4, shows relatively low levels of drift. Note that as this is a binary variable, high levels of drift in this map would indicate that the transfer price has trended in one direction (up or down) for the previous 30 days and then in the opposite direction for the following 30 days. This plot indicates that there were no such extended changes.

To summarize this example, drift increases substantially after May 2nd 1997. The increase in drift is dominated by covariate drift and the covariate drift is dominated by drift in three of the five covariates, VicPrice, VicDemand, and Transfer.

6.2 Airlines

The second example is the airlines dataset, also downloaded from the MOA dataset repository (MOA 2017). Each example in this data represents a flight, with covariates `Airline`, `Flight`. `AirportFrom`, `AirportTo`, `DayOfWeek`, `Time`, and `Length` and with a binary class indicating whether the flight arrived on time. The `DayOfWeek` has been used to partition the data into days and weeks and have not been included as a covariate in the analysis. Figure 5 shows the covariate drift from day to day. Figure 6 shows the covariate drift for the week prior to a day against the week starting with that day and is plotted daily from the seventh day. Note that the numbering starts with 4 as the first day in the data is day number 3.

The first figure shows that for the first two weeks there is a cyclical pattern in the magnitude of covariate drift, with large changes from Friday to Saturday and from Saturday to Sunday, but lower drift from Sunday to Monday and substantially lower drift between successive weekdays. However, this pattern breaks down over the following two weeks. Unfortunately we do not have the dates for which the data were collected and hence can only speculate for the reasons for this change in pattern; weather and public holidays being two potential explanations. The marginal distributions indicate that the time of day is the major contributor to drift for most of the period but that flight number overtakes it for some parts of the second half of the period.

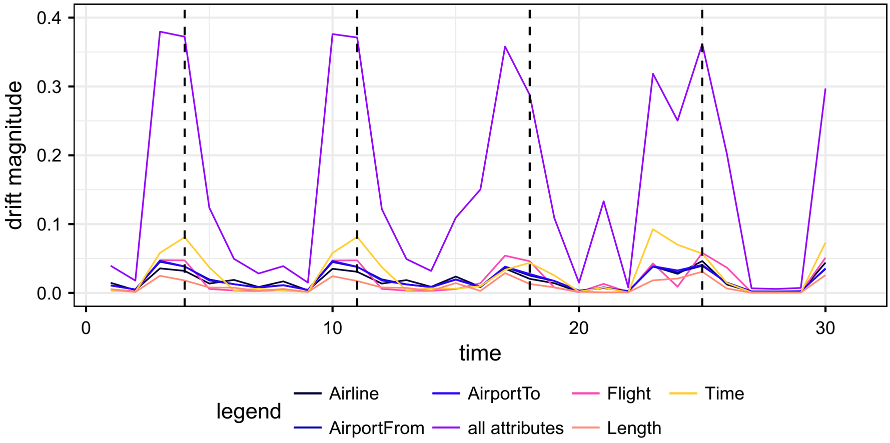

Fig. 5 Daily covariate drift for the airlines data. The dashed lines are placed between each Saturday and Sunday

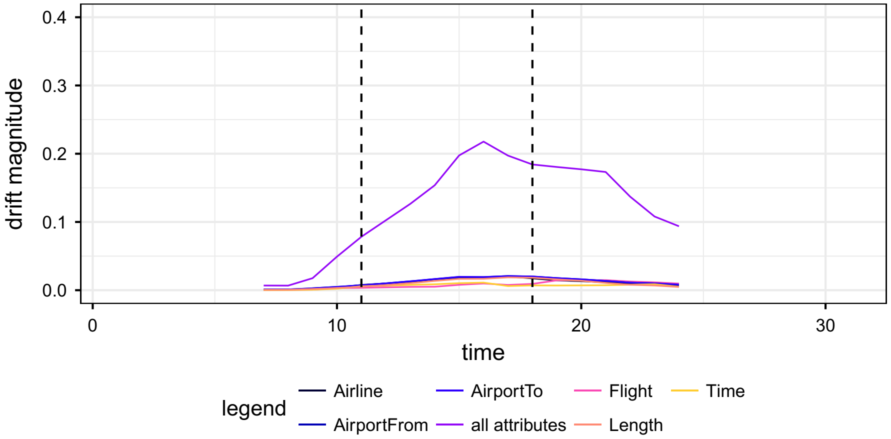

Fig. 6 Weekly covariate drift calculated daily for the airlines data. The dashed lines are placed between each Saturday and Sunday

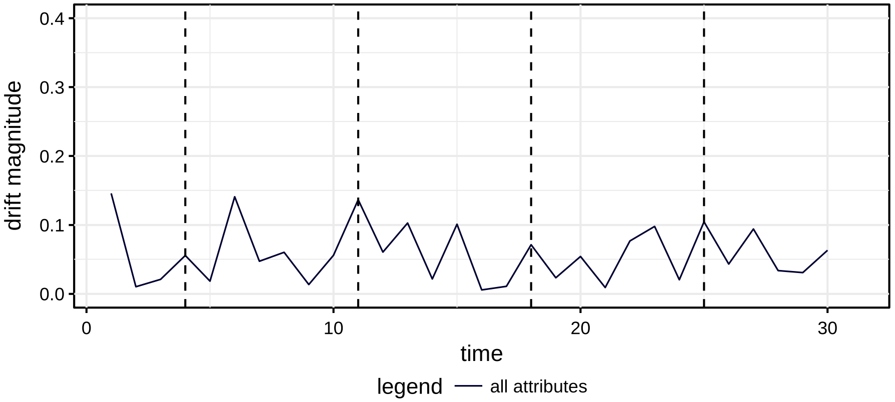

Fig. 7 Daily class drift for the airlines data. The dashed lines are placed between each Saturday an Sunday

The weekly analysis shows that while there is substantial drift from day to day, there is little drift between the first two weeks, confirming the notion that they follow a steady cycle. The inter-week drift then rises sharply. Interestingly, it is the origin and destination airports and flight lengths that change most from week to week as opposed to the time of day and flight number which dominated the inter-day drift.

Figures 7 and 8 show the daily and weekly class drift, respectively. They reveal that the class, representing on-time performance, is not subject to the same weekly cycle of drift as the covariates and that there is greatest drift in on-time performance between the second and third weeks. It is interesting to contrast the inter-week covariate drift to the inter-week class drift. The covariates start with almost no drift which then increases substantially, while the class starts with substantial drift and subsequently drops to having almost no drift. In general, these plots are revealing in that they show that the class drift for this data is quite different in nature to the covariate drift.

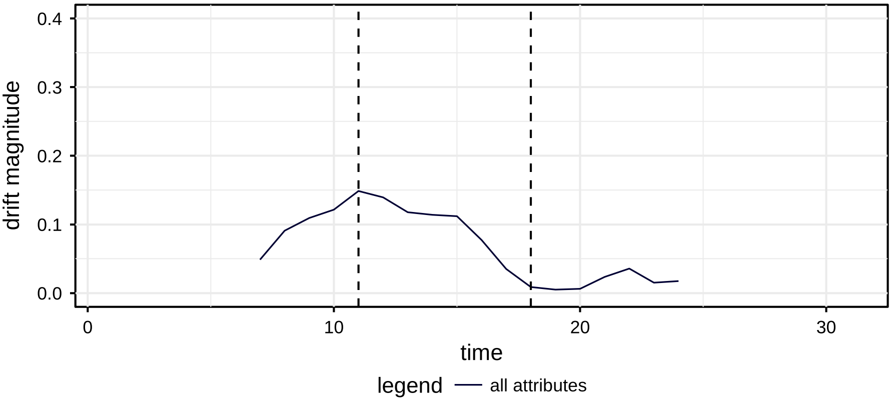

Fig. 8 Weekly class drift calculated daily for the airlines data. The dashed lines are placed between each Saturday an Sunday

This data demonstrates the importance of the granularity of the time periods used in drift analysis and the manner in which different granularities can each convey different and valuable insights. It also illustrates how it is revealing to consider each of the different forms of drift, joint, class, covariate, conditioned class and conditioned covariate. These different aspects of a distribution may each drift in different ways, and an analysis that does not consider all may miss important insights into the nature of drift in a domain.

6.3 Satellite

The final example is satellite data of land usage in France. We use Landsat-8 images acquired over the agricultural year 2013. Images were obtained through the Theia Land Data Centre (http://www.theia-land.fr/en/presentation/products). The Landsat products are orthorectified prior to their release by the USGS and then, Theia processing chains based on the algorithms described in Hagolle et al. (2015) (and cloud shadow) screening and atmospheric corrections. These corrections ensure that the values observed cover the exact same geographic areas and that they are comparable over time.

From these images, we use the multi-spectral product at a spatial resolution of 30 m (Landsat-8 band 1 to band 7) and add three additional attributes, which are indices of vegetation, water and brightness (resp. Normalized Difference Vegetation Index, NDVI, Normalized Difference Water Index, NDWI, and Brightness). An example Landsat-8 image is illustrated in Fig. 9 (Inglada et al. 2017).

In addition, we have a land-cover map for the whole year which associates a class label to each "pixel" (or line) in our database; this label map is illustrated in Fig. 10. The data was prepared by our colleagues at the CESBIO laboratory (see acknowledgements).

The `id` class represents the land usage of the point being imaged. We analyse here the drift between the images take on 5 May 2013 and 29 November 2013. These dates

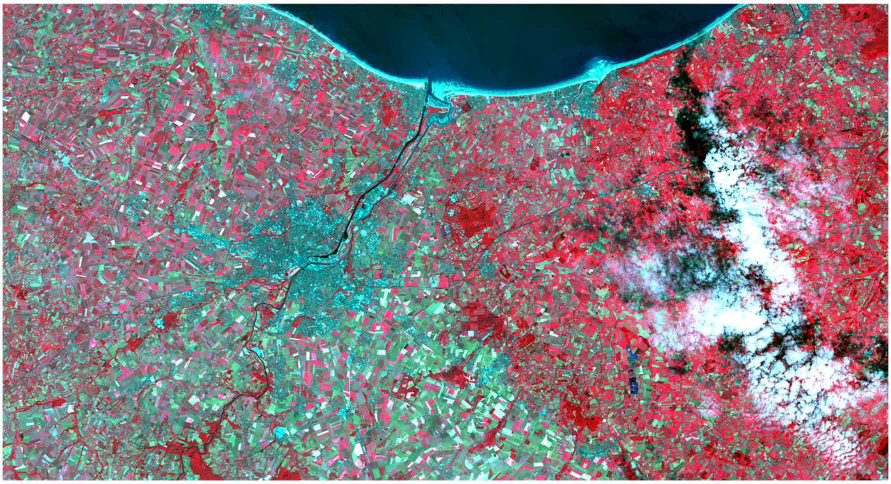

Fig. 9 Landsat-8 image taken on the 17th of July 2013—red displays near-infrared, green displays red and blue displays green (traditional false-color composite). Contains USGS/NASA Landsat Program data © 2013 processed at level 2A by CNES for THEIA Land data centre (Color figure online)

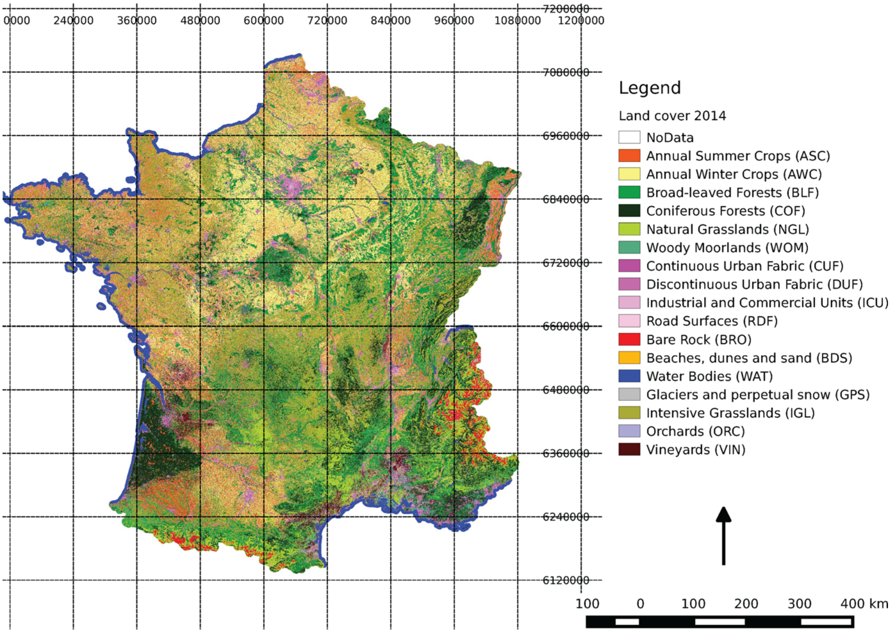

Fig. 10 Labels for the satellite dataset. © Inglada et al. (2017) available at http://dx.doi.org/10.3390/rs9010095 under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/)

in Spring and Fall were chosen as ones between which there should be expected to be substantial changes. May is generally just before the harvest of winter crops, e.g. wheat, canola, and barley (light yellow in Fig. 10).

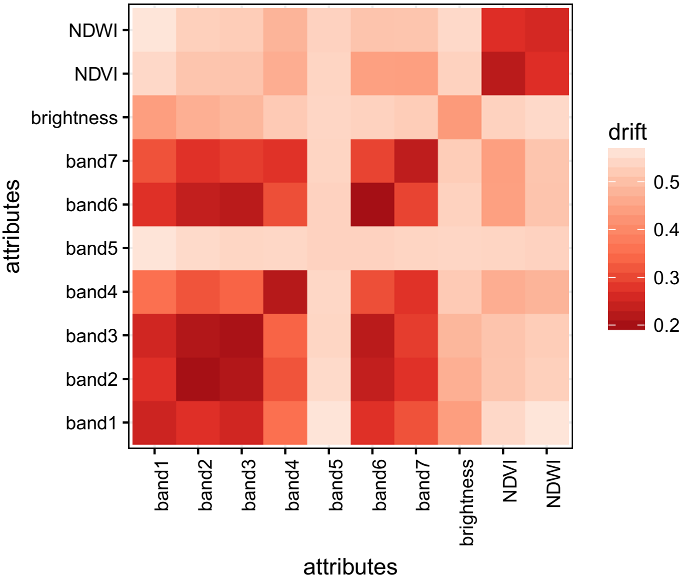

Fig. 11 Pairwise drift in the joint distribution on the satellite data between May and November

The drift magnitudes reveal that this is indeed the case. The covariate drift magnitude is 0.68, the conditional covariate drift is 0.76, the class drift is 0.00 and the conditional class drift is 0.48. There is no class drift because the land usage is determined on an annual basis and hence does not change during the year. There is nonetheless drift in the class conditioned on the covariates because when $P(Y)$ is invariant and $P(X)$ changes it follows that there must be a change in $P(Y \mid X)$.

Figure 11 is a heat map displaying the drift over each pair of attributes in the joint distribution on the satellite data. The diagonal represents univariate drift. For example, the cell at the intersection of the row and column labelled `id` gives the magnitude of the drift for the class attribute `id`. As the land usage assigned to each point does not change over the period, the drift magnitude is 0.0. The largest univariate drift is for band 5, which corresponds to near-infrared. This is explained by the fact that chlorophyll reflects near-infrared; in May, a lot of surfaces are covered by growing crops, which leads to a large amount of near-infrared being reflected. On the other hand, most crops have been harvested late November. More generally, it can be seen that each of these univariate drifts is lower than any of the bivariate drifts involving that same attribute, as our monotonicity proof in Appendix A demonstrates they must.

The drift for `NDWI` and `NDVI` is particularly interesting. The univariate drift for both these attributes considered in isolation is relatively low, but when considered in conjunction with most other attributes is high.

Figure 12 gives a heat map of the drift of each covariate conditioned on the class. The x-axis gives the classes and the y-axis gives each of the covariates. This illustrates how drift can vary greatly from class to class. For `wheat` and `rapeseed/canola`, NDVI and NDWI are changing substantially between May and November, which is

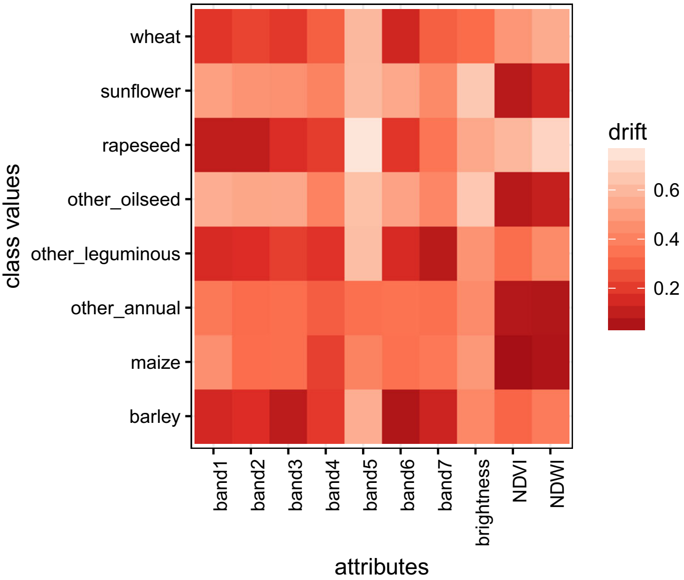

Fig. 12 Conditional marginal covariate drift of individual attributes

explained by the fact that May is the peak season for these winter crops while they have been harvested in November. As a result there are significantly changes in the NDVI—which is a proxy for plant health—and NDWI—which is a proxy for the water content of the leaves. Interestingly, `maize/corn` doesn't drift for NDVI and NDWI as these crops are growing after May and harvested before November; they thus keep the same "bare soil" reflectance.

Figure 13 provides heat maps for each pair of attributes conditioned by each class. It also illustrates the monotonicity of drift magnitude. The drift for any pair of attributes given a class must always be at least as high as the univariate drift of either of the attributes given that class. For instance, from Fig. 12, we can observe the attributes `band4` and `band7` drift the lowest among the other `band` attributes given the classes `other_oilseed` and `sunflower`. This translates to a low joint drift magnitude under the same classes in Fig. 13.

Figure 14 shows the condiitonal class drift conditioned on pairs of attributes. It might at first sight seem anomalous that there should be conditional class drift of up to 0.34 when the class is conditioned on specific pairs of attributes, but no drift when the class is considered in isolation. As explained above, this arises because the only way in which $P(X)$ can change while $P(Y)$ remains invariant is for $P(Y \mid X)$ to change. It is particularly revealing that the conditional class drift within each individual x-value is low, while for some combinations of x-values it becomes relatively high. This demonstrates the value of evaluating the drift across different combinations of attributes. We find here again high values for NDVI, NDWI, and band 5, which is explained by the difference in the agricultural season.

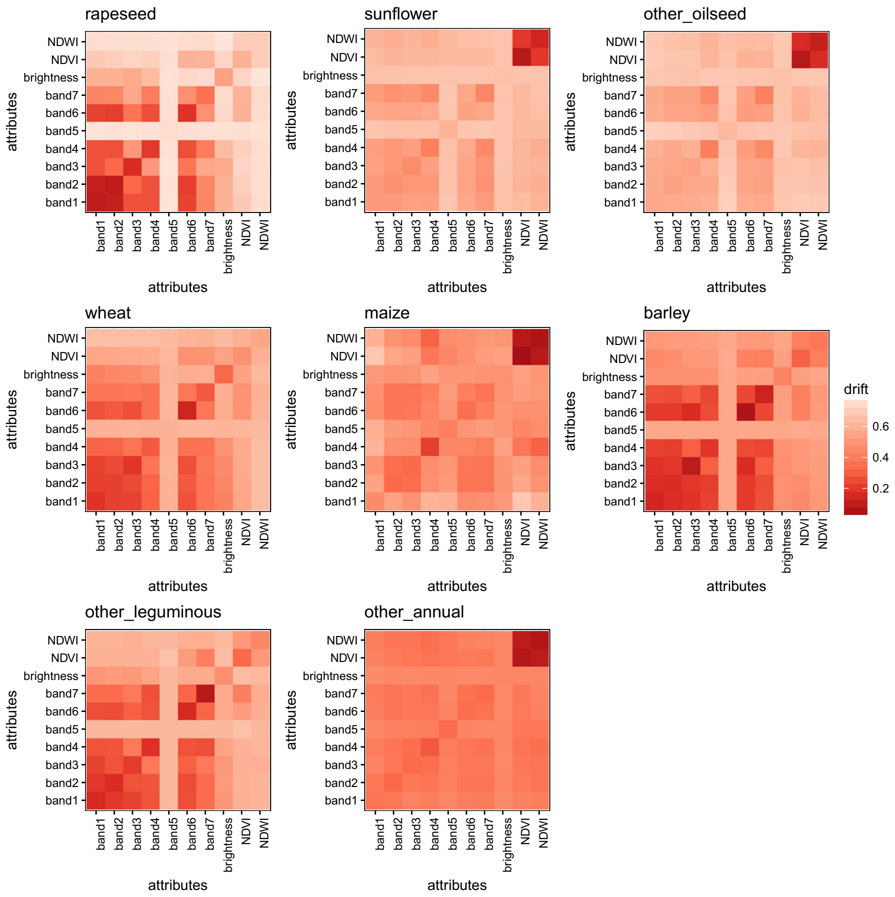

Fig. 13 Conditional marginal covariate drift for pairs of attributes

7 Related research

Prior techniques for describing concept drift (Widmer and Kubat 1996; Kifer et al. 2004; Gama et al. 2004; Baena-Garcıa et al. 2006; Nishida and Yamauchi 2007; Dries and Rückert 2009; Gama and Rodrigues 2009; Žliobaite 2010; Hoens et al. 2012; Moreno-Torres et al. 2012; Bifet et al. 2013; Gama et al. 2014; Qahtan et al. 2015) have been qualitative, utilizing terms such as *abrupt* and *gradual*. As Webb et al. (2016) argue, such qualitative descriptions are limited in that they require arbitrary specification of the boundaries between different values and cannot distinguish between different gradations along dimensions that are in reality continuous in nature, such as drift magnitude. In contrast, the current proposal provides detailed quantitative descriptions of concept drift at a fine level of granularity.

The pioneering work of Pratt and Tschapek (2003) used brushed histograms to visualize univariate drift in each of many dimensions simultaneously. Our work is distinguished by focusing on quantitative multivariate measures of drift. The visualiza-

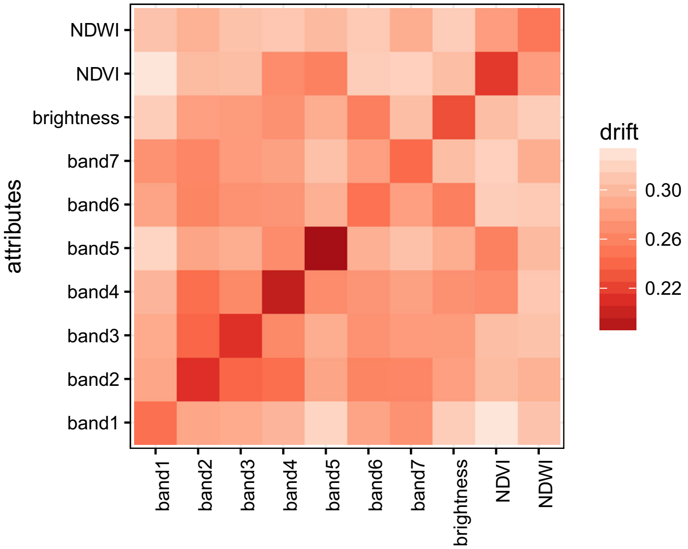

Fig. 14 Conditional class drift conditioned on pairs of attributes

tions that we develop are intended to efficiently and effectively convey these objective multivariate measurements for large numbers of combinations of dimensions.

Yao et al. (2013) develop a more complex form of visualization that relies on unsupervised learning of 'concepts.' The drift in the distributions of these 'concepts' are then visualized. In contrast, our methods directly quantify drift in the original feature space and provide an objective framework with quantitative measures that are directly comparable from one domain to another.

Drift mapping differs greatly in nature to drift detection (Gama et al. 2004; Baena-Garcıa et al. 2006; Dries and Rückert 2009; Bifet et al. 2013; Gama et al. 2014; Qahtan et al. 2015; Yu and Abraham 2017). The former seeks to describe in detail the nature of drift between specific time periods, whereas the latter seeks to identify whether or not drift has occurred at a specific point in time. Drift detection is often employed as a mechanism within online learning algorithms, while drift mapping is primarily intended as a standalone data analysis task. Whereas drift mapping is envisaged as helping us understand how different drift response mechanisms perform in the face of different forms of concept drift, drift detection is essentially one of those drift response mechanisms.

8 Conclusions and future research

Concept drift is in some senses the great elephant in the room for machine learning. The world is continually changing, but we have a dearth of techniques for understanding the nature of these changes as they apply to specific machine learning contexts. We have a growing body of sophisticated methods for learning in the context of concept drift (Gaber et al. 2005; Gama and Rodrigues 2009; Aggarwal 2009; Žliobaite 2010; Bifet et al. 2011; Nguyen et al. 2015; Brzezinski and Stefanowski 2014; Krempl et al. 2014; Gama et al. 2014; Ditzler et al. 2015). There is a need to develop a supporting body of techniques for understanding the phenomena that these methods address and thereby understanding the relative capabilities of these methods in the face of different expressions of that phenomena.

This paper proposes a new data mining task—drift mapping. The proposal builds on Webb et al's (2016) method of quantifying drift magnitude by

- revealing the importance of quantifying drift magnitude over marginals rather than through a single gross measure — both because a single global measure will become uninformative in high dimensional data and also because drift can be expected be heterogeneous across different data subspaces and it will often be critical to understand how drift differs between subspaces;
- proposing multiple techniques for communicating the complex information revealed by the maps (see Sect. 5);
- revealing the importance of interval granularity for effective drift mapping (see Sect. 6); and
- highlighting the importance of mapping all of the joint, class, covariate, conditioned class and conditioned covariate distributions.

These preliminary techniques for mapping concept drift leave substantial scope for refinement.

- It may prove useful to handle numeric data directly without requiring discretization.
- In the current work we use maximum likelihood estimates of the probability distributions. These are likely to be imprecise, adding noise to the estimates which will accumulate as dimensionality increases. Methods to address this issue are likely to be important when seeking to map high-dimensional data.
- For very high dimensional data it will not be feasible to present and consider every pairwise marginal distribution. There is a need for techniques either to identify and highlight the marginals in which the drift is most interesting, or to allow a user to explore the space of marginals in an effective manner.
- In the airlines example, inter-day and inter-week drift demonstrated very different patterns, each of which was revealing of different dynamics in the data. This well illustrates the importance of identifying informative granularity for analysis. In some domains this may be readily apparent to the relevant experts. However, there are likely to be domains where the analyst does not have access to such expertise and it would be useful to have tools to automatically identify appropriate granularities for analysis.

We have presented practical techniques for a new data analysis task—modelling and communicating the nature of drift affecting specific applications. Our case studies on three real-world datasets demonstrate that these techniques can reveal insights into the nature of specific instances of drift that cannot be obtained by any prior method.

All data analytics are necessarily after the fact. We cannot analyse drift that is yet to happen. Drift maps reveal the forms of drift that have occurred in a particular domain. This is likely to provide insight into the types of drift that are likely to occur into the future, but as our case studies reveal, just as there is drift in the underlying distributions, there is also drift in the type of drift that affects a given domain. Drift maps will help users understand how applicable historical data is to the immediate past, but cannot definitively determine their applicability to the future.

We hope that these techniques will have practical application in addressing the very real and present problem of concept drift. As a service to the community we have established an online server to which users can upload data to be analysed by our tools at http://driftmap.infotech.monash.edu.au. In the interests of reproducible research we make the software necessary to reproduce our results available at https://github.com/LeeLoongKuan/DriftMapper and https://github.com/LeeLoongKuan/DataAnalysisR. The first of these produces the drift maps in numeric form while the second creates the heat map and line plot visualizations.

Acknowledgements This work was supported by the Australian Research Council under Awards DP140100087 and DE170100037. This material is based upon work supported by the Air Force Office of Scientific Research, Asian Office of Aerospace Research and Development (AOARD) under Award Number FA2386-17-1-4033. The authors would like to thank the colleagues from CESBIO (Jordi Inglada, Arthur Vincent, Marcela Arias, Benjamin Tardy, David Morin and Isabel Rodes) for providing the Satellite dataset (data and labels).

A Proof that drift magnitude is monotone under increasing dimensionality

We here prove that total variation distance is monotone under increasing dimensionality. The proof generalizes trivially to Hellinger distance. Note that where one set of variables is conditioned on another, it is the dimensionality of the conditioned variable rather than the conditioning variables over which this monotone increase in distance applies.

Let X, Z be sets of covariates.

$$\sigma_{t,u}(X) \leq \sigma_{t,u}(X, Z)$$

$$\Updownarrow$$

$$\frac{1}{2} \sum_{\bar{x} \in \mathrm{dom}(X)} |P_t(\bar{x}) - P_u(\bar{x})| \leq \frac{1}{2} \sum_{\substack{\bar{x} \in \mathrm{dom}(X) \\ \bar{z} \in \mathrm{dom}(Z)}} |P_t(\bar{x}, \bar{z}) - P_u(\bar{x}, \bar{z})|$$

$$\Updownarrow$$

$$\sum_{\bar{x} \in \mathrm{dom}(X)} \left| \sum_{\bar{z} \in \mathrm{dom}(Z)} P_t(\bar{x}, \bar{z}) - \sum_{\bar{z} \in \mathrm{dom}(Z)} P_u(\bar{x}, \bar{z}) \right| \leq \sum_{\bar{x} \in \mathrm{dom}(X)} \sum_{\bar{z} \in \mathrm{dom}(Z)} |P_t(\bar{x}, \bar{z}) - P_u(\bar{x}, \bar{z})|$$

$$\Updownarrow$$

$$\sum_{\bar{x} \in \mathrm{dom}(X)} \left| \sum_{\bar{z} \in \mathrm{dom}(Z)} P_t(\bar{x}, \bar{z}) - P_u(\bar{x}, \bar{z}) \right| \leq \sum_{\bar{x} \in \mathrm{dom}(X)} \sum_{\bar{z} \in \mathrm{dom}(Z)} |P_t(\bar{x}, \bar{z}) - P_u(\bar{x}, \bar{z})|$$

References

Aggarwal CC (2009) Data streams: an overview and scientific applications. Springer, Berlin, pp 377–397. https://doi.org/10.1007/978-3-642-02788-8_14

Baena-Garcıa M, del Campo-Ávila J, Fidalgo R, Bifet A, Gavalda R, Morales-Bueno R (2006) Early drift detection method. In: Fourth international workshop on knowledge discovery from data streams, vol 6, pp 77–86

Bifet A, Gama J, Pechenizkiy M, Zliobaite I (2011) Handling concept drift: importance, challenges and solutions. PAKDD-2011 Tutorial, Shenzhen, China

Bifet A, Read J, Pfahringer B, Holmes G, Žliobaite I (2013) CD-MOA: change detection framework for massive online analysis. In: International symposium on intelligent data analysis. Springer, Berlin, pp 92–103

Brzezinski D, Stefanowski J (2014) Reacting to different types of concept drift: the accuracy updated ensemble algorithm. IEEE Trans Neural Netw Learn Syst 25(1):81–94

Ditzler G, Roveri M, Alippi C, Polikar R (2015) Learning in nonstationary environments: a survey. IEEE Comput Intell Mag 10(4):12–25

Dries A, Rückert U (2009) Adaptive concept drift detection. Stat Anal Data Min 2(5–6):311–327

Gaber MM, Zaslavsky A, Krishnaswamy S (2005) Mining data streams: a review. ACM SIGMOD Rec 34(2):18–26

Gama J, Žliobaite I, Bifet A, Pechenizkiy M, Bouchachia A (2014) A survey on concept drift adaptation. ACM Comput Surv 46(4):44:1–44:37. https://doi.org/10.1145/2523813

Gama J, Medas P, Castillo G, Rodrigues P (2004) Learning with drift detection. In: Brazilian symposium on artificial intelligence. Springer, pp 286–295

Gama J, Rodrigues P (2009) An overview on mining data streams, vol 206. Studies in computational intelligence. Springer, Berlin, pp 29–45. https://doi.org/10.1007/978-3-642-01091-0_2

Hagolle O, Sylvander S, Huc M, Claverie M, Clesse D, Dechoz C, Lonjou V, Poulain V (2015) Spot-4 (take 5): simulation of sentinel-2 time series on 45 large sites. Remote Sens 7(9):12242–12264. https://doi.org/10.3390/rs70912242

Harries M (1999) Splice-2 comparative evaluation: electricity pricing. Technical Report UNSW-CSE-TR-9905, University of New South Wales

Hellinger E (1909) Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen. Journal für die reine und angewandte Mathematik 136:210–271

Hoens TR, Chawla NV, Polikar R (2011) Heuristic updatable weighted random subspaces for non-stationary environments. In: Cook DJ, Pei J, Wang W, Zaiane OR, Wu X (eds) IEEE international conference on data mining, ICDM-11. IEEE, pp 241–250

Hoens TR, Polikar R, Chawla NV (2012) Learning from streaming data with concept drift and imbalance: an overview. Prog Artif Intell 1(1):89–101. https://doi.org/10.1007/s13748-011-0008-0

Inglada J, Vincent A, Arias M, Tardy B, Morin D, Rodes I (2017) Operational high resolution land cover map production at the country scale using satellite image time series. Remote Sens. https://doi.org/10.3390/rs9010095

Kifer D, Ben-David S, Gehrke J (2004) Detecting change in data streams. In: Proceedings of the thirtieth international conference on very large data bases—volume 30, VLDB Endowment, VLDB ’04, pp 180–191

Krempl G, Zliobaite I, Brzezinski D, Hullermeier E, Last M, Lemaire V, Noack T, Shaker A, Sievi S, Spiliopoulou M, Stefanowski J (2014) Open challenges for data stream mining research. ACM SIGKDD Explor Newsl 16–1:1–10

Levin D, Peres Y, Wilmer E (2008) Markov chains and mixing times. American Mathematical Society, Providence

MOA dataset repository (2017) http://moa.cms.waikato.ac.nz/datasets/. Accessed 1 Sept 2017

Moreno-Torres JG, Raeder T, Alaiz-Rodriguez R, Chawla NV, Herrera F (2012) A unifying view on dataset shift in classification. Pattern Recognit 45(1):521–530

Nguyen HL, Woon YK, Ng WK (2015) A survey on data stream clustering and classification. Knowl Inf Syst 45:535–569

Nishida K, Yamauchi K (2007) Detecting concept drift using statistical testing. In: International conference on discovery science. Springer, pp 264–269

Pratt KB, Tschapek G (2003) Visualizing concept drift. In: Proceedings of the ninth ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 735–740

Qahtan AA, Alharbi B, Wang S, Zhang X (2015) A PCA-based change detection framework for multidimensional data streams: Change detection in multidimensional data streams. In: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 935–944

Roarty M (1998) Electricity industry restructuring: the state of play. Research Paper 14, Science, Technology, Environment and Resources Group. http://www.aph.gov.au/About_Parliament/Parliamentary_Departments/Parliamentary_Library/pubs/rp/RP9798/98rp14. Accessed 1 Sept 2017

Webb GI, Hyde R, Cao H, Nguyen HL, Petitjean F (2016) Characterizing concept drift. Data Min Knowl Discov 30:964–994

Widmer G, Kubat M (1996) Learning in the presence of concept drift and hidden contexts. Mach Learn 23(1):69–101. https://doi.org/10.1007/BF00116900

Yao Y, Feng L, Chen F (2013) Concept drift visualization. J Inf Comput Sci 10(10):3021–3029

Yu S, Abraham Z (2017) Concept drift detection with hierarchical hypothesis testing. In: Proceedings of the 2017 SIAM international conference on data mining. SIAM, pp 768–776

Žliobaite I (2010) Learning under concept drift: an overview. CoRR arXiv:1010.4784