Measuring Behavioural Change of Players in Public Goods Game

Polla Fattah, Uwe Aickelin, and Christian Wagner

Abstract In the public goods game, players can be classified into different types according to their participation in the game. It is an important issue for economists to be able to measure players' strategy changes over time which can be considered as concept drift. In this study, we present a method for measuring changes in items' cluster membership in temporal data. The method consists of three steps in the first step, the temporal data will be transformed into a discrete series of time points then each time point will be clustered separately. In the last step, the items' membership in the clusters is compared with a reference of behaviour to determine the amount of behavioural change in each time point. Different external cluster validity indices and area under the curve are used to measure these changes. Instead of different cluster label comparison, we use these indices a new way to compare between clusters and reference points. In this study, three categories of reference of behaviours are used 1- first time point, 2- previous time pint and 3- the general overall behaviour of the items. For the public goods game, our results indicate that the players are changing over time but the change is smooth and relatively constant between any two time points.

Polla Fattah

School of Computer Science, The University of Nottingham, Nottingham, UK, e-mail: psxpf1@nottingham.ac.uk

Uwe Aickelin

School of Computer Science, The University of Nottingham Ningbo, China, Ningbo, e-mail: uwe.aickelin@nottingham.edu.cn

Christian Wagner

School of Computer Science, The University of Nottingham, Nottingham, UK, e-mail: pszcw@nottingham.ac.uk

1 Introduction

There are multiple studies in the economics regarding the behaviour and strategy change of players in the public goods games [9]. The players are classified into four types conditional cooperators, free riders, triangular contributors and others. The change of players behaviour means jumping from one class to another during the gameplay rounds. This jump can be considered as concept drift. As defined by Gama et al. [13], concept drift learning under is an unexpected change from the targeted future estimation due to uncalculated hidden contexts in the system. Tsymbal [31] identified two types of concept drift: sudden and gradual. This work presents a method to measure the quantity of the change occurring within populations in any two different time points.

Different methods are developed to detect drifts in the data streams such as [3, 17] and special classification methods are introduced to precisely predict items with the existence of concept drift [6, 14, 33]. Moreover measuring changes in clusters for different time points are well studied in data analysis, especially for data streams [26, 29, 34]. However, these methods aim to find overall patterns of change in the clusters rather than presenting a measure of how much change has occurred in each cluster.

External cluster validity is primarily used to check the performance of clustering algorithms by measuring the difference between ground truth labels given to the items by experts and the group in which they have been placed by a clustering algorithm [28]. This study uses external cluster validity measures like Variation of Information [24] VI to show the amount of items that jumped from one cluster to another between two consequent time points. Moreover, for comparison with these criteria we use area under the curve of Receiver Operating Characteristic ROC [5] which is originally a measure for classification algorithms to quantify change of behaviours as well. To accomplish this measurement, the items should be clustered separately in every time point. As the clustering is performed at a single time point, which eliminates the time dimension for the collected data about items, theoretically any traditional (non-temporal) clustering algorithm should be sufficient. After clustering, an external clustering validity measure can quantify the amount of changes between clusters at any two time points.

To compare our results, public goods games data were also tested using MONIC, which is a method of detecting changes among clusters in the data stream. The results show that there is a periodic change in clusters as they disappear and other clusters are emerging, but this is inconclusive as there is no indication of whether the change originated from players' strategies or from the nature of algorithm, as it reduces the effect of the old items in the cluster and removes them after two time points.

2 Background and Related Work

In the following subsections, a brief description of the public goods games and the used data sets for the experiments will be explained as well as describing the available techniques and methods to measure changes over time.

2.1 *Public Goods Games Data*

Public good is any service or resource that cannot be withheld from any individuals due to inalienable characteristics relating to citizens' rights [21]. Examples of public good resources include city parks, street lighting and roads, which are funded by the state but which are available to all. The public goods game is an experimental game that simulates real situations of public good in a lab with controlled conditions and focused purposes of conducting experiments. There are many slightly different variations of this game, but the data been used in this paper as a case study is based on the model of Fischbacher et al. [10].

There are many variations and set-ups for the public goods game experiment, However, the data which has been used in this study is collected through experiments conducted by Fischbacher et al. [10]. Their experiment for public goods game consists of two sub-experiments; P-experiment and C-experiment, both of which every participant (player) has to accomplish. In the following sections, we will explain how these two sub-experiments are conducted, and then describe the collected data which will be used in later chapters.

2.1.1 Game Set-up

Prior to each sub-experiment of P-experiment and C-Experiment, experimenters explain the rules of the game for the participants so that they understand the rules, and how their decision will affect their result and the number of points available. Participants should answer a number of control questions correctly to demonstrate their comprehension of the game. Experimenters make every effort to ensure that the players are paying attention and playing thoughtfully by rewarding them extra points for correct guesses and well-thought out decisions during the game.

In P-experiment, four players start the game with 20 points each in their private account and they can contribute any amount they deem necessary to a project which represents public good. The amount which they do not contribute in the project will be kept only for the players themselves. The collected amount from the project will be distributed among all players regardless of their contribution to the project. The amount of points each player can accrue from the project is determined by this equation:

$$PlayerShareFromProject = TotalAmountOfAllPlayersContribution * 0.4$$

So that each players total point after the game will be:

$$Player'sPoints = (20 - ContributionInTheProject) + PlayerShareFromProject$$

For each player, only one of the two contributions will be selected by the computer as their final contribution to the project. One of the four players' conditional contribution will be randomly chosen to be used as their final contribution. while for the other three players their unconditional contribution will be used. This random selection of players' contributions is one of the mechanisms that experimenters have used to make sure that players are thinking thoroughly about their decision for the contribution to the project.

When the P-experiment is completed, players start C-experiment. C-experiment is similar to a repeated sequence of unconditional contribution except this time the player, in addition to their own contribution, will be asked to guess other players' rounded average of contribution. After each round of the game, players will be notified of their total points in that particular game. The sequence length of the games can vary from one experiment to another. In this study, we will use data sets with 10 and 27 series of rounds of the game. In each round, four different random players will play the game so that players can not predict others' contributions in advance. Players will gain extra points if they make correct guesses about other players' rounded contributions. They will, therefore, not fill in the boxes randomly.

2.1.2 Data Set Attributes

To measure and classify the behaviour of players in public goods games, this study used two different data sets. These experiments are conducted on different samples of players, so the first data set has 140 players and the second data set 128 players. These data sets have the same attributes and follow exactly the same experiment procedures, except for the P-experiment length, as the first one consists of 10 rounds while the other has 27 rounds.

The attributes of the data sets can be divided into two types the temporal and non-temporal attributes. The temporal attributes are generated in the P-experiment as it contains multiple rounds and non-temporal attributes are generated in C-experiment. The following is the list of all the attributes of the data sets. Please notice that the temporal attributes are underlined:

- **Idtyp**: labels for players categories assigned by experts. The categories are: conditional contributors = 1, free riders = 2, triangle contributors = 3, and others = 4. These categories are generated depending entirely on the static data of the contribution table.
- **Idsubj**: a unique identifier for each player during both C and P experiments.

- **b0-b20**: twenty one attributes representing the contribution table for each player as their response in C-experiment to every possible rounded average of other players' contribution.
- **u**: the unconditional contribution of the player for C-experiment during the actual game.
- **Predictedcontribution**: Players' prediction about other co-players rounded average of contribution to the project.
- **Period**: the session number for P-experiment. As P-experiment for each player consists of multiple rounds, each players' playing times are recorded to keep track of the number of games played.
- **Contribution**: players' actual contribution to the project in each round of the P-experiment.
- **Belief**: players' beliefs about other players average contribution in each session.
- **Otherscontrib**: Other co-players' rounded average contribution.

2.2 External Cluster Validity

External criteria validate the results of clustering based on some predefined structures of the data which is provided from an external source. The most well-known example of structural information is labels for the data provided by experts (called true classes). The main task of this approach is to determine a statistical measure for the similarity or dissimilarity between obtained clusters and labels [16, 27]. According to the methods incorporated in the external criteria, they can be divided into three types: pairwise measures, entropy-based measures and matching based measures [35].

As mentioned previously, the four types of classification guesses evaluation are true positive, true negative, false positive and false negative. These terms are used in the terminology of external cluster validity, especially when using pairwise measures, but with slightly different meanings to enable the evaluation of clusters in the same manner as classification [35]:

- True Positives **TP**: Any two instances with the same label that are in the same cluster.
- False Negatives **FN**: Any two instances with the same label that are not in the same cluster.
- False Positives **FP**: Any two instances with different labels that are not in the same cluster.
- True Negatives **TN**: Any two instances with different labels that are not in the same cluster.

In this study we use various external cluster validity indices to determine differences between a reference of behaviour for items in a temporal data and clusters of items in each time point. The method is discussed in more detail in chapter three,

and implemented in chapter four for public goods games and chapter six for stock market data. The used criteria in the thesis are listed below:

2.2.1 Jaccard Coefficient:

This coefficient is a pairwise measure representing the degree of similarity between clusters. With this coefficient each cluster is treated as a mathematical set and the coefficient value is calculated by dividing the cardinality of the intersection of the resultant cluster with the prior cluster to the cardinality of the union between them [32]:

$$Jaccard = \frac{TP}{TP + FP + FN}$$

With a perfect clustering, when false positives and false negative equal to zero, the Jaccard coefficient value equals 1. This measure ignores the true negatives and only focuses on the true positives to evaluate the quality of the clusters [35].

2.2.2 Rand Statistic:

The Rand statistic measures the fraction of true positives and true negatives over all point pairs; it is defined as

$$Rand = \frac{TP + TN}{N}$$

Where N is the total number of instances in the data set. This measure is similar to Jaccard Coefficient, so its value equals 1 in perfect clustering [35].

2.2.3 Fowlkes-Mallows (FM) Measure:

FM define precision and recall values for produced clusters [12]

$$FM = \sqrt{prec \cdot recall} = \frac{TP}{\sqrt{(TP + FN)(TP + FP)}}$$

Where $prec = \frac{TP}{TP + FP}$ and $recall = \frac{TP}{TP + FN}$. For the perfect clustering this measure equals 1 too [35].

2.2.4 Variation of Information VI:

This index measure is based on contingency table which is a matrix with $r \times k$, where r is number of produced clusters and k is the number of externally provided clusters. Each element of this matrix contains a number of agreed instances between any two clusters of the externally provided and produced clusters. As introduced by

Meila [23], this index calculates mutual information and entropy between previously provided and produced clusters derived from the contingency table:

$$VI(C, T) = 2H(T, C) - H(T) - H(C)$$

Where C is produced clusters, T is ground truth clusters, $H(C)$ is entropy of C and $H(T)$ is entropy of T [35].

2.3 Measuring Cluster Changes in Data Streams

Spiliopoulou et al. [30] introduced the MONIC model, which finds cluster transition over accumulating data sets, providing an ageing function for clustering data that prioritizes new records over old ones and eliminates records older than two time points. Matching for clusters in one time point to the next one is carried out by passing a threshold that determines normalized maximum number of records that exist in both matched clusters in the two time points. This model defines two kinds of transitions, external and internal. In external transition clusters may survive, split, be absorbed, disappear or emerge, while in internal transition clusters can change in size, compactness or location.

According to MONIC, each cluster has a lifetime, which is the number of time points throughout which it can survive. Longer cluster lifetimes enable more predictable clustering while short lifetimes lead to volatile and unpredictable clustering.

It can be observed that this model relies on accumulated data over time to detect cluster matches, therefore it cannot be used with non-accumulated data. Moreover, it emphasises the measurement of cluster changes and cannot detect changes in cluster membership for individual items clustered over time points.

Gunnemann et al. [15] introduced a method which traces cluster evolution as change in behaviour types indicated by the value of objects (e.g. persons) in high-dimensional data sets. Different types of mapping function were introduced to map clusters according to their values in different dimensions and subspaces instead of object identifier. Using this method cluster evolutions were detected and counted in the forms of emerge, disappear, converge and diverge. Moreover, the loss and gain of dimensions of subspace clusters were calculated. This method counts the number of various changes that occur to clusters of any high dimensional data set, but it lacks to any mean by which to quantify the changes themselves; in other words, there is no indication of the quantity of change that happens to any cluster in two consecutive time points.

Hawwash et al. [18] proposed a framework for mining, tracking and validating clusters in a data stream using statistical cluster measures like cardinality, scale and density of clusters to detect milestones of clusters change and monitor the behaviour of cluster. This framework targets accumulative clustering on data streams, but instead of using fixed-time window for clustering it uses milestones to detect the

next-best clustering time. They used a linear model in their metrics, which cannot represent real-life situations. They made this concession due to time limitations and the memory complexity of higher degree models. With some enhanced models this method could be profitably used to determine critical time points in the data stream clustering and to track clusters behaviour in general using statistical measures for representative numbers pertaining to the situation of clusterings.

Kalnis et al. [20] introduced a method to discover moving objects in the snapshots of spatio-temporal data using cluster mapping function, treating clusters as sets and calculating the cardinality ratio of intersection for each two time constitutive clusters over their union; if the ratio passes a certain threshold the cluster is considered to be a moving cluster. This method detects move in overall clusters and provides visual aids enabling human experts to grasp changes in the underling data [26, 4]. This method is excellent for tracking moving cluster change [25] , but it still lacks a method to quantify the magnitude of change for overall clustering objects.

Aggarwal [1] introduced a new method to detect changes for single clusters in the data streams that also works for snapshots of data as special cases. This method uses forward and reverse time slice density estimates based on fixed length time window to calculate velocity density at time and space dimensions. By calculating velocity density three types of change can appear on the clusters in evolving data streams: 1) they may coagulate if the value passed a user specified threshold; 2) they may decay if the value does not pass the threshold; or 3) they may shift their location to another. This method is particularly germane to visually understanding the characteristics of underlying data.

The mentioned methods: are mostly designed to work with data streams or snapshots of spatio-temporal data sets. They detect changes inside data by monitoring cluster change in terms of split, absorbed, disappear and emerged etc., which is a good indication for detecting existence of change, but which does not specify the magnitude of change. Our aim is to create a simple factor (scalar) to express the magnitude of change among members of clusterings in temporal data sets.

3 Methodology

Measuring behaviour differences of items between time points requires three steps: The first step is to address time points, the second step is grouping similar behaviour and the last step is to find and measure the amount of differences between these time points.

The temporal data has to be split into separate time points. If the temporal data has discrete records of time then each timestamp can represent a single time point. If the data set has continues timestamps, then it might be converted to a discrete set of time points using fixed intervals of time windows as used by many studies like [30]. It might be preferable that the time points have similar intervals between them so that the behavioural change measure M can represent the difference between any two time points in the same data set uniformly. For illustration consider $t \in T$ and

the time intervals between $[t-1, t]$ and $[t, t+1]$ are equal which makes $m1 = \delta(t-1, t)$ and $m2 = \delta(t, t+1)$ so that $m1$ and $m2$ can represent the two defined time intervals uniformly.

The second step is grouping similar behaviours of the items in the data so that we can identify each items' category of behaviour at every particular time points. As defined by Jain [19] clustering is the task of finding groups of similar members in an item set. So that each time point is clustered using one of the clustering methods to find similarly behaved groups at each time point. The clustering algorithms which are used in the process of measuring the difference between time points in this study are Kmeans, PAM, and hierarchical clustering.

Clustering items in each time point eliminate two issues regarding items' drift between time points. These issues are minor changes of items' behaviour and shifting all of the items in the same group. These problems will be solved by clustering each time point separately without any influence of other time points. Because the clustering will ensure that the minor changes of items in the values of attributes will not affect their membership in the group and by clustering each time point's data independently, we ensure that the entire movement of a group will not affect the measures of items' behaviour change.

The last step is to find the number of items which are changed their behaviour significantly so that they can be counted as they are in other groups or using the percentage of items' behaviour change. This means finding δ function. It is also possible to use AUC of ROC to find the difference between items' clusters in any two consequent time points by using cluster labels of t and $t+1$ instead of true class labels and predicted classes by a classification model as inputs to the AUC function so that it finds the difference between these two time points.

3.1 Choosing Clustering Algorithms

As each of the produced subsets of data represents one time point of the temporal dataset so that each subset alone do not carry any information about time dimension. This means it is possible to use non-temporal clustering algorithms to cluster items in each subset of the temporal datasets.

Clustering algorithms can be categorised according to their method of finding similarities between items in the data. These categories are partitional, hierarchical, density-based, grid-based and fuzzy clustering [35]. However, the main clustering categories which we used are partition based clustering, hierarchical clustering and fuzzy algorithms. For the tests in this experiment, we used k-means and PAM as methods of partitioning clustering, hierarchical clustering with Euclidean distance and c-means as fuzzy clustering. As we aim to find similarities between items according to their distance from each other, so that we did not use density-based and grid-based clustering methods.

To find similarities between items, clustering methods can use linear distance measures like Euclidean distance or use non-linear kernels to cluster complicated

patterns in the data items. In the tests, we only used linear distance based clustering methods because the aim is to find the similarity of behaviour based on the overall proximity of the attribute values of items. For the same reason, we do not use density-based clustering like DBSCAN and grid-based clustering like STING as these methods do not depend on the mutual proximity of cluster items to a centroid which represents a behaviour category.

3.2 *Choosing External Cluster Validity Indices*

As explained before we propose using external cluster validity indices and area under the curve AUC to measure changes which might occur in the behaviour of the items between multiple time points in a temporal data. Many external cluster validity indices are available [2] to measure the validity of clusters produced by clustering methods compared with the natural partitions exist. In chapter 17 of their book Zaki et al. [35] categorised the external clustering validities into three types matching based measures, entropy-based measures and pairwise measures.

As for matching based measures external cluster validity indices are calculating the match of the clusters to the partitions. This means that the measure do not concern about individual elements' differences between clusters and partition, so that this category might not be beneficial to calculate the changes over time.

The second category of external cluster validity indices, entropy-based measures, calculates the difference of entropy between clusters and ground truth partitions. This method also does not concern about individual items in the clusters and partitions. However, we used one measure of this category which is Variation of Information VI because the entropy of the clusters might be affected by the change of items within the clusters. We also used VI for comparison with other indices.

The last category, pairwise measures, measures cluster validity by comparing the produced clusters and original labels of items' classes. As this category calculates the validity using all elements of the dataset, so that it might be the most appropriate category to be used to calculate the items' changes over time points. Three instances of pairwise measures are used in this study which they are Jaccard Coefficient, Rand Statistic and Fowlkes-Mallows Measure.

To maintain standard criteria for different external cluster validity indices so that the final result which quantifies the amount of change in each time point reflects the actual change which is happened to the groups' items regardless of the measure which is used. To make sure that the measures are standard they should follow two rules (1) the scale of the measure should be between zero and one (2) with being zero as the total change and one is the perfect match between any time point and reference of behaviour. However, not all measures are following these rules. For example, in the selected measures the VI is not bounded to any scale and zero is considered as a perfect match so that the results of this measure should be first scaled to fit in the range of [0-1] and then reversed.

3.3 Using Different Reference of Behaviours

This study considers three different references of behaviours 1) the first time point is used as a reference of behaviour for all other time points 2) The previous time point is used to be the reference of behaviour for the current time point. 3) a temporal classification method will be proposed to classify items in temporal datasets.

Each of these different references of behaviour imposes different meaning and can be used for various reasons. The first time point can be used as a reference of behaviour to quantify the progress of change which happens to the items in any later time points in the dataset. An example of that if we want to quantify the change of behaviour of players in public goods game from the first round of the game to any round of the game. Using previous time point as a reference for the current time point means we aim to measure stepwise changes of items' behaviour between any time point. This can be used to measure the stability of change over time. An example of using this method is when we want to check the stability of changes which can occur in player's behaviour between time points. Items' classes can be used as a reference of behaviour to quantify items' deviation from their generalised behaviour at any time point.

3.4 Classifying players of public goods game

To be able to use the third type of reference of behaviour, players of the public goods game have to be classified according to their temporal data to reveal their general overall behaviour. The available data sets do not contain labels for players according to their temporal behaviour so that we use the rule-based classification method which is introduced by Polla et al. [7]. This classification method consists of two steps: initial rule generation and rule optimisation.

In the first step, the experts of the area of data set provide classes for items in the data set with initial rules separating these classes from each other. The rules are expressed in the form of aggregated attributes which are derived from the temporal attributes. The aggregation ensures simplicity of the rules so that it can be understandable by human experts. The aggregation for temporal attributes can be done using one of the mathematical or statistical functions like minimum, maximum, average and standard deviation. Moreover, each rule might have a range of values to separate between adjacent classes which can be expressed as [min, max]. This range might exist due to the uncertainty of the experts of the best cutting position between two classes or disagreement among different studies [7].

The second step optimises the initial rules by finding the optimum classifier among all available ranges. After this step all [min, max] ranges of the initial rules will be replaced with a single value. This step uses the temporal attributes directly. The optimum classifier is the classifier minimises the distance (dispersion) among items of each class at all time points.

The players of public goods games are classified into four classes according to their contribution and belief attribute the classes are:

- Free Riders: players who contribute by equal or less than one point on average for all rounds or who are not contributing in most rounds. This class corresponds to the traditional category of Free Riders.
- Weak Contributors: players who contribute between 1 and 5 or they are not contributing in half of the rounds.
- Normal Contributors: players who contribute on average around 5 points. This class is strongly related to conditional contributors as it fits the same criteria.
- Strong Contributors: players who contribute more than 10 points on average. This class relates to conditional comparators and others in the classical categories.

4 Measuring Players' Strategy Change over Time

The main objective of this study is to quantify players' strategy change in the public goods game, which may lead to contribute to the understanding of the players' behaviour and present a tool for economists to measure the amount of change for different setups of their experiments. Another objective is our aim to demonstrate the ability of the proposed method to produce quantifiable measures for items' changes in the temporal data and provide interpretable results. We compare the results and findings of our method with the MONIC method which is originally used to measure cluster changes in data streams [30].

For this experiment both datasets to measure players behaviour and strategy change during the consequent rounds of the game. Players contribution attribute and their expectation for other players contribution at each time point are used by this method to find the magnitude of the change. These two datasets have different groups of players and different lengths as the first dataset is 10 rounds length and the second is 27 so that these two datasets are used separately and treated as different datasets in this experiment. We used four clusters to cluster players in each time point using kmeans, PAM, cmeans and hierarchical clustering methods as it has been selected based on our discussion in section 3.1.

As both of datasets of public goods game share the same experiment settings and setup so that it can be hypothesised that the results of the behaviour change should be consistent with regards to the length of the experiment which might affect the behaviour of players [8]. While we use all previously selected external cluster validity index in section 3.2 and AUC of ROC, however, we will depend on AUC and Rand results to compare the behaviour of players in these two different datasets.

4.1 Using Proposed Method

Different types of reference of behaviours reveal different aspects of players' strategy change. By using first time point as the reference of behaviour, we can detect drift of players' behaviour from the initial expectation and contribution. As shown in figure 1 for 10 rounds dataset and figure 3 for 27 rounds, players in both datasets are gradually drifting away from their initial game plan and expectation. This trend can be seen with all four clustering methods with different measuring methods of external cluster validity index and AUC. Results of AUC and Rand are consistent across all clustering methods so that we used AUC to calculate the linear regression of the results. The negative results of linear regression is an indication that players change their behaviour by drifting away from their original gameplay.

Fig. 1 Results of various clustering methods using first time point as reference of behaviour to calculate the amount of changes which happen to the groups of items in consequent time points in the 10 rounds public goods game dataset.

By using the previous time point as the reference of behaviour we can measure the amount of change between any two consecutive time points. This allows detection of players' behaviour transition from one time point to another. Figure 2 of the 10 rounds dataset shows that players strategy changes from one time point to another is constant which indicated by the linear regression of AUC and Rand measures. In contrast figure 4 of 27 rounds shows that the change between time points are decreasing throughout the progress of the game.

Fig. 2 Results of various clustering methods using previous time point as reference of behaviour to calculate the amount of changes which happen to the groups of items in consequent time points in the 10 rounds public goods game dataset.

At first glance, the results of 10 and 27 rounds datasets are not consistent. However, with a closer look at the results, we can detect that the players' behaviour change in 27 rounds dataset is stable without any decrease until round 10 of the game. This decrease might be due to the fact that most of the players dropped their

contribution to zero when they reached a certain round of the game. Which means no further room is left for change in the game rounds, except some players randomly start to rise their contribution again but the rise is not constant so that after round 10 we detect less change than expected.

Fig. 3 Results of various clustering methods using first time point as reference of behaviour to calculate the amount of changes which happen to the groups of items in consequent time points in the 27 rounds public goods game dataset.

We hypothesised in the previous section that the players' behaviour has to be consistent in both datasets. The results for measuring changes using first time point as the reference of behaviour are compatible in both cases as it can be noticed that the players' contribution drops gradually. The results of using previous time point as the reference of behaviour show that players strategy change are constant till round 10. In 27 rounds dataset, after round 10 most players' contribution dropped to zero so that there is no room for further change in their strategy. Hence, their strategy

decreases and their game pattern starts to be similar between any two consequent time points.

Fig. 4 Results of various clustering methods using previous time point as reference of behaviour to calculate the amount of changes which happen to the groups of items in consequent time points in the 27 rounds public goods game dataset.

The results of the proposed method for both datasets are compatible with the findings of the economists [9, 11]. However, this method provides a tool which enables them to quantify changes in players behaviour. Quantifying behaviour change is important so that they can measure the nuanced differences between various gameplay setups like the length of the rounds, the percentage of the rewards from the public project, and knowing the identity of other players.

After players were classified according to their temporal attributes which reflect their contribution behaviour over all time points, we can use the new players' classes as a reference of behaviour. As shown in the Figures 5 and 6, there are signifi-

Fig. 5 Results of various clustering methods using players' classes as a reference of behaviour to calculate the amount of changes which happen to the groups of items in consequent time points in the test dataset. The amount of change is measured by using different external cluster validity index and AUC of ROC.

cant difference between players' classes and their temporal behaviour. This difference can be seen with the low value of the behavioural change measures across all clusterings and all different external cluster validity indices external cluster validity index (less than 0.6). This indicates that the players are not always playing the same strategy. Instead, they try and explore other strategies which contribute in their learning process to identify different strategy results. However, the regression of the behavioural change for all cases are small (near zero), which indicates the difference is stable throughout all time points. This is another indication that despite their temporary strategy change, however, these changes do not affect their general playing behaviour.

Despite the sensitivity difference between external cluster validity indices, all the results of different clusterings and external cluster validity indices are similar with regression slope equal to zero. This might be an indication that using items' overall general behaviour in the temporal attributes can create more stable estimation than other two reference of behaviours on the items' behavioural change. However, each reference of behaviour can be useful for situations. It can be noticed that different reference of behaviours might reveal different aspects of the players' behaviour. Us-

Fig. 6 Results of various clustering methods using players' classes as reference of behaviour to calculate the amount of changes which happen to the groups of items in consequent time points in the test dataset. The amount of change is measured by using different external cluster validity index and AUC of ROC.

ing the first time point as the reference of behaviour will demonstrate how items are deviating from their initial behaviour. Using the previous time point shows the stability of the items during different stages of the temporal data. While using players temporal classes as reference of behaviour demonstrates items behavioural variability in various stages related to their overall behaviour across all time points.

4.2 Comparing Results with MONIC

MONIC¹ is used to gain more insights about the public goods games data sets and to compare our results with the existing methods of measuring cluster changes in different time points. The data for each time period were clustered separately using k-means with four clusters. The clustering was carried out on the main temporal attributes of the data, namely belief and contribution, then the data and cluster labels

¹ Available at <http://infolab.cs.unipi.gr/people/ntoutsis/monic.html>

of items in each consequent pair of time points are fed to the MONIC algorithm to calculate changes to clusters from one time point to another. The method calculated the number of survived, appeared and disappeared clusters, as shown in figures 7 and 8, for the ten rounds of the game.

In the 10 rounds dataset, the number of survived clusters reduced from four clusters between the first and second time points until it reached zero, while new clusters appeared in the middle of the fifth and sixth game rounds, then the number rose again until the end of the game. This might be due to the fact that players are changing their strategies and exploring new options until they ultimately settle on a certain strategic pattern. This change is consistent with our findings, as the measures slightly increase between the fifth and seventh time points, which might be an indication of players changing their strategy back to their original ones. As Keser and Winden [22] suggest, this change might be due to the players responding to the average contribution of other players in the previous round.

Fig. 7 Number of survival, appearance and disappearance of clusters between every two consequent time points for ten rounds public goods game as measured by MONIC.

The results for the 27 rounds dataset is not straightforward, as the numbers of cluster survivals, appearances and disappearances change more frequently. However, the cyclic pattern of increasing and decreasing the number of survived clusters might be an effect of changing players' strategies or due to the underlying algorithm, as it provides an ageing factor to the items.

As the MONIC algorithm was originally introduced to detect cluster changes in a data stream, it uses an ageing factor which reduces the effect of older items in the cluster and removes items older than two time points [30]. This ageing factor is essential for the algorithm to keep up-to-date with the flowing data stream and give the right results for the current status of the clusters. However, this might not be useful for public goods games data, as there is a fixed number of players which might result in the removal of players who stay in the same cluster for long time points. The effect of the ageing might not be obvious in the 10 rounds game due to the limited number of time points, but it might undermine players' strategies.

Fig. 8 Number of survival, appearance and disappearance of clusters between every two consequent time points for 27 rounds public goods game as measured by MONIC.

While the proposed method assumes a fixed number of clusters to calculate items membership change, the MONIC algorithm is a good method to have insights on the available clusters and their stability by measuring the number of survived clusters between two time points. However it does not measure the amount of items drifting from one cluster into another, which can be detected by the proposed method, as it introduces a specific ratio between each consequent time point, indicating the amount of change happening to the items in the clusters by calculating their membership change among clusters.

MONIC can be compared with the proposed method especially with the case of previous time point as a reference of behaviour because both of these methods compare the current clusters with the previous time point. The regression result for the average of cluster moves (appear, disappear and survive) is near zero (-0.00547 for 27 rounds and 0.03889 for 10 rounds of the data set), which is compatible with the proposed method results using previous time point as reference of behaviour except for 27 rounds clustered by PAM and hierarchical clustering. By comparing results from the proposed method and MONIC we can conclude that the players slightly and gradually change their cluster membership. However, the change amount of change is stable from one time point to another. The proposed method gives an exact number for the change while the MONIC presents overall clusters movement and change.

5 Summary and Conclusions

This study presents a method to quantify changes over time for items across multiple readings at different times for the same attributes using external cluster validity measures and area under the curve of receiver operating characteristic. The proposed method consists of three steps: the first step separates individual time points from the temporal data, the second step clusters each time point so that items with

the similar behaviour can be grouped together. The third step quantifies changes of items' membership in the clusters using different reference of behaviours. Two data sets are used to determine behavioural changes of players in the public goods games.

Various clustering algorithms are used to cluster items at each time point. Then these clusters are compared with the the reference of behaviours using different external cluster validity indices like VI, FM and Jaccard. Three references of behaviours are used which they are first time point, previous time point and players overall behaviour as classes for their temporal attributes. We used a rule-based temporal classification method to classify players in both data sets.

The results are compared with the MONIC method which is designed to detect changes in the number of clusters in a data stream. The results were compatible with the findings using our proposed method especially with the results of previous time point as the reference of behaviour. All of the results indicate that the players' strategy are changing and the change is subtle and gradual.

Acknowledgements The authors record their thanks to Simon Gaechter and Felix Kolle in the School of Economics at the University of Nottingham for providing us with data from the public goods experiment and taking time to explain it to us.

References

- [1] Charu C. Aggarwal. On change diagnosis in evolving data streams. *Knowledge and Data Engineering, IEEE Transactions on*, 17(5):587–600, 2005.
- [2] Olatz Arbelaiz, Ibai Gurrutxaga, Javier Muguerza, Jesús M Pérez, and Iñigo Perona. An extensive comparative study of cluster validity indices. *Pattern Recognition*, 46:243–256, 2012.
- [3] Manuel Baena-Garcia, Jose del Campo-Avila, Raul Fidalgo, Albert Bifet, Ricard Gavaldà, and Rafael Morales-Bueno. Early Drift Detection Method. *4th ECML PKDD International Workshop on Knowledge Discovery from Data Streams*, pages 77–86, 2006.
- [4] Mirko Böttcher, Frank Höppner, and Myra Spiliopoulou. On exploiting the power of time in data mining. *ACM SIGKDD Explorations Newsletter*, 10(2):3–11, dec 2008.
- [5] Andrew P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, jul 1997.
- [6] R. Elwell and R. Polikar. Incremental Learning of Concept Drift in Nonstationary Environments. *IEEE Transactions on Neural Networks*, 22(10):1517–1531, 2011.
- [7] Polla Fattah, Uwe Aickelin, and Christian Wagner. Optimising Rule-Based Classification in Temporal Data. *Zanco Journal of Pure and Applied Sciences*, 28(2):135–146, 2016.

- [8] Charles Figuères, David Masclet, and Marc Willinger. Weak moral motivation leads to the decline of voluntary contributions. *Journal of Public Economic Theory*, 15(5):745–772, 2013.
- [9] Urs Fischbacher and Simon Gächter. Social Preferences, Beliefs, and the Dynamics of Free Riding in Public Goods Experiments. *American Economic Review*, 100(1):541–556, mar 2010.
- [10] Urs Fischbacher, Simon Gächter, Simone Quercia, and Simon Gächter. The behavioral validity of the strategy method in public good experiments. *Journal of Economic Psychology*, 33(4):897–913, aug 2012.
- [11] Urs Fischbacher, Simon Gächter, and Karina Whitehead. Heterogeneous Social Preferences and the Dynamics of Free Riding in Public Good Experiments about the Centre or contact. *The American economic review*, 100(1):541–556, 2010.
- [12] EB Fowlkes and CL Mallows. A method for comparing two hierarchical clusterings. *Journal of the American ...*, 78(383):553–569, 1983.
- [13] Joao Gama, Indre Zliobaite, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A Survey on Concept Drift Adaptation. *ACM Computing Surveys (CSUR)*, 46(4):44, 2014.
- [14] Roman Garnett and Stephen J Roberts. Learning from Data Streams with Concept Drift. *Technical Report PARG-08-01, Dept. of Engineering Science, University of Oxford*, 2008.
- [15] Stephan Günnemann, Hardy Kremer, Charlotte Laufkotter, and Thomas Seidl. Tracing evolving clusters by subspace and value similarity. *Advances in Knowledge Discovery and Data Mining*, 6635:444–456, 2011.
- [16] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. Cluster validity methods: part I. *ACM Sigmod Record*, 31(2):40–45, 2002.
- [17] Maayan Harel, Shie Mannor, Ran El-Yaniv, and Koby Crammer. Concept Drift Detection Through Resampling. *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1009–1017, 2014.
- [18] Basheer Hawwash and Olaf Olfa Nasraoui. Stream-dashboard: a framework for mining, tracking and validating clusters in a data stream. *Proceedings of the 1st International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications*, pages 109–117, 2012.
- [19] Anil K Jain. Data Clustering : 50 Years Beyond K-Means. *Pattern Recognition Letters*, 31:651–666, 2010.
- [20] Panos Kalnis, Nikos Mamoulis, and Spiridon Bakiras. On discovering moving clusters in spatio-temporal data. *Advances in Spatial and Temporal Databases*, 3633:364–381, 2005.
- [21] I Kaul, I Grunberg, and MA Stern. Global public goods. *Global public goods*, 1999.
- [22] Claudia Keser and Frans van Winden. Conditional Cooperation and Voluntary Contributions to Public Goods. *The Scandinavian Journal of Economics*, 102:23–39, 2000.

- [23] Marina Meil. Comparing clusteringsan information based distance. *Journal of Multivariate Analysis*, 98(5):873–895, may 2007.
- [24] Marina Meila. Comparing clusterings by the variation of information. *Learning theory and Kernel machines: 16th Annual Conference on Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 24-27, 2003: proceedings*, page 173, 2003.
- [25] Irene Ntoutsis, Myra Spiliopoulou, and Yannis Theodoridis. Tracing cluster transitions for different cluster types. *Control and Cybernetics*, 38(1):239–259, 2009.
- [26] Irene Ntoutsis, Myra Spiliopoulou, and Yannis Theodoridis. Summarizing Cluster Evolution in Dynamic Environments. In *Computational Science and Its Applications - ICCSA 2011*, volume 6783, pages 562–577. Springer Berlin Heidelberg, 2011.
- [27] Eréndira Rendón and Itzel Abundez. Internal versus External cluster validation indexes. *International Journal of computers and communications*, 5(1):27–34, 2011.
- [28] Mohammad Rezaei and Pasi Franti. Set matching measures for external cluster validity. *IEEE Transactions on Knowledge and Data Engineering*, 28(8):2173–2186, aug 2016.
- [29] Myra Spiliopoulou, Eirini Ntoutsis, Yannis Theodoridis, and Rene Schult. MONIC and Followups on Modeling and Monitoring Cluster Transitions. *Machine Learning and Knowledge Discovery in Databases*, 8190(2013):622–626, 2013.
- [30] Myra Spiliopoulou, Irene Ntoutsis, Yannis Theodoridis, and Rene Schult. Monic: modeling and monitoring cluster transitions. *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 706–711, 2006.
- [31] Alexey Tsymbal. The problem of concept drift: definitions and related work. *Computer Science Department, Trinity College Dublin*, 2004.
- [32] Lucas Vendramin, Ricardo JGB Campello, and Eduardo R Hruschka. Relative clustering validity criteria: A comparative overview. *Statistical Analysis and Data Mining*, 4(3):209–235, 2010.
- [33] Li Xiaofeng and Gao Weiwei. Study on a Classification Model of Data Stream based on Concept Drift. *International Journal of Multimedia and Ubiquitous Engineering*, 9(5):363–372, 2014.
- [34] Di Yang, Zhenyu Guo, Elke A Rundensteiner, and Matthew O Ward. CLUES: a unified framework supporting interactive exploration of density-based clusters in streams. *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 815–824, 2011.
- [35] Mohammed J Zaki and Meira Jr. Meira. *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press, New York, 2014.