2017 International Conference on Computing, Networking and Communications (ICNC): Social Computing and Semantic Data Mining

A New Mining Method to Detect Real Time Substance Use Events from Wearable Biosensor Data Stream

Jin Wang[1,2], Hua Fang[1*], Stephanie Carreiro[3], Honggang Wang[2], Edward Boyer[3]
[1]Department of Quantitative Health Science,
University of Massachusetts Medical School, Worcester, USA
Email: {chanpaul.wang, huajulia.fang}@umassmed.edu
[2]Department of Electrical and Computer Engineering, University of Massachusetts Dartmouth, North Dartmouth, MA, USA
E-mail: hwang1@umassd.edu
[3]Department of Emergency Medicine,University of Massachusetts Medical School, Worcester, USA
Email: {Stephanie.carreiro, Edward.boyer}@umassmed.edu

***Abstract*—Detecting real time substance use is a critical step for optimizing behavioral interventions to prevent drug abuse. Traditional methods based on self-reporting or urine screening are inefficient or intrusive for drug use detection, and inappropriate for timely interventions. For example, self-report suffers from distortion or recall bias; while urine screening often detects drug use that occurred only within the previous 72 hours. Methods for real-time substance use detection are severely underdeveloped, partly due to the novelty of wearable biosensor technique and the lack of substantive clinical data for evaluation. We propose a new real-time drug use event detection method using data obtained from wearable biosensors. Specifically, this method is built upon the slide window technique to process the data stream, and a distance-based outlier detection method to identify substance use events. This novel method is designed to examine how to detect and set up the thresholds of parameters in real-time drug use event detection for wearable biosensor data streams. Our numerical analyses empirically identified the thresholds of parameters used to detect the cocaine use and showed that this proposed method could be adapted to detect other substance use events.**

***Index Terms*—Wearable biosensor; Data stream; Data Mining; Substance Use, Behavioral Intervention**

I. Introduction

According to the Substance Abuse and Mental Health Services Administration (SAMHSA), about *23.5* million persons aged 12 or older received treatment for an illicit drug or alcohol abuse problem in 2009 [1]. National Institute on Drug Abuse reports that, the use of tobacco, alcohol, prescription, and illicit drugs annually costs approximately $137 billion in direct health care expenses [2], [3]. Wearable biosensor systems are gaining increased attention in behavioral interventions for substance use due to their potential advantages in delivering near real-time treatments and their cost-effective nature in the long run.

Common methods for detecting drug use include self-report and qualitative screenings of urine for drugs of abuse (toxic screen). Drug users may under- or over-report their drug use and suffer from their recall bias [4], [5]. The urine screening is restricted by their detection time window (e.g., 72 hours for cocaine use). [6] Recently, a wearable biosensor system called iMStrong has demonstrated its ability to detectcocaine use in natural environments in clinical trials. Nevertheless, the real-time drug use event detection method remains underdeveloped.

The ultimate goal of this investigation was to develop a novel real-time drug use event detection method. Our method is built upon the data streams of surrogate markers of the Sympathetic Nervous System (abbr. SNS) activity [7] stimulated by cocaine use. Specifically, cocaine produces a catecholamine surge characterized by an abrupt increase of electrodermal activity and abnormal locomotion changewhile simultaneously causing a decrease in skin temperature due to vasoconstriction. *iMStrong* [6] used this causal relationship between cocaine use and SNS response to detect cocaine use events. Different from imStrong [6], this paper attempts to design an algorithm to automatically detect drug use events in a real-time manner as long as the drug use behavior reacts. The main contributions of this work are as follows: 1) designed a novel slide-window based outlier detection model based on the real-time data streams from the surrogate markers of SNS activity collected through portable biosensors, and automatically identify drug use events; 2) compared with the off-line methods, our method is resource friendly, as it only stores the recent data of surrogate markers of SNS activity, instead of all the data; 3) designed a comprehensive method to evaluate and select the thresholds of parameters for drug use event detection. We briefly introduced this work in Joint Statistical Meetings, 2016 [8].

The remainder of this paper is organized as follows: Section II presents related work; Section III demonstrates the design of our method, and Section IV illustrates our numerical analyses and evaluation results; Section V concludes the paper.

II. Related Work

The most commonly used drug use event detection methods include self-report and urine or blood screening. Yet, as

978-1-5090-4588-4/17/$31.00 ©2017 IEEE

demonstrated in §I, these traditional methods are inefficient in drug use event detection. Recent research adoptes wearable devices, such as mobile phone and portable biosensors, to help drug use event detection. Below introduces these mobile device based drug use event detection methods, the background of outlier analysis and slide window techniques for data stream processing.

A. Mobile Device Based Drug Use Event Detection

The existing mobile device based drug use event detection mainly focused on collecting data from real-time self-reports and biomedical surrogate markers. Ecological Momentary assessment (EMA) [9], *iHeal* [10], and *iMStrong* [6], [11] represent typical analytic methods for such data. EMA [9] is a questionnaire-based self-report method, using mobile devices to deliver timely questionnaires to patients and receive their feedbacks, but the analyses of these feedbacks are mostly off-line. Since EMA is a self-report based method, it may also have the same drawbacks as typical self-reports, such as misreports from patients. *iHeal* [10] is a biomedical system incorporating artificial intelligence, continuous biophysical monitoring, wireless connectivity, and smart-phone computation. However, its proposed dynamic Bayesian network (DBN) analyses can face real-time computational challenges if used in real time data stream processing. For example, a large number of training data may challenge its use in practice, particularly for real-time data processing. *iMStrong* collects real time data streams from biomedical surrogate markers of SNS but the real time drug use algorithm is still in development, although our preliminary parameter description method demonstrated the substantive usability of wearable biosensors for cocaine use detection [6]. All these methods are currently used off-line, and there is still a gap in pursuing near-real time drug use behavioral intervention. Based on the SNS surrogate markers used by *iMStrong* [6], this work designed and evaluated a real-time substance use event detection method, called slide-window based outlier detection method.

B. Background of Outlier Analysis and Slide Window Technique for Data Stream

The outlier analysis aims at identifying the rare data points that significantly deviate from the majority of a dataset. It has been widely used in wireless networks. Branch et al. [12] designed slide-window based in-network outlier detection for wireless sensor network. Li et al. [13] used outlier analyses to identify malicious attacker in Mobile Ad-hoc Network. Most outlier analyses focused on static data stored locally or distributed [14]. The data stream refers to the data arriving dynamically and unbounded, such as those from biosensors. The sliding window technique is usually used to process data stream. An illustration of sliding window is described in Fig. 1. The slide window consists of three parts: Window start, window width (*W*), and slide length (*SL*). For the count-based window, the window width denotes the number of data points in the current window, and the slide length equals the number of data points expiring as the window moves forward each time.

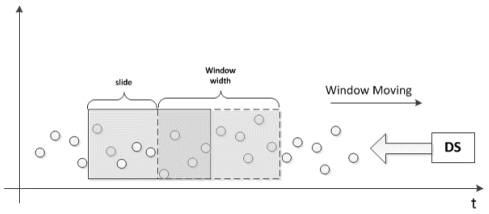


Fig. 1: Demonstration of sliding window

In outlier analyses, the mostly used method is called the distance-based outlier model, which is defined as: Assuming a dataset $\bar{X} \subset R^{N\times n}$, and $Neig^R(x_i) = \{x_j \in \bar{X} | dist(x_i, x_j) \leq R\}$ as data points in the neighborhood of x_i with radius *R*, data point $x_i \in \bar{X}$ is an outlier if $|Neig^R(x_i)| \leq k$, namely there are less than *k* data points in the neighborhood of x_i with radius *R* [15]. Outlier analysis on data stream is described as: For each data point x_i in the current window, find $Neig^R(x_i)$, and check whether $|Neig^R(x_i)| \leq k$. Angiulli et al. [16] designed two outlier query algorithms for data streams, exact query and approximate query. The exact algorithm maintained $|Neig^R(x_i)|$ for each data point in each window. Its computational complexity is $O(logN)$, and memory complexity is $O(N)$. The approximate algorithm assumed the nearest neighbors of each data point uniformly distributed along the whole window, and only kept part of safe inliers which would not be outlier until expiring. C-Abstract [17] is a neighbor-ship-based data stream pattern discovering method. It maintained $|Neig^R(x_i)|$ for each data point in each window based on the lifetime of data points, and alleviated the computation for updating $Neig^R(x_i)$ in different windows. The most recent work, LEAP [18] and COD [19], further reduces the computation of updating $Neig^R(x_i)$ in each window. LEAP split the window into different slices, and established $Neig^R(x_i)$ from the most recent slice. COD [19] uses an event queue to record possible time when the outlying status of data point may change, avoiding repetitive outlying status checking. However, how to set up the optimal parameters and their combination, such as detecting window size *W*, neighborhood radius *R*, and threshold *k*, are still unknown in different application areas. These parameters depend on specific scenarios. Our work here is the first to study how to detect and set up the thresholds of these parameters in real-time drug use event detection in wearable biosensor systems.

III. Real-time Drug Use Event Detection Algorithm

This proposed method uses three surrogate markers of SNS activity triggered by cocaine use: Electro dermal Activity (EDA), skin temperature, and locomotion (i.e., up-down, back-forth, and left-right movements). Specifically, we identify the time window of drug use events based on these dynamic surrogate markers, and illustrate the real-time substance use

event detection using slide-window based outlier analyses. Notations are displayed in Table I.

TABLE I: Notations

Notation	Description
a_i	The i^{th} attribute
X^i	The i^{th} patient
X_t^i	Observation of the i^{th} patient at time *t*
d	The number of attributes
Y_k^i	Substance use status of the i^{th} patient at the k_{th} window
W	Window width
SL	Slide length
S_i	The i^{th} slide
start	The beginning of window
$Neig^R(X_t^i)$	The neighbors of data instance X_t^i in its neighborhood with radius R
R	Radius of neighborhood
k	The threshold number of data points in neighborhood with radius R

Assuming the dynamic surrogate markers of SNS activity from patient i in the t^{th} window as $\bar{X}_t^i = \{X_t^i, X_{t+1}^i, \cdots, X_{t+W}^i\}$, $X_t^i = \{a_1, a_2, \cdots, a_d\}$ as the observation at time t, and $a_1, a_2, \cdots, a_d$ as the markers of SNS activity, i.e. skin temperature, Electra dermal activity, and locomotion, our real-time drug use event detection model is described as Eq. 1.

$$Y_t^i = f(\bar{X}_t^i | W, R, k) \tag{1}$$

in which Y_t^i, a vector with length of W, denotes the drug use status of patient i in the t_{th} window. Each element of Y_t^i denotes the drug use status of a time period. Given the parameters, LEAP [18] method was used to establish $Neig^R(X_t^i)$, which denotes the set of nearest neighbors of X_t^i. If $|Neig^R(X_t^i)| < k$, we regard X_t^i as an outlier, namely a potential drug use event. Besides, in order to efficiently maintain $Neig^R(X_t^i)$, we split the current data window into several slides with equal-size, and manage $Neig^R(X_t^i)$ at each slide, instead of updating $Neig^R(X_t^i)$ when each data point arrives. In particular, with different combinations of detecting window size W, neighborhood radius R, and the threshold number of data points k, we could obtain different detections. Our challenge and goal is to obtain the optimal parameter set and achieve more true positives and less false positives. Our parameter selection algorithm is based on the outlier ratio (α), defined as follows:

$$\alpha(W, R, k) = \frac{\#Outliers\ detected}{\#Total\ data\ points} \tag{2}$$

The outlier ratio characterizes how many data points are identified as outliers. With the worst parameter combination, all data points are regarded as outliers, and outlier ratio is 1. Otherwise, the outlier ratio would gradually decrease with better parameter combinations. Since outliers, ie., cocaine use events, are usually at a small number relative to the gigantic data streams from three surrogate markers at milliseconds, we hypothesize that the optimal parameter combinations correspond to a abrupt change point in the range of outlier ratios. With the metric outlier ratio α, we can obtain these parameters exploiting cross-validation method.

IV. Performance Evaluation

A. Evaluation Settings

This proposed slide-window based outlier detection method was implemented on Spark [20], which is capable of processing large data streams. Real cocaine use data streams generated from iMStrong were used to evaluate this method [6]. These data were collected through wearable biosensors from 15 participants, with 30 data samples per second for each patient. To be consistent and cross-validate with Ref. [6], we used the same patients as shown in Tab. II, whose cocaine use events were validated with urine screen and their self-reports. Six data attributes were used: EDA, skin temperature, and three locomotion dimensions: up-down, left-right, and back-forth. Our evaluation metrics or criteria include outlier ratio and detection accuracy. Compared to Ref. [6], we used finer time granularity (e.g., minute). The detection accuracy indicates whether our method can catch the drug use events confirmed by self-reports and/or urine screen. The evaluation is performed on Windows 7 platform with *Intel(R) Core(TM) i7-3520M 2.9 GHz CPU* and 8 *G* memory.

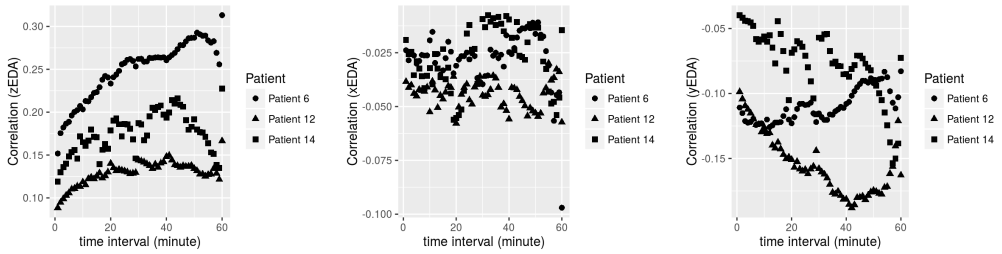
TABLE II: Dataset Description

Patient	Duration	Self-Report	Urine Screen	Drug Use Date
Patient 1	26 days	No	No	2014-03-08
Patient 12	34 days	Yes	No	2014-05-20 2014-05-23
Patient 14	41 days	No	Yes	2014-05-28

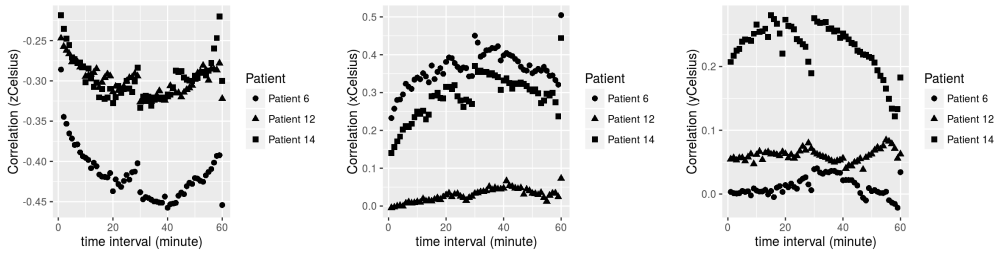
B. Results

It has been proven that cocaine use causes abruptly increase of EDA, dropping of skin temperature and obvious body locomotion changes. Although we collected up-down, left-right, and back-forth locomotions, not all are necessary in detecting cocaine use events based on our two separate empirical studies [6], [21]. We first use the correlation analysis to find out how much locomotion is correlated with cocaine use, and which locomotion parameter is the most sensitive.

The correlation magnitude and direction of locomotion with EDA and skin temperature in different time granularities are shown in Fig. 2. These figures further indicate that the up-down locomotion is more sensitive for cocaine use, while other two locomotion parameters (i.e., left-right and back-forth) are not sensitive for cocaine use over time. The correlation of up-down locomotion with EDA and skin temperature is much stronger when the time granularity is about 30 minutes. This is because the surrogate markers of SNS activity changes triggered by cocaine use do not happen immediately after drug use. Larger time granularity may smooth out abnormal surrogate markers of SNS activity, and cause false negatives. Otherwise, smaller time granularity could divide abnormal surrogate markers of SNS activity into different time episodes, leading to possible false positives. The optimal time granularity should cover multiple surrogate markers of SNS activity



(a) Correlation of locomotion with EDA



(b) Correlation of locomotion with skin temperature

Fig. 2: Sensitivity of locomotion with other attributes

response within the same time episode. Fig. 2 demonstrates that the time granularity of 30 minutes could almost cover all three attributes, up-down locomotion, EDA, and skin temperature simultaneously, and seems more appropriate for the cocaine use event detection. Given the time granularity of 30 minutes, we further used EDA, skin temperature, and up-down locomotion to detect cocaine in the following evaluation scenarios.

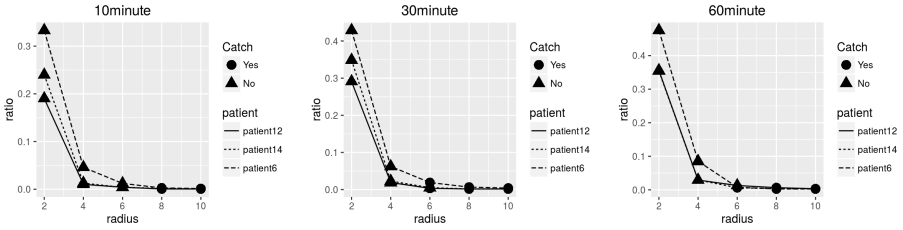
With the increase of the neighborhood radius, more data points should be covered by the neighborhood. If the parameter k, denoting the threshold number of data points in neighborhood with R, is too small, positive drug use events could be overlooked. Fig. 3 demonstrates such trend of outlier rates with k. For example, given a radius of 8, the proposed method can catch the drug use event with k of 4, yet not with k of 2. Besides, with the increase of the neighborhood radius, the outlier rate drops, namely false positives might be reduced. Especially, the outlier rate drops dramatically when R is smaller than 4, yet mild when R is larger than 4. At the radius of 4, our method can catch all drug use events. Thus, the radius of 4 seems to be appropriate for cocaine use event detection.

Then, given a fixed neighborhood radius R, the influence of k was examined. Of note, k is always less than the window width. With a fixed radius, larger k would potentially cause more false positives. As demonstrated in Fig. 4, all patients have some common pattern, namely, the outlier rate increases with the increase of k. Especially, with the window size W of 9, the outlier rate has a change point at k of 6, and the outlier rate difference between k of 4 and 6 is small. The k of 4 seems to be sufficient for cocaine use event detection with the varying length of slide windows.

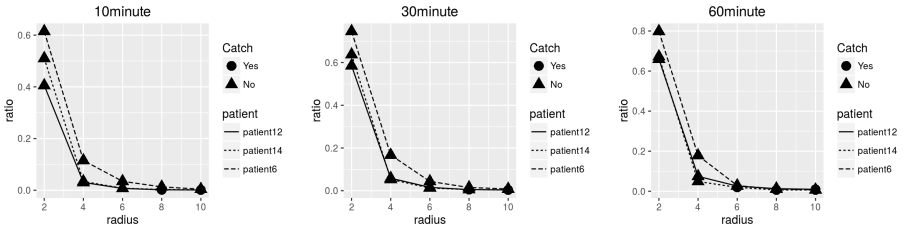
Furthermore, with the fixed neighborhood radius R of 4 and k of 4, Fig. 5 shows the variation of window width W. Large window size is supposed to reduce the false positive. Yet, larger window size also needs more computational resources. As shown in Fig. 5, the outlier rate drops significantly before W of 9, and then the outlier rate decrease mildly. So, the window size W of 9 seem to be sufficient for the cocaine use event detection. Moreover, as shown in Fig. 3, 4, and 5, by and large, the outlier rate does have a change point when window size W, neighborhood radius R, and parameter k vary. Compared with iMStrong [6], our proposed method can substantially help identify drug use events in a near-real time manner (e.g., hourly) given the elapsed time of drug users' physiological reaction to the drug, instead of conducting off-line EMA analyses of real-timely-collected data.

V. Conclusions and Future Work

This paper proposed a slide-window based outlier detection model for real-time drug use event detection and parameter selection. Unlike off-line analyses, the proposed model is designed for real-time drug use detection and could accurately catch the cocaine use events within about 30 minutes, as demonstrated with real cocaine use data streams. In fact, drug use detection accounts for two time periods: the time for collecting and averaging data over a time window and the time for outlier analyses. Since outlier analyses for each data episode are usually at the time magnitude of μs, drug use detection time equals the time granularity. Our comprehensive evaluation shows that the neighborhood radius (R) of 4 the data points in a neighborhood with a given radius (k) of 4, and the window size (W) of 9 are found to be more suitable in detecting cocaine use events. Besides, the up-down locomotion is more sensitive with EDA and skin temperature in cocaine use event detection. This model is potentially generalizable to other drug use event detection. In the future, we will test this algorithm on more real patients and would use emulation to

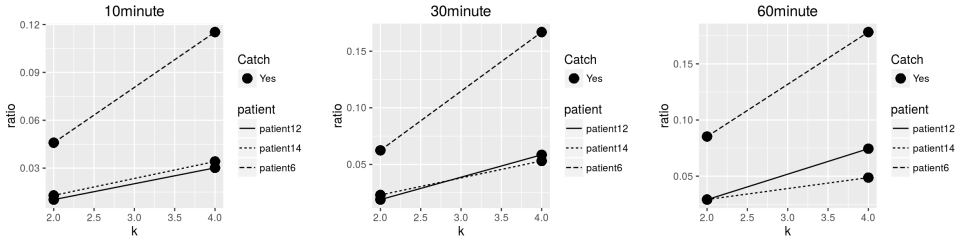


(a) W=6, K=2

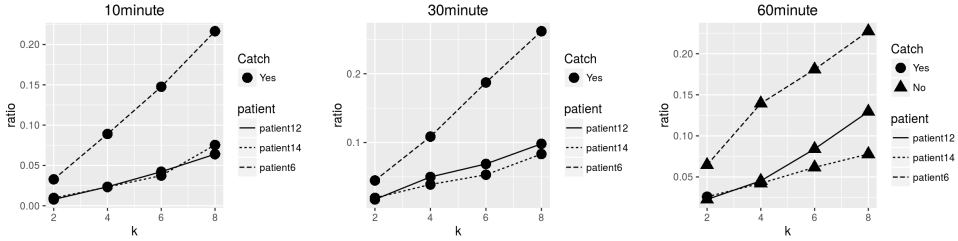


(b) W=6, K=4

Fig. 3: Selection of radius



(a) W=6, R=4



(b) W=9, R=4

Fig. 4: Selection of *k*

test our algorithms on our wearable biosensors, and integrate our model in our existing iMStrong system.

ACKNOWLEDGMENT

This research was supported by NIH grant RO1DA033323-01A1, 1UL1RR031982-01 Pilot Project to Dr. Fang, NIH R01DA033769-01 to Dr. Boyer, NIH KL2TR001454 to Dr. Carreiro, and NSF CCSS1407882, IIS1401711 and 1429120 to Dr. Wang.

REFERENCES

[1] [Online]. Available: https://www.drugabuse.gov/publications/drugfacts/treatment-statistics on May 31 2016

[2] [Online]. Available: http://drugabuse.com/library /drug-abuse-statistics/

[3] H. Fang, Z. Zhang *et al.*, "A survey on big data research," *IEEE Network Magazine*, vol. 29, no. 5, pp. 6–9, 2015.

[4] H. Fang, C. Johnson *et al.*, "A new look at quantifying tobacco exposure during pregnancy using fuzzy clustering," *Neurotoxicol Teratol*, vol. 33, no. 1, pp. 155–65, 2011.

[5] H. Fang, V. Dukic *et al.*, "Detecting graded exposure effects: A report on an east boston pregnancy cohort," *Nicotine Tob Res*, vol. 14, no. 9, pp. 1115–20, 2012.

[6] S. Carreiro, H. Fang *et al.*, "imstrong: Deployment of a biosensor system to detect cocaine use," *Journal of Medical System*, vol. 39, no. 12, p. 186, 2015.

[7] W. Vongpatanasin, Y. Mansour *et al.*, "Cocaine stimulates the human cardiovascular system via a central mechanism of action," *Circulation*, vol. 100, no. 5, pp. 497–502, 1999.

[8] J. Wang, H. Fang, S. Carreiro *et al.*, "Detecting real-time substance use from wearable biosensor data stream," *Section on Statistical Learning, Joint Statistical Meetings, American Statistical Association*, 2016.

[9] B. S. Linas, A. GenzE *et al.*, "Ecological momentary assessment of illicit drug use compared to biological and self-reported methods," *JMIR Mhealth Uhealth.*, vol. 4, no. 1, 2016.

[10] E. Boyer, R. Fletcher *et al.*, "Preliminary efforts directed toward the

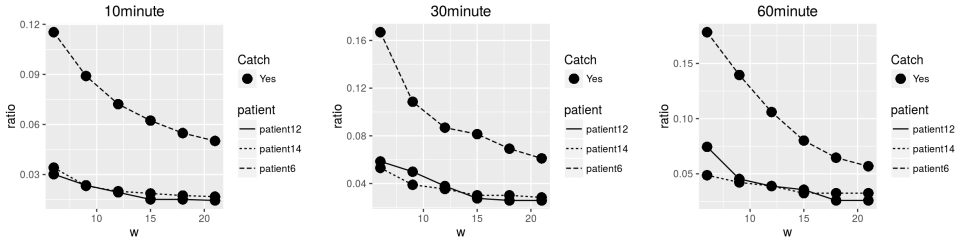


Fig. 5: Selection of window size with k=4, R=4

detection of craving of illicit substances: the iheal project," *Journal of Medical Toxicol*, vol. 8, pp. 5–9, 2012.

[11] S. Carreiro, D. Smelson *et al.*, "Real-time mobile detection of drug use with wearable biosensors: A pilot study," *Journal of Medical Toxicology*, vol. 11, pp. 73–79, 2015.

[12] J. Branch, B. Szymanski *et al.*, "In-network outlier detection in wireless sensor networks," in *2006 26th International Conference on Distributed Computing System*, 2006.

[13] W. Li and A. Joshi, "Outlier detection in ad hoc networks using dempster-shafer theory," in *10th International Conference on Mobile Data Management*, Taipei, Taiwan, May18–20 2009.

[14] C. C. Aggarwal, *Outlier Analysis*. New York: Springer, 2013.

[15] D. Georgiadis, M. Kontaki *et al.*, "Continuous outlier detection in data streams: An extensible framework and state-of-the-art algorithms," in *SIGMOD'13*, New York, New York, USA, 22–27, 2013.

[16] F. Angiulli and F. Fassetti, "Distance-based outlier queries in data streams: the novel task and algorithms," *Data Mining and Knowledge Discovery*, vol. 20, no. 2, pp. 290–324, 2010.

[17] D. Yang, E. Rundensteiner, and M. Ward, "Neighbor-based pattern detection for windows over streaming data," in *2009 EDBT*, 2009, pp. 529–540.

[18] L. Cao, D. Yang *et al.*, "Scalable outlier detection over high-volume data streams," in *2014 IEEE 30th International Conference on Data Engineering*, Chicago, IL, USA, 31– April 27, 2014.

[19] K. M. Papadopoulos, A. N. Manolopoulos *et al.*, "Efficient and flexible algorithms for monitoring distance-based outliers over data streams," *Information Systems*, vol. 55, no. 2, pp. 37–53, 2016.

[20] [Online]. Available: https://databricks.com/spark/about

[21] S. Carreiro, K. Wittbold *et al.*, "Wearable biosensors to detect physiologic change during opioid use," *J Med Toxicol*, 2016.