NONLINEAR AUDIO RECURRENCE ANALYSIS WITH APPLICATION TO GENRE CLASSIFICATION

Joan Serrà[*], *Carlos A. de los Santos, and Ralph G. Andrzejak*[†]

Universitat Pompeu Fabra
Dept. of Information and Communication Technologies
Roc Boronat 138, 08018 Barcelona, Spain

ABSTRACT

In this paper we apply nonlinear signal analysis to a music information retrieval task. More concretely, we apply the concept of recurrence plots and recurrence histograms to extract information from music audio frames. We evaluate the effectiveness of this approach with a typical genre classification framework and compare it against a baseline obtained from standard spectrum-based descriptors. The accuracy reached by the histogram-based descriptors alone does not surpass the one achieved by the spectral-based descriptors. However, we show that the combination of both descriptor sources results in consistent improvements up to 5 absolute percent points. This highlights the potential of nonlinear signal analysis for quantitative music description. In particular, it suggests that the information resulting from this approach is complementary to the information obtained through the commonly used spectral representation.

Index Terms— Music Information Retrieval, Nonlinear Time Series Analysis, Audio Recurrence, Descriptor Extraction

1. INTRODUCTION

The processing of music audio signals is key in music information retrieval (MIR) and has already led to a variety of real-world applications [1]. Many such applications are build up from low-level features of the musical content extracted from audio. These low-level features are commonly called music descriptors and, in general, are obtained from the short-time Fourier transform (STFT) [2]. For the calculation of the STFT, the audio signal is cut into short overlapping frames, a windowing function is applied to each frame, and the magnitude of the spectrum is taken. For many applications, music descriptors are fed to data mining or machine learning algorithms, which exploit the information contained in the descriptors [1, 3]. A paradigmatic example of this approach is automatic genre classification [4].

Audio content-based MIR approaches, and in particular the genre classification ones, often achieve relatively good accuracy values [3]. However, they do not reach the highest possible ones. This fact is commonly called the "glass-ceiling" phenomenon [5]. Therefore, for tackling this issue and achieving accuracies that go beyond the current "glass ceiling", alternative or complementary strategies shall be considered. In this paper, we focus on the extraction of music descriptors. More concretely, we propose the use of two types of nonlinear recurrence histograms instead of the routinely employed magnitude spectrum. While both representations are obtained from the frame signal, the crucial difference is that recurrence histograms are based on techniques from nonlinear signal analysis [6]. There is growing evidence that nonlinear signal analysis can provide useful information in the context of MIR, both for the processing of audio signals and the processing of descriptor sequences (see e.g. [7] and references therein). In particular, Terez reported that nonlinear recurrence histograms can be used for robust pitch determination [8].

Given an experimental signal, the framework of nonlinear signal analysis allows characterizing its underlying dynamics. A fundamental concept in this framework are delay coordinates [6], which can be used to reconstruct an estimate of these dynamics from the experimental signal. In this reconstruction, each state of the dynamics is represented unambiguously by one point in a multidimensional space. Therefore, to assess the similarity between different states attained at different times, one exploits the spatial proximity between points. This is formalized by so-called recurrence plots [9], where a threshold value is applied to the distance between individual points to decide whether or not two states represent a recurrence of the dynamics. Histograms of time lags between recurrences can then be used to assess the predominant time scales of the dynamics.

Importantly, with the aforementioned analysis, recurrences are taken into account regardless of whether the intervals between them are all equal or not. Accordingly, no periodicity is required for these recurrences. Furthermore, the signals can comprise different types of complicated and irregular waveforms that recur in a non-periodic way. These features can readily be detected in recurrence plots but might have a broad-band signature in the power spectrum. Moreover, for the calculation of recurrence plots, the dynamics is not assumed to be stationary [9]. Hence, recurrence plots can assess complex features that might not be discernible from the power spectrum of the signal. Music audio signals often comprise such complex features that recur in an irregular way. Therefore, recurrence plots, and nonlinear signal analysis tools in general, seem as a promising tool to characterize these signals.

2. NONLINEAR AUDIO RECURRENCE HISTOGRAMS

2.1. Delay coordinates

Let the signal $\mathbf{s} = [s_1, \dots s_{N^*}]^{\mathrm{T}}$, where $^{\mathrm{T}}$ denotes transposition, be the N^* consecutive audio samples of a given frame, obtained at a rate of F_{s}. We construct delay coordinates $\mathbf{x}_n$ by

$$\mathbf{x}_n = \left[s_n, s_{n-\tau}, \dots s_{n-(m-1)\tau}\right]^{\mathrm{T}}, \tag{1}$$

where m is the embedding dimension and τ is the time delay [6]. For $n = (m-1)\tau + 1, \dots N^*$, this yields a sequence of delay vec-

[*] Contact: `joan.serraj@upf.edu`. Partially funded by the Music 3.0 (TSI-070100-2008-318) and Buscamedia (CEN-20091026) projects.

[†] Funded by the Spanish Ministry of Education and Science (BFU2007-61710).

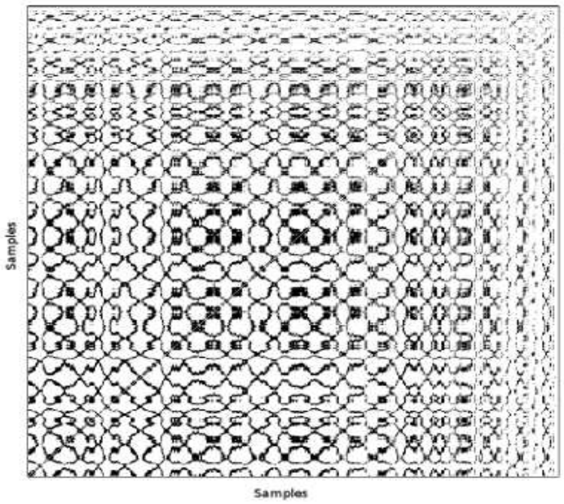

Fig. 1. Example of a recurrence plot $\mathcal{R}$ obtained from an audio frame of a blues piece. Parameters $m = 3$, $\tau = 20$, and $p = 0.3$ were used. According to the standard representation [9], $r_{1,1}$ is located in the lower-left corner.

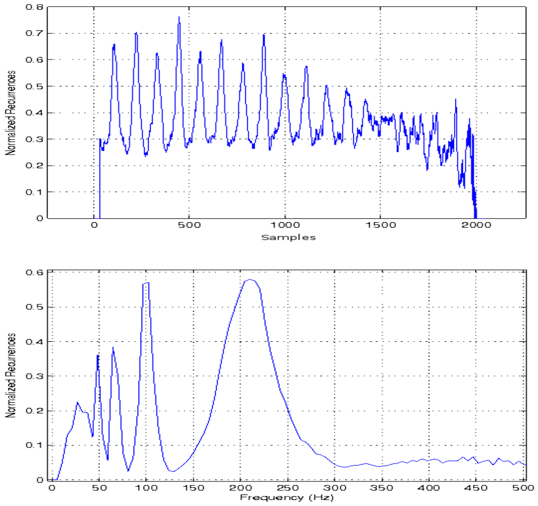

Fig. 2. Examples of $\mathbf{h}^{(\mathrm{t})}$ (top) and $\mathbf{h}^{(\mathrm{f})}$ (bottom). The bottom histogram $h^{(\mathrm{f})}$ is zoomed in the low frequency range. Same frame and parameters as in Fig. 1 were used.

tors $\mathcal{X} = [\mathbf{x}_1 \cdots \mathbf{x}_N]^{\mathrm{T}}$, where $N = N^* - (m-1)\tau$. Notice that for $m = 1$, Eq. (1) reduces $\mathcal{X}$ to the raw signal **s**. For nonlinear signal analysis, an appropriate choice of m and τ is crucial to extract meaningful information from noisy signals of finite length. Recipes for the estimation of optimal fixed values of m and τ exist [6]. However, we here opt to set them manually in order to study their effect on the final accuracies.

2.2. Recurrence plot

To assess the recurrences found in $\mathcal{X}$, we construct a recurrence plot $\mathcal{R}$ by applying

$$r_{i,j} = \Theta\left(\varepsilon - \|\mathbf{x}_i - \mathbf{x}_j\|\right), \tag{2}$$

for $i, j = 1, \ldots N$, where $\Theta(\cdot)$ is the Heaviside step function [$\Theta(z) = 0$ if $z < 0$ and $\Theta(z) = 1$ otherwise], ε is a threshold distance, and $\|\cdot\|$ is some norm [9]. We use the squared Euclidean norm and set ε to a percentage p of the mean over all possible $N(N-1)/2$ pairwise distances between samples of $\mathcal{X}$. An example of $\mathcal{R}$ can be seen in Fig. 1.

2.3. Recurrence time and frequency histograms

To quantify sample lags between recurrences we construct a normalized recurrence histogram $\mathbf{h}^{(\mathrm{t})} = [h_1^{(\mathrm{t})}, \ldots h_{N-1}^{(\mathrm{t})}]^{\mathrm{T}}$. The value for the k-th bin, representing the relative amount of recurrences at a time lag k, is calculated by

$$h_k^{(\mathrm{t})} = \frac{1}{N-k} \sum_{i=1}^{N-k} r_{i,i+k}, \tag{3}$$

for $w < k < N$. For $k \leq w$, we set $h_k^{(\mathrm{t})} = 0$. Here we follow the common practice in nonlinear signal analysis and consider only recurrences separated by a minimal sample lag w [6, 9]. Samples that are separated in time by only very short sample lags are very likely to be still spatially closer than ε. However, these close samples do not reflect a recurrence because not enough time has passed for the states to separate. In particular, the above holds for correlated signals. We therefore use a minimal sample lag $w = 30$ for all experiments (i.e. $30/F_s$ sec).

Given the recurrence time histogram $\mathbf{h}^{(\mathrm{t})}$, it is straightforward to convert the abscissa into a frequency axis, resulting in a recurrence frequency histogram $\mathbf{h}^{(\mathrm{f})}$. In particular, the following bin conversion can be used for a sample lag k:

$$k' = \underset{1 \leq l \leq N}{\arg\min} \left(\left| f_l - \frac{F_{\mathrm{s}}}{k} \right| \right), \tag{4}$$

where f_l is the central frequency for the l-th bin of the histogram. We use linearly spaced bins such that $f_l = (l-1)F_s/2N^*$. For details of the implementation and the normalization procedure for $\mathbf{h}^{(\mathrm{f})}$ we refer to [10]. Examples of $\mathbf{h}^{(\mathrm{t})}$ and $\mathbf{h}^{(\mathrm{f})}$ can be seen in Fig. 2.

3. EVALUATION METHODOLOGY

To test the capacity of recurrence time and frequency histograms to assess musical characteristics we use a common MIR evaluation framework. In particular, we use a training and testing genre classification framework [3, 4]. As music material we use the collection provided by Tzanetakis [11], which is used in a number of works for evaluating genre classification. The collection is divided into 10 genres of 100 songs each (except the reggae genre, which has 93 songs). This amounts a total of 993 audio files. These files are in WAV format and have a sampling rate $F_{\mathrm{s}} = 22050$ Hz.

In our evaluation framework, the audio is first cut into short overlapping frames. We use a frame length of $N^* = 2048$ and 50% overlap. Subsequently, the frame signal serves as input for the recurrence histogram procedures explained above. To compare with the common methodology, we also test the extraction of music descriptors from the magnitude spectrum. To calculate the STFT we use a 92 dB Blackman-Harris window [2], a frame length of $N^* = 4096$ with 50% overlap, and take the 2048 positive frequencies of the resulting magnitude spectra.

To extract information from the recurrence time and frequency histograms, as well as from the magnitude spectra, we use standard

Row	Descriptor set	Parameters			Classifiers							
		m	τ	p	1R	NN	NB	MP	RF	SVM$_{\mathrm{P}}$	SVM$_{\mathrm{R}}$	LL
1	Spectral	-	-	-	23.5	56.4	50.4	62.0	62.3	63.3	66.0	66.0
2	$\mathbf{h}^{(t)}$	3	7	0.7	24.5	49.5	49.9	51.7	59.4	54.9	57.5	56.5
3	$\mathbf{h}^{(t)}$	12	7	0.7	23.0	47.6	45.9	53.4	56.8	55.9	57.5	57.0
4	$\mathbf{h}^{(f)}$	3	1	0.3	21.0	42.8	40.2	47.0	50.6	45.2	48.7	50.3
5	$\mathbf{h}^{(f)}$	3	1	0.7	21.2	44.4	41.8	46.8	51.6	47.3	51.2	50.7
6	$\mathbf{h}^{(t)}$ + $\mathbf{h}^{(f)}$	12	3	0.7	25.7	48.4	47.4	52.5	57.6	54.0	55.5	57.3
7	$\mathbf{h}^{(t)}$ + $\mathbf{h}^{(f)}$	12	7	0.7	23.5	49.8	49.1	53.8	59.3	56.0	58.5	58.0
8	Spectral + $\mathbf{h}^{(t)}$	3	7	0.7	24.9	60.0	57.6	65.9	67.1	66.0	69.7	68.7
9	Spectral + $\mathbf{h}^{(t)}$	12	7	0.7	**26.0**	**62.2**	**57.8**	**67.7**	**69.0**	**68.5**	**70.9**	**70.5**
10	Spectral + $\mathbf{h}^{(f)}$	3	1	0.7	25.8	58.4	53.0	62.3	64.2	63.7	67.1	66.0
11	Spectral + $\mathbf{h}^{(f)}$	7	1	0.3	**26.1**	57.7	52.9	62.2	64.3	63.0	66.5	66.7
12	Spectral + $\mathbf{h}^{(t)}$ + $\mathbf{h}^{(f)}$	3	7	0.7	**26.0**	**61.1**	**58.4**	**66.0**	**68.1**	**67.4**	**70.7**	**70.1**
13	Spectral + $\mathbf{h}^{(t)}$ + $\mathbf{h}^{(f)}$	12	7	0.7	**26.3**	**63.8**	**58.6**	**66.6**	**68.6**	**68.9**	**71.5**	**69.8**

Table 1. Summary of the best classification accuracies for a given set of extracted features and parameter combination. A window $w = 30$ is used in all cases. The maximal standard deviation across trials for all classifiers was found to be 1.2%. The maximal accuracy we achieved by randomly choosing one genre for each song was 10%.

music descriptors. More concretely, we use some of the descriptors implemented in the MIRToolbox[1] [12]: Mel-frequency cepstral coefficients, chromas, brightness, roll-off, spectral centroid, spectral spread, and spectral flatness. For computing these descriptors we use the default parameters provided in the MIRToolbox, except for Mel-frequency cepstral coefficients, where we use 50 filters and 20 coefficients, brightness, where we use 3000 Hz as threshold (278 samples in the case of $\mathbf{h}^{(t)}$), and roll-off, where we use 80% of the energy as threshold. We also use statistical moments to describe histograms and spectra: mean, variance, skewness, and kurtosis.

Descriptors are processed using a bag-of-frames approach and means and variances are taken. This results in a total of 82 descriptors for each song and analysis approach, i.e. 82 descriptors based on the spectrum, 82 based on $\mathbf{h}^{(t)}$, and 82 based on $\mathbf{h}^{(f)}$. We also test the pooling of these three approaches by combining $\mathbf{h}^{(t)}$ and $\mathbf{h}^{(f)}$, spectral and $\mathbf{h}^{(t)}$, as well as spectral and $\mathbf{h}^{(f)}$, resulting in 164 descriptors each, and by combining spectral, $\mathbf{h}^{(t)}$, and $\mathbf{h}^{(f)}$, resulting in 246 descriptors.

Extracted descriptors are input to an attribute selection process before the final classification is done. For these two tasks we use the algorithms provided by the Weka[2] data mining software [13]. For attribute selection we proceed in two steps. We first apply correlation feature selection (CFS) and subsequently perform principal component analysis (PCA) [14]. A total of 30 principal components are taken after CFS: the ones covering most of the data variance. In case CFS leads to less than 30 components, PCA is applied without preliminary CFS. The attribute selection process is the same independently of how many descriptors we consider. Thus we always end up with 30 components. These components are finally normalized between 0 and 1.

For classification we use several different algorithms provided by the Weka package [13]. By considering several classifiers, we can better assess the potential accuracy improvements yielded by our approach. Hence we can be sure that the information they convey can be exploited for many classifiers and not just by a particular one. As classifiers we use the following algorithms [14]: naïve Bayes (NB), nearest neighbors (NN; denominated IBk in Weka), multilayer perceptron (MP), random forest (RF), linear logistic regression (LL; simple logistic in Weka), and support vector machines (SVM; SMO implementation in Weka). We also use the one-rule classifier (1R), which provides a classification accuracy based on the most discriminating attribute. We use 5 neighbors for NN, a learning parameter of 0.6 for MP, 100 trees for RF, the Akaike information criterion for LL, and two kernels for SVM: a polynomial function (SVM$_{\mathrm{P}}$) and a radial-basis function (SVM$_{\mathrm{R}}$) with parameter gamma set to 0.6. All other parameters are left unchanged from their default values[2]. The classification accuracies for these algorithms are evaluated 10 times with 3-fold cross validation and the mean value is taken [14]. The classification setup remains the same for both the histogram-based descriptors and the spectrum-based descriptors, and also for combinations of them.

4. RESULTS AND DISCUSSION

We first evaluate the baseline accuracy for spectral descriptors (Table 1, first row). We see that a maximal accuracy of 66% is achieved, both for the LL and SVM$_{\mathrm{R}}$ classifiers. We then evaluate the accuracies for descriptors extracted from the recurrence histograms for different values of the parameters m, τ, and p. In pre-analysis, parameters were set empirically, based on visual inspection of histograms and recurrence plots for frames of arbitrarily selected songs [10]. After this visual inspection, combinations of $m \in [3, 7, 12]$, $\tau \in [1, 3, 7]$, and $p \in [0.2, 0.3, 0.7]$ were used for final testing. The best accuracies found for $\mathbf{h}^{(t)}$, $\mathbf{h}^{(f)}$, and their combinations are reported (Table 1, rows 2 to 7). No drastic change in the accuracy was observed for alternative parameter combinations. For more details on our parameter assessment we refer to [10].

We see that the accuracies for histogram approaches do not surpass the ones achieved by our baseline spectral descriptors. However, once information from spectral descriptors is combined with recurrence histograms we observe a consistent accuracy increment (Table 1, rows 8 to 13). This holds particularly when $\mathbf{h}^{(t)}$ is included. Improvements are more accentuated for classifiers that show a low accuracy when only spectral descriptors are used (see the columns

[1] Version 1.3: `https://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/mirtoolbox`

[2] Version 3.6.2: `http://www.cs.waikato.ac.nz/ml/weka`

NN and NB, row 1 vs. rows 8, 9, 12, and 13). For those classifiers that performed best with the spectral descriptors we still get an absolute accuracy improvement of 4 to 5% (see the columns LL and SVM$_\mathrm{R}$, row 1 vs. rows 8, 9, 12, and 13).

In general, poorer accuracies are obtained from approaches including information from $\mathbf{h}^{(\mathrm{f})}$. Sometimes, the addition of descriptors based on $\mathbf{h}^{(\mathrm{f})}$ to the baseline does not result in any improvement at all. This is somehow to be expected, since the information in $\mathbf{h}^{(\mathrm{f})}$ is concentrated in a small part of the histogram [the lowest frequency bins, see Eq. (4) and Fig. 2]. Thus, we hypothesize that, due to the linear spacing of frequency bins an excessive compression of the recurrence information is taking place.

On the other hand, when information from $\mathbf{h}^{(\mathrm{t})}$ is added to spectral-based descriptors, accuracies for all classifiers raise without any exception. More importantly, this holds for different combinations of the parameters underlying the calculation of the histograms. In general, no excessive fine-tunning of the parameters is needed to achieve an accuracy improvement. In particular, for the results presented here, no exhaustive grid-search was done [10]. One might expect that with such an exhaustive search the accuracies might even increase further. However, the risk of overfitting the parameters to the data may be higher [14]. In any case, parameters that we found to yield the best accuracy values remain to be validated on further music collections and classification tasks.

It is also important to note that the bins of $\mathbf{h}^{(\mathrm{t})}$ represent sample lags instead of frequencies. However, the extraction of quantitative information from $\mathbf{h}^{(\mathrm{t})}$ is done by means of common descriptors originally developed for frequency-based representations. Thus, this might not be the most effective way of extracting the information contained in $\mathbf{h}^{(\mathrm{t})}$. Once can envision that new ways to quantify the information in $\mathbf{h}^{(\mathrm{t})}$ can increment the accuracy improvement reported here.

5. CONCLUSIONS AND FUTURE WORK

In this paper we test the use of nonlinear recurrence analysis for quantitative music description. More concretely, we propose the use of recurrence time and frequency histograms for the extraction of information from a music audio frame. We evaluate the accuracy of our approach with a typical genre classification framework and compare against a baseline of common descriptors obtained from a spectral representation.

Although the accuracies reached by the histogram-based descriptors alone do not surpass the accuracies achieved by the spectral-based descriptors, we show that the combination of both sources can result in substantial improvements. This highlights the potential of nonlinear recurrence analysis for quantitative music description. In particular, it suggests that the information resulting from this process is complementary to the information obtained from the common spectral representation. Overall, our results underline that considering complementary strategies to the common audio processing chain is a promising direction to potentially overcome (or at least reduce) the “glass-ceiling” of MIR systems [3, 5].

As future work we plan to validate these results with different music collections, not only with the genre classification task, but also considering other common MIR classification tasks [3]. Furthermore, a deeper understanding of the information overlap between histogram-based and spectral-based descriptors is needed. A more comprehensive study of the effect of parameters m, τ, and p is also left for further research. In addition, we plan to use specific quantification measures for $\mathbf{h}^{(\mathrm{t})}$. These measures should be suitable for temporal or sample lag-based representations. Finally, we also want to study the effect of a nonlinear spacing of the frequency bins of $\mathbf{h}^{(\mathrm{f})}$, so that its information is less compressed in a few bins.

6. ACKNOWLEDGMENT

The authors thank George Tzanetakis for sharing with them the music collection used in this study.

7. REFERENCES

[1] M. Casey, R. C. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney, “Content-based music information retrieval: current directions and future challenges,” *Proceedings of the IEEE*, vol. 96, no. 4, pp. 668–696, 2008.

[2] J. O. Smith III, *Spectral audio signal processing*, Center for Computer Research in Music and Acoustics (CCRMA), Stanford University, USA, 2010, Online resource: `https://ccrma.stanford.edu/~jos/sasp`.

[3] J. S. Downie, “The music information retrieval evaluation exchange (2005–2007): a window into music information retrieval research,” *Acoustical Science and Technology*, vol. 29, no. 4, pp. 247–255, 2008.

[4] N. Scaringella, G. Zoia, and D. Mlynek, “Automatic genre classification of music content: a survey,” *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 133–141, 2006.

[5] J. J. Aucouturier and F. Pachet, “Improving timbre similarity. How high is the sky?,” *Journal of Negative Results on Speech and Audio Sciences*, vol. 1, no. 1, 2004.

[6] H. Kantz and T. Schreiber, *Nonlinear time series analysis*, Cambridge University Press, Cambridge, UK, 2nd edition, 2004.

[7] J. Serrà, X. Serra, and R. G. Andrzejak, “Cross recurrence quantification for cover song identification,” *New Journal of Physics*, vol. 11, pp. 093017, 2009.

[8] D. E. Terez, “Robust pitch determination using nonlinear state-space embedding,” in *Proc. of the IEEE Int. Conf. on Audio, Speech, and Signal Processing (ICASSP)*, 2002, vol. 1, pp. 345–348.

[9] N. Marwan, M. C. Romano, M. Thiel, and J. Kurths, “Recurrence plots for the analysis of complex systems,” *Physics Reports*, vol. 438, no. 5, pp. 237–329, 2007.

[10] C. A. de los Santos, “Nonlinear audio recurrence analysis with application to music genre classification,” M.S. thesis, Universitat Pompeu Fabra, Barcelona, Spain, 2010, Available online: `http://mtg.upf.edu/node/1803`.

[11] G. Tzanetakis and P. Cook, “Musical genre classification of audio signals,” *IEEE Trans. on Speech and Audio Processing*, vol. 5, no. 10, pp. 293–302, 2002.

[12] O. Lartillot and P. Toiviainen, “MIR in Matlab II: a toolbox for musical feature extraction from audio,” in *Proc. of the Int. Conf. on Music Information Retrieval (ISMIR)*, 2007, pp. 127–130.

[13] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, “The Weka data mining software: an update,” *ACM SIGKDD Explorations*, vol. 1, no. 1, pp. 10–18, 2009.

[14] I. H. Witten and E. Frank, *Data mining: practical machine learning tools and techniques*, Elsevier, Amsterdam, The Netherlands, 2nd edition, 2005.