German Medical Data Sciences: Bringing Data to Life
R. Röhrig et al. (Eds.)

doi:10.3233/SHTI210059

Semantic Anomaly Detection in Medical Time Series

Sven FESTAG [a,1] and Cord SPRECKELSEN [a]
[a] *Institute of Medical Statistics, Computer and Data Sciences, Jena University Hospital*

Abstract. The main goal of this project was to define and evaluate a new unsupervised deep learning approach that can differentiate between normal and anomalous intervals of signals like the electrical activity of the heart (ECG). Denoising autoencoders based on recurrent neural networks with gated recurrent units were used for the semantic encoding of such time frames. A subsequent cluster analysis conducted in the code space served as the decision mechanism labelling samples as anomalies or normal intervals, respectively. The cluster ensemble method called cluster-based similarity partitioning proved itself well suited for this task when used in combination with density-based spatial clustering of applications with noise. The best performing system reached an adjusted Rand index of 0.11 on real-world ECG signals labelled by medical experts. This corresponds to a precision and recall regarding the detection task of around 0.72. The new general approach outperformed several state-of-the-art outlier recognition methods and can be applied to all kinds of (medical) time series data. It can serve as a basis for more specific detectors that work in an unsupervised fashion or that are partially guided by medical experts.

Keywords. Unsupervised Machine Learning, Cluster Analysis, Electrocardiography, Anomaly Detection, Time Series

1. Introduction

The surge of digital patient data and the increased use of data-driven methods in medical research and care expedite the need for fast, accurate and automatic anomaly detection in chronologically ordered sequences of medical measurements (medical time series). The present work focuses on semantic anomaly recognition in medical time series data with the help of deep learning approaches. The basic idea is to use an autonomous and semantic encoding strategy that is trained to map similar intervals of time series (time frames) to neighbouring points. In this context, the term "semantic" refers to the characteristic of the encoding to "integrate the entire correlation structure" [1] among all time frames. Thus, the encoder learns to preserve certain relations between time frames while mapping them to a low-dimensional space. This automatic encoding of information must not be confused with the explicit manual modelling of associations as known from Semantic Web applications.

Due to the large amount of data measured at an intensive care unit for every patient, a manual review of such time series is infeasible. Hence, systems that automatically extract valuable information are needed. One building block of these systems is a

[1] Corresponding Author: Sven Festag, Institute of Medical Statistics, Computer and Data Sciences, Jena University Hospital, Bachstr. 18, 07743 Jena, Germany; E-mail: sven.festag@med.uni-jena.de

mechanism for anomaly detection that autonomously identifies time frames diverging from the norm. Depending on the application, "normal" either means physiologically unsuspicious or is defined by a certain signal pattern specified by medical experts.

The main goal of the project was to develop and evaluate a mechanism for the detection of anomalous time frames that are specific to certain signals like the electrical activity of the heart (ECG) or the blood pressure.

To determine sequential anomalies, qualitative and semantic analyses are needed [2]. The methodological basis for our project is the combination of denoising autoencoders (DAEs) built of recurrent neural networks (RNN) and density-based clustering algorithms. In the first phase of this approach, a DAE encodes time frames into low-dimensional semantic representations. The second phase applies a cluster analysis to this low-dimensional embedding space. The extracted clusters are then interpreted as groups referring to "normal" or "anomalous" signals, respectively. In the present work "anomalous time frames" are defined as rare and unusual sequences. A remarkable aspect of the described method is that it can be trained in a fully unsupervised manner and does not depend on human expert knowledge.

For the training and evaluation of the anomaly detection approach real-world medical data collected at hospitals were used. We employed data sets curated by the MIT Lab for Computational Physiology (PhysioNet). The assessment of the newly developed method is based on a data set that has been subdivided into "anomalous" and "normal" by medical experts. Thus, the performance can be quantified exactly and compared against state-of-the-art outlier detection methods.

The overall research question answered in this paper can be stated as follows: "Can recurrent autoencoders be used in combination with density-based cluster analyses to detect anomalous intervals of real-world medical time series in an unsupervised fashion?"

2. Related Work

According to Hodge and Austin there are three different general approaches to anomaly recognition [3].

1. Determining outliers without prior information
2. Modelling normality and abnormality
3. Modelling only normality

Approaches 2 and 3 require labelled training data, i.e. the data have to be classified manually into "anomalous" and/or "normal" before training. Labelled data is always a scarce resource and particularly hard to get for the anomaly detection task. The first difficulty is that there is no precise definition of "anomaly" that holds true for all domains. The classification of data as anomalous requires domain knowledge and gives latitude to the classifying expert [2]. Furthermore, anomalous data points or sequences appear rarely in real-world data sets due to their very nature.

In the literature, one finds many approaches for the recognition of anomalies in medical signals. Most of these methods are developed for the use in alarm systems for critical care monitoring. Imhoff and Kuhls gave an extensive summary of available approaches in 2006 [4]. Most of the described early systems for time series analysis rely on pure statistical approaches like dynamic linear models, autoregressive models and self-adjusting thresholds. However, also some machine learning approaches including support vector machines and basic artificial neural networks (NN) trained in

a supervised fashion were applied at that time. More recent approaches for general anomaly detection rely on recurrent NNs trained to predict future values from a previous course. The difference between prediction and actual value is then used as a criterion for anomaly [5,6]. In the last few years, many approaches leveraged a special type of NNs, called autoencoder or replicator NN, in order to learn to detect anomalies in a semi-supervised fashion [7,8].

Autoencoders (AE) encode the input data by transforming it into a lower dimensional latent space before it is decoded back to the original dimensionality. The training aims at minimising the residual vectors, i.e. the differences between the inputs and outputs of an AE. In general, these autoencoders are trained on normal samples only (type 3). The approach is based on the assumption that (unseen) normal data should be reproduced relatively well when transformed by the network. In contrast, new anomalous data become apparent by large residual vectors, as their latent attributes deviate from those of normal ones [8]. In contrast to the previously described methods, these approaches can make their decisions on the basis of current sequence data without the need for any predictions.

Many of these techniques are based on simple feedforward AEs and are tailored to detect single data points of fixed dimensionality. To be able to handle sequences of data points RNNs are needed in the encoding part of an AE. The focus of our work is on Gated Recurrent Unit (GRU) RNNs used to this end. They achieve state-of-the-art results when applied to tasks with sequential data [9].

In order to detect outliers without having labelled training data at hand, an RNN-based AE can be combined with a clustering mechanism. By clustering a large set of encoded sequence samples, different classes of “anomalous” and “normal” ones can be distinguished in the latent space. Such a model can be trained fully unsupervised and thus, belongs to type 1 of Hodge’s and Austin’s classification.

Several publications such as [11] build upon this idea and describe the use of AEs to transform samples in a more discriminative latent space before clustering. Normal samples can then be identified as points lying in large and dense clusters, while anomalies appear in smaller groups or as noise points.

3. Methods

Our approach can be divided into two parts, 1) the semantic encoding of medical signals and 2) the subsequent cluster analysis in the generated low-dimensional space. This section also follows this subdivision.

3.1. Denoising Autoencoder for Semantic Encoding of Time Series

The inputs of the utilised AE are short intervals of a digitised physiological signal collected by a bedside monitor. Such intervals can be interpreted as vectors with a

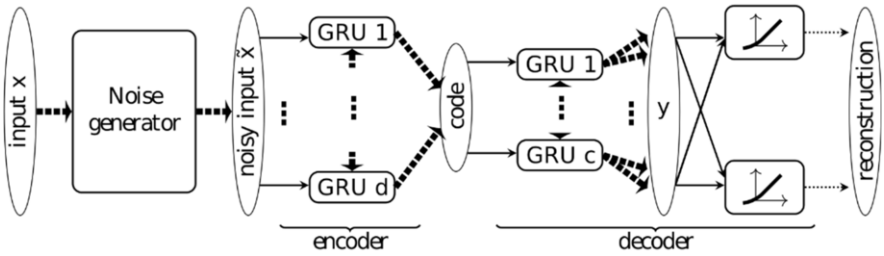

Figure 1. Topology of the denoising autoencoder used for the semantic encoding.

dimensionality d depending on the duration and digitisation frequency. For example, a snippet taken from a single ECG lead with a length of 10 s and a digitisation frequency of 125 Hz can be represented by a vector of 1250 dimensions. The task of the AE or, more precisely, of the included encoder is to transform these pieces into a low-dimensional latent space.

If an AE with high capacity is trained to extract low-dimensional features from the inputs, there is a risk that it learns to memorise which code belongs to which input signal without meaningfully distributing the codes in the latent space [12]. In such a case the learned encoding function cannot be used as a feature extractor. Alain and Bengio could prove that denoising training, i.e. a training based on noisy versions of the inputs, leads to autoencoders that implicitly learn the data-generating distribution from which the training points are sampled [13]. This means that denoising AEs map training samples to codes that preserve information of the training data distribution. The ability of DAEs to incorporate these "semantics" of the original vectors into the distribution of corresponding codes led to the name "semantic encoding". We used an additive isotropic Gaussian noise with zero mean as the noise model.

The DAE trained to conduct the semantic encoding consists of RNN-based encoder and decoder and its topology is depicted in Figure 1. In the first step a Gaussian noise is added to the input of size d. Afterwards, the noisy input is fed element wise into a bidirectional RNN made from GRU cells [14]. The last outputs of both directions are concatenated to get the code. To ensure a fixed code dimensionality c, both directions work with states of size c/2. A similar bidirectional GRU RNN is used as the decoder. In contrast to the encoder RNN, the full sequence of concatenated outputs is used. The sequence is interpreted as a single vector and trimmed to size d by dropping the last elements. Afterwards, it is passed to a final dense feedforward layer consisting of d neurons that compute leaky ReLU activations. The initial states of all RNNs as well as all biases are initialised with zeros. The weight matrices are initialised in accordance with the Glorot Uniform strategy.

After training is finished, only the encoder part is needed for the semantic encoding of sequences.

3.2. Cluster Analyses in Semantic Space for Anomaly Detection

After the transformation of (medical) time series into representations that lie in a low-dimensional semantic space, a cluster analysis is needed to detect anomalous or, more precisely, rare and unusual sequences. The reasoning is that by the encoding important

features are extracted or generated as combinations of original features. Moreover, the placement of these feature vectors in the latent space preserves information of the original data distribution.

Since one cannot make any assumption on data distributions of unknown sets of medical time series, we decided to use a density-based clustering approach. In contrast to k-means and many other distance-based methods, density-based clustering does not rely on assumptions about cluster shapes. The so-called Density-Based Spatial Clustering of Applications with Noise (DBSCAN) finds an automatically determined number of clusters with arbitrary shapes. Roughly speaking, DBSCAN identifies core points, i.e. points with a dense neighbourhood, and connects these with their surroundings in order to define clusters. It depends on three parameters: a distance function, an upper bound on this distance ε, and a minimum number of points minPts, which in combination define dense neighbourhoods. At the end of the clustering procedure there might be points which are not assigned to any cluster and are interpreted as noise points. For a detailed description of the algorithm see [15].

One problem of this clustering method is that the parameter values must be set manually and that they have a strong influence on the outcome. To reduce this shortcoming, we utilised a cluster ensemble approach that integrates many clusterings based on different parameter combinations into one consensus clustering: The Cluster-based Similarity Partitioning (CSPA) suggested by Strehl and Ghosh [16]. As in a grid search, DBSCAN is performed several times with different parameter values. Based on these clusterings a similarity matrix over all input vectors is computed. Afterwards, graph partitioning is conducted to find the final consensus clusters. In derogation from the original line of action described by Strehl and Ghosh, we used Normalised Spectral Clustering instead of METIS for the graph partitioning. The reason for this is that METIS aims at generating sub-graphs/clusters of similar size. In the context of outlier detection this is inexpedient, as clusters are likely to be of different magnitudes. A comprehensive description of Normalised Spectral Clustering is given in [17]. The only parameter that must be set for Spectral Clustering is the number of partitions/clusters to be found.

4. Experiments and Results

For the training and evaluation of our approach we used freely-available real-world data taken from the MIMIC-III Waveform Database Matched Subset (MIMIC Waveform) [18] and from the training set of the PhysioNet/Computing in Cardiology Challenge 2015 (PhysioNet/CinC) [19]. Time series extracted from the first set were used to train and test the DAE, while the second set was utilised to evaluate the performance of the CSPA working on time series embedded by the previously trained DAE.

4.1. Evaluation of Denoising Autoencoders for Semantic Encoding of Time Series

Every training or test sample corresponds to a time frame of 10 s taken from an ECG signal (measured in Einthoven's bipolar lead II). As these waveform data have a frequency of 125 Hz, samples are vectors comprised of 1250 dimensions. For the training set, ECG signals of 242 patients were used. The test set consisted of data collected from 21 different patients. Since several waveform intervals were taken from

every patient, the training set comprises around 1400000 and the test set approx. 120000 vectors. To exclude bad signals that were generated due to technical errors or during calibration, samples showing at least one of the following characteristics were omitted: NaN value included, value larger than 3mV or smaller than -3mV included, no negative value for more than 1.5 second, interval of only negative values for more than 0.2 seconds. This cleaning was not used on test data.

The DAE was set up as described in section 3.1. It allowed inputs of size d = 1250 and produced codes with length c = 126. For the minimisation of the mean squared error (MSE) between inputs and reconstruction the Adam optimiser with a learning rate of 0.001 was chosen. The mini-batch size was set to 2300 and the noise parameter of the isotropic Gaussian noise to 0.1. During 200 training epochs the error on the test set dropped continuously from an initial MSE of 0.048 to a final one of 0.0300.

Experiments where LSTM cells or plain feedforward decoders instead of GRUs were used led to worse results.

4.2. Evaluation of Cluster Analyses in Semantic Space for Anomaly Detection

The unlabelled data sets used for the previously described experiment cannot be applied to quantify performances of the cluster analysis on the anomaly detection task. For this reason, another set that includes labelled data was used for the trials. It was built upon the training set of the PhysioNet/CinC challenge. All included ECG series had been identified by real-time bedside monitors as containing pathologic behaviour in the last 10 seconds and thus, had triggered an alarm. Afterwards, a team of medical experts has labelled the sequences as true or false arrhythmia alarms, respectively. We extracted the last interval of 10 seconds from every lead II ECG signal and sampled it down to a frequency of 125 Hz in order to generate 1250 dimensional vectors. Vectors containing NaN values were excluded, as they lead to undefined intermediate results. In total there were 428 signals labelled false alarms and 281 marked as true ones. The test set was enriched by 3000 random unlabelled samples from the MIMIC Waveform training set to include more information into the cluster analysis. The reasoning for this approach is that the more semantically encoded samples are given, the more information about their correlations is contained in the cluster space. For the evaluation, however, only the labelled samples were considered.

To evaluate the performance of CSPA on the task of discriminating between true alarms (normal) and false alarms (outliers) the ensemble method was applied as described in section 3.2. Cosine distance was used for the DBSCAN analysis which was repeated for all parameter combinations $(\varepsilon, \text{minPts}) \in \{n/100 \mid n \in \mathbb{N} \wedge n \leq 100\} \times \{m \in \mathbb{N} \mid 3 \leq m \leq 100\}$. To remove useless clusterings, only results with one normal group and one noise group were retained for the ensemble computation. Furthermore, in the described experiment, noise clusters had to contain at least 50% and at most 70% of the labelled samples for the clustering to be considered a valid result. This range was chosen, as the contamination ratio in PhysioNet/CinC set is around 0.6 if false alarms are considered the anomalous samples. The contamination ratio and the number of expected normal clusters are the only hyperparameters that have to be set manually before this kind of ensemble clustering. The final cluster-based similarity partitioning on all retained results led to the following two groups. The noise group comprised 310 false and 120 true alarms, while the normal group included 118 false alarms and 161 true ones. This corresponds to an adjusted Rand index (ARI) of 0.1051, a BCubed

Precision of 0.5626 and a BCubed Recall of 0.5637. For the two-class anomaly detection, this signifies a precision and recall of around 0.72.

Stacked DAEs or DAEs with bigger code dimensions led to faster convergence and better performance during training, but CSPA performed worse on the generated embeddings.

Two existing state-of-the-art outlier detection methods, Local Outlier Factor [20] and Isolation Forest [21], led to ARI scores of 0.0076 or 0.0165, respectively, when applied to the same (unembedded) data set.

5. Discussion

The findings of our project prove that the new semantic anomaly detection approach based on denoising GRU-autoencoders in combination with an ensemble of DBSCAN clusterings is suited for the task of unsupervised anomaly detection in medical time series. The method even outperformed two state-of-the-art outlier detection procedures on real-world clinical data. Nonetheless, these approaches have specific merits as one of them avoids a binary classification and instead introduces a "degree of being an outlier" [20] while the other exhibits favourable computational performance [21].

A surprising result of the trials is that the anomaly detection works better in a semantic space generated by a single DAE with a rather small code dimension than in a space computed by a stacked DAE or a DAE with larger code dimensions. It is likely that RNN-based DAEs which have very high capacity lead to overly complex representations that shadow the important information needed for the anomaly detection. However, we expect the right code dimensionality to be domain dependent.

The division of the presented detection approach into two phases has an advantage regarding the computation cost. The cost-intensive training of a DAE can be conducted on a large data set of medical signals in a first phase. Afterwards, the trained DAE can be used repeatedly to encode new and small sets of the same signal before they are clustered to detect anomalies. In contrast to DAE training, encoding and clustering do not require large computation power.

The presented method can be useful in diagnosis and patient monitoring.

6. Conclusion

This research has shown that it is possible to detect anomalous time frames in an unsupervised way by a new semantic anomaly detection approach based on recurrent denoising autoencoders and density-based cluster analyses. In this dichotomous system a recurrent DAE serves to reduce dimensionality and to preserve semantic information. The subsequent cluster analysis is needed in order to detect dense clusters corresponding to "normal" samples and noise points corresponding to anomalies.

Future studies could address the extension of the semantic anomaly detection approach for the recognition of anomalies with variable lengths. Due to the used RNN, the implementation of this extension is easily possible. Since the data sets used throughout the experiments contained only ECG signals, further tests could be carried out on different time series types.

Acknowledgments

Simulations were partially performed with computing resources granted by RWTH Aachen University under project nova0025. SF acknowledges the valuable feedback by members of the Chair of Computer Science 5, RWTH Aachen University.

The authors state that they have no conflict of interests.

References

[1] D.B. Skillicorn, Outlier Detection Using Semantic Sensors, in: IEEE Int. Conf. Intell. Secur. Inform., Arlington, VA, 2012: pp. 42–47. doi:10.1109/ISI.2012.6284089.

[2] C.F.G. Schendera, *Datenqualität mit SPSS*, De Gruyter, Berlin, 2007. doi:10.1524/9783486710694

[3] V.J. Hodge, and J. Austin, A Survey of Outlier Detection Methodologies, *Artif. Intell. Rev.* **22** (2004), 85–126. doi:10.1007/s10462-004-4304-y.

[4] M. Imhoff, and S. Kuhls, Alarm Algorithms in Critical Care Monitoring, *Anesth. Analg.* **102** (2006), 1525–1537. doi:10.1213/01.ane.0000204385.01983.61

[5] L. Bontemps, V.L. Cao, J. McDermott, and N.-A. Le-Khac, Collective Anomaly Detection Based on Long Short-Term Memory Recurrent Neural Networks, in: Future Data Secur. Eng., Can Tho City, 2016: pp. 141–152. doi:10.1007/978-3-319-48057-2_9

[6] P. Malhotra, L. Vig, G. Shroff, and P. Agarwal, Long Short Term Memory Networks for Anomaly Detection in Time Series, in: 23rd Eur. Symp. Artif. Neural Netw., Bruges, 2015: pp. 89–94.

[7] C. Baur, B. Wiestler, S. Albarqouni, and N. Navab, Deep Autoencoding Models for Unsupervised Anomaly Segmentation in Brain MR Images, in: Brainlesion: Glioma Mult. Scler. Stroke Trauma. Brain Inj., Granada, 2018: pp. 161–169. doi:10.1007/978-3-030-11723-8_16

[8] J.T.A. Andrews, E.J. Morton, and L.D. Griffin, Detecting Anomalous Data Using Auto-Encoders, *Int. J. Mach. Learn. Comput.* **6** (2016) 21–26.

[9] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, Recurrent Neural Networks for Multivariate Time Series with Missing Values, *Sci. Rep.* **8** (2018). doi:10.1038/s41598-018-24271-9

[10] R. Salakhutdinov, and G. Hinton, Semantic hashing, *Int. J. Approx. Reason.* **50** (2009) 969–978.

[11] E. Tzoreff, O. Kogan, and Y. Choukroun, Deep Discriminative Latent Space for Clustering, *ArXiv E-Prints*. (2018). http://arxiv.org/abs/1805.10795

[12] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, Cambridge, MA, 2016.

[13] G. Alain, Y. Bengio, and S. Rifai, Regularized Auto-Encoders Estimate Local Statistics, *ArXiv E-Prints*. (2012). http://arxiv.org/abs/1211.4246v1

[14] K. Cho, B. van Merrienboer, Ç. Gülçehre, F. Bougares, H. Schwenk, and Y. Bengio, Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation, *ArXiv E-Prints*. (2014). http://arxiv.org/abs/1406.1078v1

[15] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, San Francisco, CA, 2011.

[16] A. Strehl, and J. Ghosh, Cluster Ensembles - a Knowledge Reuse Framework for Combining Multiple Partitions, *J. Mach. Learn. Res.* **3** (2002) 583–617.

[17] J. Shi, and J. Malik, Normalized cuts and image segmentation, *IEEE Trans. Pattern Anal. Mach. Intell.* **22** (2000) 888–905. doi:10.1109/34.868688

[18] B. Moody, M. Craig, A. Johnson, T. Kyaw, G. Moody, M. Saeed, and M. Villarroel, *The MIMIC-III Waveform Database Matched Subset*, physionet.org, 2017. doi:10.13026/C2294B

[19] G.D. Clifford, I. Silva, B. Moody, Q. Li, D. Kella, A. Shahin, T. Kooistra, D. Perry, and R.G. Mark, The PhysioNet/Computing in Cardiology Challenge 2015: Reducing false arrhythmia alarms in the ICU, in: Comput. Cardiol. Conf. CinC, Nice, 2015: pp. 273–276.

[20] M.M. Breunig, H.-P. Kriegel, R.T. Ng, and J. Sander, LOF: identifying density-based local outliers, *ACM SIGMOD Rec.* **29** (2000) 93–104.

[21] F.T. Liu, K.M. Ting, and Z.-H. Zhou, Isolation Forest, in: Proc. Eighth IEEE Int. Conf. Data Min., Pisa, 2008: pp. 413–422.