論文内容の要旨

論文題目：Unsupervised Anomaly Detection within Non-Numerical Sequence Data
(非数値系列データにおける教師無し異常検出)

氏名：シテファン　ヤン　スクドラレク
Stefan Jan Skudlarek

The present thesis examines a particular aspect of the general problem of *anomaly detection*, which has developed into an important paradigm of data mining during the past two decades. Anomaly detection tries to detect erroneous data output by a source by defining it as a deviation from the normal output.

The conventional approach to anomaly detection processes training data guaranteed to have been generated by a normal source state in order to build a profile of the normal system output. Clean training data, however, may be difficult to acquire, depending on the application. Therefore, several methods for detecting anomalies within an ensemble of data sequences or records, the majority of which is supposed to be normal, have been devised. Such approaches are usually subsumed under the category of *unsupervised anomaly detection*, although the term is sometimes used to refer to anomaly detection as defined above when discriminating it from the anomaly detection approach using training data to deduce a model of both the normal and the erroneous output.

Because unsupervised anomaly detection as defined above, although not ignored, has been studied much less exhaustively than (supervised) anomaly detection, especially with respect to real-world non-numerical sequence data, we decided to choose *Unsupervised Anomaly Detection within Non-Numerical Sequence Data* as the topic of our research, examining the following scenario:

1. We are given a set $\mathcal{S}$ of n sequences $x^{b_1}, x^{b_2}, \ldots, x^{b_{n-1}}, x^{b_n}$, the lengths being a multiple of the minimum length b, with $x \in \mathcal{X} = \{a_1, a_2, \ldots, a_{Z-1}, a_Z\}$, with the alphabet size Z not fixed in advance. The index $i \in \{1, 2, \ldots, n-1, n\}$ of the sequences may either be assigned at random or indicate the temporal order of sequence generation.

2. We suppose that the majority of $1-\rho$ ($0 \leq \rho \leq \rho_{\max} = 0.33$) of the sequences was generated by one stationary normal source N (one-class scenario), while the remaining share ρ of the sequences was generated by one or more stationary abnormal sources A. Note that this scenario includes the case of $\rho = 0$ (No anomalous data). The indices of the anomalous sequences may be random with respect to the overall index range, or cover a limited subsection of the range.

3. The task is to derive a measure for the normality of each sequence.

We present two distinct approaches to the above problem.

The first approach fuses together the set of sequences $\mathcal{S}$ into a single global sequence of length g. It uses a function called the average index difference to respectively generate a numerical value associated with every single symbol within the global sequence.

We introduce the average index difference function, which calculates the average of the index differences between a symbol or subsequence found at a particular index j and symbols or subsequence of identical value within the global sequence. We prove the convergence of the function to an expected value dependent only on the global index j but not on the symbol generation probability for the case of stationary ergodic symbol generation.

We present two algorithms based on this function.

The first algorithm exploits the fact that, if the abnormal sequences happen to be clustered within a subsection of the global sequence, the output value of the average index difference function is reciprocally related to the likelihood of the symbol to be representative of the anomalous data compared to the normal data and to have been generated within an anomalous sequence. This is because most of the identical symbols will be generated close to the index j, thus significantly decreasing the average index difference function value. The percentage c of symbols within a sequence featuring an average index difference below a certain threshold τ_{th} is used for final classification.

Because the first algorithm is hampered by supposing the abnormal sequences to be consecutively clustered within the overall sequence, we conceive the second algorithm, which extends the original average index difference function, allowing for an arbitrary location of the anomalous sequences within the global sequence. Furthermore, the algorithm extends the function to subsequences of symbols, the maximum length $M_{\max}$ of which was set via an information theoretic criterion. It compares index differences between neighboring occurrences $\Delta\left(x^{M+1}\right)$ to a ξ multiple of the empirical mean value $\overline{\Delta}\left(x^{M+1}\right)$ of those index differences, in order to identify gaps between anomalous blocks prior to the calculation of the average index difference.

Besides conceiving the algorithms, our contribution consists of showing how suitable settings for all the parameters both for the case of stationary ergodic generation and i.i.d. generation can be derived by theoretical considerations. We also deduce bounds for the computational cost of both algorithms, showing how the average index difference of every symbol within the sequence of length g can be computed in a time linear with g, and evaluate the performance of the two algorithms using both computer security-related real world data and artificial data, comparing our results to those of previous methods. Calculating the curves of the respective receiver operating characteristic, we deduce the thresholds from the set of scalar values returned by the algorithms by means of robust statistics.

Our second approach to the problem computes the matrix of pairwise distances of the set of sequences $\mathcal{S}$ by mapping them into a numerical space via a suitable kernel function, turning the scenario into a spatial classification problem.

The algorithm presented works as follows: First we map the sequences of $\mathcal{S}$ into a vector space using a suitable kernel, such that the vectors calculated from the sequences output by a stationary source will form a hypersphere. After calculating the matrix of pairwise distances of the vectors, we select a sequence close to the center of the hypersphere of the normal data as a representative of the normal data. This is done by using the distance matrix to calculate the radius β necessary to cover a share of θ of the n sequences for any sequence within $\mathcal{S}$. The sequence with minimum β is chosen as representative of the normal data. Finally, the sequences are classified according to their distance from the representative sequence.

We theoretically explain the choice and parameter setting of the kernel function used for our experiments, the

so-called spectrum kernel. Using the structural similarities between the kernel and a probabilistic suffix tree, we show how the optimized depth of the tree may be used for setting the dimensional parameter of the spectrum kernel. This parameter regulates the subsequence length for mapping. Moreover, we explain the setting of the key parameter of the algorithm, θ. We also bound the computational complexity of the algorithm, and evaluate the performance of the algorithm using both real world data and artificial data, comparing our results to those of previous methods.

When compared to previous methods, the two approaches presented stand out by the theoretical foundation of their parameter settings, which contrasts with the trial-and-error settings employed by most unsupervised anomaly detection algorithms. Moreover, experiments show equal or superior performance for real-world non-numerical sequence data. The first approach also establishes a unique method for comparing a particular block with all the other blocks in the data set by employing the average index difference function, which was first defined in a different context.