Streszczenie

W związku z narastającą ilością danych przetwarzanych przez współczesne systemy informatyczne i pomiarowe, nastaje konieczność opracowania skalowalnych metod uczenia maszynowego analizujących dane zdarzeniowe. O ile możliwe jest przetworzenie tego rodzaju danych przez klasyczne metody, ich użycie wymaga konwersji danych do postaci zawierającej redundancję. Kluczowe jest zatem rozwijanie algorytmów dedykowanych do danych zdarzeniowych.

Celem tej rozprawy jest przeanalizowanie dwóch różnych podejść do tematu uczenia maszynowego dla danych zdarzeniowych. Teoria procesów punktowych jest często wykorzystywana w neurobiologii obliczeniowej do analizy impulsów, jednakże jej użycie w szerszym kontekście nadzorowanej klasyfikacji szeregów czasowych jest zaskakująco rzadko spotykane. Z drugiej strony, impulsowe sieci neuronowe cieszą się dużym zainteresowaniem badaczy, mimo iż ich proces uczenia zwykle wymaga symulowania stanu całej sieci w każdej chwili czasowej, co nie jest efektywnym wykorzystaniem zasobów obliczeniowych. Ponadto, nie jest jasne w jaki sposób konwersja sygnału do postaci zdarzeniowej wpływa na działanie wyuczonego modelu.

Wyżej wspomniane problemy zaadresowano przy pomocy metod analizy statystycznej oraz symulacji numerycznych. Zaproponowano reguły klasyfikacji procesów punktowych przy pomocy metod bayesowskich, a następnie przeanalizowano zbieżność algorytmu do ryzyka bayesowskiego w funkcji liczby przykładów uczących. Sprawdzono również wpływ efektów brzegowych na działanie klasyfikatora opartego o jądrowy estymator gęstości. W kontekście impulsowych sieci zrównoleglono algorytm obliczania wyjścia neuronu reagującego na czas wystąpienia impulsu wejściowego, co znacząco skróciło czas uczenia sieci. Ponadto rozszerzono ów model o możliwość obserwowania i generowania wielu zdarzeń. Zdefiniowano również funkcję kosztu, która umożliwia uczenie sieci syjamskiej bezpośrednio w dziedzinie zdarzeń.

Praktyczną stosowalność tych metod zweryfikowano w kontekście trzech problemów badawczych: identyfikacji botów w mediach społecznościowych, wykrywaniu artefaktów podczas detekcji cząstek promieniowania kosmicznego oraz kategoryzacji pojazdów drogowych. W każdym z tych scenariuszy sprawdzono wpływ reprezentacji sygnału w postaci zdarzeniowej oraz hiperparametrów procesu uczenia na działanie gotowego modelu. Wybór tak zróżnicowanych zastosowań dowodzi dużej uniwersalności opracowanych metod.

30.04.2024

Pabian Mateusz

Abstract

With the rising volume of data processed by modern measurement and IT systems there is a need for the development of scalable machine learning solutions analyzing event data. While existing methods can be adapted to process such types of data, doing so introduces redundancy. And so, algorithms that are dedicated to sparse, event-based data representations are required.

The aim of this dissertation is to discuss the applicability of two different approaches to machine learning from event data. The classical theory of point processes is used extensively in computational neuroscience to describe spiking patterns; nevertheless, its application to supervised temporal sequence classification is surprisingly under-developed. Conversely, the artificial spiking neural networks (SNN) are more widely used. However, these models are usually trained by simulating the state of the entire network over time. For efficiency, the training procedure should instead consider the network state only at event occurrence. Furthermore, it is unclear how the signal-to-spike conversion process impacts the trained model performance when extracting event sequences from analog and digital data.

The aforementioned knowledge gaps are addressed using statistical analysis and numerical simulations. Classification rules for point processes are proposed based on the Bayes theory of classification. This algorithm is subsequently analyzed in terms of the rate of convergence to the optimal Bayes risk as the number of training examples increases. The impact of boundary effects on the kernel classifier performance is also assessed. In the scope of the SNN framework an existing single-spike time-to-first-spike layer computation is parallelized to achieve a significant speed-up compared to the original formulation. This model is further extended to process multiple input events and generate multiple output events. Lastly, modifying the training objective leads to the Siamese SNN model – the first-ever end-to-end training of a Siamese network in the spiking domain.

The practical implications of this research are evaluated in three supervised machine learning tasks: social media bot detection, cosmic ray imaging artifact rejection, and in-roadway sensor vehicle type identification. In all scenarios the impact of event sequence preprocessing (including signal-to-spike conversion) and hyperparameter choice is assessed. Selecting such distinct case-studies highlights the versatility of the methods developed in this thesis.

30.04.2024

Pabian Mateusz