Instytut Podstaw Informatyki
Polskiej Akademii Nauk

Julian Zubek

Metody integracji wieloskalowej informacji w sztucznych systemach uczących się

Rozprawa doktorska
przygotowana pod kierunkiem
dr hab. Dariusza Plewczyńskiego, prof. UW.

Warszawa 2016

Abstract

Julian Zubek, *Methods for integration of multiscale information in artificial learning systems*. Doctoral dissertation supervised by Dariusz Plewczyński, Institute of Computer Science, Polish Academy of Sciences, Warsaw 2016.

A multiscale classifier is a classifier integrating heterogeneous information coming from multiple conceptual levels of description. The need for such models is present in many disciplines operating on hierarchically nested data such as educational research, geographical research or modern systems biology. For example, macromolecules interactions in biology can be modelled simultaneously on the level of individual atoms, functional groups of atoms, or whole molecules. Each of these levels corresponds to different kinds of information which needs to be integrated in the final predictor.

To fully utilise multiscale information more specialised modelling techniques are necessary. This thesis focus on constructing multilevel classifier ensembles in which structure of the classifier reflects structure of the problem. A formal way of description of such systems is proposed. The most central notion is *level of description* defined as set of attributes possible to encode numerically on a given level. Classification is expressed as a sequence of transformations from input features to class label. Within this sequence of transformation machine learning algorithms are used to define parametrised transformations which depend on the available training data. Various forms of data preprocessing or result aggregation constitute an integral parts of the process. The whole classification system is presented in the form of computation graph, which makes it easy to trace the dependencies between data sources and classifiers.

Descriptions of the same phenomena on different levels are not necessary equivalent. Very often numbers of observations and attributes available on different levels vary. Some levels offer precise information on the population distribution while others describe only a general tendency. This notion is very similar to degrees of freedom as understood in statistics. The more degrees of freedom the more reliably parameters of the distribution can be estimated. This can be linked to notions of data complexity and classifier complexity from machine learning. Analysing this kind of data complexity is especially important in multilevel modelling where multiple estimations are made.

To measure data complexity more accurately a technique called complexity curve is introduced. The core idea behind it is that the information contained in a data set can be approximated by a smaller subset. Quality of the approximation is measured using Hellinger distance. Complexity curve is a plot presenting Hellinger distances between distributions induced by the subsets of different sizes and the distribution induced by the full set. It is demonstrated that under certain assumptions complexity curve can be treated as an upper bound of variance error component of the trained

classifier. Experiments with 81 diverse data sets show that complexity curve can contribute to explaining performance of specific classifiers on these sets. Complexity curve is also an effective heuristic for determining optimal sample size for inference. When applied to data pruning scenario it allows to reduce training time of complex classifiers multiple time without sacrificing accuracy.

Another issue, raising from the hierarchical nature of the modelled data, is the construction of balanced schema for evaluating multiscale classifiers. It is demonstrated that splitting the sample randomly into training and testing set may introduce bias into the procedure. To solve this problem multiscale dependencies between objects can be modelled as a graph and this information can be used in splitting procedure. A few graph-based splitting procedures are proposed and evaluated. Among them deterministic greedy procedure allow to utilise available data most efficiently.

To demonstrate applicability of the proposed methodology to practical biological problems, a multiscale classifier predicting protein-protein interactions is developed. It integrates information on direct protein residue contacts coming from 3D protein complexes to construct features for predicting binary protein interactions. Such two-stage approach is demonstrated to be superior than standard sequence-based methods operating in a single scale (AUC ROC 0.62–0.67 vs 0.56–0.57). Unfortunately those results does not translate into larger data sets of different characteristics. Low performance of already established methods and methodological issues regarding classifier evaluation suggests that the problem of protein interaction prediction may be ill-posed.

Streszczenie

Julian Zubek, *Metody integracji wieloskalowej informacji w sztucznych systemach uczących się*. Rozprawa doktorska przygotowana pod kierunkiem Dariusza Plewczyńskiego, Instytut Podstaw Informatyki Polskiej Akademii Nauk, Warszawa 2016.

Klasyfikatorem wieloskalowym nazywamy klasyfikator, który wykorzystuje do przewidywania różnorodną informację pochodzącą z wielu skal opisu. Potrzeba tego rodzaju modeli występuje we wszystkich dziedzinach, które operują na hierarchicznie zagnieżdżonych danych, między innymi w badaniach edukacyjnych, geograficznych oraz we współczesnej biologii systemów. Na przykład w biologii oddziaływania makromolekuł mogą być analizowane jednocześnie na poziomie pojedynczych atomów, funkcjonalnych grup atomów i całych cząsteczek. Każdy z tych poziomów wiąże się z innego rodzaju informacją i musi zostać uwzględniony przy całościowej predykcji.

Żeby wykorzystać w pełni wieloskalową informację są potrzebne wyspecjalizowane techniki modelowania. Ta rozprawa skupia się na metodach budowy wieloskalowych komitetów klasyfikatorów, w których struktura klasyfikatora odzwierciedla strukturę rozwiązywanego problemu. Proponowany jest formalny język opisu tego rodzaju systemów. Głównym pojęciem w ramach tego systemu jest *poziom opisu* definiowany jako zbiór atrybutów możliwych do zakodowania numerycznie na danym poziomie. Klasyfikacja jest wyrażona jako ciąg transformacji od cech wejściowych do wynikowej etykiety klasy. W ramach tego ciągu transformacji algorytmy uczenia maszynowego są stosowane do zdefiniowania sparametryzowanych transformacji zależnych od danych wejściowych. Różne formy wstępnej obróbki danych lub agregacji wyników stanowią istotne fazy opisywanego procesu. Pełen system klasyfikacyjny jest przedstawiany w postaci grafu obliczeń, dzięki któremu można łatwo śledzić zależności pomiędzy zbiorami danych a klasyfikatorami.

Opisy tych samych zjawisk na różnych poziomach nie muszą być równoważne. Bardzo często liczby obserwacji i atrybutów dostępnych na różnych poziomach są różne. Niektóre poziomy opisu dają szczegółową informację na temat rozkładu wartości, podczas gdy inne opisują jedynie ogólne tendencje. Pojęciowo jest to bardzo zbliżone do stopni swobody znanych ze statystyki. Większa liczba stopni swobody próbki pozwala na bardziej wiarygodną estymację parametrów rozkładu. Ma to również związek z pojęciami złożoności danych i złożoności klasyfikatora stosowanymi w kontekście uczenia maszynowego. Analiza tego rodzaju złożoności ma szczególne znaczenie dla wielopoziomowych modeli, w których stosuje się wielokrotną estymację.

W celu dokładniejszego pomiaru złożoności danych wprowadzana jest technika nazywana krzywą złożoności. Główną ideą stojącą za tą metodą jest obserwacja, że informacja zawarta w dowolnym zbiorze danych może być przybliżona poprzez jego mniejszy podzbiór. Jakość takiego przybliżenia jest mierzona odległością Hellingera.

Krzywa złożoności to wykres przedstawiający odległości Hellingera pomiędzy rozkładami prawdopodobieństwa indukowanymi przez podzbiory różnych rozmiarów a rozkładem indukowanym przez pełen zbiór. Wykazuje się, że pod pewnymi założeniami krzywa złożoności może być traktowana jako górne ograniczenie składowej błędu wariancji klasyfikatora. Eksperymenty z udziałem 81 zróżnicowanych zbiorów danych potwierdzają jej użyteczność w objaśnianiu skuteczności różnych klasyfikatorów trenowanych na tych zbiorach. Krzywa złożoności jest również efektywną heurystyką, pozwalającą dobrać optymalny rozmiar próbki danych do wnioskowania. Zastosowana do przycinania rozmiaru danych pozwala wielokrotnie skrócić czas treningu złożonych klasyfikatorów bez szkody dla ich skuteczności.

Kolejną kwestią, wynikającą z hierarchicznej natury modelowanych danych, jest konstrukcja zrównoważonej procedury ewaluacji klasyfikatora. Wykazuje się, że czysto losowy podział na część treningową i testową może doprowadzić do zaburzenia procedury. Na potrzeby rozwiązania tego problemu wieloskalowe zależności pomiędzy obiektami mogą być modelowane jako graf, który następnie jest wykorzystywany jako źródło dodatkowej informacji w procedurze podziału. Testowanych jest kilka grafowych metod podziału. Pośród nich deterministyczny algorytm zachłanny pozwala wykorzystać dostępne dane najbardziej efektywnie.

Dla zademonstrowania użyteczność opisywanej metodologii w praktycznych problemach biologicznych, opisywana jest konstrukcja wieloskalowego klasyfikatora przewidującego oddziaływania białko-białko. Wykorzystuje on informacje o bezpośrednich kontaktach reszt aminokwasowych w łańcuchach białek, pochodzące z bazy trójwymiarowych kompleksów, dla skonstruowania cech przydatnych w przewidywaniu binarnych oddziaływań białek. Takie dwupoziomowe podejście okazuje się bardziej skuteczne od standardowych metod bazujących na sekwencji białka i operujących tylko w jednej skali (AUC ROC 0.62–0.67 vs 0.56–0.57). Niestety te wyniki nie dają się powtórzyć na zbiorach większych rozmiarów o innej charakterystyce. Niska skuteczność znanych metod przewidywania oddziaływań oraz trudność metodologiczne związane z ewaluacją klasyfikatora sugerują, że sam problem przewidywania oddziaływań między białkami może być źle sformułowany.