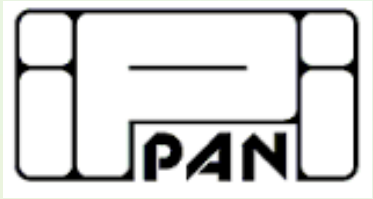
Big Data i ich eksploracja: spojrzenia z perspektywy statystyki i uczenia maszynowego



Jacek Koronacki
Instytut Podstaw Informatyki, PAN

Jerzy Stefanowski
Instytut Informatyki, Politechnika Poznańska

Poznań, 22 kwietnia 2016

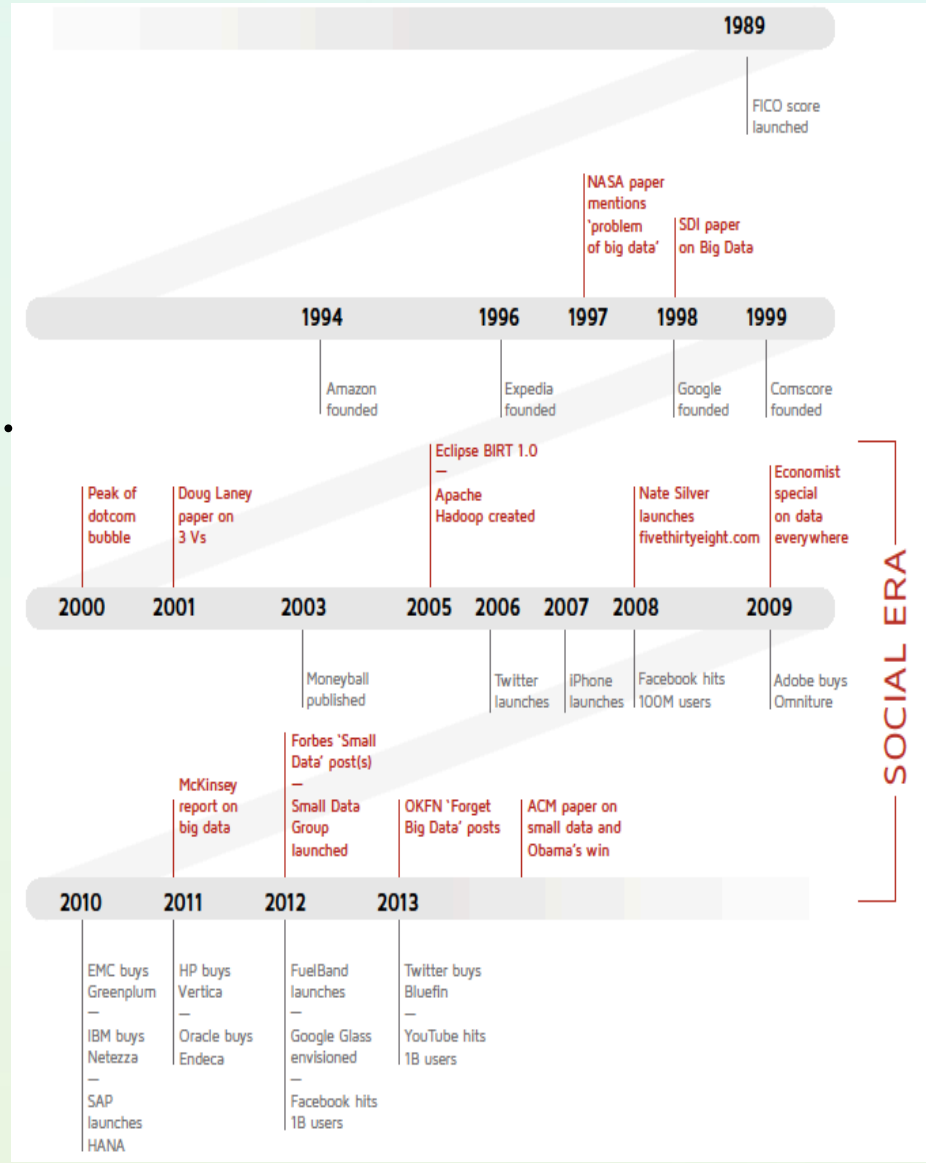
Kilka uwag historycznych

Rosnące rozmiary danych vs. Big Data

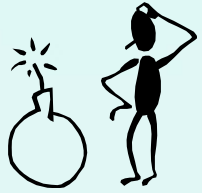
Massive Data mining, Very Large Databases – inne spojrzenia + wcześniejsze

"Big Data" – jako termin na amer. konf. pojawia się pod koniec lat 90tych:

- J.Mashey seminarium SGI, 1998
- S. Weiss, N.Indrukhya: Predictive data mining. A practical guide 1998
  - Rozdział 1.1. Big Data. – duża część rozdziału 1 dyskutuje też problemy „Massive data".
- Doświadczenia wielkich projektów badawczych – NASA



Charakterystyka Big Data – połączenie niejednorodnych i złożonych źródeł danych



3 V → „High volume, velocity and variety” [Doug Laney 2001]

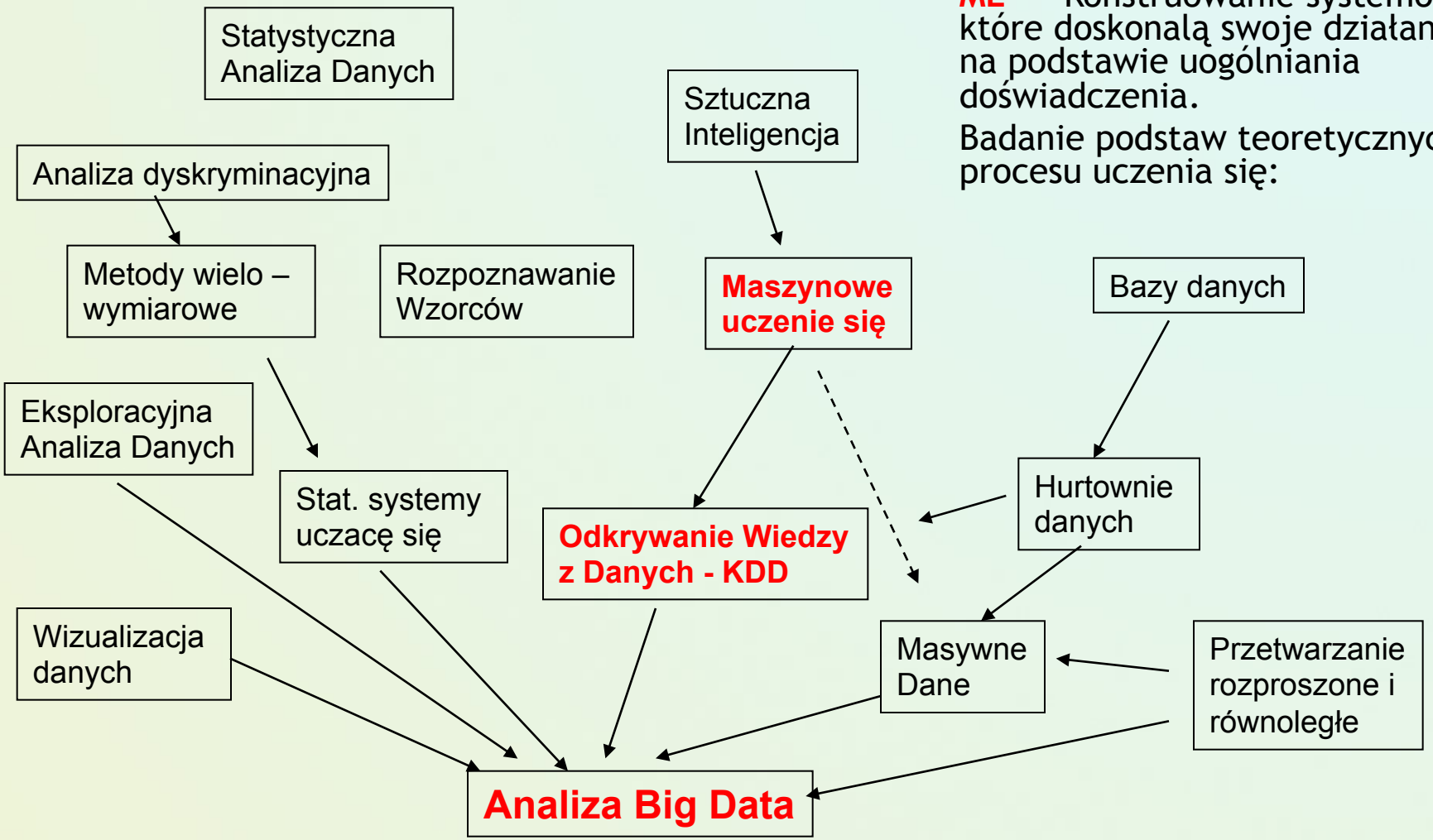
Kolejne V’s stopniowo dodawane (Veracity, Value, …)

“Big Data” – to dane, których skala, zróżnicowanie i złożoność wymaga nowych technologii i algorytmów w celu odkrycia wartościowej wiedzy … [J.Gama 2015]

HACE Theorem: Big Data starts with large-volume, **h**eterogeneous, **a**utonomous data sources with distributed and decentralized control, and seeks to explore **c**omplex and **e**volving relationships among data [Xindong Wu et al. 2013]

Spójrz: A de Mauro, M. Greco, M. Grimaldi: What is Big Data? A consensual definition and a review of key research topics. Proc. 4th conf. on Integrated Information 2014.

Powiązania dziedzin + ML vs. KDD/DM



ML --- Konstruowanie systemów, które doskonalą swoje działanie na podstawie uogólniania doświadczenia.

Badanie podstaw teoretycznych procesu uczenia się:

KDD – proces poszukiwania nowych, interesujących, potencjalnie użytecznych (i zrozumiałych) wzorców z danych – G.Piatesky-Shapiro (1989)

KDD/Data Mining vs Analiza Big Data

Zagadnienie	Tradycyjne Data Mining	Analiza Big Data
Dostęp do pamięci	Centralna pamięć operacyjna, łatwiejsze wielokrotne operacje odczyt / zapis	Dane często rozproszone Minimalizowanie zapamiętanych elementów i dostępu do nich
Architektura oblicz.	Centralna pojedyncza jednostka (skalowalna)	Rozproszone przetwarzanie Grona (clusters) słabszych komputerów
Dane	Dobra strukturalizacja (rel. DB), jednorodne, statyczne / integracja DW	Zróżnicowane źródła; brak struktury; Zmienne / dynamika i czas
Jakość danych	Dobrze przygotowane lub popr. wiele technik korekcji Udokumentowane pochodzenie Wiarygodne próbkowanie Wyselekcjonowane „dobre dane”	Słaba jakość danych, niepewność i niedokładność; Słabo dokument. pochodzenie i pre-processing; Użyteczne dane mogą być połączone z wieloma bezużytecznymi
Bezpieczeństwo i prywatność	Nie są wymagane Proste metody „anonimizacji”	Krytyczny problem Współdzielenie danych; łączenie danych
Przetwarzanie danych	Klasyczne (batch); może być off-line Brak konieczności próbkowania Prędkość nie tak krytyczna	Możliwość – wymagania on-line; szybkość; Wydajność alg. ma znaczenie Dane nie mieszczą się w pamięci Kompresja i próbkowanie danych
Analiza rezultatów	Rozwinięte metody oceny wyników oraz wizualizacji	Niebezpieczeństwa odkrycia nieznaczących rezultatów Trudności wizualizacji

Więcej w N.Japkowicz, J.Stefanowski: A Machine Learning Perspective on Big Data Analysis (2016)

Big Data – znane metody w innym kontekście?

Standard metod Data Mining

- Klasyfikacja nadzorowana
- Regresja / ANN
- Analiza skupień
- Asocjacje (zbiory częste, reguły asocjacyjne)
- Wzorce sekwencji
- Szeregi czasowe
- Wykrywanie anomalii i obserwacji nietypowych
- Statystyka opisowa
- Statystyka wielowymiarowa
- Dekompozycja macierzy (PCA, MDS,…)
- …

Klasyfikacja i predykcja

- Drzewa decyzyjne
- Reguły
- Naive Bayes
- K-NN
- Regresja logistyczna
- Sztuczne sieci neuronowe
- Analiza dyskryminacyjna
- Metoda wektorów wspierających SVM
- Zespoły klasyfikatorów
- …

Spójrz: P. Tan, M. Steinbach, V. Kumar: An Introduction to Data Mining

T.Morzy: Eksploracja danych, PWN

J.Berman: Principles of Big Data Analysis

Large Scale Machine Learning

- Large *n*, large *p*, large *k* (number of outputs)
  - Czasami inne zależności (large *p*, smaller *n*,*k*)

"The computing resources available do not grow faster than the volume of data"

Needs for linear time learning algorithms

Leon Bottou

  - Ideal O(*pn*+*kn*)

- Wielu badaczy – poprawa czasu obl. algorytmów:

  Np. Linear regression -> O(p) , k-nearest neighbour, SVM, … → O(n)

  K-means clustering, …

Spójrz do: F.Bach: Stochastic gradient methods for machine learning 2013

Michael I. Jordan, np., On statistics, computation and scalability, 2013.

A.Grey: Machine learning on massive data

Drzewa klasyfikacyjne i … (regresji …)

80's

ID3

CART

CHAID

C4.5

…

Oblique trees

90

Ensembles

Mega-induction,
Sampling, splits

Bagging

Boosting

Random Forests

Ivotes

KDD generalizations

end. 90

Data pre-sorting / reduce disc acc.

Approx. representations
Samples, Histograms

Rotation Forests

Gradient Boosted Machines

SLIQ
SPRINT
…

BOAT
CLOUDS
SPIES

Parallel versions

Incremental trees
ID4, ID5R, ITL

Massive incremental tress
Hoeffding trees, **VFDT**

MOA,
SAMOA

Stream ensembles
Online …/ block based

Data streams + drifts
VFDT -> CVFDT; **VFDTc**

Rozwój implementacji w MapReduce, Hadoop / SPARK

Big Data z punktu widzenia uczenia maszynowego

Nowe problemy badawcze

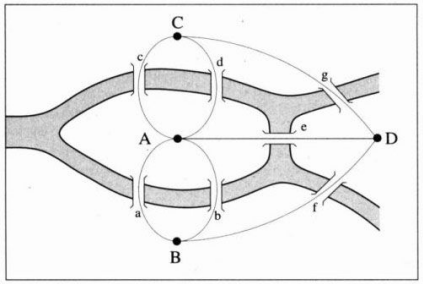
- Analiza grafów
- Social networks
- Integracja lub przetwarzanie on-line różnorodnych reprezentacji danych
- Eksploracja danych strumieniowych
- Analiza danych czasowo-rozproszonych
- Obliczenia mobilne (IoT)
- Wizualizacja danych
- Privacy data mining
- Data trust + provenance
- …

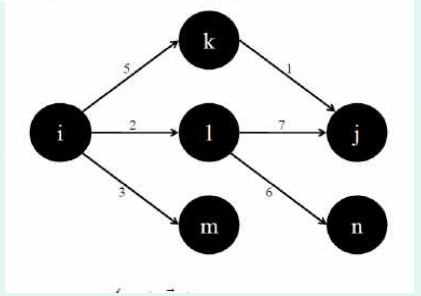
Inne problemy

- Interakcja z ekspertem
- Ocena wiedzy
- Etyka analizy Big Data
- Wpływ na społeczeństwo
- …

Sprawdź: N.Japkowicz and J.Stefanowski (Eds), Big Data Analysis: New Algorithms for a New Society, Springer (2016).

Eksploracja grafów





Grafy - Lata 80

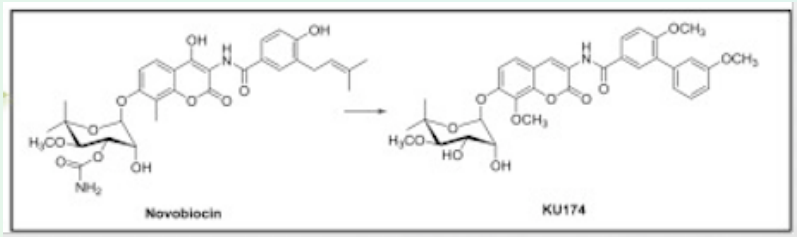
- Kilkadziesiąt wierzchołków

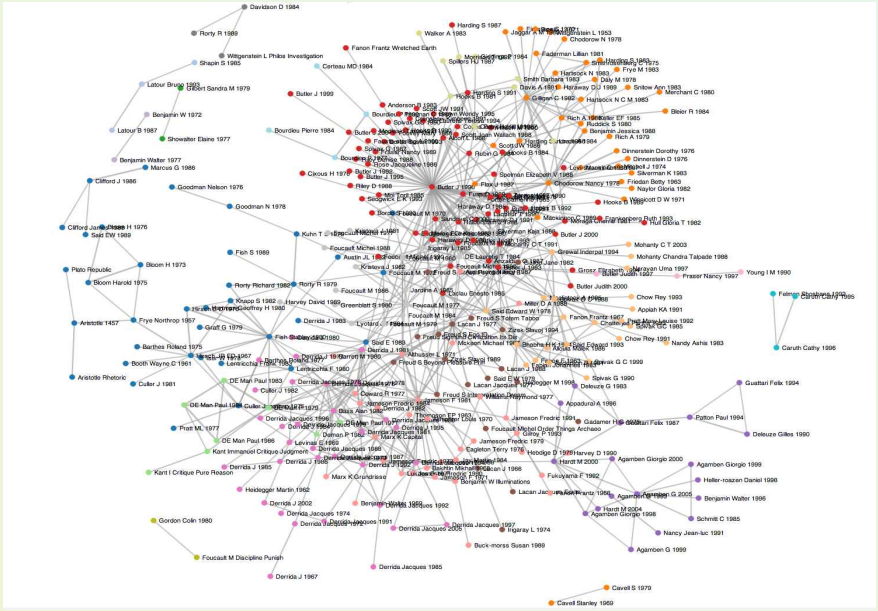
Lata 90

- Setki wierzchołków

Obecnie (Internent, social networks)

- Tysiące, setki tysięcy, miliony





Name	Weighted	Directed	Vertices	Edges
calls	✓	✓	7,786,471	33,292,508
condmat-collab	✓		17,216	110,544
dblp-cite	✓	✓	15,963	344,373
dblp-collab	✓		367,725	2,088,710
disease-g	✓		399	15,634
disease-p	✓		437	81,158
hepth-cite	✓	✓	8,249	335,028
hepth-collab	✓		8,381	40,736
huddle	✓		4,243	997,008
patents-cite	✓	✓	1,461,714	32,418,457
patents-collab	✓		1,162,227	5,448,168
sms	✓	✓	5,016,746	11,598,843

Strumienie danych (data streams)

- Źródła – generują dane w postaci nieprzerwanego strumienia
  - Trudności z przechowywaniem wszystkich elementów danych
  - Konieczność przetwarzania szybkiego strumienia w ograniczonym czasie
  - Przyrostowe działanie algorytmów
- Zmienność rozkładów danych → niestacjonarne środowiska

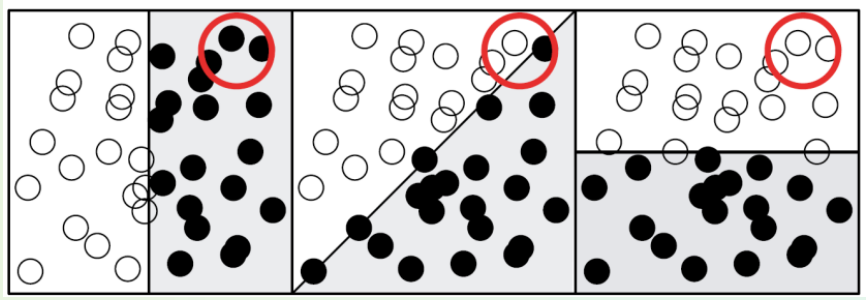
	Statyczne	Strumień
No. odczytów	wielokrotne	pojedyncze
Czas	nie jest krytyczny	ograniczony
Uzycie PAO	b. elastyczne	ograniczone
Rezultaty	dokładne	przybliżone
Przetwarz. Rozproszne	zwykle nie	tak

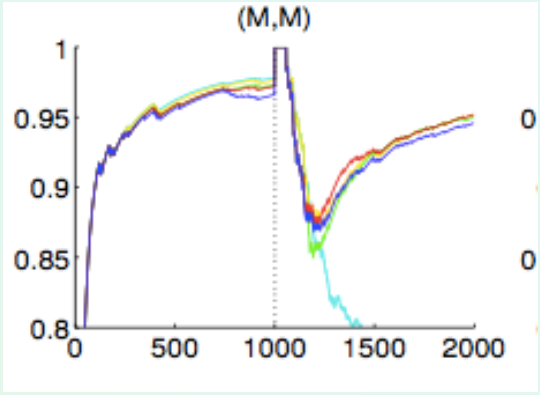
Nowe wyzwania:

- Próbkowanie danych, tworzenie zastępczych reprezentacji (histogramy)
- Grupowanie danych
- Analiza zbiorów częstych, wzorców sekwencyjnych
- Złożone reprezentacje danych

Ubiquitous Data Mining + Mobile Analytics + Internet of Things

Zmienne strumienie danych (data streams)





Zmienność danych (także definicji klas) wraz z upływem czasu tzw. concept drift

Zmiany te mają negatywny wpływ na trafność klasyfikacji

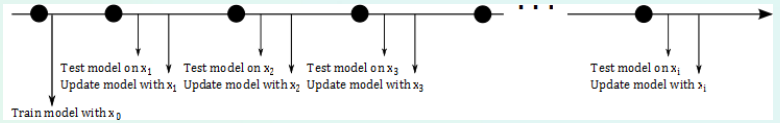
Statyczne algorytmy eksploracji danych nie mogą być stosowane!

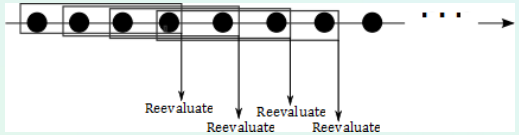
Nowe wymagania do algorytmów – wydajne obliczeniowo + zdolności reakcji na zmiany

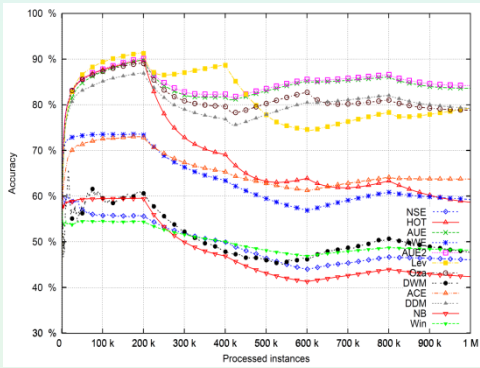
Intensywny rozwój metod wykrywania zmian oraz nowych algorytmów uczących klasyfikatory strumieniowe!

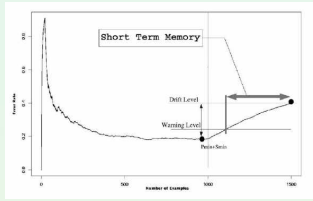
Klasyfikacja zmiennych strumieni

- Zarządzanie pamięcią i “zapominanie”
  → sliding? windows
- Inne spojrzenie na ocenę
  - Miary oceny
  - Interleaved test-then-train or prequential techniques
- Pełny lub częściowy dostęp do prawdziwych etykiet
  - Częściowo-nadzorowane podejścia (micro-clustering)
  - Strategie aktywnego uczenia się
- Wykrywanie zmian→ drift detection methods
  - Trudniej w nieetykietowanych danych









Spójrz na prace przeglądowe

- J.Gama: Knowledge discovery from data streams. CRC, (2010)
- G. Krempl, et al. Open challenges for data stream mining research. SIGKDD Explorations, 16(1):1–10 (2014).
- G.Ditzler, M.Roveri, C.Alippi, R.Polikar: Learning in nonstationary environments – a survey IEEE Comput. Intel. (2015)

Specjalizowane zespoły klasyfikatorów (ensembles)

Block-based ones:

Streaming Ensemble Algorithm (SEA) – Street & Kim 2001

Accuracy Weighted Ensemble (AWE) Wang et al 2003

- BWE - Deckert 2011, Weighted Aging Ensemble (Wozniak et al 2013)

Learn++.NSE – Polikar et al. 2011

Others , e.g., EAE (Jackowski 2014)

Recurring concepts

- CCP – Katakis et al. 2010
- RCD – Goncalvas et al. 2013
- FAE – Diaz et al 2015

Block processing also in some semi-supervised or novel class detection – Masud et al. 2009; Farid et al 2013

On-line (instance based)

WinNow, Weighted Majority Alg. – Littlestone 1988, L & Warmuth 1994

Dynamic Weighted Majority (DWM) – Kolter & Maloof 2003 → AddExp (2005)

On-line bagging and on-line boosting [Oza]

- BagADWIN,
- Leverage bagging – Bifet et al. 2007
- Using in DDD (Minku, Yao)

Hoeffding Option Trees (HOT)

UFFT (Gama at al. 2005)

ADACC – Jaber 2013

Boosting classifiers for drifting concepts – Scholtz & Klinkenberg 2007 + more

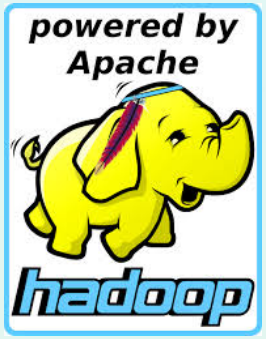
Hybrid approaches

ACE – Nishida 2009, OBWE ,

AUE → OAUE [Brzeziński, Stefanowski]

Rozwiązania obliczeniowe

- Przetwarzanie rozproszone i równoległe
  - Hadoop i MapReduce
    → Spec. software Apache (Pig, Hive,Hbase)
  - Spark
- NoSQL bazy danych (Google, Facebook, Amazon,…)
  - Google BigTable
  - Dynamo, Casandra, MongoDB,…
- Nowe środowiska programowe
  - Mahout (scalable machine learning)
  - Spark MLBase / Mlib
  - Vowpal Wabbit
  - h2o
  - MOA → SAMOA
  - …



Naturalne przyspieszenie obliczeń Random Forests / lecz nie każdy algorytm da się tak przyspieszyć

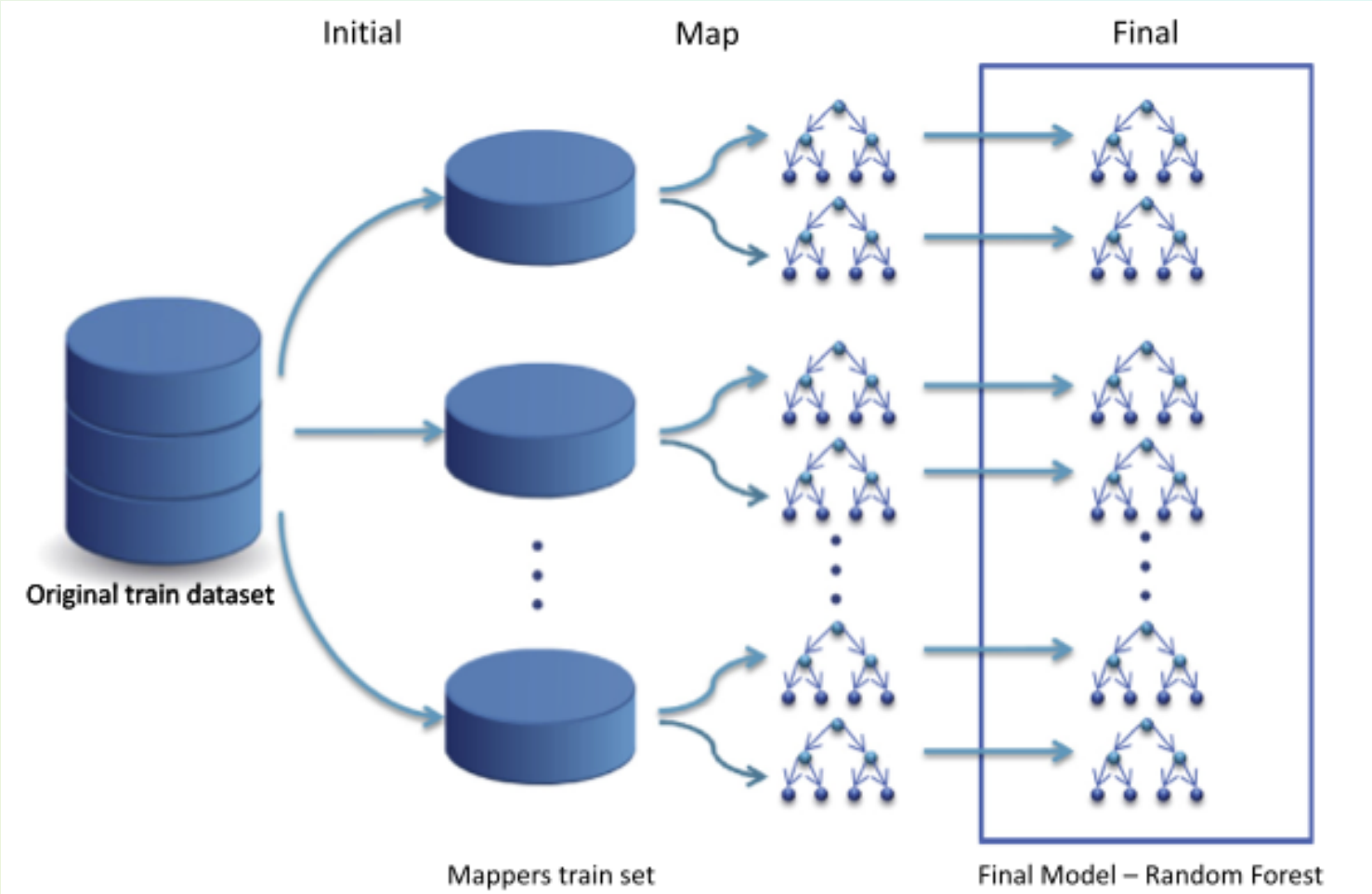
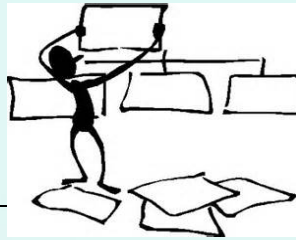


Fig. 3. A flowchart of how the building of the Random Forest is organized in Mahout.

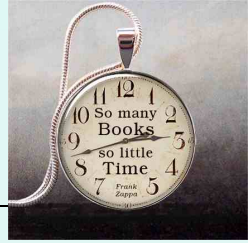
Za S.Rio in. On the use of MapReduce for imbalanced big data using Random Forest 2014

Inne aspekty algorytmów Big Data



- Konstruowanie wersji przyrostowych dla rozwiązań statycznych
- Inne metody losowania
  - Bag of little bootstraps [Kleiner in 2012]
- Nowe podejście do wstępnego przetwarzanie danych,
  - np., outliers detection; nowe metody dekompozycji macierzy [M.Jordan i in. 2012]
- Sparse, incomplete, linked data repositories
- Uczenie się z niezbalansowanych danych
- Uczenie sie częściowo nadzorowane
  - Active learning, … [Krempl 2015]
- Rozwój systemów rekomendacyjnych
- Intensywniejsze wykorzystywanie repozytorium wiedzy dziedzin.
- Privacy data mining [Matwin]
- Aspekty etyczne wykorzystania algorytmów
- …

Czy zmieniamy zasady analizy danych?



Mit Big Data – „N = całość / N = All”

- Nawet popularne media społecznościowe obejmują tylko ograniczoną część ludności - Ograniczenia danych Twitter (Tufekci)
- „N = całość” - często pobożne życzenie lub iluzja, a nie rzeczywistość!

Lekcje ze statystycznej analizy danych nie mogą być zapomniane!

Doświadczenia statystyczne nt. reprezentatywności i obciążenia prób

- Niepoprawnie pozyskane próby prowadzą do błędnych predykcji
- Im więcej danych, tym zwiększona szansa na pozorne zależności i pomyłki „false positive” [M.Jordan; T.Hastie]

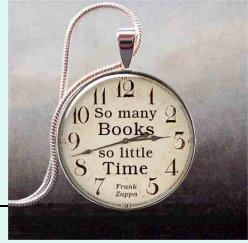
Pomijanie dodatkowych źródeł wiedzy i powierzchowność analiz na podstawie zbyt prostych danych

Analiza stopnia zażyłości osób na podstawie połączeń telefonicznych (D. Boyd)

Więcej dyskusji w:

Boyd, D., Crawford, K.: Critical questions for Big Data. Information, Communication and Society (2012).
Harford, T.: Big Data: are we making a big mistake? Financial Times, March 28, 2014.
Tufekci, Z.: Big Data: Pitfalls, methods and concepts for an emergent field. SSRN (March 2013)

Transparentność analizy Big Data



Zagrożenia stosowania algorytmów wobec ludzi

- Wiemy co jutro będziesz robić
  - Profilowanie osób i działania wobec osób przed ich akcją

- Ponadto problem dostępu danych
  - Kto wykorzystuje i do czego moje dane?
  - Kto jest właścicielem danych?
  - Powtórne wykorzystywanie danych.

- Przejrzystość wykorzystywania danych i stosowanych algorytmów
  - Badacze vs firmy komercyjne
  - Lecz kto ma to kontrolować?

"Big Data ethics is huge, messy and personal topic, which cannot be easily settled" [Kord Davis]

Big Data z perspektywy statystycznych systemów uczących się

- Metody statystyczne stanowią podzbiór ogromnej wagi i niezbywalny metod uczenia maszynowego
- Jakkolwiek np. statystyczne meta-analizy są rozwijane, to w obszarze Big Data rozwój metod uczenia statystycznego dotyczy głównie analizy danych masywnych z jednego źródła
- Niezbywalność metod uczenia statystycznego wynika z mocy wyjaśnień/rozwiązań, których dostarczają, oraz ich metodologicznej przejrzystości i czystości

Big Data z perspektywy statystycznych systemów uczących się, c.d.

- Statystyk dobrze wie, że potrzebuje powtarzalnych danych i że dane, którymi dysponuje, w najlepszym razie mogą odpowiedzieć tylko na pewne, dobrze określone pytania – narzuca to specyficzny reżim analizowania tych danych

- Metody statystycznego uczenia dostarczają modeli przyczynowo-skutkowych, gdy to osiągalne (możliwe), i pozwalają zadowolić się podejściem algorytmicznym/predykcyjnym/behawiorystycznym, gdy głębsze poznanie jest nieosiągalne

Big Data z perspektywy statystycznych systemów uczących się, c.d.

- Paradoksalnie, to rozwój technik komputerowych pozwolił statystykom wyjść z Tukeyowskiego więzienia eksploracyjnej analizy danych (EDA), nad bramą którego widniał napis: „Pozwólmy przemówić danym – niech same mówią za siebie”

- W 1979 roku nie tak słynny jak John Tukey, za to bardziej radykalny i też znakomity statystyk, William Eddy napisał:

”The data analytic method denies the existence of "truth"; the only knowledge is empirical.

[...] If we can make without models, I think we should.”

Big Data z perspektywy statystycznych systemów uczących się, c.d.

- Dzisiaj, w powodzi Big Data, niektórzy badacze powiadają właściwie to samo: mając Big Data nie musimy już wyjaśniać, wystarczy znać korelacje, które pozwalają przewidywać

  por. odnośne uwagi w: N.Japkowicz and J.Stefanowski (Eds.), Big Data Analysis: New Algorithms for a New Society, Springer (2016), rozdz. 1 i 2

- Każdy pretekst jest dobry, by opowiedzieć się za głupotą, ale lepiej chcieć pozostać rozumnym i starać się rozumieć (a nie tylko przewidywać)

Big Data z perspektywy statystycznych systemów uczących się, c.d.

Gdy mówimy o metodach statystycznego uczenia w kontekście danych masywnych warto najpierw rozróżnić dwie sytuacje:

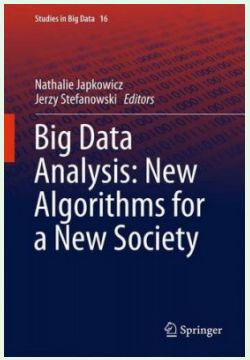
- wielkiej liczby obserwacji (*large n*) i małej liczby zmiennych (cech) objaśniających (opisujących każdą obserwację – *small p*)
- wielkiej liczby cech opisujących każdą obserwację (*large p*) i niekiedy jednocześnie małej liczby obserwacji (*small n large p*)

UWAGA: Od tego slajdu ograniczamy się do problemów uczenia pod nadzorem

Big Data z perspektywy statystycznych systemów uczących się, c.d.

W przypadku wielkiej liczby obserwacji (*large n*) i małej liczby zmiennych (cech) objaśniających niezbędne i stosunkowo łatwe do przeprowadzenia może być podzielenie zadania na mniejsze podzadania, rozwiązanie podzadań i na tej podstawie zbudowanie rozwiązania dla całego zadania

por. odnośne uwagi w: N.Japkowicz and J.Stefanowski (Eds.), Big Data Analysis: New Algorithms for a New Society, Springer (2016), podrozdz. 2.4



Big Data z perspektywy statystycznych systemów uczących się, c.d.

W przypadku zadań o wielkiej liczbie cech i zwłaszcza problemów znanych jako *small n large p problems* można ich omówienie podzielić na trzy (jakkolwiek nierozłączne!) części:

- metody oparte na podejściu Monte Carlo
- metody z regularyzacją (karą za złożoność modelu)
- metody bayesowskie

UWAGA: Lapidarne omówienie zasygnalizowanych trzech typów metod uczenia statystycznego oraz krótki wykaz literatury można zaleźć w oddzielnej prezentacji


Dziękujemy za uwagę

Pytania lub komentarze?

Kontakt:
Jacek.Koronacki@ipipan.waw.pl
Jerzy.Stefanowski@cs.put.poznan.pl