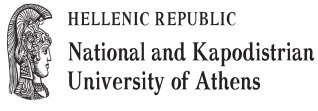
Solving Complex Machine Learning Problems with Ensemble Methods
ECML/PKDD 2013 Workshop

Ioannis Katakis, Daniel Hernández-Lobato, Gonzalo Martínez-Muñoz and Ioannis Partalas

National and Kapodistrian University of Athens
Universidad Autónoma de Madrid
Université Joseph Fourier



September 27th, 2013

Introduction to Ensemble Methods

- Deal with the construction and combination of multiple learning models
- Goal: obtain more accurate and robust predictions than single models
- Useful to tackle many learning problems of practical interest:
  - Recommendation systems [Koren and Bell, 2011]
  - Weather forecasting [Gneiting and Raftery, 2005]
  - Real-time human pose recognition [Shotton et al., 2011]
  - Feature selection [Abeel et al., 2010]
  - Active Learning [Abe and Mamitsuka, 1998]
  - Reverse-engineering of biological networks [Marbach et al., 2009]
  - Concept drift [Wang et al., 2003]
  - Credit card fraud detection [Bhattacharyya et al., 2011].

Ensemble Approach: there and back again

- The combination of opinions is rooted in the culture of humans
- Formalized with the *Condorcet Jury Theorem*:

Given a jury of voters

Assume independent errors. Let p be the prob. of each being correct and L the prob. of the jury to be correct.

- $L \to 1$, for all $p > 0.5$ as the number of voters increases

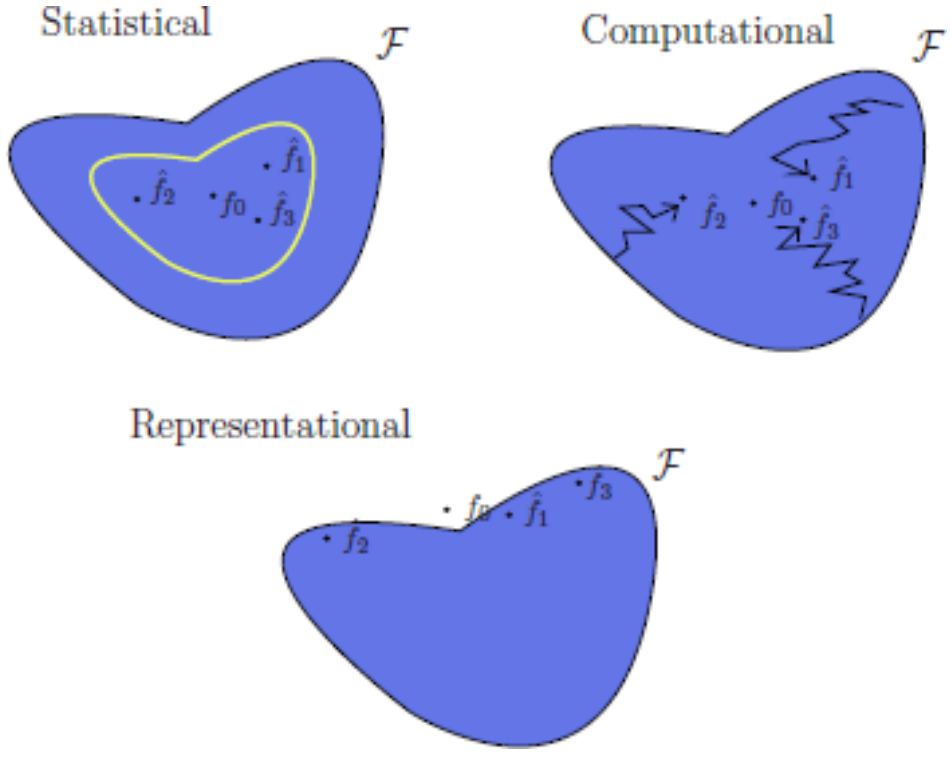
Nicolas de Condorcet (1743-1794),

French mathematician

Why to use ensembles?

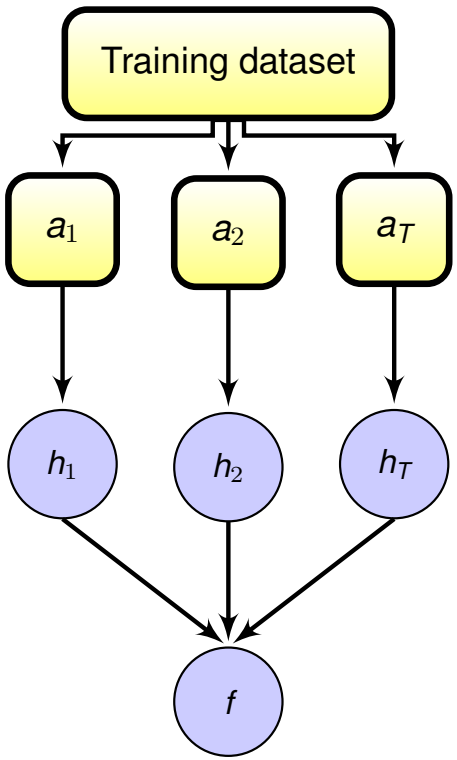
Three main reasons [Dietterich, 2000]:

- Statistical
  - Not sufficient data to find the optimal hypothesis
  - Many different hypothesis with limited data
- Representational
  - Unknown functions may not be present in the hypotheses space
  - A combination of present hypotheses may expand it
- Computational
  - Algorithms may get stuck in local minima



Ensemble framework

- A training dataset $D = \{(x_n, y_n)\}_{n=1}^{N}$
- A set of inducers $A_T = \{a_i(\cdot)\}_{i=1}^{T}$
- A set of models $H_T = \{h_i(\cdot)\}_{i=1}^{T}$
  - For classification: $h_i : \mathcal{X} \mapsto \mathcal{Y}$, $\mathcal{Y} = \{1 \dots K\}$ for K classes
- An aggregation function f
  - e.g. $f(x, H) = \frac{1}{T} \sum_{i=1}^{T} h_i(x)$



Particular Details of Ensemble Methods

- Ensemble construction
  - Homogeneous Ensembles
    - Different executions of the same learning algorithm
    - Manipulation of data
    - Injecting randomness into the learning algorithm
    - Manipulation of the features
  - Heterogeneous Ensembles
    - Different learning algorithms
- Diversity
  - Plays a key role on ensemble learning
  - No single definition of diversity
- Combination methods
  - Majority voting
  - Weighted Majority voting
  - Stacked Generalization
- Ensemble Pruning

Success Story 1: Netflix prize challenge

- Dataset: 5-star rating on 17770 movies
- 480189 users

Belkor's Pragmatic Chaos

Blended hundreds of models from three teams

Ensemble

Used variant of Stacking

Leaderboard

Showing Test Score. Click here to show quiz score

Display top 20 leaders.

Rank	Team Name	Best Test Score	% Improvement	Best Submit Time
Grand Prize - RMSE = 0.8567 - Winning Team: BellKor's Pragmatic Chaos				
1	BellKor's Pragmatic Chaos	0.8567	10.06	2009-07-26 18:18:28
2	The Ensemble	0.8567	10.06	2009-07-26 18:38:22
3	Grand Prize Team	0.8582	9.90	2009-07-10 21:24:40
4	Opera Solutions and Vandelay United	0.8588	9.84	2009-07-10 01:12:31
5	Vandelay Industries !	0.8591	9.81	2009-07-10 00:32:20
6	PragmaticTheory	0.8594	9.77	2009-06-24 12:06:56
7	BellKor in BigChaos	0.8601	9.70	2009-05-13 08:14:09
8	Dace	0.8612	9.59	2009-07-24 17:18:43
9	Feeds2	0.8622	9.48	2009-07-12 13:11:51
10	BigChaos	0.8623	9.47	2009-04-07 12:33:59
11	Opera Solutions	0.8623	9.47	2009-07-24 00:34:07
12	BellKor	0.8624	9.46	2009-07-26 17:19:11

Success Story 2: KDD cup

- Annual data mining competition[1]
- KDD cup 2013: Predict papers written by given author.
- KDD cup 2009: Customer relationship prediction.

KDD cup 2013

The winning team used Random Forest and Boosting among other models combined with regularized linear regression.

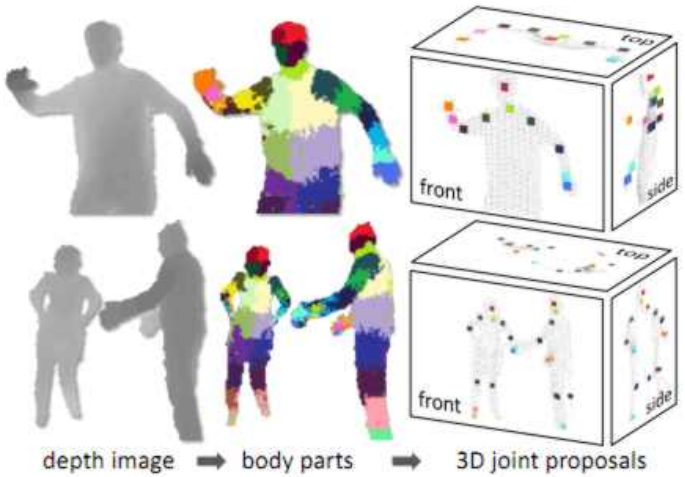
KDD cup 2009

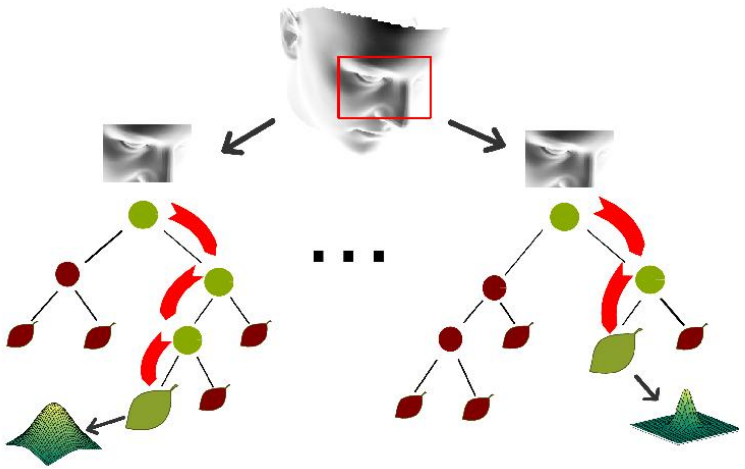
- Library of up to 1000 heterogeneous classifiers.
- Ensemble pruning to reduce the size.

[1] http://www.sigkdd.org/kddcup/index.php

Success Story 3: Microsoft Xbox Kinect

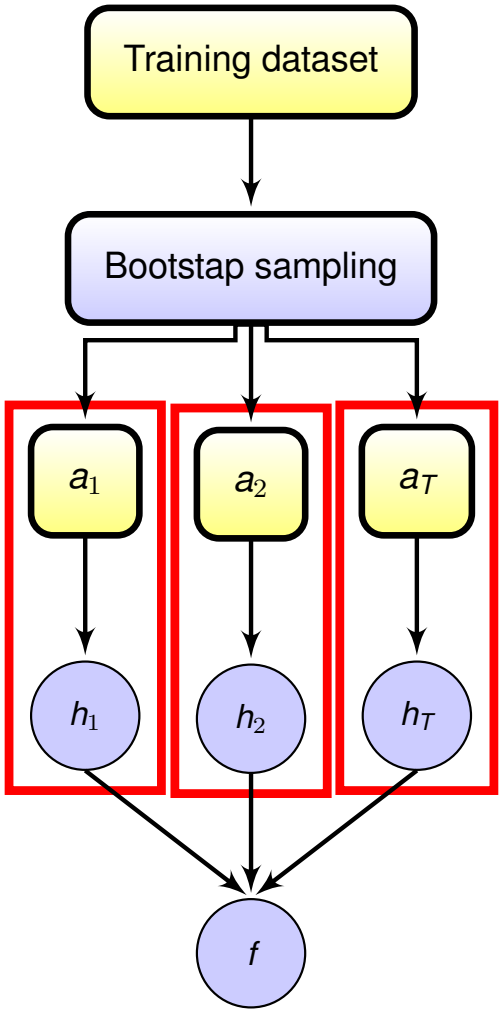
- Computer Vision
- Classify pixels into body parts (leg, head, etc)
- Use *Random Forests!* [Shotton et al., 2011]





Large Scale Ensembles

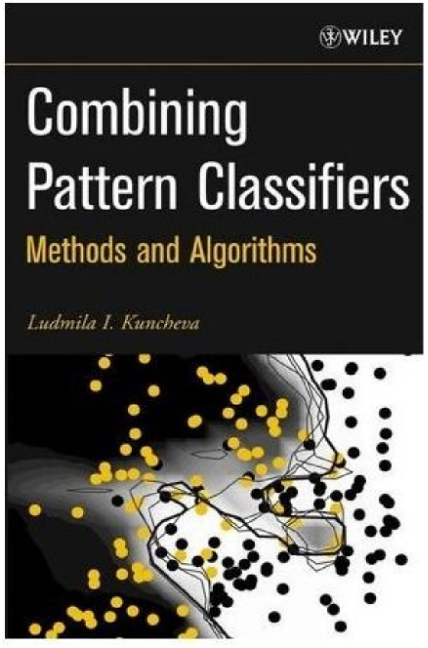
- Ensembles are well suited for large-scale problems
- Training is easily parallelized
- Non-sequential algorithms can be invoked
  - e.g. Bagging and Random Forests
- Ensembles can be coupled with frameworks for distributed computing
  - MapReduce (Google), Hadoop (Apache, open source)
  - Mahout: machine learning and data mining library
  - Pig: high-level platform for Hadoop programs

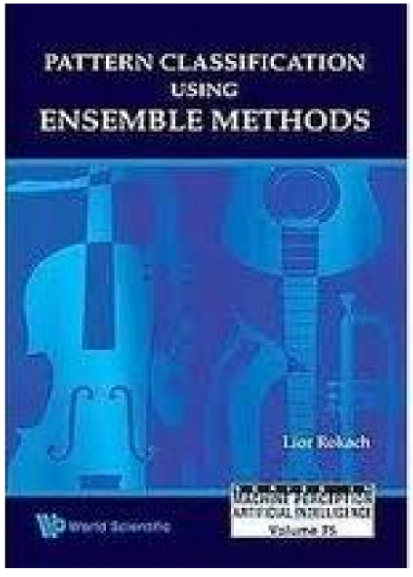


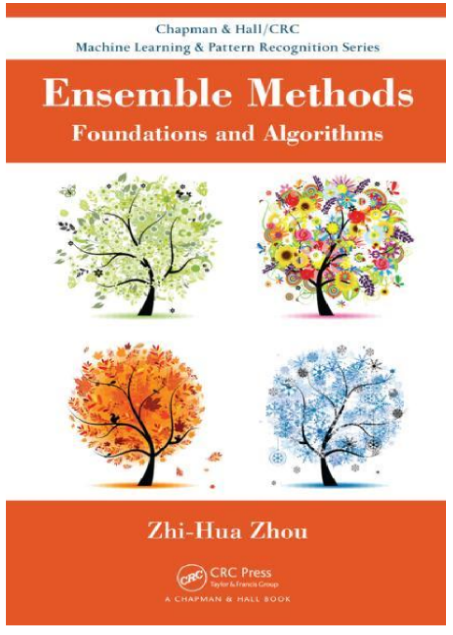
Examples of these include [Basilico et al., 2011, Lin and Kolcz, 2012].

Books and Tutorials

- Kuncheva, 2004
- L. Rokach, 2009
- Z.H. Zhou, 2012







- Ensemble-based classifiers [Rokach, 2010]
- Ensemble-methods: a review [Re and Valentini, 2012]

Advanced Topics in Ensemble Learning ECML/PKDD 2012 Tutorial[2]

[2] https://sites.google.com/site/ecml2012ensemble/

Schedule of the Workshop

10:45 - 12:15 - Session A

- COPEM - Overview
- Invited talk by Prof. Pierre Dupont
- Local Neighborhood in Generalizing Bagging for Imbalanced Data

12:15 - 13:45 - Lunch break

13:45 - 15:15 - Session B

- Anomaly Detection by Bagging
- Efficient semi-supervised feature selection by an ensemble approach
- Feature ranking for multi-label classification using predictive clustering trees
- Identification of Statistically Significant Features from Random Forests

15:15 - 15:45 - Coffee Break

15:45 - 17:15 - Session C

- Prototype Support Vector Machines: Supervised Classification in Complex Datasets.
- Software Reliability prediction via two different implementations of Bayesian model averaging.
- Multi-Space Learning for Image Classification Using AdaBoost and Markov Random Fields.
- An Empirical Comparison of Supervised Ensemble Learning Approaches.

17:15 - 17:30 - Coffee Break

17:30 - 19:00 - Session D

- Clustering Ensemble on Reduced Search Spaces
- An Ensemble Approach to Combining Expert Opinions
- Discussion and Conclusions

Some numbers...

Submissions

- Submitted: 22 papers
- Accepted: 11 papers
- Ratio: 50%

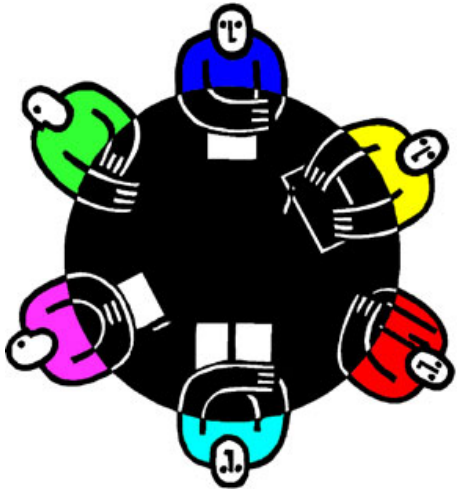
Reviews

- Each paper got at least 2 reviews (16 papers).
- Some papers got 3 reviews (6 papers).

Authors from *13* different countries

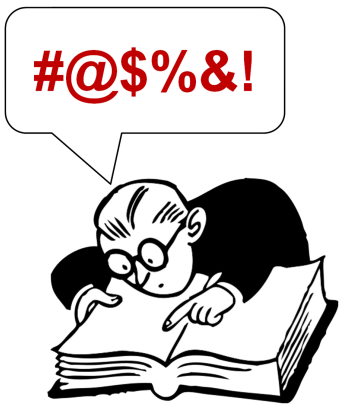
Thanks: Programme Committee!

Massih-Reza Amini	University Joseph Fourier (France)
Alberto Suárez	Universidad Autónoma de Madrid (Spain)
José M. Hernández-Lobato	University of Cambridge (United Kingdom)
Christian Steinruecken	University of Cambridge (United Kingdom)
Luis Fernando Lago	Universidad Autónoma de Madrid (Spain)
Jérôme Paul	Université catholique de Louvain (Belgium)
Grigorios Tsoumakas	Aristotle University of Thessaloniki (Greece)
Eric Gaussier	University Joseph Fourier (France)
Alexandre Aussem	University Claude Bernard Lyon 1 (France)
Lior Rokach	Ben-Gurion University of the Negev (Israel)
Dimitrios Gunopulos	National and Kapodistrian Univ. of Athens (Greece)
Ana M. González	Universidad Autσnoma de Madrid (Spain)
Johannes Furnkranz	TU Darmstadt (Germany)
Indre Zliobaite	Aalto University (Finland)
José Dorronsoro	Universidad Autónoma de Madrid (Spain)
Rohit Babbar	University Joseph Fourier (France)
Jesse Read	Universidad Carlos III de Madrid (Spain)



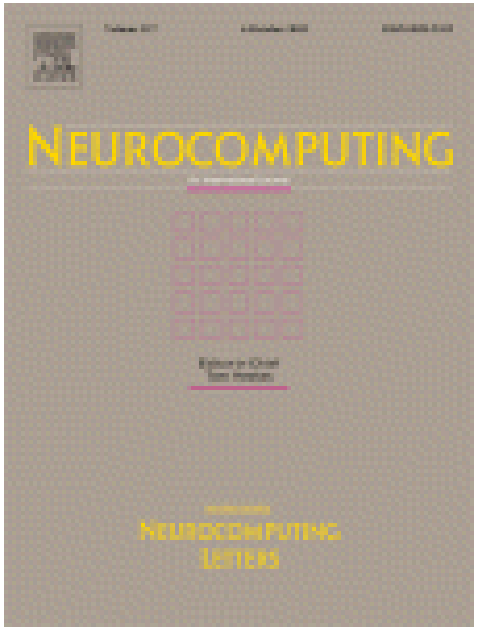
Thanks: External Reviewers!

Aris Kosmopoulos	NCSR “Demokritos” (Greece)
Antonia Saravanou	National and Kapodistrian Univ. of Athens (Greece)
Bartosz Krawczyk	Wrocław University of Technology (Poland)
Newton Spolaôr	Aristotle University of Thessaloniki (Greece)
Nikolas Zygouras	National and Kapodistrian Univ. of Athens (Greece)
Dimitrios Kotsakos	National and Kapodistrian Univ. of Athens (Greece)
George Tzanis	Aristotle University of Thessaloniki (Greece)
Dimitris Kotzias	National and Kapodistrian Univ. of Athens (Greece)
Efi Papatheocharous	Swedish Institute of Computer Science (Sweden)



Special Issue in Neurocomputing

After the workshop a selection of the presented papers will be invited to submit an extended and revised version for a **Special Issue of the Neurocomputing journal**.



Naoki Abe and Hiroshi Mamitsuka. Query learning strategies using boosting and bagging. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pages 1–9, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc. ISBN 1-55860-556-8.

Thomas Abeel, Thibault Helleputte, Yves Van de Peer, Pierre Dupont, and Yvan Saeys. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics*, 26(3):392–398, 2010.

Justin Basilico, Arthur Munson, Tamara Kolda, Kevin Dixon, and Philip Kegelmeyer. Comet: A recipe for learning and using large ensembles on massive data. In *IEEE International Conference on Data Mining*, 2011.

Siddhartha Bhattacharyya, Sanjeev Jha, Kurian Tharakunnel, and J. Christopher Westland. Data mining for credit card fraud: A comparative study. *Decis. Support Syst.*, 50:602–613, February 2011. ISSN 0167-9236.

Thomas G. Dietterich. Ensemble methods in machine learning. In *Multiple Classifier Systems: First International Workshop*, pages 1–15, 2000.

Tilmann Gneiting and Adrian E. Raftery. Weather Forecasting with Ensemble Methods. *Science*, 310(5746):248–249, October 2005.

Yehuda Koren and Robert M. Bell. Advances in collaborative filtering. In *Recommender Systems Handbook*, pages 145–186. 2011.

Jimmy Lin and Alek Kolcz. Large-scale machine learning at twitter. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, pages 793–804, 2012.

Daniel Marbach, Claudio Mattiussi, and Dario Floreano. Combining Multiple Results of a Reverse Engineering Algorithm: Application to the DREAM Five Gene Network Challenge. *Annals of the New York Academy of Sciences*, 1158:102–113, 2009.

Matreo Re and Giorgio Valentini. Ensemble methods: a review. In *Advances in Machine Learning and Data Mining for Astronomy*, pages 563–594. Chapman and Hall Data Mining and Knowledge Discovery Series, 2012.

Lior Rokach. Ensemble-based classifiers. *Artificial Intellignce Review*, 33(1-2):1–39, 2010.

Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake. Real-Time Human Pose Recognition in Parts from Single Depth Images. June 2011.

Haixun Wang, Wei Fan, Philip S. Yu, and Jiawei Han. Mining concept-drifting data streams using ensemble classifiers. In *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 226–235. ACM Press, 2003.

Let the workshop begin!