UNIVERSITAS SCIENTIARUM SZEGEDIENSIS

Research Group on Artificial Intelligence

Pletykaalapú gépi tanulás teljesen elosztott környezetben

Hegedűs István

Jelasity Márk
témavezető

Szegedi Tudományegyetem
MTA-SZTE Mesterséges Intelligencia Kutatócsopot

Motiváció

- Az adat adatközpontokban gyűlik
- Költséges tárolás és adatfeldolgozás
  - karbantartás, infrastruktúra, biztonság
- Korlátozott hozzáférés
  - még kutatók számára is
- De az adatot az eszközeink állítják elő

Motiváció – ML Alkalmazások

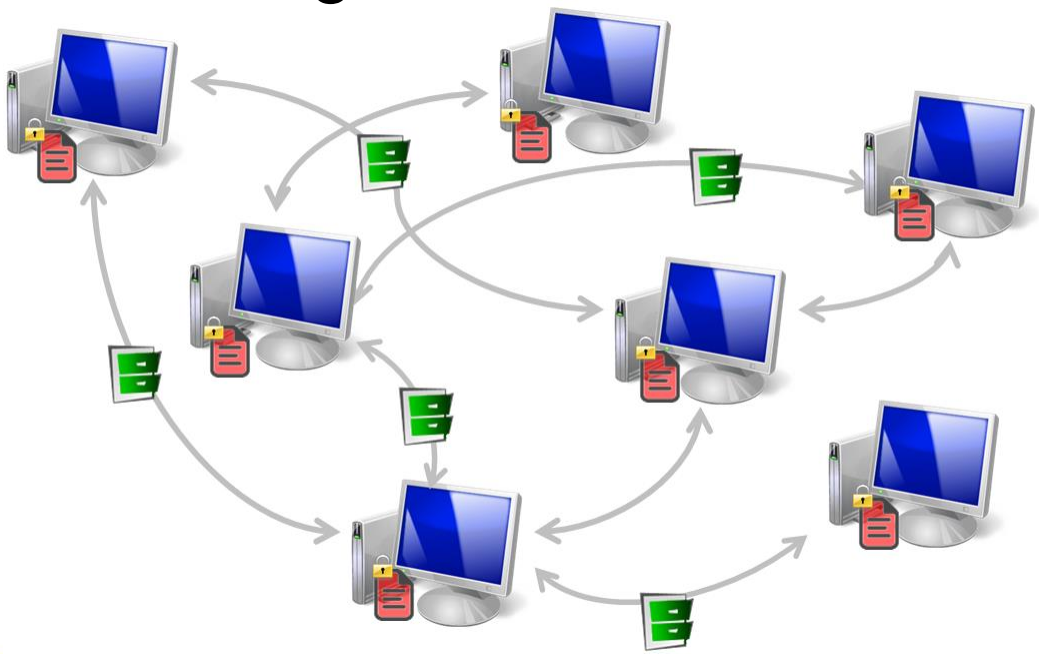
- Személyre szabott lekérdezések
- Ajánlórendszerek
- Dokumentum klaszterezés
- Spam szűrés
- Kép szegmentálás

Pletykaalapú tanulás

- ML általában egy optimalizálási probléma
- A lokális adat nem elegendő

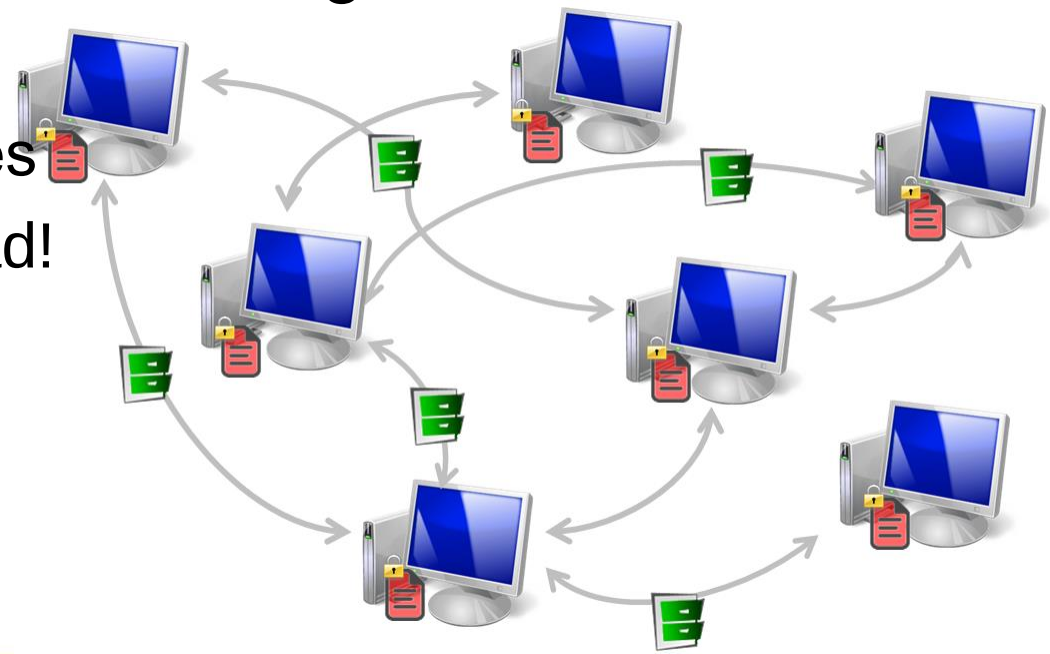
Pletykaalapú tanulás

- ML általában egy optimalizálási probléma
- A lokális adat nem elegendő
- A modellt a eszközök küldözgetik és frissítik



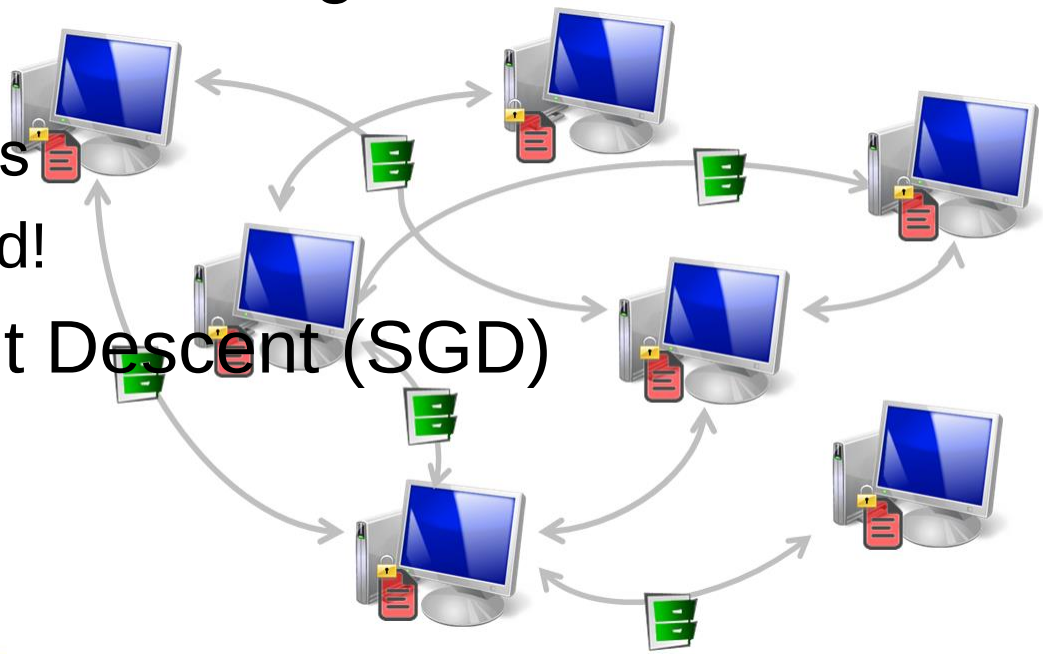
Pletykaalapú tanulás

- ML általában egy optimalizálási probléma
- A lokális adat nem elegendő
- A modellt a eszközök küldözgetik és frissítik
  - Véletlen séta
  - Példánkénti frissítés
  - Adat helyben marad!



Pletykaalapú tanulás

- ML általában egy optimalizálási probléma
- A lokális adat nem elegendő
- A modellt a eszközök küldözgetik és frissítik
  - Véletlen séta
  - Példánkénti frissítés
  - Adat helyben marad!
- Stochastic Gradient Descent (SGD)



SGD

- Célfüggvény

$$w = \arg\min_w J(w) = \frac{1}{n}\sum_{i=1}^{n} \ell(f_w(x_i), y_i) + \frac{\lambda}{2}\|w\|^2$$

SGD

- Célfüggvény

$$w = \arg\min_w J(w) = \frac{1}{n}\sum_{i=1}^{n} \ell(f_w(x_i), y_i) + \frac{\lambda}{2}\|w\|^2$$

- Gradiens módszer

$$w_{t+1} = w_t - \eta_t\left(\frac{\partial J}{\partial w}\right)$$

$$= w_t - \eta_t\left(\lambda w + \frac{1}{n}\sum_{i=1}^{n} \nabla \ell(f_w(x_i), y_i)\right)$$

SGD

- Célfüggvény

$$w = \arg\min_w J(w) = \frac{1}{n}\sum_{i=1}^{n} \ell(f_w(x_i), y_i) + \frac{\lambda}{2}\|w\|^2$$

- Gradiens módszer

$$w_{t+1} = w_t - \eta_t\left(\frac{\partial J}{\partial w}\right)$$

$$= w_t - \eta_t\left(\lambda w + \frac{1}{n}\sum_{i=1}^{n} \nabla \ell(f_w(x_i), y_i)\right)$$

- SGD, az adat online feldolgozható (példánként)

$$w_{t+1} = w_t - \eta_t(\lambda w + \nabla \ell(f_w(x_i), y_i))$$

SGD

- Célfüggvény

$$w = \arg\min_w J(w) = \frac{1}{n}\sum_{i=1}^{n} \ell(f_w(x_i), y_i) + \frac{\lambda}{2}\|w\|^2$$

- Gradiens módszer

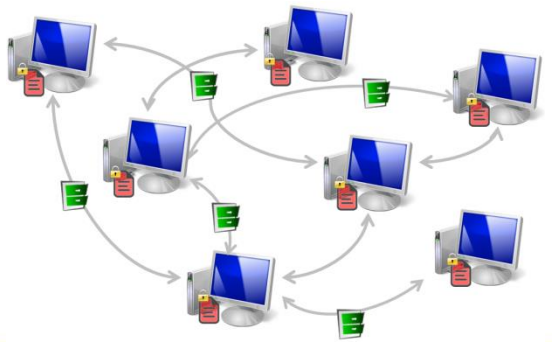
$$w_{t+1} = w_t - \eta_t\left(\frac{\partial J}{\partial w}\right)$$

$$= w_t - \eta_t\left(\lambda w + \frac{1}{n}\sum_{i=1}^{n} \nabla \ell(f_w(x_i), y_i)\right)$$

- SGD, az adat online feldolgozható (példánként)

$$w_{t+1} = w_t - \eta_t(\lambda w + \nabla \ell(f_w(x_i), y_i))$$

- Pletykaalapú tanulás

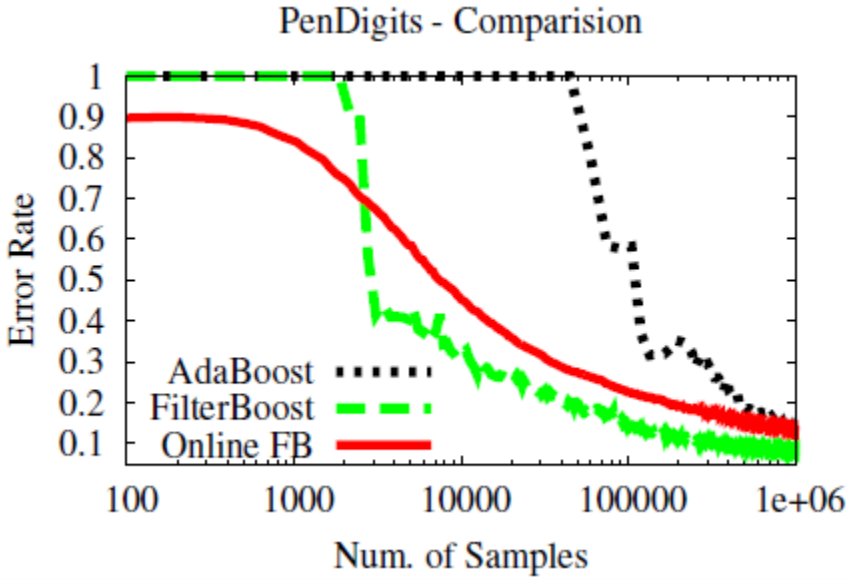


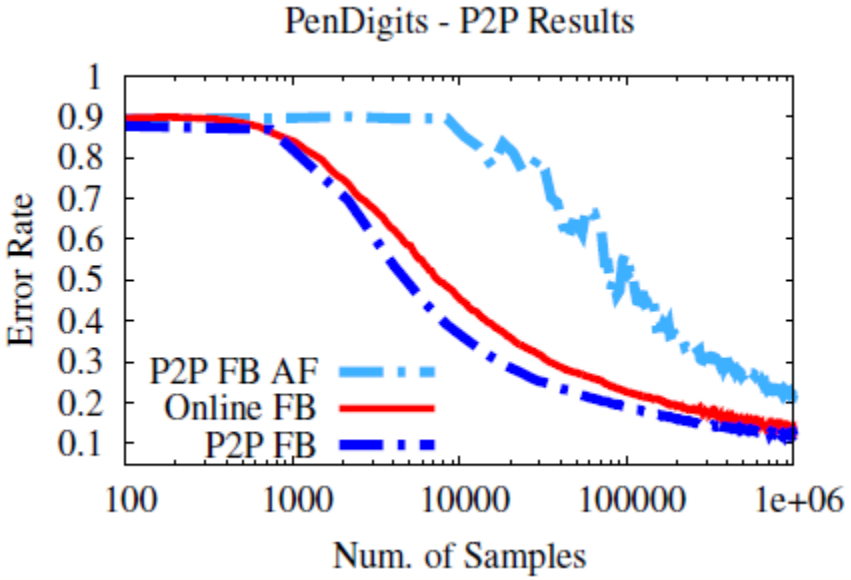
Pletykaalapú tanulás

- SGD-alapú gépi tanuló módszerek használhatók
  - Logistic Regression
  - Support Vector Machines
  - Perceptron
  - Artificial Neural Networks
- Tanító adat soha sem hagyja el az eszközt
- A tanult modell lokálisan használható, további kommunikációs költség nélkül

Boosting

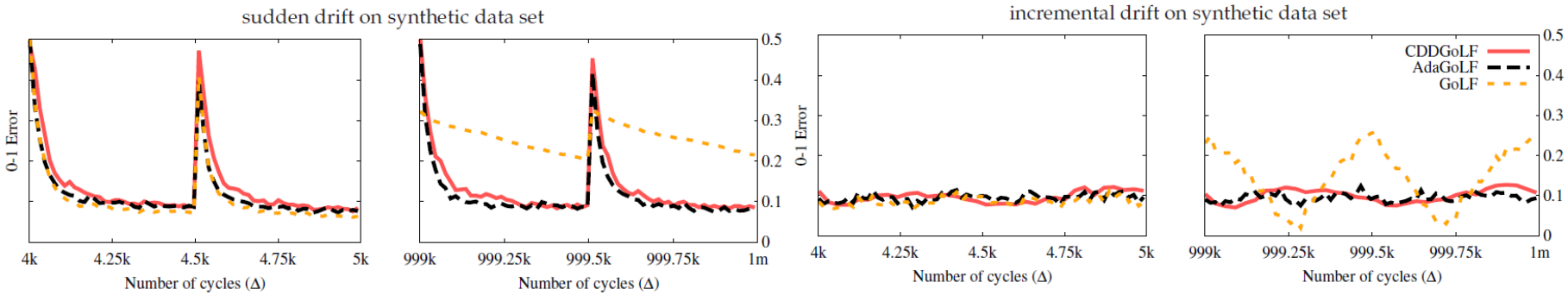
- Boosting módszer online tanulók segítségéve
- Online FilterBoost algoritmus
- Versenyképes az AdaBoost-hoz képes





Fogalomsodródás kezelése

- Két adaptív tanuló módszer
  - Modell életkor eloszlás karbantartásával
  - Modell teljesítmény monitorozással
- Fogalomsodródás kezelés és detekció



Szinguláris felbontás

- SGD alapú alacsony rangú mátrix közelítés

$$J(X,Y) = \frac{1}{2}\|A - XY^T\|_F^2 = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n}(a_{ij} - \sum_{l=1}^{k} x_{il}y_{jl})^2$$

- Egy változat, amely az SVD-hez konvergál
- Felhasználható
  - Ajánlórendszerekhez
  - Dimenzió redukcióhoz
- Az érzékeny adat nem hagyja el az eszközt
- IEEE P2P'14 legjobb cikk díj

Konklúzió

- Egy módszer lett ajánlva a teljesen elosztott gépi tanulás megvalósítására
- Egy pletykaalapú keretrendszer lett bemutatva különféle tanuló algoritmusokkal
  - Logistic regression, SVM, Perceptron, Boosting, SVD
- A fogalomsodródás kezelésének megoldásával

Kapcsolódó publikációk

	3. fejezet	4. fejezet	5. fejezet	6. fejezet
CCPE 2013 [6]	●	○	○	○
EUROPAR 2012 [3]	○	●		
SASO 2012 [4]	○		●	
SISY 2012 [2]	○		●	
ACS 2013 [5]	○		●	
P2P 2014 [9]	○			●
EUROPAR 2011 [1]	○			
ICML 2013 [7]	○			
ESANN 2014 [8]	○			
TIST 2016 [11]	○			○
PDP 2016 [12]	○			
PDP 2016 [10]	○			○

Kérdések (Alberto Montresor)

What are the advantages of executing your approach not in completely decentralized systems (like P2P networks), but instead in a cluster of distributed machines. This should be answered for all the proposed techniques.

Kérdések (Kiss Attila) I.

In these algorithms, nodes exchange model parameters. While this is better than sharing personal data, it is well-known that exchanging such information can still leak some sensitive information about the data used to compute these parameters/gradients. In machine learning, the most popular notion of privacy is differential privacy, which gives strong probabilistic guarantees. Differential privacy can be achieved by adding noise to various quantities: either the data itself, the model updates, the objective function, or the output (see e.g. C. Dwork. Differential privacy: A survey of results. In Proceedings of the 5th International Conference on Theory and Applications of Models of Computation, pages 1-19, 2008.)Could the algorithms in the thesis be extended merits and drawbacks in terms of convergence rate and communication cost?

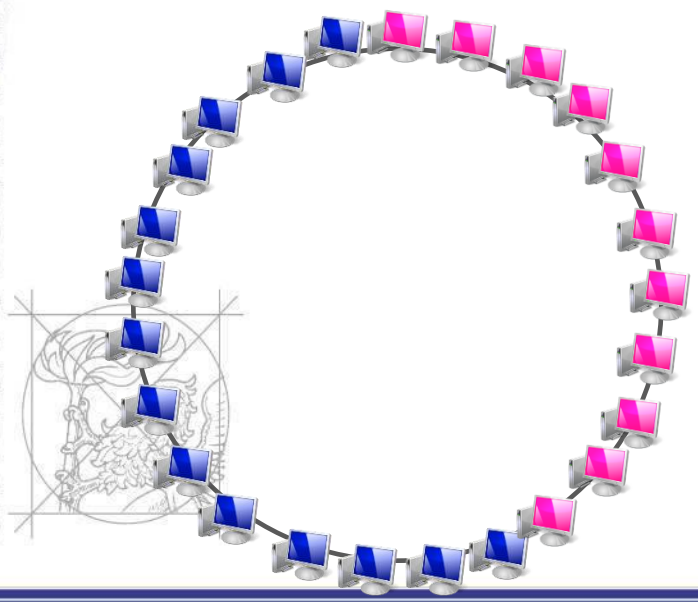
Kérdések (Kiss Attila) II.

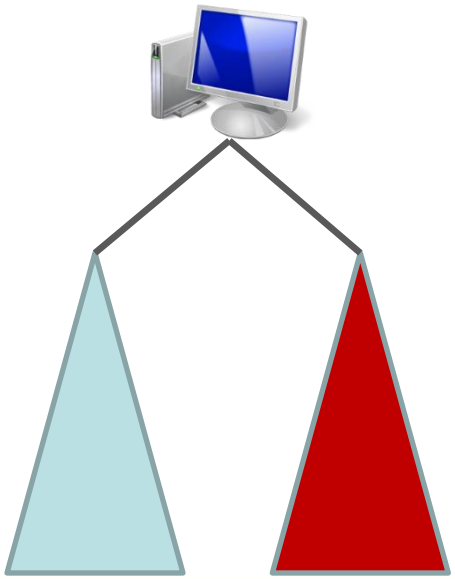
The author assumes that the homogenous network graph reflects the similarity between nodes (i.e., neighbors in the network graph have similar objectives). However, in practical scenarios, nodes could be different, one node can store larger or more reliable data than the other nodes, communicates faster, has more computing capacity or providing more useful information. This requires strategies to discover good peers and combining this information with the algorithms in the thesis to obtain more efficient decentralized protocols. What could be a good trade-off between exploration and exploitation in peer discovery to improve decentralized learning?

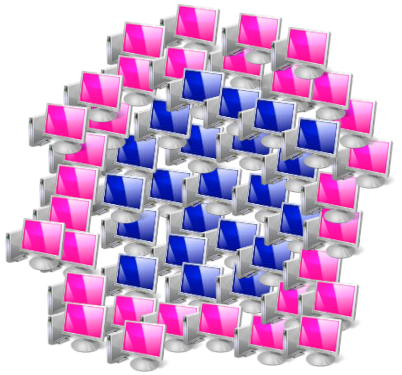
Kérdések (Kiss Attila) III.

What is the impact of the network topology on the convergence speed of the algorithm in the thesis? How does this speed depend from the usual graph parameters e.g. from clustering coefficient of the network in general or in special cases?

Topológia függő adateloszlások







Kérdések (Kiss Attila) IV.

Could the author give negative cases, machine learning methods in the field of classification, clustering or association rules, where gossip based approach is not applicable?