Gossip-Based Machine Learning in Fully Distributed Environments

István Hegedűs

Supervisor

Dr. Márk Jelasity

MTA-SZTE Research Group on Artificial Intelligence

and the University of Szeged

PhD School in Computer Science

University of Szeged

Summary of PhD Thesis

Szeged

2016

Introduction

Data mining and machine learning algorithms are present in our digital life even if we do not recognize this. For example, these algorithms are the spam filtering methods in our e-mail client, the automatic moderation in a social site or in a blog, the autocompletion mechanisms in a web search engine or in a text editor and the recommendations on a movie or on a web-shop site. Other algorithms help us or protect our health in a medical application or just amuse us in a game. They can recognize spoken words and handwritten letters. Without a detailed description of its workings, let us quote a definition from Tom M. Mitchell for machine learning [13]:

> "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E"

On the basis of this definition, the goal is to learn from the examples. These examples are usually generated by us when we drag and drop a mail into the spam folder, when we click on a search result or an advertisement on a website or in a mobile application. These examples can form a database that can be used to train the machine learning algorithms to extract information from the data or identify patterns in the databases. This information and these patterns can help us to understand how things work around us, identify and predict the trends and optimize the traffic in a city.

Although the huge amount of data is usually stored on servers, clusters or clouds owned by firms, it was generated on our devices (PCs, laptops, tablets, smart phones, wearable devices, sensors in a smart house, etc.). The processing of this data is an even more challenging task due to its size and restricted accessibility. Parallel computing techniques can still handle the size problem, but as time goes by, the data accumulates and the companies need more and more resources to store and process it. They have to have bigger and bigger servers and data farms, which have sufficient space to store our data, sufficient computational capacity to serve our queries and sufficient memory to run data mining tasks. Moreover, the access to the collected data is often forbidden even to researchers. The motivation for using fully distributed (i.e. peer-to-peer (P2P)) machine learning algorithms partly comes from the above-mentioned reasons. Since the data is generated on the device that we use in our everyday life, these devices should work collectively and solve the data processing tasks together. Machine learning algorithms should be also trained on our devices as well in a distributed manner. The other motivation for the P2P algorithms is the privacy issue. Nobody wants their search history, private photos from a cloud or the list of movies

Table 1: The relationship between the chapters and the corresponding publications (where ● and ○ refer to the basic and the related publications, respectively).

	Chapter 3	Chapter 4	Chapter 5	Chapter 6
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]	○			○

that you saw ever to be leaked out just because a hacker has found a backdoor in a server. Distributed methods allow us to keep our private data in our device instead of uploading our files into data centers, but they can still build a machine learning model based on the data. Here, we demonstrate a possible method for fully distributed machine learning.

The thesis is organized as follows. First, we give an overview in Chapter 2, which summarizes the necessary background, includes an introduction to supervised learning, an outline of the applied system model and the data distribution; and we introduce the fully distributed algorithms through some examples. Later, in chapters 3 – 6, we present the main parts of this thesis, where we introduce a fully distributed learning scheme that can be applied with any data modeling algorithm that can be trained in an online manner. We present several learning algorithms that can provide these kinds of online models, including sophisticated models. Afterwards we present state-of-the-art machine learning algorithms such as boosting and matrix factorization. Then, we describe a modification of the framework for achieving higher efficiency and capability for concept drift handling. In Table 1, we present the corresponding publications [1].

[1]The implementations of the proposed algorithms in this thesis are available online at: https://github.com/isthegedus/Gossip-Learning-Framework

Summary of the Thesis Results

As we mentioned in the Introduction, the data that contains the interesting information comes from our daily used electronic devices. And we cannot access this accumulated data, since it is stored in private servers. These issues motivated us in this study to develop fully distributed methods to process the data that is presented by the participants.

The main goal of this thesis is to present a possible way of machine learning on fully distributed data without collecting and storing it in a central location. Here, we expect that a huge number of computational units solve machine learning tasks together and communicate with each other via message passing only. We proposed a gossip-based framework to handle this learning problem. Within this framework, various algorithms can be applied. In Chapter 3 we described this framework and possible instantiations of several learning algorithms. Here, we focused on the basic supervised learning methods, such as the Logistic regression, Support Vector Machines and Artificial Neural Network. Then in the later chapters we presented more sophisticated methods and applications of the framework. We presented a boosting method, to improve the performance of the classification algorithms, a technique for handling the concept drift of the data, and we proposed an unsupervised matrix factorization method on the distributed data. Finally, it should be mentioned that the framework supports privacy preservation, the data processing is performed without moving any local data that is stored on the personal device.

Thesis 1: Gossip-Based Machine Learning

Algorithm 1 Gossip Learning Framework

```
1: x, y                                   ▷ local data
2: currentModel ← initModel()
3: loop
4:     wait(Δ)
5:     p ← selectPeer()
6:     send currentModel to p
7: end loop

8: procedure ONRECEIVEMODEL(m)
9:     m.updateModel(x, y)
10:    currentModel ← m
11: end procedure
```

We proposed gossip learning as a generic framework used to learn models, based on stochastic gradient search, on fully distributed data in large scale P2P systems. The basic idea behind gossip learning is that many models perform random walks in the network, while being updated at every node they visit. In this framework all SGD based learning method can be instantiated. The nodes in the network run the same algorithm (which is presented in the Algorithm 1), thereby performing the collective learning. When a node joins to the network it initializes a so-called local model. After that periodically (whose period is represented by Δ) sends the local model to one of its neighbors. When a node receives a model, it updates and stores it. To manage the neighbors of the nodes, we used the NEWSCAST peer sampling service, which can propose addresses of uniform randomly selected peers from the network. If the peer sampling service works properly, the models in the network will perform random walks. Figure 1 presents this system and the distributed learning, where the green models take walks on the communicational links and the data, shown in red, never leaves the node.

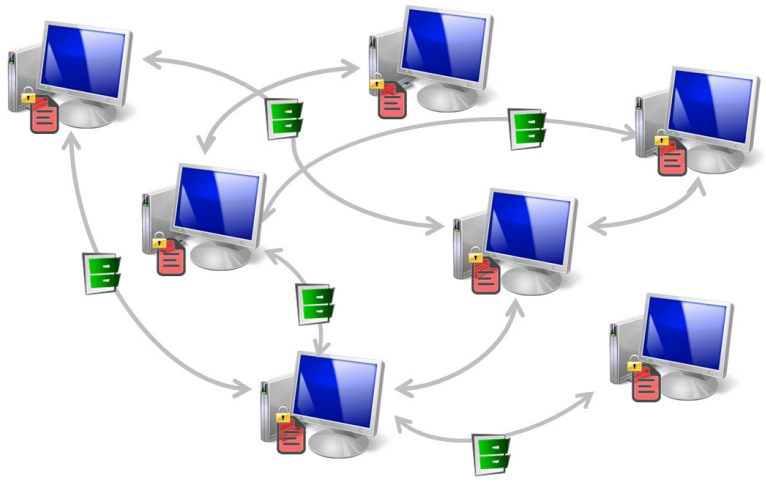

Figure 1: Gossip-based learning

In this framework, every algorithm can be used that can process the data as a stream. Furthermore, the algorithms that are based on the stochastic gradient descent (SGD) optimization technique and optimize a convex objective function will converge to the optima of the problem, as long as the peer sampling service is working properly. Hence the models take random walks in the networks and will be updated using uniform randomly selected data samples. We proposed numerous instantiations of gossip-based learning algorithms in this framework like Pegasos SVM, Logistic Regression and ANN.

The framework makes it possible to compute predictions locally at every node in the network at any point in time without additional communication cost. Furthermore, it has an acceptable message complexity: each node sends one message in each gossip cycle.

The proposed framework supports privacy preservation, since the data never leaves the node that stores it. The private data can be observed only by sending specific models for a node and monitoring its results.

Main contributions:

- A fully distributed learning framework (GOLF);
- Numerous implemented algorithms;
- The models can be used locally for every node;
- The framework supports privacy preservation;
- The corresponding paper is: [6]
  Róbert Ormándi, István Hegedűs, and Márk Jelasity. Gossip learning with linear models on fully distributed data. *Concurrency and Computation: Practice and Experience*, 25(4):556–571, 2013

Thesis 2: Fully Distributed Boosting

We demonstrated that the above-described gossip learning framework is suitable for the implementation of a multi-class boosting algorithm. The boosting technique allows a machine learning algorithm to have higher representation power by applying an appropriate weighted learning and voting mechanism.

To achieve this, we proposed a modification of the original FILTERBOOST which allows it to learn multi-class models in a purely online way, and we proved theoretically that the resulting algorithm optimizes a suitably defined negative log likelihood measure. The significance of this result is that a state-of-the-art machine learning technique from the point of view of the quality of the learned models is available in fully distributed systems. The Figure 2 tells us that the distributed algorithm is competitive with the other solutions.

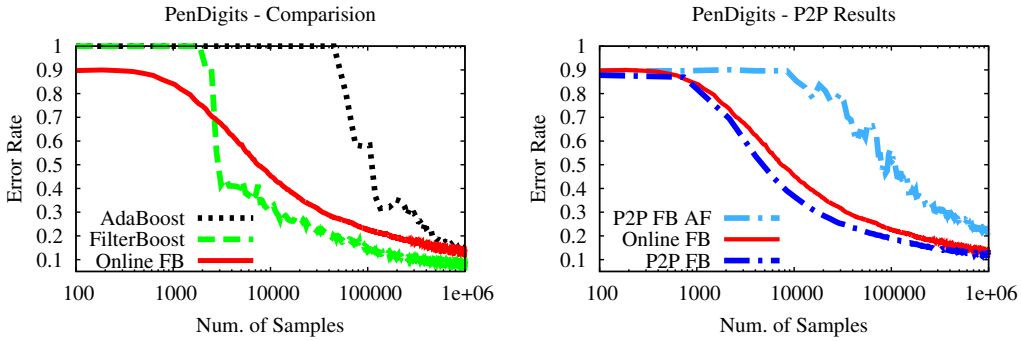

Figure 2: Comparison of boosting algorithms (left-hand side) and P2P simulations (right-hand side). FB and AF stand for FilterBoost and the "all failures" scenario, respectively.

We also pointed out the lack of model diversity as a potential problem with GOLF. We provided a solution that is effective in preserving the difference between the best model and the average models. This allowed us to propose spreading the best model as a way of benefitting from the large number of models in the network.

Main contributions:

- An implementation of a fully distributed multi-class boosting method;
- An online version of the FILTERBOOST and the theoretical derivation;
- The model diversity preservation aspect of the GOLF;
- The corresponding paper is: [3]
  István Hegedűs, Busa-Fekete Róbert, Ormándi Róbert, Jelasity Márk, and Kégl Balázs. Peer-to-peer multi-class boosting. In Christos Kaklamanis, Theodore Papatheodorou, and Paul Spirakis, editors, *Euro-Par 2012 Parallel Processing*, volume 7484 of *Lecture Notes in Computer Science*, pages 389–400. Springer Berlin / Heidelberg, 2012

Thesis 3: Handling Concept Drift

Here, we proposed adaptive versions of our GOLF framework: ADAGOLF and CDDGOLF. With these methods, the framework is capable of handling the change in the data patterns that we want to learn. This can happen due to different reasons. For example, the set of users of an application may change, or external factors (such as the weather) can vary. This has an effect on how people react to, it can influence the kind of movies they want to watch; it can affect their mobility patterns, and so on. People can also behave differently during their normal daily routine, or during a demonstration, for example.

In the case of ADAGOLF, the adaptivity is implemented through the management of the age distribution of the models in the network, ensuring that there is sufficient diversity of different ages in the pool. This method results that in the network there will be young (adaptive) and old (high performance) models. CDDGOLF also restarts some of the models, but this decision is based on the performance history of the model. This method can detect the occurrence of the concept drift as well.

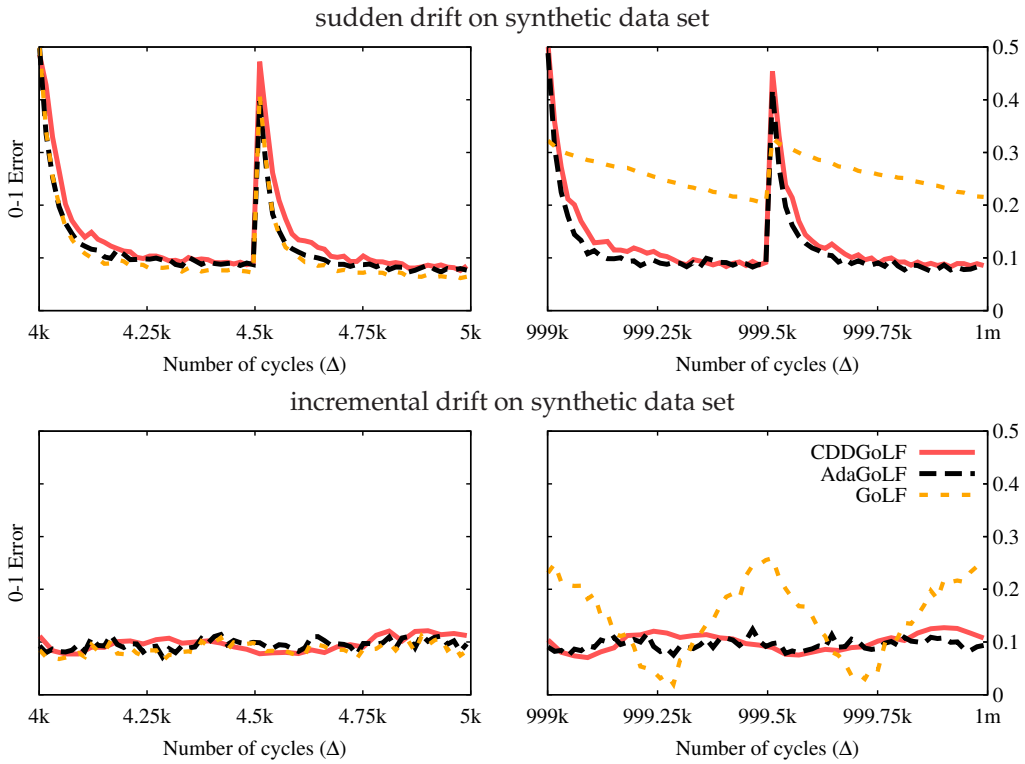

Figure 3: The burn in effect as a motivation for adaptivity.

In Figure 3, the phenomenon of concept drift can be seen. That is, the basic machine learning

algorithms cannot handle the change of the data, so after a certain time (update steps) they will burn in (cannot adapt to the change of the data distribution). Our proposed methods help the learning algorithms to avoid the above-mentioned problem.

Our main conclusion is that in those scenarios where the sampling rate from the underlying distribution is low relative to the speed of drift, our solutions clearly outperform all the baseline solutions, approximating the "God's Eye view" model, which represents the best possible performance.

We also showed that our algorithms can be enhanced to deal with (or rather, be robust to) higher sample rates as well, although in this case purely local model building can also be sufficient.

Main contributions:

- Two adaptive learning mechanism for GOLF, based on model restarts
  - One of them maintains the age distribution of the models
  - The other resets the models that have low performance;
- Drift handling and detection capabilities;
- High performance on low sampling rate;
- The corresponding papers are: [2, 4, 5]

  István Hegedűs, Ormándi Róbert, and Jelasity Márk. Gossip-based learning under drifting concepts in fully distributed networks. In *2012 IEEE Sixth International Conference on Self-Adaptive and Self-Organizing Systems*, SASO'12, pages 79–88. IEEE, 2012

  István Hegedűs, Lehel Nyers, and Róbert Ormándi. Detecting concept drift in fully distributed environments. In *2012 IEEE 10th Jubilee International Symposium on Intelligent Systems and Informatics*, SISY'12, pages 183–188. IEEE, 2012

  István Hegedűs, Róbert Ormándi, and Márk Jelasity. Massively distributed concept drift handling in large networks. *Advances in Complex Systems*, 16(4&5):1350021, 2013

Thesis 4: Singular Value Decomposition

Here, we proposed an SGD algorithm with an update rule that solves the problem of the low-rank decomposition of a matrix in a fully distributed manner. Additionally, we proposed a modification that has stable fix points only in the SVD of a matrix A. The output of the algorithm for rank k are two matrices X and Y that contain scaled versions of the first k left and right singular vectors of A, respectively. Matrices X and Y are unique apart from the scaling of the columns. This problem can be solved by optimizing the following objective function

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

where $A \in \mathbb{R}^{m\times n}, X \in \mathbb{R}^{m\times k}, Y \in \mathbb{R}^{n\times k}$, and the gradients, for the update rule, are the partial derivatives

$$\frac{\partial J}{\partial X} = (XY^T - A)Y, \quad \frac{\partial J}{\partial Y} = (YX^T - A^T)X. \tag{2}$$

The importance of this result can be verified by the wide range of problems that are solved by the matrix factorization technique. These problems are the recommender systems, where the A matrix contains the ratings of the users for the items and we want to approximate the missing ratings in the matrix; dimension reduction, where the rows of the A are the training instances in a learning problem and the method results in a compact representation of the instances; graph clustering, where A is the normalized affinity matrix; and the Latent semantic indexing (LSI) method that is used for extracting key terms from text data.

Algorithm 2 P2P low-rank factorization at node i

```
1: a_i                                   ▷ row i of A
2: initialize Y
3: initialize x_i                        ▷ row i of X
4: loop
5:     wait(Δ)
6:     p ← selectPeer()
7:     send Y to p
8: end loop
9: procedure ONRECEIVEY(Ỹ)
10:    Y ← Ỹ
11:    (Y, x_i) ← update(Y, x_i, a_i)
12: end procedure
13: η                                    ▷ learning rate
14: procedure UPDATE(Y, x_i, a_i)
15:    err ← a_i − x_i Y^T
16:    x_i' ← x_i + η · err · Y
17:    Y' ← Y + η · err^T · x_i
18:    return (Y', x_i')
19: end procedure
```

In Algorithm 2, we present the modified version of the GOLF that can solve the distributed matrix decomposition problem. Matrices A and X are private, that is, only the node which stores

a given row has access to it. The matrix Y performs random walk in the network and gets updated by the nodes and the local row of the X will be updated as well. Moreover, a version of the matrix Y is available in full at all nodes.

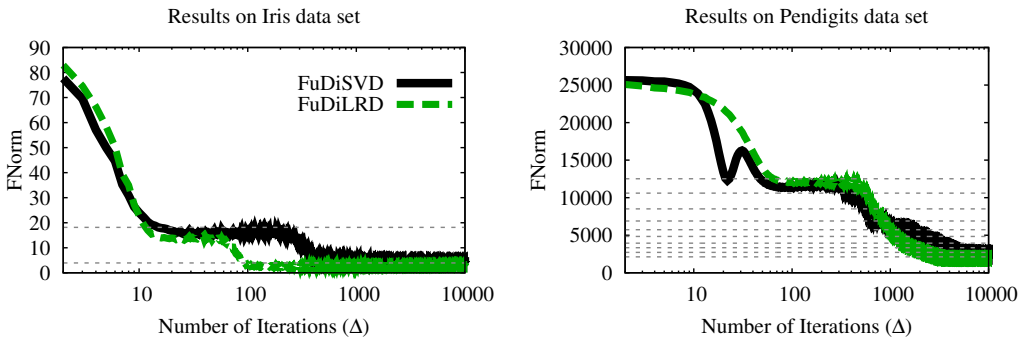

Figure 4: Convergence on the real data sets. Error is based on the Frobenius norm. Horizontal dashed lines in top-down order show the FNORM value for the optimal rank-i approximations for $i = 1, \ldots, k$.

Through experimental evaluation we studied the convergence speed of the algorithm (see Figure 4), and we also showed that it is competitive with other gradient methods that require more freedom for data access. We also demonstrated the remarkable robustness of the method in extreme failure scenarios.

Main contributions:

- An SGD based matrix low-rank decomposition technique in GOLF;
- A method that converges to a solution that corresponds to the singular value decomposition;
- The node related parts of matrices (the sensitive data) never leave the nodes;
- The corresponding paper is: [9]
  István Hegedűs, Márk Jelasity, Levente Kocsis, and András A. Benczúr. Fully distributed robust singular value decomposition. In *Proceedings of the 14th IEEE Fourteenth International Conference on Peer-to-Peer Computing (P2P)*, P2P'14. IEEE, 2014

References

[1] Róbert Ormándi, István Hegedűs, and Márk Jelasity. Asynchronous peer-to-peer data mining with stochastic gradient descent. In *Proceedings of the 17th international conference on Parallel processing - Volume Part I*, Euro-Par'11, pages 528–540, Berlin, Heidelberg, 2011. Springer-Verlag.

[2] István Hegedűs, Lehel Nyers, and Róbert Ormándi. Detecting concept drift in fully distributed environments. In *2012 IEEE 10th Jubilee International Symposium on Intelligent Systems and Informatics*, SISY'12, pages 183–188. IEEE, 2012.

[3] István Hegedűs, Busa-Fekete Róbert, Ormándi Róbert, Jelasity Márk, and Kégl Balázs. Peer-to-peer multi-class boosting. In Christos Kaklamanis, Theodore Papatheodorou, and Paul Spirakis, editors, *Euro-Par 2012 Parallel Processing*, volume 7484 of *Lecture Notes in Computer Science*, pages 389–400. Springer Berlin / Heidelberg, 2012.

[4] István Hegedűs, Ormándi Róbert, and Jelasity Márk. Gossip-based learning under drifting concepts in fully distributed networks. In *2012 IEEE Sixth International Conference on Self-Adaptive and Self-Organizing Systems*, SASO'12, pages 79–88. IEEE, 2012.

[5] István Hegedűs, Róbert Ormándi, and Márk Jelasity. Massively distributed concept drift handling in large networks. *Advances in Complex Systems*, 16(4&5):1350021, 2013.

[6] Róbert Ormándi, István Hegedűs, and Márk Jelasity. Gossip learning with linear models on fully distributed data. *Concurrency and Computation: Practice and Experience*, 25(4):556–571, 2013.

[7] Balázs Szörényi, Róbert Busa-Fekete, István Hegedűs, Ormándi Róbert, Jelasity Márk, and Kégl Balázs. Gossip-based distributed stochastic bandit algorithms. In *Proceedings of The 30th International Conference on Machine Learning*, volume 28(3) of *ICML'13*, page 19–27. JMLR Workshop and Conference Proceedings, 2013.

[8] Árpád Berta, István Hegedűs, and Róbert Ormándi. Lightning fast asynchronous distributed k-means clustering. In *22th European Symposium on Artificial Neural Networks*, ESANN 2014, pages 99–104, 2014.

[9] István Hegedűs, Márk Jelasity, Levente Kocsis, and András A. Benczúr. Fully distributed robust singular value decomposition. In *Proceedings of the 14th IEEE Fourteenth International Conference on Peer-to-Peer Computing (P2P)*, P2P'14. IEEE, 2014.

[10] Árpád Berta, István Hegedűs, and Márk Jelasity. Dimension reduction methods for collaborative mobile gossip learning. In *2016 24th Euromicro International Conference on Parallel, Distributed, and Network-Based Processing (PDP)*, pages 393–397, Feb 2016.

[11] István Hegedűs, Árpád Berta, Levente Kocsis, András A. Benczúr, and Márk Jelasity. Robust decentralized low-rank matrix decomposition. *ACM Trans. Intell. Syst. Technol.*, 7(4):62:1–62:24, May 2016.

[12] István Hegedűs and Márk Jelasity. Distributed differentially private stochastic gradient descent: An empirical study. In *2016 24th Euromicro International Conference on Parallel, Distributed, and Network-Based Processing (PDP)*, pages 566–573, Feb 2016.

[13] Tom M. Mitchell. *Machine Learning*. McGraw-Hill, New York, 2 edition, 1997.