# Applications of Support Vector-Based Learning

Róbert Ormándi

The supervisors are
*Prof. János Csirik* and *Dr. Márk Jelasity*

*Research Group on Artificial Intelligence of the University of Szeged and the Hungarian Academy of Sciences*

PhD School in Computer Science
University of Szeged



Summary of PhD Thesis

Szeged
2013

# Motivation

More and more data is accumulated around us. That is, various tools become easily available for managing very large-scale data, while the storage of data is getting cheaper and cheaper. This phenomenon—although the problem of machine learning has long been considered fundamental—continuously increases the need for machine learning algorithms that work *properly* on specific tasks, and operate *efficiently* in unusual settings (like in fully distributed computational environments), since without these algorithms we simply cannot extract useful information from the data. That is, without the appropriate application of machine learning algorithms, we are not able to utilize that large amount of data. However, achieving these goals is still challenging, since the inappropriate adaptation of a learning algorithm to a specific task can yield models that are far from the optimal ones. On the other hand, the naive adaptation of the algorithms can result in unexpected effects within the system that applies them (like huge, unbalanced load in a distributed system).

This thesis deals with the so-called *support vector-based learning algorithms*. The algorithms belonging to this particular family were designed around a central idea called *maximal margin heuristic*. The basic idea of support vector-based learning is pretty simple: find a good learning boundary while maximizing the margin (i.e. the distance between the closest learning samples that correspond to different classes). This basic idea has many practical applications and several formalisms.

The aim of this thesis is to present various approaches which help achieve the above men-

Table 1: The relation between the chapters of the thesis and the referred publications (• denotes the *basic* publications, while ○ refers to related publications).

| | Chapter 3 | Chapter 4 | Chapter 5 | Chapter 6 | Chapter 7 |
|---|---|---|---|---|---|
| ICDM 2008 [4] | • | | | | |
| TSD 2010 [6] | | • | | | |
| EUROPAR 2010 [7] | | | • | | |
| WETICE 2010 [5] | | | • | | |
| EUROPAR 2011 [8] | | | | • | • |
| CCPE 2012 [9] | | | | | • |
| SASO 2012 [3] | | | | ○ | ○ |
| SISY 2012 [1] | | | | ○ | ○ |
| EUROPAR 2012 [2] | | | | ○ | ○ |
| ICML 2013 [10] | | | | ○ | ○ |

tioned goals (appropriate adaptation of the algorithms in terms of both algorithmic and system model aspects) related to the support vector-based learning algorithms. That is, we investigate the *adaptivity* of support vector-based learners to various tasks and computational environments. The main "take-home message" of the thesis can be summarized as follows. One can achieve significant improvements through applying powerful basic ideas, like the maximal margin heuristic of support vector-based learners, as building-blocks, and careful system design.

As our system model, we focus on the so-called *fully distributed environment*. This system model consists of a large number of units, called *nodes*, that are capable of local computation and can communicate with other nodes using the network. We assume that the system consists of a potentially very large number of nodes, typically personal computing devices such as PCs or mobile devices. We do not assume any homogeneity between the nodes. They can be different in their computational power, operating system, or any other characteristics. Additionally, we assume that the communication is done by message passing between connected nodes without any central control (fully distributed aspect). That is, every node can send messages to every other node, provided the address of the target node is available. We assume that messages can have arbitrary delays, and messages can be lost as well (failure scenarios). In addition, nodes can join and leave at any time without warning (churn), thus leaving nodes and crashed nodes are treated identically. Leaving nodes can join again, and while offline, they may retain their state information.

The thesis begins with an introductory chapter (Chapter 2), which summarizes the necessary background information related to support vector machines and fully distributed sys-

tems. The main part of the thesis (Chapter 3-7) can roughly be divided into two distinct parts. In the first part (Chapters 3-5), we investigate the *algorithmic adaptivity* of support vector machines. That is, we focus on how we can adapt the basic idea of support vector-based learners, the maximal margin heuristics, to develop efficient algorithms to a wide-range of applications, like time series forecasting (in Chapter 3), domain adaptation (in Chapter 4), and recommender systems (in Chapter 5).

In the second part of the thesis (Chapters 5-7), we gradually turn to examine the *system model aspect of adaptivity*. In this part, our main question is how we can implement support vector-based learning algorithms in the fully distributed setting. In Chapter 6, we propose a gossip-based support vector implementation called P2PEGASOS, then, in the last chapter (Chapter 7), we improve significantly the convergence speed of P2PEGASOS, while we theoretically analyse the algorithm as well.

Each chapter of the thesis is based on at least one accepted publication. Tab. 1 shows the relation among the chapters of the thesis and the more important[1] publications.

[1]For a complete list of publications, please visit the corresponding section of my website: `http://www.inf.u-szeged.hu/~ormandi/papers`

# Summary of the Thesis Results

Here we overview the main goals and results of this thesis by giving a brief summary of each chapter (Chapters 3-7). During the summarization, we reveal our contributions as well by giving an itemized list of the key contributions of the given chapter.

## VMLS-SVM for Time Series Forecasting

In Chapter 3, we investigated the problem of time series forecasting. That is, how we can extend the Least Squares SVMs to become more suitable algorithms for predicting the future values of a time series. Our proposal (the VMLS-SVM algorithm) introduced a weighted variance term in the objective function of the LS-SVM. This idea is based on the preliminary observation which says that if we have two time series forecasting models with the same prediction error, the one with the smaller variance results in a better overall performance aside from overfitting. The proposed method can be considered as the generalization of the LS-SVM algorithm, which keeps all the advantages of the original algorithm, like the applicability of the kernel-trick, unique and linear solution. However, it introduces a new hyperparameter which makes the fine-tuning of the algorithm more complicated.

The first part of the chapter briefly overviewed the related approaches and introduced the basic properties of the original LS-SVM approach, then it gave a detailed description of the proposed method. It proposed a mechanism for handling the above mentioned increased

complexity of parameter setting as well. A thorough empirical evaluation closed the chapter which points out that with appropriate parameter tuning the proposed method achieves a significantly better performance against a numerous state-of-the-art baseline algorithms measured on three different, widely used benchmark datasets.

The key contributions and their effects are:

- The theoretical and algorithmic introduction of the VMLS-SVM algorithm.
- Parameter optimization mechanism for the VMLS-SVM algorithm.
- The proposed algorithm outperforms significantly a number of baseline approaches on three widely used benchmark datasets.

The results of this chapter are based on our recent work published in [4].

## Transformation-based Domain Adaptation

This chapter (Chapter 4) investigated the problem of domain adaptation on an opinion mining task. We proposed a general framework, called DOMAIN MAPPING LEARNER (DML) and two instantiations of this general idea: the first one is based on SVMs as source model and the second one applies the Logistic Regression (LR) as source model. The main idea of the general approach can be summarized as follows: it models the relation between the source and target domain by applying a model transformation mechanism which can be learnt by using labeled data of a very limited size taken from the target domain.

In the chapter, we briefly overviewed the related approaches for the task of domain adaptation. Then, we formally defined the problem and proposed our general, transformation-based approach, the DML algorithm. We introduced the novel instantiation of this abstract algorithm, based on SVMs and LR methods. Our experiment evaluations validated that our approach is capable of training models for the target domain which use a very limited number of labeled samples taken from the target domain. This phenomenon is even true when we have enough samples, but the baseline methods cannot generalize well. From this evaluation, we also concluded that the SVM-based instantiation is a more suitable choice since it is more robust than the LR-based variant.

The key contributions and their effects are:

- The transformation-based formalism of the problem of domain adaptation.

- The introduction of the general algorithm.
- The two instantiations of the general idea based on SVM and LR based models.
- The proposed algorithms result in models which have a better performance than the direct method (baseline), and two other state-of-the-art baseline algorithms (SCL and SCL-MI).

The results presented in this chapter are mainly based on our earlier work [6].

## SVM Supported Distributed Recommendation

In this chapter (Chapter 5), we dealt with the problem of distributed recommendation. The goal of this chapter is twofold, first we described and motivated the process of inferring ratings from implicit user feedbacks, then we introduced two heuristic approaches (direct and time-shift based approaches) to overcome the problem. Here we presented an interesting use-case of the SVMs. We use them to validate our inferring heuristics indirectly and proved that—without any ground truth—the inferred dataset has some interesting properties, and the one generated based on the improved heuristics variant (time-shift based variant) has a more interesting inner structure. Second, we introduced novel overlay management protocols which support the implementation of user-based collaborative filtering (CF) approaches in fully distributed systems, while keeping low the overall network load.

In the chapter, first we reviewed both the centralized and distributed CF approaches. Then, we introduced our inferring heuristics, described the validation methodology through learnability, and performed this validation applying the SMO SVM solver. Later, we pointed out that most of the CF dataset has almost power-law in-degree distribution which can cause serious issues in distributed setting. To overcome this problem, we introduced a bunch of overlay management approaches (random sampling based kNN approach, T-MAN based variants (GLOBAL, VIEW, BEST and PROPORTIONAL) and their randomized variants) and performed a thorough empirical evaluation against each other and a real world overlay management protocol called BUDDYCAST. We can draw multiple conclusions: the aggressive peer sampling can cause untreatable network load, while the proposed T-Man based approach with the GLOBAL peer selection strategy is a good choice considering that it has a fully uniform load distribution with an acceptable convergence speed.

The key contributions and their effects are:

- The inferring heuristics (direct and time-shift based approaches).
- Learnability based indirect validation technique.
- The FileList.org inferred recommender dataset.
- Overlay management approaches: random sampling based kNN approach, T-MAN based variants (GLOBAL, VIEW, BEST and PROPORTIONAL) and their randomized variants.
- The GLOBAL peer selection based overlay management tool provides a good trade-off between the convergence speed and network load.

The results of this chapter are mainly based on our earlier works [5, 7].

## P2PEGASOS—A Fully Distributed SVM

This chapter (Chapter 6) focused on the problem of fully distributed data mining. That is, our goal here was to propose an SVM-based algorithm which performs in a fully distributed network realizing good quality models with an affordable communication complexity while assuming as little as possible about the underlying communication model. Our proposal, the P2PEGASOS algorithm, introduces a conceptually simple, yet powerful SVM implementation. The key of our contribution is that many models perform a random walk over the network while updating themselves applying an online update rule.

At the beginning of the chapter, we carefully defined the system and data model. Then, we overviewed the basic concept of the Pegasos SVM solver, and the related fully distributed machine learning approaches. After a detailed algorithmic description of our proposal, we turned to evaluate our approach. Here we considered a number of baseline algorithms (mainly centralized SVM implementations) and investigated the speed of the convergence in various scenarios including ones with extreme network failures. These experimental evaluations show that our approach is robust against various network failures, while provides reasonable models with affordable network load.

The key contributions and their effects are:

- The P2PEGASOS algorithm.
- Local voting mechanism for improving the prediction performance.
- The algorithm shows amazing convergence properties even in scenarios with extreme network failures.

The results presented in this chapter are based on those presented in our previous work [8].

## Speeding Up the Convergence of P2PEGASOS

In Chapter 7, we continued to investigate the fully distributed setting and proposed a mechanism which increases the convergence speed of P2PEGASOS with almost an order of magnitude. These algorithms are referred to as P2PEGASOSMU and P2PEGASOSUM. The idea of the proposed algorithms is based on applying an ensemble component by introducing a mechanism which averages the models that "meet" on a certain node. We demonstrated that in the Adaline model our proposal behaves exactly as if an exponentially increasing number of model would be collected and voted through the prediction, but in our case the model requires only a constant size. We pointed out that in the case of the P2PEGASOS algorithm, the exact equality is not true, however, the behavior of the approach is pretty similar. A convergence proof of P2PEGASOSMU was provided as well.

The chapter began with a motivating section of fully distributed data, then the related work of ensemble learning was briefly discussed. The detailed algorithmic description of the proposed mechanism followed with the discussion of the Adaline model, the Pegasos model and the convergence proof of P2PEGASOSMU algorithm. The empirical evaluation shows that the improved approaches can result in almost an order of magnitude faster convergence speed than the original P2PEGASOS algorithm, while they keep all the advantages of the method. We pointed out that P2PEGASOSMU maintains more independence between the models, hence this is the favorable variant between the two proposals.

The key contributions and their effects are:

- The merging mechanism for the P2PEGASOS algorithm (resulting in the P2PEGASOSMU and P2PEGASOSUM algorithms).
- The convergence proof of P2PEGASOSMU.
- The algorithms show even faster convergence speed than the original P2PEGASOS algorithm while keeping all the advantages of the original one.

The presented results are mainly based on our previous paper [9].

## Overall Summary of the Thesis Results

The key results of this thesis are the following:

- The theoretical and algorithmic introduction of the VMLS-SVM algorithm presented in Chapter 3, originally published in [4].
- Parameter optimization mechanism for the VMLS-SVM algorithm introduced in Chapter 3, originally proposed in [4].
- The transformation-based formalism of the problem of domain adaptation defined in Chapter 4, originally published in [6].
- The introduction of the general DML algorithm introduced in Chapter 4, originally introduced in [6].
- The two instantiations of the general idea DML algorithm based on SVM and LR based models presented in Chapter 4, originally proposed in [6].
- The inferring heuristics (direct and time-shift based approaches) introduced in Chapter 5, originally published in [5].
- Learnability based indirect validation technique based on SMO presented in Chapter 5, originally proposed in [5].
- Overlay management approaches: random sampling based kNN approach, T-MAN based variants (GLOBAL, VIEW, BEST and PROPORTIONAL) and their randomized variants introduced in Chapter 5, originally published in [7].
- The P2PEGASOS fully distributed SVM learner presented in Chapter 6, originally proposed in [8].
- Local voting mechanism for improving the prediction performance of P2PEGASOS introduced in Chapter 6, originally presented in [8].
- The merging mechanism for the P2PEGASOS algorithm (resulting in the P2PEGASOSMU and P2PEGASOSUM algorithms) proposed in Chapter 7, originally introduced in [9].
- The convergence proof of P2PEGASOSMU presented in Chapter 7, originally published in [9].

# References


[1] István Hegedűs, Nyers Lehel, and Ormándi Róbert. 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.

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

[3] 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.

[4] Róbert Ormándi. Variance minimization least squares support vector machines for time series analysis. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, pages 965–970, Washington, DC, USA, 2008. IEEE Computer Society.

[5] Róbert Ormándi, István Hegedűs, Kornél Csernai, and Márk Jelasity. Towards inferring ratings from user behavior in bittorrent communities. In *Proceedings of the 2010 19th IEEE International Workshops on Enabling Technologies: Infrastructures for Collaborative Enterprises*, WETICE '10, pages 217–222, Washington, DC, USA, 2010. IEEE Computer Society.

[6] Róbert Ormándi, István Hegedűs, and Richárd Farkas. Opinion mining by

transformation-based domain adaptation. In *Proceedings of the 13th international conference on Text, speech and dialogue*, TSD'10, pages 157–164, Berlin, Heidelberg, 2010. Springer-Verlag.

[7] Róbert Ormándi, István Hegedűs, and Márk Jelasity. Overlay management for fully distributed user-based collaborative filtering. In *Proceedings of the 16th international Euro-Par conference on Parallel processing: Part I*, EuroPar'10, pages 446–457, Berlin, Heidelberg, 2010. Springer-Verlag.

[8] Róbert Ormándi, István Hegedűs, and Márk Jelasity. Asynchronous peer-to-peer data mining with stochastic gradient descent. In *17th International European Conference on Parallel and Distributed Computing (Euro-Par 2011)*, volume 6852 of *Lecture Notes in Computer Science*, pages 528–540. Springer, 2011.

[9] Róbert Ormándi, István Hegedűs, and Márk Jelasity. Gossip learning with linear models on fully distributed data. *Concurrency and Computation: Practice and Experience*, pages n/a–n/a, 2012.

[10] Balázs Szörényi, Róbert Busa-Fekete, István Hegedűs, Ormándi Róbert, Márk Jelasity, and Balázs Kégl. Gossip-based distributed stochastic bandit algorithms. In *Proceedings of The 30th International Conference on Machine Learning (ICML), 3rd Cycle*, volume 28 of *JMLR: Workshop and Conference Proceedings*, pages 19–27. JMLR: W&CP, 2013.