Collaborative Classification over P2P Networks

Odysseas Papapetrou Wolf Siberski Stefan Siersdorfer

L3S Research Center, Hannover, Germany
{papapetrou, siberski, siersdorfer}@L3S.de

ABSTRACT

We propose a novel collaborative approach for distributed document classification, combining the knowledge of multiple users for improved organization of data such as individual document repositories or emails. The approach builds on top of a P2P network and outperforms the state of the art approaches in collaborative classification.

Categories and Subject Descriptors: H.2.8 [**Database Management**]:Database Applications–*Data Mining*
General Terms: Algorithms

1. INTRODUCTION

Supervised classification is relevant for a variety of applications, such as email spam filtering, focused crawling, structuring Web directories, and social bookmarking. For addressing the cold-start problem, but also to improve classification quality, collaborative classification solutions are proposed in the literature, e.g., collaborative tagging and spam detection, where information from different users is aggregated to constructing better machine learning models. A naive approach based on sharing training samples directly among users to obtain larger training sets is prohibitive, since it ignores privacy, security, and copyright aspects of the user's personal information sources, and also leads to high network costs for the participants.

To address these issues, our approach, called Collaborative SVM (**CSVM**), is instead based on exchanging classification models. To save network resources, the models are reduced to their most significant components, rendering the approach scalable, even for mobile devices. We describe the approach using Support Vector Machines as local classifiers, and a P2P network as infrastructure, although it can be generalized to a wide range of classifiers and to different network architectures. The experimental evaluation and comparison with state-of-the-art approaches validate the efficiency and effectiveness of CSVM.

2. METHODOLOGY

CSVM combines dimensionality reduction, model sharing, and meta-model construction, to realize scalable distributed classification with an excellent cost/quality tradeoff. The participating nodes are connected in a P2P network, with each peer carrying its own training set (with varying size and quality). The algorithm consists of the following steps, repeated at regular intervals:

- Every peer computes a local classification model using its own training set.
- Peers reduce their local models to the most significant components, and exchange them with a small number of selected neighbors.
- Each peer merges the received models with its own model to construct a more powerful meta classifier, taking reliability weights into account.

The resulting meta classifiers exhibit higher quality than the local classifiers, and can be used at each node for classification. We now describe the algorithm elements in detail.

Local classification models. The training data at each user node (e.g., a set of emails manually classified as spam or ham) is used to compute the local classification model. In this work we use linear Support Vector Machines (SVMs) for computing the local classifiers, the state-of-the-art in supervised learning. We also test our model using Reduced SVMs (RSVMs), which enable trading quality with classification efficiency. Other approaches, like, e.g., Fisher's discriminant, are also applicable.

Model Reduction. After the local classifiers are constructed, the users need to exchange their classification models. To save network resources, each model is reduced before transmission. Different model reduction techniques are possible, including hashing, ℓ_1 regularization, and feature selection. CSVM uses the technique of [4], which reduces the model to the requested number of dimensions by keeping only the model components with the highest absolute values. We have selected this approach because it incurs very low computational costs and has a favorable cost/quality tradeoff. Other approaches are equally applicable.

Model exchange and meta-model construction. The participating peers periodically exchange their reduced models with selected neighbors. To boost the classification quality without imposing additional computation overhead to the peers, each peer combines all models received from its neighbors to construct a single meta-model. In this work we merge SVM models using weighted averaging. It can be shown that the combined meta-model yields the same results as executing each SVM classifier individually, and combining the results afterwards. The weights represent the size of the training sets used for constructing each model, so that low-quality models have a small impact on the classification quality. Other weighing schemes, such as trust or cross-validation scores, and other ensemble methods, such as bagging or stacking, could also be integrated.

Copyright is held by the author/owner(s).
WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0637-9/11/03.

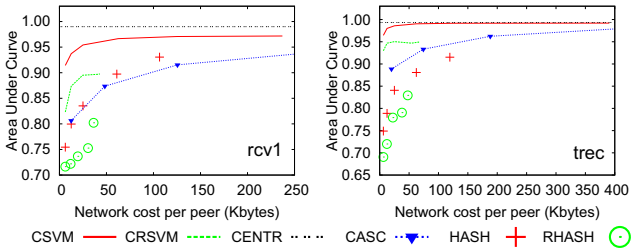

Figure 1: Classification quality

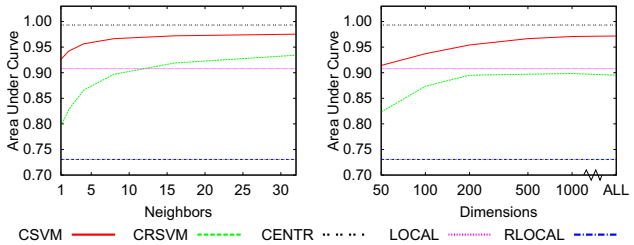

Figure 2: Influence of (a) number of neighbors, and, (b) number of dimensions.

3. EXPERIMENTAL EVALUATION

CSVM can employ different classifiers for computing the local models. For our experiments, we used standard SVMs (denoted as **CSVM**) as well as Reduced SVMs [3] (**CRSVM**). These variants were compared with their non-collaborative counterparts (**LOCAL** and **RLOCAL**) where each peer uses only its local classifier built solely on its own training set. As gold standard, we use a non-distributed SVM classifier (**CENTR**) trained on the union of the training sets of all peers. Note that CENTR has practical constraints, e.g., network and computational cost, as well as privacy issues, and therefore cannot be applied in real-life. Furthermore, the approach was compared with the state-of-the-art methods in distributed and collaborative classification, Cascade RSVM (**CASC** [1]) and **HASH** [2].

We report results for a network of 100 peers, built over an unstructured P2P network. Since the algorithm's accuracy and efficiency is orthogonal to the network characteristics, the results also apply to larger and differently formed networks. The experiments were conducted on two standard web classification datasets, the Reuters Corpus Volume I (**rcv1**), and the TREC 2007 spam corpus (**trec**), using the standard features and ground truth (Available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm/` and `http://plg.uwaterloo.ca/~gvcormac/spam/`). Each peer was assigned 25 positive and 25 negative examples for training. The remaining documents were used for testing. Efficiency was measured with computational overhead and network cost (transfer volume) and effectiveness with Area Under the ROC Curve (AUC).

All algorithms were compared with respect to their quality/cost ratio, i.e., which quality can be achieved with a given network cost budget. CSVM and CRSVM were initialized with 8 neighbors per peer, and the number of dimensions per model was determined from the network budget. HASH and RHASH were also configured to keep the same cost budget.

Fig. 1 depicts the AUC measure in correlation to the network requirements, for the two datasets. CSVM substantially outperforms all other distributed algorithms, and closely approximates the quality of CENTR with a very small network cost. CRSVM is inferior to CSVM, but still outperforms HASH and CASC in its cost range. We expect RSVM to show its strengths only with very large local training sets. Another limitation of CRSVM and of all the RSVM-based algorithms also becomes apparent from these results: the underlying RSVM already trades classifier accuracy for efficiency, limiting the possibility of a fine-grained control of the desired cost/quality trade-off.

Fig. 2(a) shows the influence of the number of neighbors to CSVM and CRSVM. The presented results are for the rcv1 collection, and a model reduction to 500 dimensions. Both CSVM and CRSVM clearly outperform the corresponding non-collaborative approaches. As expected, the benefit of collaboration increases with the neighborhood size. Interestingly, the improvement is significant even for small neighborhood sizes. In particular, CRSVM with just 4 neighbors yields a performance improvement of more than 10% compared to RLOCAL. Similarly, CSVM with 4 neighbors achieves a performance increase of more than 5% compared to LOCAL. Adding more neighbors per peer further improves classification quality at a slower rate.

Fig. 2(b) shows the influence of the number of dimensions. The results correspond to CSVM and CRSVM configured with 8 neighbors per peer. Similar to the case of the neighborhood size, the number of dimensions does not need to be high: both CSVM and CRSVM yield already substantial benefits with 500 dimensions, achieving a classification quality almost equal to the unreduced models. The experiments with trec had the same qualitative outcome.

4. CONCLUSIONS AND FUTURE WORK

In this paper, we have presented CSVM, a collaborative classification algorithm built on top of a P2P network. The experimental results confirm that CSVM substantially outperforms the state-of-the-art collaborative classification techniques while keeping network costs negligible. Our approach offers the additional advantage that network load can be controlled in a precise and flexible way, allowing for an optimal utilization of network resources. While default configurations for CSVM provide already very good results, we now work towards dynamically tuning the system parameters (number of dimensions and number of neighbors) to achieve maximum accuracy for given network constraints.

Acknowledgments. This work is partially supported by the FP7 EU Project GLOCAL (contract no. 248984)

5. REFERENCES

[1] H. H. Ang, V. Gopalkrishnan, S. C. H. Hoi, and W. K. Ng. Cascade RSVM in P2P networks. In *ECML/PKDD*, 2008.

[2] J. Attenberg, K. Weinberger, A. Dasgupta, J. Langford, A. Smola, and M. Zinkevich. Collaborative email-spam filtering with consistently bad labels using feature hashing. In *CEAS*, 2009.

[3] Y.-J. Lee and O. L. Mangasarian. RSVM: Reduced support vector machines. In *SDM*, 2001.

[4] D. Mladenić, J. Brank, M. Grobelnik, and N. Milic-Frayling. Feature selection using linear classifier weights: interaction with classification models. In *SIGIR*, 2004.