ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 2, February 2015

Measuring the Importance of Collaborative Learning

Deepika.S

PG Student, Department Of Computer Science Engineering, Parisutham Institute of Technology and Science,

Thanjavur, Tamilnadu, India

Abstract*:* The problem of online multitask learning for solving multiple related classification tasks in parallel, aiming to classify every sequence of data received by each task accurately and efficiently. First, to meet the critical requirements of online applications which is highly efficient and scalable classification, this solution can make immediate predictions with low learning cost is needed. Second, classical classification methods, it is the batch or online, often encounter a dilemma when applied to a group of tasks i.e., a process of single classify model trained on the entire collection of data from all tasks may fail to capture characteristics of individual task; on other process, a model trained independently on individual tasks may suffer from the insufficient training data. To overcome these kinds of challenges we have proposed a collaborative online multitask learning method, which will learn a global model over the entire data of all tasks.

Keywords*:* Multitask learning, Global Model, collaborative Model, Scalable Classification.

I. INTRODUCTION

The classical machine learning methods are often formulated as a single task learning problems which is defined for learning one task at a time. The multitask learning aims to solve the multiple related learning tasks in parallel. Many other problems are essential in multitask learning which are often broken into single learning tasks which are then solved individually by the classical learning methods. The Multitask learning has been extensively studied in machine learning and the data mining over a past decade. In which the Empirical findings have demonstrate the advantages of multitask learning over single task learning across a variety of application domain. The classical multitask learning methodology often makes two assumption. First process is to assumes there is one primary task and the other related tasks are simply secondary process whose training data are exploited by multitask learning to improve primary task. The classical multitask learning approach focuses on learning the primary task without caring how the other tasks are learned. Second process is the classical multitask learning problem is often studied in a batch learning settings, which assumes the training data for all tasks which are available. In this paper, we investigate the problem of online multitask learning, in which they differs from the classical multitask learning in the two aspects. First process is to improve the learning performance of all tasks instead of focusing on a single primary task. Second process, we have frame the multitask learning problem in an online learning setting by assuming that the data for each task arrives sequential. This batch learning techniques, in online learning methods learn over a sequence of data by processing each sample. At the each round, the learners first receives one instance, makes the prediction and received in a true label. In which the error of information is then used to update the learning process. In the early study on this work was first motivated by need to classify online user-generated content. In this process each individual exhibits uniqueness and also shares certain characteristics with the others in a group. It is then desirable to develop an efficient and scalable classifier that can solve individual task by adapting to the global knowledge shared by all the users. Consider the real problem of micro blog sentiment analysis on the group of users where the goal is to classify micro-blog posts generated by each user into the several emotional or non-emotional categories in a near real-time manner. When solving this kind of problem by classical machine learning techniques in which a single global classification model that trained on the entire collection of the data from all users may fail to capture the peculiarity of individual users and thus it often works poorly. We propose a novel collaborative online multitask learning (COML) [8] technique to attack the aforementioned challenges.

II. RELATED WORKS

This work is closely related to the two groups of research in machine learning and the data mining i.e., online learning and multitask learning. We briefly survey this representative work in each area. Thus the Online learning has been extensively. In the batch learning methods, which is assumes all training samples to be available before the learning task begins, then online learning algorithms incrementally build a model from the stream of samples, thus allowing for simple and fast model update. They are naturally capable of dealing with large datasets and the problems whose data arrive sequential. The origin of the online learning can date back to well-known Perceptron algorithms, which then updates the model weight by moving it to a tad closer to each misclassified samples. Descendant Perception like methods employ has more sophisticated update strategies. Typically, a variety of the online learning algorithms have been proposed based on the maximum margin learning principle that has been successfully applied to the batch mode learning.

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 2, February 2015

Particularly, a Relaxed Online Maximum Margin(ROMMA) [10] algorithm repeatedly chooses the hyper-plane that correctly classifies existing training examples with the maximum margin. Then the family of Passive-Aggressive (PA) algorithms [7] maintains a trade-off between the amount of progress which made on each training round and information retained from previous rounds. In particular, thus the PA algorithm updates the model whenever a new example is misclassified or when its classification score is smaller than the predefined margin. Thus, Empirical studies shows that the maximum margin based on online learning algorithms are generally more effective than the Perceptron algorithm. Thus the above online learning algorithms in general belong to the family of first-order online learning techniques.

In Recent years have witnessed emerging studies on exploring second order information for online learning. Then the second order online learning algorithms that improves upon the Perceptron like methods which include the second order Perceptron (SOP), thus confidence weighted (CW) learning and successors. The confidence weighted learning algorithm maintains the probabilistic measure of confidence in each component of its weight vector using a Gaussian distribution. The weighted distribution is updated by minimizing the Kullback-Leibler divergence between a new weight distribution and the old one under the constraint that the probability of correct classification is greater than the threshold. AROW, which it stands for "adaptive regularization of weighted vectors", which softens the hard constraint in confidence weighted learning as the regularizes. It has been also uses an improved update strategy, leading to the extra robustness in the case of non-separable data.

The AROW has been demonstrated to show the state of-the-art performance for the typical online learning tasks. The problem of jointly solving several related learning tasks by leveraging the commonality among the tasks has been studied in a machine learning community under the guise of multitask learning.

Thus the relationship of the tasks has been modelled in a number of ways. The pioneering work assumed in which there is one primary task and the other secondary tasks, which are solely used to improve the learning of the primary task. Eugenio et al. it introduced multitask kernel [2] and considered batch multitask learning as a regularized optimization problem. Ando et al. it formulated multitask relations by enforcing predictive functions for the different tasks to belong to the same hypothesis set. Kang et alit studied the problem of multitask learning of shared feature representations among tasks, while simultaneously determining "with whom" in each task should share. Some other studies tried to explore the underlying spectral dependencies among tasks. In the authors used for feature hashing to solve multitask learning problem. Then for each task, they has minimized the interaction between its parameter vector and the combination of other tasks' parameter vectors. Thus the existing batch multitask learning studies, the work is closer to the online multitask learning methodology. Then the online multitask learning problem was first addressed in [1], who has assumed a very general setting where in the tasks are related by a global loss function and then the goal is to reduce the cumulative loss (for all tasks involved) over all rounds of the online algorithm. Following the same line of thought, thus the studies in formulated by the multitask learning problems as the online learning with expert advice.

Thus Regret bounds are given under the assumption that there is a set of best experts who perform well on the entire set of all tasks. Saha et alit proposed to learn the task models as well as the task relatedness in a coherent ways.

III. PROPOSED SYSTEM

Thus the motivation of our solution is two-fold. First, as the tasks often exhibit varying patterns, it is neither practical nor effective to learn a single global model for classification. Second process, it is also not always possible to learn a good classification model for each task since training data available for a single task are often limited. For such cases, it is reasonable to pool together from data across many related tasks.

Building a Global Model by Online Learning:

The first step of the collaborative online multitask learning is to built a global classification model to exploit the commonality among tasks. We adopt the online passive aggressive (PA) framework to build a global model using data collected from all users.

Learning the Collaborative Models:

The critical step of our collaborative online multitask learning is to apply the existing global model to collaboratively learn the each of the K individual user models. Using the same PA formulation, the goal is to learn a classification model for the k user.

The next step is to use the shared information learned by the global mode is to enhance each individual learning model. We formulate the collaborative learning model as a convex optimization problem that minimizes the deviation of the new weight vector from the prior collaborative one and the global one.

Extending Confidence Weighted Learning:

A current trend in online learning research is to use parameter confidence information to guide online learning process. Confidence weighted learning algorithms [9] have been shown to perform well on many tasks. we extend the proposed collaborative online multitask learning with the confidence weighted hypothesis.

Global Model:

It learns a single classification model from all task data by applying the PA/AROW algorithm. At each online learning round, the algorithm receives a training sample from each task, and uses that sample to update its weight vector.

Personal Model:

It employs the PA/AROW algorithm to train a personal classification model for each task only using its own data. In other words, every task is associated with a personalized classification model.

Simple model:

It simply switches between the Global and Personal models according to their cumulative error counts in

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 2, February 2015

previous online learning rounds. In particular, at each round, it sets its weight vector to that of the best model (Global or Personal), i.e., one with the least cumulative errors to-date. Benchmarking against this method is important as it will show whether the proposed COML algorithm is more effective than a naive combination.

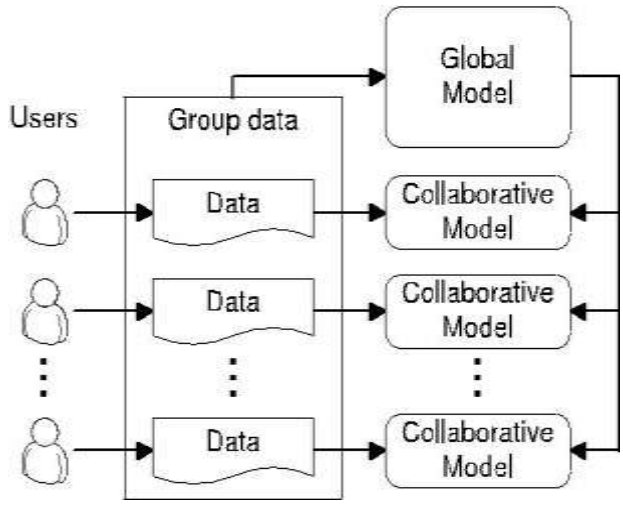

Fig.1 Multitask Model

Synthetic Dataset:

We used a synthetic dataset designed to show that solving multiple related tasks jointly outperforms the solution that treats each task in isolation. The goal is to discriminate the two classes (positive class and negative class) in a two-dimensional plane with non-linear decision boundaries. The results of COML are proven to be statistically significant in Figure1.When additional covariance information is incorporated, CW-COML is still better than the AROW based methods. In general, the proposed collaborative online multitask learning algorithms [3] achieve lower cumulative error rates for all of the five tasks, which show that they are effective in learning problems with a common shared representation across multiple related tasks. The failure of the Global model is consistent with our intuition that learning related tasks via a single model is inappropriate as it ignores the individual task characteristic. A naive combination of global and individual models is also ineffective, as indicated by the suboptimal performance of the Simple method.

First, the proposed COML consistently beats the other online learners in terms of error rate and F1-measure. In particular, in accordance to the results from the synthetic dataset, learning tasks collaboratively outperforms the baselines Global model, Personal model, and a simple combination of either.

Second, the performance of the proposed collaborative online multitask learning [6] methods is better than that of the two batch learning algorithms (MTFL and TRML). It should be noted that compared to online learners who update models based only on the current sample, batch learning methods have the advantage of keeping a substantial amount of recent training samples, at the cost of storage space and higher complexity. COML does not store recent training samples. It only uses the current training sample and a simple rule to update the model. In contrast, batch learning algorithms need to keep a certain number of recent training samples in memory, leading to extra burden on storage and complexity. What's more, both MTFL and TRML needs to solve an optimization problem in an iterative manner. For practical applications involving hundreds of millions of users and features, the batch learning algorithms are no longer feasible, while online learners remain highly efficient and scalable.

Computational methods are widely used in bioinformatics to build models to infer properties from biological data. In this experiment, we evaluate several methods to predict peptide binding to human MHC (major histo-compatibility complex) [4] class I molecules. It is known that peptides binding to MHC-I molecules plays a crucial role in the vertebrate immune system. COML is slightly slower than the original PA algorithm. This is expected since COML has to update one additional global model. However, for a group of users, only one group model is needed. The extra computational cost is trivial compared to the combined cost to update every user model. Therefore, the proposed COML algorithm is efficient and applicable to solving large-scale problems. The simple combination of Global model and Personal model-Simple model is able to approach the best model between Global and Personal.

IV. ARCHITECTURE DIAGRAM

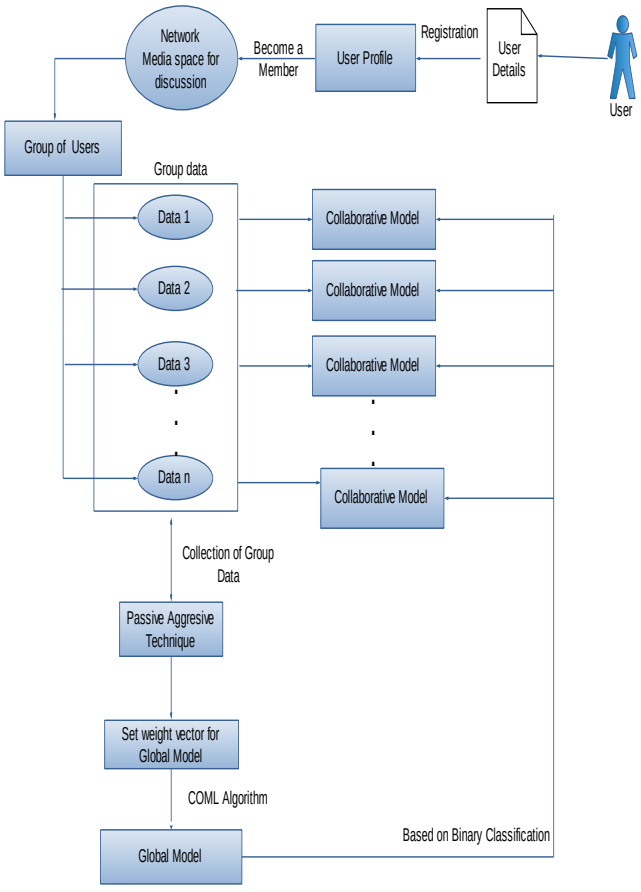

Fig.2 Architecture diagram

DOI 10.17148/IJARCCE.2015.4225

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 2, February 2015

V. DATA FLOW DIAGRAM

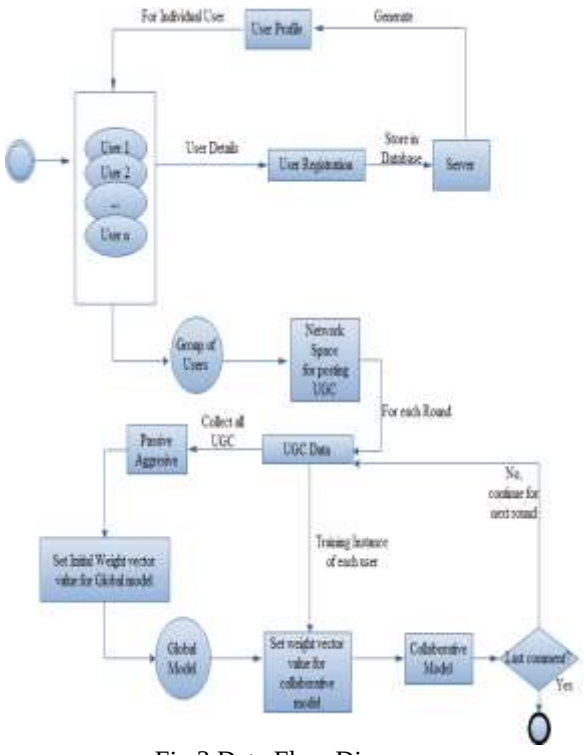

Fig.3 Data Flow Diagram

VI. PERFORMANCE EVALUATION

The experimental results demonstrate that our algorithms are both effective and efficient for three real-life applications in Figure2, including online spam email filtering, MHC-I binding prediction, and micro-blog sentiment detection task [5] in collaborative learning.

VII. CONCLUSION

A collaborative online multitask learning method that is able to take advantage of individual and global models to achieve an overall improvement in classification performance for jointly learning multiple correlated tasks. We showed that it is able to outperform both the global and personal models by coherently integrating them in a unified collaborative learning framework. The experimental results demonstrate that our algorithms are both effective and efficient for three real-life applications, including online spam email filtering, MHC-I binding prediction, and micro-blog sentiment detection task. Although the collaborative online multitask learning algorithm was firstly designed to solve the UGC classification problem, it has potential applications outside of the domains studied here. We hope to be able to extend our experiments to a more substantial size dataset and also to more applications. Our methods assume uniform relations across tasks. However, it is more reasonable to take into account the degree of relatedness among tasks in Figure3. How to incorporate hierarchies and clusters of tasks is also worthy of further study. In conclusion, our collaborative online multitask learning method is a significant first step towards a more effective online multitask classification approach.

ACKNOWLEDGEMENT

My sincere thanks to my guide Prof.S.Jenifer Asst.Professor Department of Computer Science and Engineering, Parisutham Institute of Technology and Science, Thanjavur for her help and guidance.

REFERENCES

[1] G. Cavallanti, N. Cesa-Bianchi, and C. Gentile, "Linear algorithms for online multitask classification," *J. Mach. Learn. Res.*, vol. 11, pp. 2901–2934, Oct. 2010.
[2] T. Evgeniou, C. A. Micchelli, and M. Pontil, "Learning multiple tasks with kernel methods," *J. Mach. Learn. Res.*, vol. 6, pp. 615–637, Apr. 2005.
[3] A. Saha, P. Rai, and S. Venkatasubramanian , "Online learning of multiple tasks and their relationships," in *Proc. 14th Int. Conf. AISTATS*, Ft. Lauderdale, FL, USA, 2011, pp. 643–651.
[4] L. Jacob and J.-P. Vert, "Efficient peptide-MHC-i binding prediction for alleles with few known binders," *Bioinformatics*, vol. 24, no. 3, pp. 358–366, Feb. 2008.
[5] G. Li, S. C. H. Hoi, K. Chang, and R. Jain, "Micro-blogging sentiment detection by collaborative online learning," in *IEEE 10th ICDM*, Sydney, NSW, Australia, 2010, pp. 893–898.
[6] G. Li, K. Chang, S. C. H. Hoi, W. Liu, and R. Jain, "Collaborative online learning of user generated content," in *Proc. 20th ACM Int. CIKM*, 2011, pp. 285–290.
[7] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer, "Online passive-aggressive algorithms," *J. Mach. Learn. Res.*, vol. 7, pp. 551–585, Mar. 2006.
[8] R. Caruana, "Multitask learning," *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, Jul. 1997.
[9] M. Dredze, K. Crammer, and F. Pereira, "Confidence-weighted linear classification," in *Proc. 25th ICML*, Helsinki, Finland, 2008, pp. 264–271.
[10] Y. Li and P. M. Long, "The relaxed online maximum margin algorithm," in *NIPS*, 1999, pp. 498–504.

BIOGRAPHY

S.Deepika received B.E (CSE) from Dhanalakshmi srinivasan Engineering college, Anna University in 2013. Currently persuading M.E – Computer Science Engineering in Parisutham Institute of Technology and Science.