Progressive Learning via Rearrangement of Noisy Labels

Adlene Ebenezer P, Shantanu Fartyal, Manish Prakash, Osama Habib, Aditya Siddharth

Abstract: ***In the recent years the innovation in machine learning has scaled up to a whole another level. Large scale learning problems need a vast variety of labels which can be collected at a low cost. Crowdsourced data offers a really low cost but it comes with a lot of noise, this means that the data collected cannot be trusted and hence can degrade the performance. Among the noisy labels, some labels can be really important. To tackle the difficulty of noisy labels and degraded performance, we offer to propose and actualize, a framework including POSTAL (Progressive Stochastic Learning of Noisy Labels), a new innovation for rearrangement of labels. Our framework gives a double arrangement. One, it sorts all the labels from the reliable one to noisy one. Two, progressively feed the data into the machine to learn.***

Index Terms: POSTAL, Noisy Labels, Progressive Learning

I. INTRODUCTION

Crowdsourced data may be a worldview wherever the knowledge is required in several pools that are much spread. Crowdsourced data provides its shoppers a good deal of benefits, for instance, vast topics and less cost. Highlights such unbelievable as these pull in additional clients to obtain large amount of data at cheap price. Despite the very fact that the crowdsourced data has been broadly labelled, regardless it neglects to capture the important information and are often noisy. There are a pair of problems that must be talked. The principal issue includes noisy labels. The crowdsourced data often comes with a lot of noise which degrades the performance of the machine. Concerns, for instance, begin in the very fact that the crowdsourced knowledge are collected from amateur and unreliable sources. The second issue with respect to the crowdsourced data is as per the degrade in the performance of the machine. As the utilization of crowdsourced data is expanding, so the performance of large-scale machines are at risk. Furthermore, among the labels that are put on crowdsourced platforms, are mostly noisy. Such a situation requires an innovation called stochastic learning, by which the noisy labels will be sorted from the reliable region to the noisy label. In this manner, when all the labels are arranged then it can be given into the system for progressive learning.

Revised Manuscript Received on May 07, 2019.

Adlene Ebenezer P Asst Professor, Department of Computer Science and Engineering, SRM IST, Ramapuram

Shantanu Fartyal, Computer Science and Engineering , SRM IST, Ramapuram

Manish Prakash, Computer Science and Engineering , SRM IST, Ramapuram

Osama Habib, Computer Science and Engineering , SRM IST, Ramapuram

Aditya Siddharth, Computer Science and Engineering , SRM IST, Ramapuram

In this paper, we have explained how noisy labels collected from crowdsourced platforms such as AMT(Amazon Mechanical Turk) and CrowdFlower can be used in a way which is efficient and at the same time cost-effective . The proposed system uses the low-cost labels collected from the crowdsourced platforms and sorts or rearranges them from the most reliable label to the noisy or un-reliable one. This labels can then be used for machine learning and this will increase the efficiency and hence the result will always be of high quality. This technique is also effective because it will decrease the learning time required by the machine.

II. LITERATURE SURVEY

In this section the paper describes the review of current market that was performed. We could find few applications that performs almost the same functions as this application

Existing System

[1] Bo Han , Ivor W. Tsang, Ling Chen, Celina P. Yu, and Sai-Fu Fung, IEEE 2018. Large-scale learning problems require a plethora of labels that can be efficiently collected from crowdsourcing services at low cost. However, labels annotated by crowdsourced workers are often noisy, which inevitably degrades the performance of large-scale optimizations including the prevalent stochastic gradient descent (SGD). Specifically, these noisy labels adversely affect updates of the primal variable in conventional SGD. To solve this challenge, we propose a robust SGD mechanism called progressive stochastic learning (POSTAL), which naturally integrates the learning regime of curriculum learning (CL) with the update process of vanilla SGD. Our inspiration comes from the progressive learning process of CL, namely learning from “easy” tasks to “complex” tasks. Through the robust learning process of CL, POSTAL aims to yield robust updates of the primal variable on an ordered label sequence, namely, from “reliable” labels to “noisy” labels. To sum up, POSTAL using screening losses ensures robust updates of the primal variable on reliable labels first, then on noisy labels incrementally until convergence. In theory, we derive the convergence rate of POSTAL realized by screening losses. Meanwhile, we provide the robustness analysis of representative screening losses. Experimental results on UCI1 simulated and Amazon Mechanical Turk crowdsourcing data sets show that the POSTAL using screening losses is more effective and robust than several existing baselines.

[2]Chen Gong, Dacheng Tao,Stephen J. Maybank, Wei Liu, Member, Guoliang Kang, and Jie YanG, VOL. 25, NO. 7, July 2016 Semi-supervised image classification aims to classify a large quantity of unlabelled images by typically harnessing scarce labelled images. Existing semi-supervised methods often suffer from inadequate classification accuracy when encountering difficult yet critical images, such as outliers, because they treat all unlabelled images equally and conduct classifications in an imperfectly ordered sequence. In this paper, we employ the curriculum learning methodology by investigating the difficulty of classifying every unlabelled image. The reliability and the discriminability of these unlabelled images are particularly investigated for evaluating their difficulty. As a result, an optimized image sequence is generated during the iterative propagations, and the unlabelled images are logically classified from simple to difficult. Furthermore, we associate each kind of features with a teacher, and design a multi-modal curriculum learning (MMCL) strategy to integrate the information from different feature modalities. In each propagation, each teacher analyses the difficulties of the currently unlabelled images from its own modality viewpoint. A consensus is subsequently reached among all the teachers, determining the currently simplest images (i.e., a curriculum), which are to be reliably classified by the multi-modal learner.

III. PROPOSED SYSTEM

Proposed the utilization of mechanism that avoids using the crowdsourced labels directly and hence saves the machine output from degrading. The utilization of the arranged labels, i.e., crowdsourcing the labels and sorting them from the trusted region to noisy region.

In this paper, we propose an in-depth method to solve the problems related with using crowdsourced noisy labels. We tend to address security and honesty problems associated with the crowdsourced data and the time wastage that is associated with the data. The proposed system tackles all these problems and does not increase the cost of the process.

The proposed system takes in crowdsourced noisy labels and arranges them. First, this ensures better quality of data being provided to the machine and hence the quality of data can be kept in check. Second, by integrating this the machine will be able to optimize the learning process. This method optimizes the vibrational inequalities in the crowdsourced data.

IV. SYSTEM ARCHITECTURE

The figure represents a procedural format of the activities that are involved in the process.

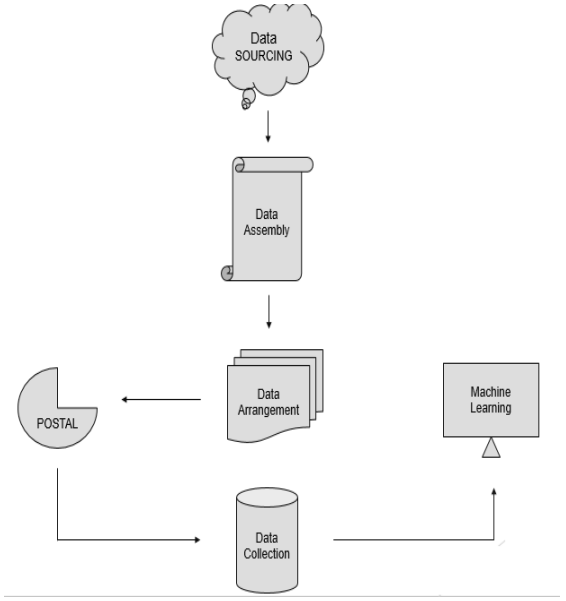

Fig.4.1. System Architecture diagram

Algorithm used to process the Data are as follows:

Decision tree-based classifiers : Decision Tree algorithm belongs to the family of supervised learning algorithms. ... The general motive of using Decision Tree is to create a training model which can use to predict class or value of target variables by learning decision rules inferred from prior data(training data) . Support Vector Machine : support-vector machines (SVMs, also support-vector networks) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis.

Certain data process algorithm is mention in fig 4.2.

V. IMPLEMENTATION

The project is focussed on two angles: data arrangement and machine learning, by progressively feeding arranged data into the machine. These two viewpoints are performed thorugh the calculations which works when summoned and these are as per the following:

- Data Arrangement
- Progressive Learning

Data Arrangement:

It compares labels (reliable or noisy) and rearranges labels that already exist within the knowledge set or are noisy. The sorting method moves the noisy or unreliable labels the end og the knowledge set hence ensures that the machine always receives correct information.

- Divide the crowdsourced data into blocks or “chunks.”
- Use POSTAL algorithm to identify the noisy labels.
- Use these values to sort the labels from reliable to noisy one.

- Replace the duplicate knowledge with a relevance the thing already exists.
- Once the arrangement is done use the arranged data for machine learning.

Progressive Learning:

Progressive learning ensures that the machine learning is optimized and use of noisy labels are avoided at all times. This learning mechanism saves the machine performance from degrading due to unreliable labels. The steps are followed in this manner:

- Arrangement of labels.
- Proper update of the knowledge set.
- Progressive learning of the data ensuring that the noisy labels are avoided at all times.

POSTAL(Progressive Stochastic Learning):

Algorithm 1 PrOgressive STochAstic Learning (POSTAL)

Input: $\lambda \geq 0$, b, the max number of epochs T_{max}, the initial learning rate η_0, the step size μ, the loss function $r(\mathbf{w}; \{\mathbf{x}_i, y_i\})$, the regularizer $\rho_\lambda(\mathbf{w}) = \frac{\lambda}{2}\|\mathbf{w}\|^2$, and the training set $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^n$

Initialize: $t = 0$, $\tilde{\mathbf{w}}^{(0)}$ randomly, the dynamic threshold $D_{th} = 1$ by the max-margin principle
for $T = 1, 2, \ldots, T_{max}$ **do**
  Preprocess: $\mathbf{w}^{(tmp)} = \tilde{\mathbf{w}}^{(epoch-1)}$ and shuffle n data points in $\mathcal{D}$ stochastically
  for $k = 1, \ldots, n$ **do**
    Select: $\{\mathbf{x}_{it}, y_{it}\}$ from $\mathcal{D}$, $it \in \{1, ..., n\}$
    Curriculum: $z_{it}(\mathbf{w}^{(tmp)}) = (\langle \mathbf{w}^{(tmp)}, \mathbf{x}_{it} \rangle + b) y_{it}$
    If $z_{it}(\mathbf{w}^{(tmp)}) \geq D_{th}$:
      Update: $t = t + 1$ and $\eta = \eta_0(1 + \lambda\eta_0 t)^{-1}$
      Compute: $Q = \partial_{\mathbf{w}} r(\mathbf{w}^{(tmp)}; \{\mathbf{x}_{it}, y_{it}\})$
      Update: $\mathbf{w}^{(new)} = \mathbf{w}^{(tmp)} - \eta(\lambda \mathbf{w}^{(tmp)} + Q)$
      Assign: $\mathbf{w}^{(tmp)} = \mathbf{w}^{(new)}$
  end
  Assign: $\tilde{\mathbf{w}}^{(epoch)} = \mathbf{w}^{(tmp)}$
  Update: $D_{th} = D_{th} - \mu\sqrt{T}$
end
Output: $\tilde{\mathbf{w}}^{(T_{max})}$

VI. MODULE IDENTIFICATION

The system contains four essential modules which build up the whole network and carry out the necessary tasks and these are as follows:

- Data Sourcing
- Data Categorization
- Data Arrangement
- Progressive Learning

Data Sourcing:

Large scale machine learning projects require vast variety of data labels. These labels are often collected from crowdsourced data platforms such as AMT(Amazon Mechanical Turk). These platforms store data shared by amateurs which are very cheap to obtain. The major drawback of the crowdsourced data is that it is often noisy.

Data Categorization:

The crowdsourced data is the passes through the POSTAL algorithm and then categorized as reliable or noisy. This helps in finding the labels which store information that is really important for the machine. Data Categorization is one of the most important part of the process because the performance of the system will be dependent on the reliable data.

Data Arrangement:

The categorized data is then arranged or sorted from the most reliable one to the noisy one. Arrangement of data ensures that the noisy or unreliable data is always at the end of the knowledge set and hence the machine performance is always high. Noisy labels contain distorted data and should never be put in the system.

Progressive Learning

This is the final part of the mechanism. This ensures that the machine learning is progressive and the knowledge set is updated time to time to maintain the reliable label at top of the knowledge set. Progressive learning is always important because it is time efficient and leaning curve of the machine is in check. ur results indicated that the previous strategy performed slightly higher, though grouping opposite categories was less time overwhelming once making ready the dataset. in addition, analysis indices like the performing loss, precise match accuracy, and MAP were measured additionally to the coaching and prediction times. Our model performed well in terms of accuracy and time consumption, with few convolutional layers as a results of adding dropout and BN layers. in addition, a example for our system was developed to demonstrate the feasibleness of the model in terms of analyzing CCTV pictures. totally different models, together with progressive models like VGG and Res internet, were compared to our custom-built model. we have a tendency to trained individual categories and classified positive and negative categories.

VII. FUTURE ENHANCEMENT

We proposed a framework that could manage the difficulties of quality and performance in machine learning. In spite of the fact that the framework manages different issues of securities in crowdsourcing data, with the expanding advances, there will requirement of mechanism to identify reliable data labels. Categorization and arrangement of data is an easy and cost-effective method to tackle this issue of inequalities in data. Alongside this, the sorting and progressive machine learning should be smoothened and upgraded in future. We also plan to make the arranged data available for others to use and hope that platforms like AMT(Amazon Mechanical Turk) implement this mechanism.

VIII. CONCLUSION

. We projected a framework to arrange the crowdsourced data. Assessed existing platforms and distinguished problems with the current framework primarily based on performance and cost associated with the process.

In this work, we discover the difficulties associated with using crowdsourced data and propose a mechanism which uses POSTAL mechanism to sort the data from reliable to noisy. The proposed mechanism is cost effective and saves the machine performance from degrading due to the noisy labels. The proposed mechanism is robust and in the near future, we will explore possibilities to eliminate noisy labels completely.

ACKNOWLEDGMENT

This paper would not have been possible without our guide Ms Adlene Ebenezer P. We would also like to show our deepest gratitude to our HOD and Dean.

REFERENCES

1. Q. Li, F. L. Ma, J. Gao, L. Su, and C. Quinn, "Crowdsourcing high quality labels with a tight budget," in Proc. WSDM, 2016, pp. 237–246.
2. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in Proc. CVPR, Jun. 2009, pp. 248–255.
3. R. Krishna et al., "Visual genome: Connecting language and vision using crowdsourced dense image annotations," Int. J. Comput. Vis., vol. 123, no. 1, pp. 32–73, 2017.
4. C. Yan, H. Xie, S. Liu, J. Yin, Y. Zhang, and Q. Dai, "Effective uyghur language text detection in complex background images for traffic prompt identification," IEEE Trans. Intell. Transp. Syst., vol. 19, no. 1, pp. 220–229, Jan. 2018.
5. J. Vuurens, A. P. de Vries, and C. Eickhoff, "How much spam can you take? An analysis of crowdsourcing results to increase accuracy," in Proc. SIGIR Workshop CIR, 2011, pp. 21–26.
6. P. Wais et al., "Towards building a high-quality workforce with mechanical turk," in Proc. NIPS Workshop Comput. Soc. Sci. Wisdom. Crowds, 2010, pp. 1–5.
7. I. Mitliagkas, C. Caramanis, and P. Jain, "Memory limited, streaming PCA," in Proc. NIPS, 2013, pp. 2886–2894.
8. A. Cotter, O. Shamir, N. Srebro, and K. Sridharan, "Better mini-batch algorithms via accelerated gradient methods," in Proc. NIPS, 2011, pp. 1647–1655.
9. Y. Nesterov, "Efficiency of coordinate descent methods on huge-scale optimization problems," SIAM J. Optim., vol. 22, no. 2, pp. 341–362, 2012.
10. A. Agarwal, D. P. Foster, D. Hsu, S. M. Kakade, and A. Rakhlin, "Stochastic convex optimization with bandit feedback," SIAM J. Optim., vol. 23, no. 1, pp. 213–240, 2013.
11. Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in Proc. ICML, 2009, pp. 41–48.
12. M. P. Kumar, B. Packer, and D. Koller, "Self-paced learning for latent variable models," in Proc. NIPS, 2010, pp. 1189–1197.
13. Q. Tao, Q.-K. Gao, D.-J. Chu, and G.-W. Wu, "Stochastic learning via optimizing the variational inequalities," IEEETrans. Neural Netw. Learn. Syst., vol. 25, no. 10, pp. 1769–1778, Oct. 2014.
14. L. Bottou, "Large-scale machine learning with stochastic gradient descent," in Proc. COMPSTAT, 2010, pp. 177–186.
15. Q. V. Le, J. Ngiam, A. Coates, A. Lahiri, B. Prochnow, and A. Y. Ng, "On optimization methods for deep learning," in Proc. ICML, 2011, pp. 265–272.

AUTHORS PROFILE

Adlene Ebenezer P Asst Professor, Department of Computer Science and Engineering, SRM IST, Ramapuram

Shantanu Fartyal, Computer Science and Engineering , SRM IST, Ramapuram

Manish Prakash, Computer Science and Engineering , SRM IST, Ramapuram

Osama Habib, Computer Science and Engineering , SRM IST, Ramapuram

Aditya Siddharth, Computer Science and Engineering , SRM IST, Ramapuram