Paper Critique - 5

Intro of Data Mining - Fall, 2021

Student

Aissata Diallo (Undergraduate) & Sheng Kai Liao (Graduate)

Date:11/28/2021

Paper:

Unifying Domain Adaptation and Domain Generalization for Robust Prediction across Minority Racial Groups

Critique: (Eddy)

In the machine learning domain, domain adaptation and domain generalization are really useful to develop robust and effective machine learning models in real-world applications since the real world datasets sometimes contain various biases due to the difference of samples and the differences of data collection procedures. Therefore, according to Khoshnevisan et. al, this paper provides a novel model, multi-source adversarial domain separation model (MS-ADS) framework, which is able to train data from different domains and even generalized into other unseen domains.

The object function of this paper is to train a machine learning model which can predict the sepsis and septic shock regarding the datasets from different hospitals which may contain bias due to not only having different data collection workflows but also having different proportions of racial population of patients. In order to train the model with different domain's dataset, the basic concept of this paper is to transfer the local representation from each dataset into one global feature space. In order to achieve this, authors firstly extract the features of the different domain datasets with a local RANN model which is trained with it and implement a global RANN model to convert all local representations data into global representations. By doing so, the models are able to learn the global representations which contain information from different domain's datasets.

In their experiments, they use Three EHR datasets that each document different hospital's data. The experiments are defined as three phases, domain adaptation across the three dataset, domain generalization to unseen dataset and unseen racial group across all datasets. For the domain adaptation experiment, they compared their model with several VRNN models which trained with one or all datasets and a multi-VRADA model which trained with three datasets, and the result shows that their model has stable and good performance over all RANN models whether trained with single dataset or all datasets. Furthermore, in individual dataset testing, their model still gained competitive performance compared to the RANN model which trained with the specific dataset. Also, in domain generalization experiments, they trained their model and the baseline with two datasets and tested them with the third dataset to test the adaptation from the two. Still, their model has better performance than other baseline. And finally, they trained their model and the baselines on one particular dataset with specific racial patient data and tested with the other hospital's datasets with same or other racial. The result turned out that their model has outstanding performance over all baseline.

In short, this paper provides a multi-source adversarial domain separation model (MS-ADS) framework to deal with the systemic bias and covariate shift from different domain dataset and the data point among different types of samples. According to their result, their model not only has better performance than most of the baselines but also has stable performance over all experiment sets. In my opinion, this model could be referred to as a solution while the developer

needs to deal with multi-datasets training and as an alternative to achieve global optimization with different but similar datasets.

Paper:

Diversity-aware k-median: Clustering with fair center representation

Critique: (Eddy)

Nowadays, precision calculation and algorithm fairness are getting important for machine learning applications in the real world. For example, critical decision-making systems include determining credit score for a consumer, insurance risk score computing, pre-screen resume application, dispatching patrols for predictive policing and so on. To implement machine learning model to above examples, it is essential to make sure that the model is trained with minimized bias and discrimination. According to Thejaswi et al., this paper introduced a novel problem for diversity-aware clustering in which they tried to minimize the cost of clustering in order to obtain a specified minimum number of cluster centers to solve the problem.

In their paper, they focused on addressing two application scenarios in the real world, committee selection and News-articles summarization. Committee selection means that in the real world, we often select committees to represent the underlying population data for our task. In this way, the data is usually biased within specific scenarios. News-articles summarization means that sometimes the news and articles resource still have bias as well. It is important to ensure that your data resources are diverse enough to minimize the bias. In order to address these two application scenarios, they introduce a novel formulation of diversity-aware clustering with representation constraints which is not only able to minimize the fraction of the data in specific aspects.

In their experiment, they implemented their method into several real-world dataset from the UCI machine learning repository and compared their method with a baseline which is basically a local-search algorithm without any representation constraints. For evaluation, they employed a new scale, price of diversity (POD), which is the ratio of increasing cost based on the solutions obtained from the baseline and their methods. According to their observations, while the minority fraction of the dataset increased, the POD increased, which means that the clustering model required more cost of clustering in order to solve the problem. Also, they report the running time for solving the problem. Based on their research, there is no significant difference between their methods but for particular datasets, both methods have efficient performance. Yet, despite this increase in time, there is no significant improvement in the cost of the solutions. Finally, authors analyze the relaxed objective which focuses on the behavior of their methods with different representation constraints. According to their result, while the constraints increase, the POD increases really small. And in all dataset experiment cases, the POD are close to zero which means solutions that have very few constraint violations and their clustering cost is almost as low as in the unconstrained version.

In short, this paper provides a novel formulation of diversity-aware clustering which is able to avoid under-representations of the clustering and minimize the bias in the dataset. In addition, they conducted several experiments with real-world datasets to verify their

assumptions. In my opinion, I think this paper could be used to estimate the feasibility of the machine learning objective function and the time of finding solutions based on certain datasets.

Paper:

Poisoning Attacks on Algorithmic Fairness

Critique (Aissata)

Based on research in adversarial machine learning it has been determined that the performance of machine learning models can be seriously compromised by injecting even a small amount of poisoning points into training data. Even though the effects of poisoning attacks on the model accuracy have been studied, there is still more evaluation that needs to be done on its potential effect on other model performance metrics. In this research, Solans et al. show an optimization framework for poisoning attacks against algorithm fairness, and create a gradient based poisoning attack that is purposefully created for the classification disparities among different groups in the data. Through their research Solans et al. are able to empirically show that the attacks are effective not only in the white box setting in which the attackers have access to the targeted model, but also in more difficult black box scenario where the attacks are optimized against a substitute model that is then transferred to the target model. The goal of this paper was to show the potential harm that an attacker is able to cause to machine learning if the attacker is able to manipulate and misclassify the training data.

In the experiment the researchers use aforementioned notation to formulate an optimal poisoning strategy. This is done in order to maximize a loss in function on a set of validation samples. Next, for attacking algorithmic fairness they define an objective function that is used in terms of loss validation, which allows for the compromise of algorithmic fairness without affecting the classification accuracy significantly. Furthermore, they generate a gradient based attack algorithm that allows them to solve the given optimization problem. They then proceeded to define attack initialization in the figures represented by the role the poisoning samples play. The initialization of points on the attacked pointes was found to be useless even though it is correctly classified by the algorithm because the point did not start an attack at all. They followed this up with gradient computation which computed various classifiers. Finally they used the White-Box and Black-Box Poisoning attacks with the assumption that the attacker has full knowledge of the attacked system, including the training data, the feature representation, the learning, and classification algorithm. The white box had access to the targeted systems while the black box could have been used against different classifiers by using the same algorithm. With this method they were able to find that the effect of any attack increased with the number of poisoned samples.

Overall, Solans et al. were able to demonstrate that attackers can actually alter the algorithmic fairness properties of a model even if there exists disparities in the training data set. The researchers hope to introduce a novel way to set off adversarial attacks targeting algorithmic fairness in different scenarios. They also believe that investigating such vulnerabilities will help design more robust algorithms and counter measures in the future. I enjoyed reading this research

and I think this would also be useful in fields such as cybersecurity especially when it comes to vulnerability detection.