Machine Learning Homework-1 (Goutham Deepak)

For the assignment, I have chosen three research papers that talks about different aspects of active learning. These papers come up with innovative approaches to improve active learning. The first paper is **"From Theories to Queries: Active Learning in Practice"**. It focuses on bridging the gap between the theoretical models in active learning and practical implementation. Many challenges are faced when active learning is implemented in practical problems. The author has come up with strategies that can be used to solve them. The second paper I chose was **"Improving Generalization with Active Learning"**. It introduced a novel active learning approach using two neural networks to improve the efficiency of the model. The third paper I chose was **"Learning Active Learning from Data"**. In this paper active learning was treated as a regression problem. So this allowed the model to learn more efficient strategies for querying.

Paper: From Theories to Queries: Active Learning in Practice

Link: https://proceedings.mlr.press/v16/settles11a/settles11a.pdf

Problems Trying to Solve:

- In traditional machine learning, the performance of the models improves with the amount of labeled data used for training. But obtaining labeled data isn't easy. It is costly, time consuming and requires expert knowledge to label. Randomly selected samples with labels can be inefficient as there can be redundancy. Sometimes they may not provide new information and so they become inefficient.
- Many theoretical models of active learning have been studied, but their implementation in real world situations isn't straightforward. So, the gap between theory and practice must be studied and considered during implementation. Various practical challenges will be faced when deploying these methods in real life problems.

How Does It Solve the Problem?

The paper talks about different ways to address the problems of active learning to make it more efficient for real world applications.

Different Active Learning Strategies:

- **Uncertainty Sampling**: This method focuses on labeling examples where the model is most unsure about its predictions.
- **Query by Committee (QBC)**: Multiple models are trained on the same data with variations. All these models will vote on the correct label and if they disagree, then the example is sent to the oracle for labeling.
- **Expected Model Change**: It selects examples that would cause the most significant shift in the model's predictions. So by using these, the model learns the most from this data.
- **Expected Error Reduction**: The model picks up examples that it thinks will help reduce its mistakes in future predictions. So this new data that is chosen will improve the model.

Different practical issues should also be handled so that the data is reliable, and the model can query accurately and learn and for these solutions.

These issues were discussed, and the solution was given. Like the labelers may not always be reliable, as people can make mistakes when labeling data. The solution discussed in the paper was multiple annotators can be used for the same data and the aggregated response can be taken to obtain a more reliable label. Instead of querying the oracle one example at a time, different methods can select groups of examples. By doing them in batches time and resources will be saved.

Cost can also play a role. Data labelling is costly and time consuming. This should be factored into the decision-making process so that selected examples are informative and cost-effective.

The paper also talks about combining active learning with approaches like semi-supervised learning so that both labeled and unlabeled data can be used.

Novelties and Contributions:

- **Bridges the Gap Between Theory and Practice**: The paper provides a detailed review of active learning, and the challenges faced when implementing it in real-world scenarios. Many papers I have come across don’t discuss this in detail as they tend to stick to theoretical implementation.
- **New Approaches**: Different approaches are introduced to help solve real-world challenges like noisy oracles, varying label costs, and batch-wise selection. This is important as in practical scenarios, data will be imperfect and limited.
- **Domain-Specific**: It explains about how active learning has been applied to practical problems in different domains like NLP, Computer Vision, Medical Imaging and Healthcare, Speech and Audio Processing, Finance, and other applications where the constraints have to be unique.

Downsides

- **Lack of Depth**: Although it covers different strategies and challenges, it doesn't go into depth on any particular algorithm. This could have been explained with detailed experiments or by comparing the performances of different algorithms.
- **Theoretical**: Although the aim of the paper was to bridge the gap between theory and practice, some models remain very theoretical and weren’t given much real-world insights. I think that more practical case studies could have been discussed for specific issues.
- **Need for Experimental Results**: More empirical data could have been used to obtain results. There was no empirical data provided to validate if these approaches would work in real-time scenarios.

Paper: Improving Generalization with Active Learning

Link: https://link.springer.com/article/10.1007/bf00993277

Problem Trying to Solve:

- Acquiring labeled samples is expensive, time-consuming, and requires expert knowledge. So new alternative methods are required that are more efficient in terms of time and resources and can still achieve high accuracy.
- When limited number of labeled samples are available, then the model will fail to generalize. So new methods have to be incorporated to make the model generalize better.
- In active learning querying is done by the model when it is uncertain about anything. This process is very time-consuming and inefficient. If the model does not query enough points, it may learn the wrong labels. This will result in poor learning outcomes.

To solve these problems, the paper introduces a novel approach to develop an innovative active learning strategy.

How Does It Solve the Problem?

The paper proposes an innovative solution using active learning with a method called selective sampling. This method has two different neural networks to define the region of uncertainty that will help the model decide which data points to query for labeling. One is the most general network, and the other is the most specific network.

The models starts with a small set of labeled examples to initialize both the most general and the most specific networks.

- **The Most General Network:** It is a lenient approach that is highly inclusive in classification. It tries to cover as many points as possible in the positive class to make sure that no positive examples are excluded.
- **The Most Specific Network:** This is a strict approach that is very conservative in classification. It includes points in the positive class only if it is very confident about them. False positives are minimized as much as possible.

The region of uncertainty is the area where the two networks disagree. In this if one network classifies a point as positive and the other classifies it as negative then it is sent for labelling to the oracle. So now the model will avoid wasting time and resources on points it is already confident about and just focus on these. These labels are added to the training set and the two networks are retrained. So this will help reduce the region of uncertainty.

When the process repeats, the regions of uncertainty will become smaller. So the model becomes more confident. Now the model will reach the optimal level of generalization accuracy. When this is at an acceptable level where further querying does not significantly improve the model's performance, the process stops.

Novelties and Contributions

- **SG Network**: It uses the most general and most specific networks to identify the region of uncertainty and determine which data points need to be queried. By querying only the doubtful datapoints not all of them will be sent to the oracle. So this saves a lot of time.

- **Real-Time Application**: The paper also has experimented on how SG networks can be used to improve learning efficiency in various real-time scenarios like:
  - **Triangle Learning Problem**: A 2D problem designed to distinguish points between two non-connected triangles.
  - **Real-Valued Threshold Function**: It involves a threshold function in a higher-dimensional space with 25 input variables. The goal was to determine whether the sum of the inputs was above or below a certain threshold
  - **Power System Security Analysis**: This was a practical real-world problem. The goal was to assess the stability of an electrical power system under different load parameters to see if the system was secure or at risk of overload

  In all of these, the results showed that the SG-network was effective to approximate the region of uncertainty which lead to a high accuracy.

Downsides

- **Computational Complexity**: There are two separate neural networks that has to be maintained. For high-dimensional data or complex models, it would become very resource-intensive.
- **Noisy Data**: The method assumes that the data will always be without noise which isn’t practical in real time scenarios.
- **Scalability**: As the size of data increases, the process of querying and labeling could become too much especially when large datasets are used.

Paper: Learning Active Learning from Data
Link: https://proceedings.neurips.cc/paper/2017/file/8ca8da41fe1ebc8d3ca31dc14f5fc56c-Paper.pdf

Problems Trying to Solve:

- Active learning methods rely on fixed rules to decide which data points should be labeled next. It may work well for simple models but will fail with more complex models like deep neural networks. They also do not learn from past experiences or improve when more data is added and labeled. They continue to follow the same rules regardless of effectiveness.
- Labeling data is time-consuming and expensive especially in fields like medicine. Existing models are not efficient in selecting the most informative examples. So this can lead to labeling with unnecessary costs.
- Traditional methods are usually designed for specific domains, and they are not generalized. For instance, active learning strategies optimized for image recognition may not work well for text classification.

How Does It Solve the Problem?

The paper introduces an approach called Learning Active Learning (LAL)**.** In this active learning is treated like a regression problem. The goal is to predict the benefit of querying data points based on past experiences. So by doing this, the model will learn how to effectively query data. A prediction is done on how much labeling each example will improve the model's performance.
There are two types of LAL:

- **LAL-Independent**: It is a simpler version that is trained once on a dataset and then used without further updates. It is more efficient and used in environments where data is static.
- **LAL-Iterative**: It is a more advanced version that updates the model continuously as more data is labeled. It is used in environments where the data is dynamic, that is it changes every time. This makes it adaptive.

By learning from past experiences, LAL helps to save cost and time when labeling. Since LAL uses regression models, the model learns general patterns from past data that make it query more efficiently. Generalization is also better in different domains. An LAL model trained on image data might perform well on text data if there are similarities between them.

Novelties and Contributions:

- **Using Active Learning as a Regression Problem**: LAL is used to allow active learning strategies to be learned from data rather than being predefined by fixed rules. So now more efficient querying can be done based on past experiences.
- **Multiple Variants**: By using the two variants of LAL, different needs are taken into account. LAL-Independent is computationally cheaper, while LAL-Iterative is a more dynamic approach that adapts continuously.
- **Not Data-Specific**: Different types of data can be used. This method can be implemented in image recognition, natural language processing, medical, etc. I find this intriguing as it can be very useful in interdisciplinary projects.

Downsides

- **Computationally Expensive**: Since most data is dynamic, the LAL-Iterative model is the most efficient solution. However, it is computationally very expensive, especially when large datasets are used or there are complex models. So, when computational resources are limited and quick decisions are needed this model may not be effective.
- **Quality of Data**: To train the regression model, high-quality data is required. If the training data does not cover a wide range of scenarios, then the model won't be able to generalize on unseen data.
- **Overfitting**: Regression model always has the risk of overfitting. So, the model might perform well on the training data but many not generalize new examples.