Good morning(afternoon), everyone! We are CS520 Group 23 who reviewed and are going to give a presentation about the paper named 'Computing Rule-Based Explanations by Leveraging Counterfactuals'. My name is Haeun Suh and I’ll be today’s speaker.

The paper our group reviewed is related to 'Data Provenance', one of the topics we studied in class, and aims to introduce a new advanced explanation system that can be applied to sophisticated machine learning models used for high-stakes decisions. We will proceed with our presentation in the order of the index presented in this slide. This order is like the way the paper is developed. It proceeds in the following order: motivation, which presents the need for the algorithm, terminology, which establishes the main concepts, algorithm, which introduces the algorithm, evaluation, which introduces the results of evaluating the algorithm, and limitation, which specifies the limitations of the paper. In conclusion, we will briefly talk about our group's impressions and conclusions about the paper.

Before starting, let us talk about the concept 'Data Provenance’. Through the class, we learned that Data Provenance is a Metadata describing the origin and creation process of data. However, here we would like to explain Data Provenance as more straightforward through a slightly modified example from the example included in the paper we reviewed.

For example, let us assume that a customer named Jesse visits a bank to apply for a mortgage loan. Let us say the banker, Laura, takes over the relevant documents from her and enters them into the bank's computer system. The bank operates an automated loan screening system, so Laura only needs to check the screening results based on the information she entered. However, unfortunately, according to Laura's screening results, Jesse was found to be ineligible for the loan.

Here, depending on the specific system the bank operates, you can think about how Laura would report the results. For example, Laura could explain that Jessie was rejected because she had features that put her in a group that did not allow her to apply for a mortgage loan. In another case, Jessie was rejected because she does not have the group's features to be approved, and Laura might reply that she could be approved for a mortgage loan if she achieved those features.

However, in any case, Laura or the bank must be able to explain that the loan reviews in our system were conducted in a reasonable and fair manner. For example, in the first case, the customer's interest is in getting approved for a mortgage loan, so the reason for the rejection may not be very important. On the other hand, in the second case, Jesse can question the system because her friend Paul had the same features as her and was approved for a mortgage loan. The bank's customers, including Jesse, are not interested in how overly complex or limited-access electronic screening systems work. We are only interested in ensuring that the system operates in a reasonable and fair manner. Non-IT employees, including the banker Laura, are also not interested in the system's algorithms; they only want to know on what basis the system works. This is because the system is already so complex and processes so much data that people cannot control it.

Let us recall once again the definition of data provenance mentioned earlier. When considered in relation to the case, we can now guess that a rational algorithm is needed that encompasses the origins and application logic of a system that applies complex machine learning, such as a loan screening system. With this goal in mind, the authors of the paper we reviewed developed a more efficient and persuasive algorithm and introduced it in the paper.

In summary, this paper is a method of mixing rule-based explanations and counterfactual explanations in a way that can improve the efficiency of the rule-based explanation system while overcoming the incompleteness of the counterfactual explanation system under the premise that the duality theorem is established. We are introducing a new algorithm.

We found that there were several concepts in the paper that were relatively unfamiliar or needed to be clarified in the development of the paper. Representative examples include duality theorem, rule-based explanation, and counterfactual explanation. Before proceeding with the explanation of the algorithm, we will briefly explain these concepts.

Before we begin the explanation, let us recall the example we looked at in the last step. We have seen that a bank can have two types of explanation systems.

In the first case, the system extracts the most common features (i.e., minimum number of) of features from customers who have already been rejected for a loan and checks whether the loan applicant meets the conditions. Although it does not confirm how customers are approved for a loan, the reasons for being rejected are clear and logical. This system judges the value of new data based on a set of the most common rules. Such a system can be said to be a rule-based explanation system.

According to the rule-based explanation system, when the three properties of Relevance, Global Consistency, and Interpretability are satisfied for a specific instance, it becomes a system that can explain that instance. Relevance means that the instance must be relevant to the rule, which means that the rule must already contain the same value as the instance. Global Consistency means that all instances corresponding to the rule must be consistent. For example, in a loan review system, this means that all instances included in the rule must have been rejected for loans. Finally, Interpretability states that rules should be the most common and simple and, therefore, have a small cardinality. In other words, the redundant rule should be excluded as much as possible.

One more concept that should be mentioned is the consistency of the data. In general, because it is realistically difficult to check all characteristics, the scope of consistency checking in the paper is limited to the database, and this property is referred to as data consistency.

On the other hand, the counterfactual explanation system is an explanation system of a more purposeful nature. Like the second case in the previous example, through counterexamples, i.e., cases where the screening was passed, you can specify the features that a specific customer needs to change to pass the screening next time. The main purpose of the counterfactual explanation system is to identify counterfactual instances by checking whether the judgment result changes when the value of the instance included in the existing rule is changed. Therefore, a customer may be judged as unable to pass the loan screening unless he or she has an exception value for that feature, that is, a counterfactual instance.

To be the counterfactual instance, it is necessary to satisfy two properties. First, the instance must be feasible and plausible with respect to the original instance value. Feasibility imposes constraints on the new values, while plausibility imposes constraints on how the new values in the counterfactual instance differ from the target instance. The authors propose to express these two features as PLAF predicates, which are composed of a conjunction of predicates for each feature of an instance. Second, using a function that calculates the difference from the old value to the new value of each feature with respect to the target instance, PLAF constraints are scored and ranked by their distance from the target instance.

As seen earlier, the two explanation systems have their own pros and cons. The authors offer a new approach that integrates the two systems, but a rationale is needed to make it possible. In the paper, the author introduced the 'Duality and Duality Theorem' and proved the validity of the algorithm to be presented in the future by proving that the two systems have 'Duality' that complements each other. So what does ‘Duality’ mean? In a mathematical sense, 'dual' is close to meaning a symmetrical relationship. In this regard, the author devoted part of the space to explaining the concepts of 'duality' and 'duality Theorem' to prove that rule-based explanations and counterfactual explanations are, in fact, symmetrical and combinable systems.

In the paper, the duality of the two systems is proven by proving the related lemmas and Theorem. However, we can check duality in a more intuitive way.

Through the data consistency mentioned by the author and the properties that a counterfactual instance must have, we were able to confirm that the corresponding instances consist only of common features. This means that despite the differences between some features, the two corresponding instances are symmetric, and each feature forms a one-to-one relationship. Since we know that a one-to-one relationship has an inverse function, we can easily verify that two instances, or two systems, are a dual relationship. Assume that you are matching different data schemas. Schemas that have the same characteristics in a one-to-one relationship are virtually identical and can be combined in any direction to produce equivalent results. In conclusion, the rule-based explanation system and counterfacutal explanation system are dual relationships, so they can be used in combination.

Following the discussion, the paper introduces three algorithms, GeneticRule, GeneticCF, and GreedyCF. The postfix in each name suggests the system the algorithm is using. In other words, the GeneticRule algorithm applies only Rule-Based, and the other two algorithms apply Counterfactual. Of these, GeneticRule was introduced as the base algorithm for the other two algorithms.

The explanation of individual algorithms cannot be covered here due to time constraints, but if you are interested, please refer to the summary that we will upload later. First, the way each algorithm works is the pseudocode included in this slide.

First, GeneticRule is the base algorithm of the other two algorithms and is responsible for finding rule combinations from a given data set. Repeat crossover,mutation, and sort by fitness score until finding K candidates that are consistent on both the D and s samples of the data set from the more general INST space.

The other two algorithms are basically extension algorithms to Genetic Rule and show an advanced form that includes new rules in existing rules by introducing a counterfactual explanation system through the CFRules function and consistentCF function. The difference lies in the cardinality of the candidate rules. GreedyRuleCF differs from GeneticRuleCF in that it tries to achieve minimum cardinality by including only the optimal fit among candidates.

Here, a selectFittest function is used to classify the candidates, which is used in each algorithm to classify the candidates according to the degree of consistency and leave the optimal combination. However, in the case of the GreedyCF algorithm, it does not use this method because it uses a method of sorting according to cardinality.

In the paper, the authors mentioned the details and results of the experiments they conducted to verify the effectiveness of the algorithm they developed. As in the case of algorithms, it is difficult to describe the specific details of individual experiments here, so only the results of the experiments will be described.

The four data sets shown in the table above were used to evaluate the algorithm. As you go from left to right, you can see that the dataset becomes more complex and atypical. Note that the counterfactual explanation model applied to the newly introduced algorithm was considered a black box and was reflected in the form of borrowing the GeCo model after evaluating 13 existing models. Evaluation of the model can be found at the GitHub address above. The CFRules and consistentCF included in the algorithm were not implemented by themselves but were applied utilizing GeCo.

Additionally, the authors strengthened the effectiveness of the verification by including a comparison with existing systems Anchor and MinSetCover in the evaluation.

The first assessment performed was on quality in terms of consistency and interpretability. In other words, the evaluation concerns the effectiveness of the algorithm. The authors evaluated three algorithms and two benchmark algorithms on the synthetic classifier version, which refers to the rule itself, and then performed the same evaluation on the real classifier version.   
When synthetic classifiers were tested, the consistency level of each algorithm was evaluated against the same cardinality of the classifiers.

As can be seen in Figure 1 on the left, the three newly developed algorithms are superior to the two remaining benchmark algorithms in all cases, and no inconsistency is reported except for the most complex case. Among these, only the GeneticRule algorithm, which did not utilize the Couterfactual explanation system, showed a relatively low level of consistency. On the other hand, the benchmark system shows inconsistency in all cases except for the results shown by MinSetCover in the simplest case comparison. Additionally, as cardinality increases, this problematic trend increases exponentially.

As can be seen in Figure 2 on the right, the tests applied to the Real Classifier test were somewhat more realistic than the previous results and a low level of consistency was reported for the three newly developed algorithms. Among them, the Genetic algorithm that was not leveraged with a counterfactual explanation showed a low level of consistency like that of the benchmark algorithm. One notable point is that unlike GeneticRuleCF, GreedyRuleCF does not report any redundancy.

The second evaluation performed was in terms of run-time comparison, that is, efficiency.

According to the results of the synthetic classifier shown in Figure 3, the three newly developed algorithms were more efficient than the remaining two benchmark algorithms in all cases. Also, according to the results, an increase in cardinality, or an increase in complexity, affected runtime, but the counterfactual explanation leveraged algorithm, which can add one or more rules in one iteration, tended to be less affected.

The results of such run-time tests appear in a more complex form when testing real classifiers that utilize actual datasets. The authors analyzed that, in addition to cardinality, that is, the complexity factor, each algorithm's ability to handle data, such as whether one-hot encoding can be handled as a single function or strong consistency verification ability, affected execution time. In this case, what is particularly noteworthy is that the performance of Greedy Rule CF drops very sharply as the dataset becomes more complex. Despite the superior effectiveness of GreedyRuleCF, analyzed previously, its performance appears to largely offset the gains on large, complex datasets. The authors believe that the main reason is the strong consistency verification ability of counterfactual explanations.

Finally, the authors included an attempt to analyze the previously evaluated test results in more detail using microbenchmarks.

First, the authors tested the pure run time of Classifier and GeCo for each dataset. Consequently, the authors judged that the counterfactual explanation system had the greatest impact and suggested the need for improvement. Second, tests were conducted to check the number of candidate rules discovered for each algorithm developed for each different data set. According to this, when the dataset is relatively simple, algorithms leveraged by counterfactual explanations focus on minimizing data redundancy, and in the opposite case, the degree of search increases sharply by focusing on finding consistent candidates. Finally, tests were performed to determine to what extent GeCo was called for each dataset. Based on their results, the authors concluded that GeneticRuleCF is more ideal than GreedyRuleCF for large and complex datasets.

In addition to previous evaluations, the authors acknowledged the limitations of the results and mentioned future improvement plans. In addition to the parts described in this paper, we would like to add some opinions we have discovered regarding the feasibility and effectiveness of the algorithm.

The authors presented the limitations of the algorithm they developed, and the improvements needed in the future, organized into the five items above. However, I think these five elements can be largely summarized in the 'generality' or ‘versatility' aspect. Constraints represented by data consistency are useful for comparison with existing systems but limit the verification of whether the algorithm created is an effective explanation system. Additionally, data are not binary and are mutable. Above all, when data consistency is not assumed, the premise for the existing correlation proven through the Duality Theorem becomes unstable. In addition, the need for a more developed counterfactual explanation model is that the introduced algorithm relies heavily on the black-box counterfactual explanation model, so it will be difficult to evaluate its value until the counterfactual explanation model, the corresponding algorithm, is established. In summary, the evaluation of the paper's algorithm can only be properly performed if the same performance and effect can be expected even when using a more realistic dataset, a more effective counterfactual explanation model must be introduced, and the existing premise of Duality must be maintained.

So far, through this paper, we have looked at the algorithm introduced Rule-Based Explanations by Leveraging counterfactuals as a new Explanation model for complex machine learning suitable for high-risk. Our conclusion is that although the algorithm has many limitations, it is meaningful to see that a more consistent and efficient approach to complex data automation can exist. Despite the ever-accelerating black boxization of machine learning models, if a valid system can be maintained and developed while maintaining clear data provenance, the development of a better automated learning system will not be far away.

Thank you for listening to our presentation so far. Well then, I hope you have a nice day.