Abstract

Similarity-based inference has been widely used for recognition. The principle behind the similarity-based inference is that similar objects will share common properties. In machine learning, similarity-based inference is employed through various methods such clustering, k-nearest neighbors, etc. In addition, similarity-based inference is useful for controlling confounding factors in statistical causality inference.

There are several issues in using similarity-based inference in practice. The principles of the inference are applicable if the representation of objects and similarity measure used for this representation are ideal. In case these factors are not ideal, there has the inconsistency of the similarity measured based on the objects' representation with the similarity of objects' target values. In addition, in analogy-based causality inference, similar causes play the role of reference factors for assessing the relation between the cause of interest with effects. Hence, the main issue here is how to choose good similar causes for accurately recognizing confounding factors.

This work aims to solve the issues mentioned above through verifying the proposed hypothesis that conservation of diversity in selecting models and data samples can help to effective solve these issues. As such, we enrich the knowledge about the diversity preservation in machine learning.

We demonstrate issues in similarity-based inference through specific studies. The first one regards to measure the similarity between materials for effectively predicting materials' formation energies. The second one regards to control polypharmacy-induced confounding in assessing the cause of drug adverse reaction. Through these studies, we can evaluate the likelihood of our proposed hypothesis. In both studies, we focus on model interpretation and explanation based on model performance.

In the first study, we address the problem that most materials' descriptors in vector space are not ideal for representing materials for predicting formation energy, which induces the roughness of the energy surface. Hence, the similarity of materials measured based on their presentation is not consistent with the similarity of their energies. In this situation, finding an appropriate similarity measure for these descriptors may help to improve the performance of similarity-based learning models in approximating the energy surface. We hypothesize that to effectively approximate the energy function, similarity measures need to preserve the distinction of two objects in comparison with the third one. We propose a protocol for verifying this hypothesis that incorporates various methods for investigating the roughness of energy surface and similarity measures. In addition, we also proposed a method for estimating the loss of distinction of two objects in comparison with the third one when using similarity measures. The experimental results show the high likelihood of our proposed hypothesis. Furthermore, we establish general principles for effectively using similarity measures for mining materials data, which do not depend on any specific learning method.

In the second study, we concentrate on an important problem in post-marketing pharmaceutical surveillance that is drug-adverse reaction causality assessment. The main issue here is to deal with confounding factors induced by polypharmacy in the treatment. In this study, we employ reference sets constructed based on the analogy criterion -- one of nine Bradford Hill criteria to control confounding factors. This criterion states that similar drugs may cause similar adverse events. We propose a novel model, called the analogy-based active voting, for effectively assessing causal relations between drugs and adverse events. This model mimics the analogy criterion by a voting

process of similar drugs. In this context, each drug is represented by a set of its associated adverse events extracted from electronic medical records. The diversity of these sets induce the conflict in voting of similar drugs, which plays an importance role for eliminating non-causal drug-adverse reaction pairs. This case study demonstrates the importance of diversifying reference in analogy-based causality inference.

Keywords: Similarity-based inference, diversity preservation, similarity measure, confounding, analogy-based causality inference.