

Cost-Sensitive Learning with Data Acquisition

Victor Sheng

Leonard N. Stern Business School, New York University

Abstract:

Cost-sensitive learning is one of the most important topics in machine learning and data mining, and attracts high attentions in recent years. It takes costs into consideration in the learning process. Many real-world applications involve different types of costs. Two most important types of costs are identified as misclassification costs and data acquisition costs. The misclassification cost is used to measure the consequence of different misclassifications. The data acquisition cost is used to measure the cost of acquiring additional information, which would improve the performance of classification models. The first part (also the major part) of my research integrates cost-sensitive learning and data acquisition. We studied three different data acquisition paradigms (attribute value acquisition, label acquisition, and example acquisition) in cost-sensitive framework, with the goal of minimizing the sum of the two costs.

We focus on acquiring attribute values for test examples in the first paradigm. The simple scenario of this paradigm is the medical diagnosis. When a patient sees his doctor, the doctor suggests him which medical tests and in what order to be performed. We propose three categories of test strategies to determine which medical tests and in what order to be performed under the corresponding policies. Label acquisition is well known as a specific form of active learning. Most previous works of active learning do not unify the misclassification cost and label acquisition cost. We unify the two costs and develop a cost-sensitive active learning algorithm, which achieves a smaller total cost with fewer numbers of examples than other active learning algorithms. Example acquisition occurs in many real-world applications. To avoid acquire more examples than necessary, we investigate and develop two strategies: complete attribute strategy and partial attribute strategy. Partial attribute strategy further reduces the example acquisition cost by acquiring a partial example which contains values of a subset of attributes.

Biography:

Dr. Victor S. Sheng is an associate research scientist in the Information Systems at Stern Business School, New York University, after he received his Ph.D. in 2007 from The University of Western Ontario in Computer Science. He won the NSERC (Natural Sciences and Engineering Research Council of Canada, like NSF in U.S.) Postdoctoral Fellowship and the NSERC postgraduate scholarship for his research in data mining and machine learning. His research results have been published in the competitive conferences (KDD, ICML, AAAI, ECML, ICDM, etc.) and journals (TKDE) in these fields. He has published about thirty papers. He is the reviewer of the journals (JML, TKDE, etc.) and conferences.