Nitesh V. Chawla

Director, Data Inference Analysis and Learning Lab
Director, interdisciplinary Center of Network Science and Applications (iCeNSa)
Department of Computer Science and Engineering
University of Notre Dame, IN 46556, USA

Topic

Mining in distribution sensitive environments: Tackling the problem of class imbalance and predictive uncertainties.

Motivation

Modern intelligence is presenting exacting needs on the cycle of learning from data to knowledge discovery to action. While we are in the abundance of data age, this proliferation is also the *Achilles heel* in the processing and communication of information. *Modern data* is huge, relentless, deeply skewed, ill-suited, noisy and riddled with error and ambiguity. Models for knowledge discovery in the real world face the pervasive and compelling problem of irregularities in data distribution. Decisions that are optimal in expected utility can be vulnerable to catastrophic failure, and value functions that reflect the discontinuities of the real world pragmatics can quickly become intractable. Surprises can happen in uncertain environments. The class distributions may not be the same, with the class of interest being rare. The training and testing distributions can differ. The costs of making mistakes or benefits from making correct predictions may also not be constant and can evolve due to operational reasons. Various real-world applications, including but not limited to finance (credit default), direct marketing, scientific simulations, medicine, spam and fraud detection, and network security, exemplify the problem of imbalance in class distribution (the class of interest is relatively rare and is accompanied with a higher cost of error). This ubiquity of the problem in applications and challenge for data mining algorithms has propelled the problem of learning from unbalanced data sets in the ``Top 10 Challenging Problems in Data Mining Research." The Technological Roadmap of European Network of Excellence in Machine Learning has also identified incorporation of costs during learning among the most relevant topics for future machine learning research.

It is also difficult to compare or evaluate on data confidently when recorded under different circumstances and at different times. The data may have experienced a population drift or selection bias between training and testing. Optimal decisions, while they can maximize some evaluation measure or efficiency in static environments, can result in fragility for complex, uncertain, and rapidly changing problems. Reflecting on the work of French mathematician Henri Poincare on predicting the motion of planets. The solutions to the planetary motions were very sensitive to the initial conditions. This phenomenon, termed as the Butterfly Effect, implies that small changes in the input conditions can lead to exponential changes in the final predictions. Similarly, one can envision sensitivity of data mining algorithms when applied to the real-world --- class distribution may change, costs of making errors may change, features may get biased by selection, etc. A number of applications, as mentioned in the beginning, posit a requirement that the data mining models developed be robust to changes in distribution and/or at least gauge an impact of the same. Such emerging properties of data are providing exciting opportunities for fundamental and applied research. The challenges are pivotal in a variety of real-world applications requiring a literal embracing of the task of ``mining needles in haystack". The two fundamental problems that this tutorial will tackle are: mining in the presence of class Imbalance (class of interest is relatively rare and more costly to make errors on), and a framework for evaluating and monitoring classifiers under changing distributions. The first half of the tutorial will introduce the problem of class imbalance, address the scope of solutions available, present the appropriate metrics for evaluating performance, discuss the applications with case studies, highlight some of the open problems, and make a connection learning from imbalanced datasets to cost-sensitive learning. The second half of the tutorial will present the challenges and solutions to evaluating classifiers under scenarios of changing distributions. The tutorial offers a confluence of theory and applications highly relevant and pervasive problems in machine learning.