Scaling Unsupervised Risk Stratification to Massive Clinical Datasets

Zeeshan Syed, University of Michigan, USA
Ilan Rubinfeld, Henry Ford Hospital, USA

ABSTRACT

While rare clinical events, by definition, occur infrequently in a population, the consequences of these events can be drastic. Unfortunately, developing risk stratification algorithms for these conditions requires large volumes of data to capture enough positive and negative cases. This process is slow, expensive, and burdensome to both patients and caregivers. This paper proposes an unsupervised machine learning approach to address this challenge and risk stratify patients for adverse outcomes without use of a priori knowledge or labeled training data. The key idea of the approach is to identify high-risk patients as anomalies in a population. Cases are identified through a novel algorithm that finds an approximate solution to the k-nearest neighbor problem using locality sensitive hashing (LSH) based on p-stable distributions. The algorithm is optimized to use multiple LSH searches, each with a geometrically increasing radius, to find the k-nearest neighbors of patients in a dynamically changing dataset where patients are being added or removed over time. When evaluated on data from the National Surgical Quality Improvement Program (NSQIP), this approach successfully identifies patients at an elevated risk of mortality and rare morbidities. The LSH-based algorithm provided a substantial improvement over an exact k-nearest neighbor algorithm in runtime, while achieving a similar accuracy.

Keywords: Anomaly Detection, Locality Sensitive Hashing, Massive Datasets, Nearest Neighbors, Risk Stratification

1. INTRODUCTION

Many important clinical conditions affect only a small number of patients within a population. While these cases occur infrequently, the consequences of these conditions can be drastic. For example, the rate of a wide range of serious complications, ranging from coma to bleeding requiring transfusion, was well below 1% in the National Surgical Quality Improvement Program (NSQIP) data sampled at over 200 hospital sites (Khuri, 2005). Fewer than 2% of the patients undergoing general surgery at these sites died within 30 days of the procedure.

Identifying patients at risk of such rare but potentially serious outcomes is challenging. For existing algorithms to risk stratify patients, a key problem to be overcome is the reliance

DOI: 10.4018/jkdb.2011010103
on labeled training data. To characterize differences between high- and low-risk patients, these algorithms require a large number of positive and negative examples. For events that occur infrequently, collecting enough positive examples (i.e., where patients experience events) requires monitoring a large number of patients. This process is slow, expensive, and burdensome to both caregivers and patients.

Recent work has focused on addressing this challenge using unsupervised machine learning (Syed & Guttag, 2010). In contrast to existing methods, which attempt to develop models for individual diseases using a priori knowledge or labeled training data, this work attempts to identify high-risk patients as anomalies in a population (i.e., patients lying in sparse regions of the feature space). The hypothesis underlying this work is that patients who differ the most from other patients in a population are likely to be at an increased risk. In earlier studies on patients admitted with acute coronary syndrome and on patients undergoing inpatient surgical procedures, unsupervised anomaly detection was able to successfully identify individuals at increased risk of adverse endpoints in both populations (Syed & Rubinfeld, 2010). In some cases, this approach outperformed other machine learning methods such as logistic regression (LR) and support vector machines (SVMs) that used additional knowledge in the form of labeled examples. This result was due to supervised methods being unable to generalize for complex, multi-factorial clinical events when only a small number of patients in a large training population experience these outcomes. An associated advantage of unsupervised anomaly detection was that it provided a single, uniform approach that could identify patients at risk of many different adverse outcomes.

Subsequent work in this area, investigating the relative merits of different anomaly detection methodologies to identify high-risk patients, has shown that classification-based, nearest neighbor-based, and clustering-based techniques are all able to successfully identify patients at increased risk following surgery (Syed, Saeed, & Rubinfeld, 2010). The best results in this case were obtained using a k-nearest neighbor approach. This approach assumes that normal data instances lie in dense neighborhoods, while anomalies occur far from their closest neighbors. The anomaly score for a patient using this method is defined as the distance from the patient to its k-th nearest neighbor in the dataset.

The k-nearest neighbor approach has two key advantages. First, it is non-parametric and does not make any assumptions regarding the generative distribution for the data. Instead, the k-nearest neighbor approach is a purely data driven method. This makes it appropriate for capturing complex cases. Second, this method is generally robust to noise, since the likelihood that an anomaly will form a close neighborhood in the dataset is low. Despite these advantages, a notable limitation of unsupervised anomaly detection using k-nearest neighbors is the computational complexity of this approach. Finding the neighbors of a patient may involve computing the distance to all other patients in the dataset.

Several variants of the basic k-nearest neighbor algorithm have been proposed to improve efficiency. In the past few years, researchers have proposed the use of spatial index structures (e.g., KD-trees, R-trees, or X-trees) (Chandola, Banerjee, & Kumar, 2009). While these approaches work well for low-dimensional data, they suffer from the curse of dimensionality and are worse than the basic k-nearest neighbor algorithm when the data has as few as 10 or 20 features (Breunig, Kriegel, Ng, & Sander, 2000). Methods to partition the feature space (e.g., into hyper-rectangles) are also similarly affected by high dimensionality and are exponential in the number of dimensions (Knorr & Ng, 1999). Recent work on improving the efficiency of the basic k-nearest neighbor algorithm has looked at pruning sufficiently randomized data to reduce average complexity (Bay & Schwabacher, 2003). This approach iteratively calculates the nearest neighbors for a data instance and sets the anomaly threshold to the score of the weakest anomaly found in a working group of size n. Using this pruning
Related Content

Pattern Formation Controlled by External Forcing in a Spatial Harvesting Predator-Prey Model
Feng Rao (2011). *Interdisciplinary Research and Applications in Bioinformatics, Computational Biology, and Environmental Sciences* (pp. 214-221).
[www.igi-global.com/chapter/pattern-formation-controlled-external-forcing/48378?camid=4v1a](www.igi-global.com/chapter/pattern-formation-controlled-external-forcing/48378?camid=4v1a)

State-of-the-Art Neural Networks Applications in Biology
[www.igi-global.com/article/state-of-the-art-neural-networks-applications-in-biology/105598?camid=4v1a](www.igi-global.com/article/state-of-the-art-neural-networks-applications-in-biology/105598?camid=4v1a)
Pharmacovigilance: Basic Concepts and Applications of Pharmacoinformatics
www.igi-global.com/chapter/pharmacovigilance-basic-concepts-applications-pharmacoinformatics/64081?camid=4v1a

Text Mining for Biomedicine
www.igi-global.com/chapter/text-mining-biomedicine/23052?camid=4v1a