

Research Statement

Amit Dhurandhar

My primary research interests are in machine learning, data mining, fuzzy systems and pattern recognition. However, the thrust areas of my research have been machine learning and data mining. Considering the large amounts of data that is collected everyday in various domains such as health care, financial services, astrophysics and many others, there is a pressing need to convert this information into knowledge. Machine learning and data mining are both concerned with achieving this goal in a scalable fashion. The main theme of my work has been to analyze and better understand prevalent classification techniques and paradigms which are an integral part of machine learning and data mining research, with an aim to reduce the hiatus between theory and practice.

Importance of studying classification problems: Classification is and has been for a long time the central problem of focus in machine learning research. The reason being that classification problems arise consistently in a wide spectrum of real life applications. For example, financial firms need to classify transactions as fraudulent or safe. Another example is gene classification in biology where genes need to be classified based on their functionality. These are just some of the numerous applications in which classification problems arise.

I will now provide a brief overview of the classification problems I have grappled with, along with my contributions below. I would like to mention here that though my advisor's primary research area is Databases with focus on approximate query processing which is significantly different from what I have been working on, his detached (from the field) perspective actually helped me in doing something original.

1. Primary Research

Classification Model Selection: Machine learning and data mining researchers have developed a plethora of classification algorithms to tackle classification problems. Unfortunately, no one algorithm is superior to the others in all scenarios and neither is it totally clear as to which algorithm should be preferred over others under specific circumstances. Hence, an important question now is, what is the best choice of a classification algorithm for a particular application? This problem is termed as classification model selection and is a very important problem in machine learning and data mining. The primary focus of my research has been to propose a novel methodology to study these classification algorithms accurately and efficiently in the non-asymptotic regime. In particular, we propose a moment based method where by focusing on the probabilistic space of classifiers induced by the classification algorithm and datasets of size N drawn independently and identically from a joint distribution (iid), we obtain efficient characterizations for computing the moments of the generalization error. Moreover, we can also study model selection techniques such as cross-validation, leave-one-out and hold out set in our proposed framework. This is possible since we have also established general relationships between the moments of the generalization error and moments of the hold-out-set error, cross-validation error and leave-one-out error. Deploying the methodology we were able to provide interesting explanations for the behavior of cross-validation. The methodology aims at covering the gap between results predicted by theory and the behavior observed in practice.

Future Work: In this framework we have thus far characterized the Naive Bayes Classifier, Random Decision Trees and the K Nearest Neighbor algorithm. The above mentioned model selection measures have also been characterized for these algorithms. In the near future I would like to analyze other classification algorithms in this framework maintaining scalability of the analysis. As a more long term goal it would be interesting to see just how far such kind of analysis can be pushed to study not only classification problems but also regression problems over finite sample sizes.

Collective Classification in Statistical Relational Learning: The other part of my research has been in Statistical Relational Learning (SRL) which is an upcoming sub-area in machine learning and data mining. Collective classification is one of the important tasks performed by researchers working in this sub-area. In collective classification instances are classified considering the class labels (sometimes also attributes) of related instances. In my research I have compared the two classification paradigms, collective classification and independent classification (i.e. traditional classification). In particular, I have provided necessary conditions under which one should be preferred over the other. Moreover, I have derived distribution free bounds for the collective classification setting where unlike independent classification correlations between data points have to be considered.

Future Work: The derived bounds are worst case and hence assume maximum dependence between interacting data points. In the future I would like to tighten these bounds by taking into account the "amount" of dependence between data points. This is definitely a challenging and interesting prospect since the current distribution free bounds for non-iid data assume independence at some coarser level of granularity.

In my current research I have suggested certain simple baseline algorithms to evaluate more sophisticated collective classification algorithms. I would like to further investigate such simple yet effective algorithms. Sampling in the relational domain is another interesting problem that grabs my attention. Unlike the i.i.d. case it is not clear in the relational setting as to how data should be sampled. At some point I would like to delve into other interesting problems in SRL such as entity resolution, link prediction which have wide range of applications (e.g. in social networks).

2. Other Research

Other than the above problems I have also worked in fuzzy systems where I have come up with an analytical expression for the choquet integral. The choquet integral is a very generic aggregation function. Popular aggregation functions such as mean, median, majority voting etc. are special cases of this aggregation function. I have worked in bioinformatics on the problem of multiple protein sequence alignment. We used suffix trees to find out the alignment and the Blosum matrix to evaluate alignment scores. In my undergrad I was a leading member of a team that developed a character recognition algorithm (and software) for scanned documents containing Devanagiri script. Devanagiri script has many more characters than the Roman script and what makes the problem of recognition truly challenging is the fact that combinations of multiple characters are also allowed. The developed algorithm was integrated in a commercial software, Chitrakan which is copyrighted by Centre for Development of Advanced Computing (C-DAC).

In the future I hope to use my well honed technical skills and innovative ability to attack important problems not just in machine learning/data mining but in other areas of interest such as bioinformatics, computational neuroscience, computational geometry and algorithms.