Int J Comput Vis
DOI 10.1007/s11263-013-0630-3

Active Rare Class Discovery and Classification Using Dirichlet Processes

Tom S. F. Haines · Tao Xiang

Received: 30 September 2012 / Accepted: 9 May 2013

Abstract Classification is used to solve countless problems. Many real world computer vision problems, such as visual surveillance, contain uninteresting but common classes alongside interesting but rare classes. The rare classes are often unknown, and need to be discovered whilst training a classifier. Given a data set active learning selects the members within it to be labelled for the purpose of constructing a classifier, optimising the choice to get the best classifier for the least amount of effort. We propose an active learning method for scenarios with unknown, rare classes, where the problems of classification and rare class discovery need to be tackled jointly. By assuming a non-parametric prior on the data the goals of new class discovery and classification refinement are automatically balanced, without any tunable parameters. The ability to work with any specific classifier is maintained, so it may be used with the technique most appropriate for the problem at hand. Results are provided for a large variety of problems, demonstrating superior performance.

Keywords Active learning · Rare class discovery · Classification

1 Introduction

Classification is an important technique, key to solving innumerable problems in areas such as computer vision. A training set is collected, and a domain expert labels each exemplar in the set with the desired (discrete) answer. The relationship between the exemplars and the labels is then learnt by a classification algorithm, such that the answer can be estimated for future exemplars. As domain experts are not cheap the greatest expense often lies in the labelling step. In many real-world computer vision problems, such as visual surveillance, computer-aided diagnoses for medical imaging and image segment labelling, the proportion of exemplars in different classes is imbalanced—the majority belong to uninteresting background classes whilst the interesting classes have few exemplars. This imbalance can dramatically increase the labelling cost, as many more exemplars have to be labelled by the domain expert to have a reasonable chance of including all the rare classes. Furthermore, the interesting minority is often unknown in advance. To give examples:

- In the Sloan Digital Sky Survey most of the survey images of galaxies and quasars capture known phenomena, whilst unusual phenomena, that could be evidence of new science, constitute only 0.001 % of the total data set (Pelleg and Moore 2004).
- When detecting buildings from aerial/satellite imagery the percentage of positive examples for one data set (Maloof et al. 2003) is less than 5 %. Buildings can come in many shapes and materials, and for military scenarios buildings may be camouflaged—deliberately designed to look like something else entirely.
- Figure 1 demonstrates the inherent imbalance in the faces data set (Huang et al. 2007). This data set has been constructed by extracting face images from news articles on the Internet over a 12 month period—it shows the classical power law bias, with a few people dominating the headlines whilst the vast majority get few mentions. As a sampling of current media interest new classes can appear at any time, when events push a previously unknown individual into the news. Class discovery thus will never cease.

T. S. F. Haines (✉) · T. Xiang
School of Electronic Engineering and Computer Science,
Queen Mary University of London, London E1 4NS, UK
e-mail: thaines@eecs.qmul.ac.uk

T. Xiang
e-mail: txiang@eecs.qmul.ac.uk

Published online: 23 May 2013