Initial Plan: Predicting future crime activity utilizing machine learning and Flickr metadata

James White
Cardiff University
CM2303
Supervised by: Steven Schockaert
Moderated by: Irena Spasic

February 3, 2014

1 Project Description

1.1 Introduction to problem

Flickr photos and social media in general, are increasingly used to understand and analyse what is happening in the world around us. For this purpose, it is especially useful that many Flickr photos have explicit metadata associated with them, such as geographical coordinates, a time stamp, or (in the case of Flickr) descriptive tags.

Predictive policing is an incredibly current affair, which aims to forecast criminal activity prior to the possibility of it happening. Data mining techniques are utilized for large sets of historical data, in an attempt to identify trends and patterns that can pre-empt crime. Machine learning is a form of artificial intelligence which focusses on trying to educate or learn from some training dataset. This particular approach is therefore especially useful in attempting to make future predictions based on historical data.

The aim of this project is to make predictions regarding general and more specific types of crime in a particular demographic. These predictions will be made via the deployment of machine learning tools that consider large sets of historical crime data and associated Flickr metadata for the specified region. Predictions that are made using a training dataset can be compared to real life test data in order to analyse the accuracy of the technique used.

1.2 Proposed Solution

Initially the focus will be on making predictions using solely historical crime statistics, later followed by the incorporation of Flickr metadata taken from photos within the specified region. This will allow for analysis regarding whether social media data can influence the accuracy of machine learning predictions. By implementing a feature selection on tags that are scraped from Flickr, it will be possible to dispose of redundant and potentially invaluable metadata.

Predictions will be calculated through building a regression model that identifies a dependent variable amongst the input data. Acquiring data in a format that is suitable to create feature vectors for machine learning is therefore essential. Regression analysis is a process that is best used with numerical values, in order to expose relationships between variables and calculate an expected outcome for some unforeseen case. For this reason, the crime statistics that are collected must be converted into numerical representatives suitable for analysis.

It is hoped that correlations between specific types of crime in an area and commonly appearing Flickr tags for the region will give a greater estimate in predicting results. For example, a region that has a high usage of the tags ‘Porsche’ and ‘BMW’ may strengthen predictions regarding vehicle crime in the area. Alternatively, common tags such as ‘drinking’ and ‘clubbing’ in an area could reinforce predictions regarding likeliness of anti-social behaviour.

2 Aims & Objectives

2.1 Core Objectives

To gather sufficient representative crime statistics that will form the basis for a predictive model. Having datasets that are coherent and extensive are vital for successful machine learning. It is possible a significant data cleanse will have to occur in order to dispose of invaluable and irrelevant data.

To scrape a vast collection of metadata tags taken from photos on Flickr within a specified region. These tags will be used as extra input for the machine learning process. Flickr has a detailed API that consists of a collection of callable methods useful for this function.

To perform feature selection to acquire Flickr tags that will be most relevant to the dataset. This will aid in removing potentially redundant or inconsistent tags that could disrupt the accuracy of machine learning. Only commonly appearing tags will be utilized for the machine learning process.

To build a regression model that estimates the number of crimes committed in a particular demographic for a forthcoming period. This model will be based on historical data regarding all types of crime committed over a period of time for the specified region. Later inclusion of Flickr metadata will be used in attempt to gage more accurate results.

To build a regression model that estimates the count of a specific type of crime in a particular area for a forthcoming period. This model will be based only on historical data regarding the chosen specified crime type and associated Flickr meta tags.

To analyse whether the incorporation of Flickr metadata improves the accuracy of predictive output. By assessing predictions made when using meta tags and without using such data, it will allow for comparison regarding the impact it has made.

To produce an effective analysis of results that draws upon some conclusion regarding the overall success and results identified through carrying out the research.

2.2 Desirable Objectives

To identify crime shifts and paradigms when comparing street-level predictions with neighbouring streets and neighbourhoods. The analysis of data regarding a locations surroundings could help draw conclusions regarding transposition of crime activity.

To visualize predictions in a manner that is more simple to digest than statistical figures. This would help to make results easier to consume and identify obvious relationships or trends.

3 Work Plan

Week Commencing	Task
03/02/2014	Research and reading into data mining techniques, in particular machine learning tools provided with 'Weka'. Exploration of the Flickr API and its associated documentation.
10/02/2014	Gathering of crime statistics data that is suitable for training and testing. This data will most likely need to be cleansed and structured in a suitable manner.
17/02/2014	Begin implementation of regression model for analysing crime data and making future predictions of general crime.
24/02/2014	Continue implementation of regression model and collect results regarding crime predictions.
03/03/2014	Modify existing regression model to make predictions regarding specific types of crime in specific regions.
10/03/2014	Begin implementation of program to scrape tags from Flickr for photos taken within a specified demographic. ***Review Meeting with Steven Schockaert.***
17/03/2014	Create feature selection process for removing Flickr tags from a dataset that are invaluable, and isolating potentially useful and commonly appearing tags.
24/03/2014	Build regression model that incorporates historical crime data in conjunction with tags remaining after feature selection.
31/03/2014	Finalise implementation and collect results regarding crime prediction when using statistics and Flickr tags.
07/04/2014	Analyse the impact of making predictions when utilizing metadata from Flickr, in comparison to without. Conclude as to whether the information adds anything to accuracy of predictions by comparing to test data. ***Review Meeting with Steven Schockaert***
14/04/2014	Perform analysis of results in comparison to test data. Also perform testing with finite results on smaller pieces of software developed (such as flickr tag scraper and feature selection).
21/04/2014	Collate documents and ensure documentation is presented in a uniform manner.
28/04/2014	Ensure report is complete, covering all required sections and spend time checking grammar and perfecting.
05/05/2014	Hand in final report.