INFO411/INFO911
Data Mining and Knowledge Discovery

Project 2

Instructions:

This tasks is a real-world data mining problem. You are required to prepare a set of presentation slides which must include (1) the full name and student number of each student in the group, the contribution (in percent) of each group member, (2) a description of the task, (3) your proposed data mining approach and methodology; (4) the strengths and weaknesses of your proposed approach; (5) the performance measures that can evaluate your data mining results; (6) the results a brief discussion and a conclusion.

Below is the recommended structure of your slides:

- Introduction (define the problem and the goal)
- Methods (propose approaches, and discuss their strengths and weaknesses)
- Results (Figures and tables of data analysis)
- Discussion (discovered knowledge from data mining)

Task: Vehicular Data Mining (fraud detection)

Background:

The mining of vehicular data has numerous practical applications such as in traffic control, planning of road networks, flow prediction, intelligent driver navigation, and many more. Vehicular Data Mining is one of the fundamental data mining problems which focuses on discovering patterns by analysing vehicular data (i.e. location data). One popularly used dataset in this context is the DRAWDAD roma taxi driver dataset. The dataset is available from: http://crawdad.org/roma/taxi/20140717/

Several studies have been conducted on this dataset. A good starting point to understanding Vehicular Data Mining can be found in:
Cristian Chilipirea, Andreea Petre, Ciprian Dobre, Florin Pop, Fatos Xhafa. Enabling Vehicular Data with Distributed Machine Learning. In *Transactions on Computational Collective Intelligence XIX*, Vol. 9380, pp. 89-102, 2015.

Definition of the task:

Analyse the CRAWDAD roma taxi driver dataset for possible cases of fraud. Find cases by which taxi drivers drive from one location to another location that is neither using the shortest distance route nor the fasters route nor acceptable alternate routes (i.e. the customer is taken for a ride). We call such cases “fraud”. There are two main challenges associated with this task:

1. Find ways by which the starting point and the end point of a journey can be identified.
2. Identify what the acceptable routes (i.e. shortest or fastest) would be then compare the route taken to those routes.

There are a number of ways to approaching this task. Starting points for this task can be found in i.e.:
http://maperitive.net/
http://gis.stackexchange.com/questions/134268/mapping-raw-gps-data-to-openstreetmaps-node-way-relation

Questions to be answered are:

1. How many cases of fraud can be detected in this dataset?
2. Are some drivers more likely to commit fraud than others?
3. How common is fraud among the various drivers?

4. What are the possible data mining approaches to answering these questions?

Requirements:

1. Present a general description of the dataset and present the general properties of the dataset as far as it is relevant for this task.
2. Investigate into possible approaches to this task. Briefly explain those approaches.
3. Among the approaches that you identified, select the one that you think will work best. Explain why you think this approach will work best.
4. Present and explain the answers to afore mentioned questions.
5. Summarize: What are the most interesting things that you discovered while working on this project?