Analysis of road traffic accidents

Edited by Peter Flach

Sol-Eu-Net *IST-1999-11495 (2000-2003)*

Hampshire County Council (HCC, UK) is currently being sponsored by the AA Foundation for Road Safety Research and the CSS (formerly the County Surveyors Society) to carry out a research project ""ROAD SURFACE CHARACTERISTICS AND SAFETY". In particular, HCC wants to obtain a better insight into how the characteristics of accidents may have changed over the past 20 years as a result of improvements in highway design (direct effect) and in vehicle design (indirect effect). Part of this work is an analysis of a copy of the DETR STATS19 (Personal Injury Road Accident) Database for 1979 to 1999 to identify trends over time in the relationships between recorded road-user type/injury, vehicle position/damage, and road surface characteristics. The STATS19 Database records details about the accident, together with the vehicles and casualties involved, which in principle can be easily discerned by the police either at the scene of the accident, or when they are reported to the police at a later date after the accident. The details recorded are: accident time, date and location; vehicle type, location and manoeuvre; driver details; casualty details and severity.

This Phase II Industrial Project can be seen as the culmination point of the SolEuNet data mining workpackage because:

1. The dataset is large (1.5GB) and multi-relational.
2. The task concerns exploratory mining rather than simple prediction.
3. A large group of project partners (7) has contributed.

One of the major aims of this data mining project has been to test the collaborative data mining methodology developed by the SolEuNet project. The CRISP-DM and RAMSYS methodologies have been followed as far as practically possible. Zeno has been used extensively as collaboration platform, and SumatraTT has been used for pre-processing.

Kick-off workshop

A major factor contributing to the success of this data mining project was the kick-off workshop organised on June 9-10 2002 in Bristol to initiate the work. The aim of the workshop was to work on CRISP-DM phases 1: Business Understanding and 2: Data Understanding. An important secondary aim was to create a genuinely collaborative group spirit, and to promote the use of SolEuNet tools such as Zeno2 and SumatraTT. The workshop was attended by 15 representatives of 7 SolEuNet partners (BRI, CTU, FHG, IJS, KUL, LIACC, UOXF.BL). In addition, John Bullas from Hampshire County Council (the end-user) attended both days and took an active part in the workshop. On Sunday 8 June the program consisted mostly of presentations on Business Understanding (John Bullas, HCC), Data Understanding (Shaomin Wu & Thomas Gaertner, BRI), Zeno2 (Angi Voss, FHG), SumatraTT (Petr Miksovsky & Petr Aubrecht, CTU), and Knowledge Management (Mitja Jermol, IJS). On Monday 9 June the program consisted mostly of data mining sessions followed by discussions.

The involvement and participation of the end-user in the workshop was an important factor which made the workshop into a success. Many business and data understanding issues have been clarified during the workshop and after, as the end-user is actively participating in discussions on Zeno2. The workshop has also been successful in establishing Zeno2 and SumatraTT as collaborative platforms. A timetable and initial division of work among partners was also agreeed.

Another workshop was organised by Jozef Stefan Institute on 14 October 2002 and involved the end-user's participation.

Preliminary results

A range of data mining techniques has been applied by the participating partners. Space does not permit to do them all justice. We mention a few highlights:

- An innovative visualisation method pioneered by Leuven animated the development of accidents by location over time, and helped pointing out data quality issues regarding grid references.

- Prague used association rules to find associations between road numbers and particular classes of accidents (the end-user was particularly pleased with format and contents of findings which are now being evaluated by local domain experts).
- Ljubljana has used text mining technology and subgroup discovery to determine common kinds of accidents.
- Bristol has used dynamic subgroup discovery which has highlighted certain data quality issues.

End-user testimonial

»Analysing road safety data is a highly exploratory process which to a large extent depends on asking the right questions. While domain experts have mostly been using statistical techniques for this analysis, the project has so far been very successful in highlighting how a very large dataset can be approached and analysed from a range of novel perspectives.

The combination of a pool of datamining experts and domain experts has generated considerable synergy enabling associations, previously beyond the ability of the domain experts, to be explored and developed. This ranges from innovative visualisation of locational data over time, methods to identify data quality problems, to the use of advanced data mining methods to establish patterns and rules beyond the normal constraints of the clustering techniques commonly used in this realm.

Feedback by local domain experts is currently being obtained to assess the full value of the new findings in the real world, but the analysis of the STATS19 Database performed so far by the Sol-Eu-Net consortium holds considerable promise for the application of these technologies to other databases currently analysed with long established and limited repertoires of processing tools.« (John Bullas, Hampshire County Council)

The following partners contributed to the work:

Bristol (Peter Flach, Shaomin Wu, Thomas Gärtner, Simon Rawles), Jozef Stefan Institute (Nada Lavrac, Branko Kavsek, Peter Ljubic, Marko Grobelnik, Dunja Mladenic, Mitja Jermol), Leuven (Hendrik Blockeel, Jan Struyf, Gert Sclep, Stefan Raaymakers, D. Krzywania), Porto (Luis Torgo, Rita Ribeiro), Prague (Petr Miksovsky, Petr Aubrecht, Martin Kejkula, Jan Rauch, J. Burian), Oxford (Steve Moyle), Bonn (Angi Voss).