Preprocessing and Clustering large-scaled Data-Mining problem instances

Éric D. Taillard*, Laura H. Raileanu*, Philippe Waelti*

*HEIG-Vd, University of Applied Sciences of Western Switzerland
Route de Cheseaux 1, Case Postale, CH-1401 Yverdon-les-Bains
{philippe.waelti,eric.taillard,laura.raileanu} at heig-vd.ch

The increasing availability of data in our information society has led to the need for valid tools for its modelling and analysis. Data-Mining is the process designed to explore large amounts of business, family, or institution data in order to discover interesting models and patterns. The process can be decomposed in three main steps: acquisition, preprocessing and analysis of data. For each of the step we developed efficient and robust *C* libraries to manage, organise, transform and analyse huge amounts of data.

The first step involves the managing of large amount of data.

Second step modules includes the management of the missing values, the equal-width and equal-depth binning techniques for data smoothing as well as some different types of normalization. In the same phase, data reduction provides a convenient way to reduce the number of observations through a simple random selection or by using stratified sampling.

Third stage module is realised through an unsupervised clustering algorithm which has been modelised using a well-known p-Median location-allocation problem (see [2]). As Data-Mining often implies large amounts of data, the practical approach to analyse observations through a p-Median problem (shown *NP*-Hard) requires very efficient and robust methods. To solve large-scaled instances, the developed approach is to decompose the global problem in a set of subproblems in order to decrease the complexity of their resolution. To carry out these decompositions and subsequent optimizations phases, relatively elaborate methods like *POPMUSIC* [5, 6], respectively *CLS* [5] have been used.

The p-Median framework developed through a *C* library has shown to be very efficient to solve classic large-scaled p-Median instances based on the TSPLIB. Solution quality as well as computation times for a 11849 entities problem (known as RL11849) are shown in table 1 on the following page. Best previously known solution have been obtained using a parallel implementation of the *VNS* algorithm [1] for a time of 256000 seconds (time adjusted according to execution platforms) per instance. Results are based on a single execution of our implementation.

In the scope of the DeProLo project, granted by the *University of Applied Science of Western Switzerland*, Data-Mining and p-Median libraries as well as a Data-Mining graphical user interface are freely available on the project website at the following address: `http://deprolo.heig-vd.ch`

Table 1: Numeric results for RL11849, *POPMUSIC*(r, *CLS*(q)). Column p provides the number of clusters, Column r provides the size of sub-problems in *POPMUSIC* frame, Gap is provided in % above best solution value previously published, q refers to the number of iterations performed by the *CLS* procedure used as optimizer in *POPMUSIC* frame.

p	r	Best published	*CLS*(10)		*CLS*(100)	
		(from [1])	Gap [%]	t [s]	Gap [%]	t [s]
100	12	5855395.00	0.23	323.86	0.09	2692.42
200	14	4017110.50	**-0.08**	230.11	0.25	826.89
300	15	3210784.00	0.04	155.24	**-0.08**	475.20
400	16	2712334.50	**-0.00**	155.13	**-0.10**	389.58
500	16	2367523.00	0.06	193.24	0.13	271.74
600	17	2125355.50	0.04	207.77	**-0.28**	278.76
700	17	1932731.75	**-0.06**	197.56	**-0.05**	277.53
800	18	1775417.62	0.07	239.98	**-0.06**	277.28
900	18	1644025.75	**-0.02**	283.50	**-0.15**	292.55
1000	18	1531481.88	**-0.00**	321.62	**-0.30**	321.60

References

[1] T. Crainic, M. Gendreau, P. Hansen, and N. Mladenović (2004): "Cooperative parallel variable neighborhood search for the p-median". Init Journal of Heuristics, **10** (3), 293–314.

[2] S. L. Hakimi (1965): "Optimum distribution of switching centers in a communication network and some related graph theoretic problems". Init Operations Research, **13**, 462–475.

[3] G. Reinelt (1995). TSPLIB. `http://www.iwr.uni-heidelberg.de/groups/comopt/software/TSPLIB95`.

[4] E. D. Taillard (2003): "Heuristic methods for large centroid clustering problems". Init Journal of Heuristics, **9**,51–73.

[5] E. D. Taillard and S. Voss. POPMUSIC (2001): "Partial optimization metaheuristic under special intensification conditions". In: C. Ribeiro and P. Hansen, (eds.): it Essays and surveys in metaheuristics. Kluwer Academic Publishers, 613–629.