

Chapter 8

Case Study

Mining The Welfare Services Database

An extensive performance evaluation of ApproxMAP verifies that ApproxMAP is both effective and efficient. Now, we report the results on a real data set of welfare services accumulated over a few years in North Carolina State.

We started this research on sequential data mining in order to answer a policy question with our social welfare service data. What are the common patterns of services given to children with substantiated reports of abuse and neglect? What are its variations? Using the daysheet data of services given to these children, ApproxMAP confirmed much of what we knew about these children. These findings gave us confidence in the results. It further revealed some interesting unknown patterns that sent our clients back to investigate.

8.1 Administrative Data

There are three administrative databases used in this analysis. Most of the data comes from the NC social workers' daysheet data. It records activities of social workers in NC for billing purposes. County DSS (Department of Social Services) have numerous funding sources that are combined to pay social workers' salary. In order to properly bill the correct funding source, each social worker is required to report on their 40 hour work week. The daysheet data is a timesheet of the social worker's 40 hour week workload indicating what services were given when, to whom, and for how long. Time is accounted for in minutes. The data gives a fairly accurate picture on the various services given to clients each month. Therefore, we can convert the daysheet data into monthly services given to clients.

Now we need to identify the clients of interest. We can identify children who had a substantiated report of abuse and neglect using the abuse and neglect report database. And then, using the foster care database, we can further split the children with substantiated reports into those that were placed in foster care and those that were not. The children who

were never placed in foster care received very little services in comparison to those who were placed. Thus, the interesting patterns were found in those who were placed in foster care. Here we report on our results from children who had a substantiated report of abuse and neglect and were placed in foster care.

There were 992 such children. Each sequence starts with the substantiated report and is followed by monthly services given to each child. The follow up time was one year from the report. In summary we found 15 interpretable and useful patterns.

8.2 Results

The most common pattern we detected was

$$\langle\langle(RPT)(INV,FC)\overbrace{(FC)\cdots(FC)}^{11}\rangle\rangle$$

In the pattern, *RPT* stands for a *Report*, *INV* stands for an *Investigation*, and *FC* stands for a *Foster Care service*. In total, 419 sequences were grouped together into one cluster, which gave the above consensus pattern. The pattern indicates that many children who are in the foster care system after getting a substantiated report of abuse and neglect have very similar service patterns. Within one month of the report, there is an investigation and the child is put into foster care. Once children are in the foster care system, they stay there for a long time. Recall that the follow up time for the analysis was one year so 12 months in foster care means the child was in foster care for the full time of analysis. This is consistent with the policy that all reports of abuse and neglect must be investigated within 30 days. It is also consistent with our analysis on the length of stay in foster care. The median length of stay in foster care in NC is about 1 year with many children staying in foster care for longer.

Interestingly, when a conventional sequential algorithm is applied to this data set, variations of this consensus pattern overwhelm the results, because roughly half of the sequences in this data set followed the typical behavior shown above approximately.

The rest of the sequences in this data set split into clusters of various sizes. Another obvious pattern was the small number of children who were in foster care for a short time. One cluster formed around the 57 children who had short spells in foster care. The consensus pattern was as follows.

$$\langle\langle(RPT)(INV,FC)(FC)(FC)\rangle\rangle$$

There were several consensus patterns from very small clusters with about 1% of the sequences. One such pattern of interest is shown below.

$$\langle\langle(RPT)(INV,FC,T)(FC,T)\overbrace{(FC,HM)}^8(FC)(FC,HM)\rangle\rangle$$

In the pattern, *HM* stands for *Home Management services* and *T* stands for *Transportation*.

There were 39 sequences in the cluster. Our clients were interested in this pattern because foster care services and home management services were expected to be given as an "either/or" service, but not together to one child at the same time. Home management services were meant to be given to those who were not placed in foster care. Thus, this led us to go back to the original data to see if indeed many of the children received both services in the same month over some time. Our investigation found that this was true, and lead our client to investigate this further in real practice. Was this a systematic data entry error or was there some components to home management services (originally designed for those staying at home with their guardian) that were used in conjunction with foster care services on a regular basis? If so, which counties were giving these services in this manner? Such an important investigation would not have been triggered without our analysis because no one ever suspected there was such a pattern.

It is difficult to achieve the same results using the conventional sequential analysis methods because with *min_support* set to 20%, there is more than 100,000 sequential patterns and the users just cannot identify the needle from the straws.