Mining the Knowledge Mine

The Hot Spots Methodology for Mining Large Real World Databases

Graham J. Williams (CMIS/Cbr)
and Zhexue Huang (CMIS/Cbr)

Presented at the Tenth Australian Joint Conference on Artificial Intelligence, Perth, Western Australia, December 1997. Published in Lecture Notes in Artificial Intelligence, Springer-Verlag.

Report Number CMIS 97/31
December 1997

CSIRO Mathematical and Information Sciences
GPO Box 664, Canberra ACT 2601
Telephone: (+61 2) 6216 7042
Facsimile: (+61 2) 6216 7112
Email: Graham.Williams@cmis.csiro.au
www.cmis.csiro.au/Graham.Williams

Mining the Knowledge Mine:
The Hot Spots Methodology for Mining Large Real World Databases

Graham J. Williams and Zhexue Huang
Cooperative Research Centre for Advanced Computational Systems
CSIRO Mathematical and Information Sciences
GPO Box 664 Canberra 2601 Australia
Email: Graham.Williams@cmis.csiro.au
Phone: (+61 2) 6216 7042
Fax: (+61 2) 6216 7111

Abstract

As databases grow in size and complexity the task of adding value to the wealth of data becomes difficult. Data mining has emerged as the technology to add value to enormous databases by finding new and important snippets (or nuggets) of knowledge. With large training sets, however, extremely large collections of nuggets are being extracted, leading to much “fools gold” amongst which to fossick for the real gold. Attention is now being directed towards the problem of how to better focus on the most precious nuggets. This paper presents the hot spots methodology, adopting a multi-strategy and interactive approach to help focus on the important nuggets. The methodology first performs data mining and then explores the resulting models to find the important nuggets contained therein. This approach is demonstrated in insurance and fraud applications.

1 Introduction

With the rapid increase in the use of databases together with the dramatic increases in storage capacities and performance many businesses today maintain extremely large databases. NRMA Insurance Ltd, one of Australia’s largest general insurers has, for example, many millions of active insurance policies with several million claims being made against those policies each year. Australia’s Health Insurance Commission (HIC) maintains a record of every visit every person in Australia has ever made to a medical practitioner since 1975 (Viveros, Nearhos and Rothman 1996).

To be effective and competitive such organisations must continually update their understanding of their core business. This often involves monitoring trends in their client base and attempting to understand and hence predict changes. Although the data is often available to perform such analyses traditional (statistical) approaches are reaching their limits (Mallows and Pregibon 1996). Statistics is not alone in being hamstrung with extremely large and complex datasets. Traditional

Machine Learning and Knowledge Acquisition techniques often work well with small datasets but struggle when applied to more than a few thousand records. Run-times of days and weeks are common and the results are often so complex to be of little cognitive use. Such approaches often rely on sampling the data and identifying trends from those samples—necessitated by a lack of processing power. High performance computers now provide the opportunity to analyse all (or at least more) of the data. Hitherto hidden, but important, nuggets of information may now be discovered.

Data Mining (or exploratory data analysis with large and complex datasets) brings together the wealth of knowledge and research in Statistics and Machine Learning for the task of discovering new nuggets of knowledge in very large databases (Fayyad, Piatetsky-Shapiro and Smyth 1996). Visualisation and database research also play a central role in Data Mining. Together, these technologies have demonstrated that they can help businesses better understand their data, to rapidly identify changing trends in the data, and to discover insights that were previously missing.

Data Mining, however, is giving rise to its own peculiar difficulties. When Data Mining techniques are applied to very large databases, the amount of "discovered knowledge" itself can be beyond the capabilities of any person to comprehend and analyse. With a focus on the discovered knowledge rather than on building accurate models of the data, we are again lost in the wealth of information.

In this paper we present the **hot spots** methodology which tackles the specific problem of coming to terms with the large amounts of knowledge that can be discovered from large databases. In essence, we develop the idea of *mining the results of data mining*—turning data mining on itself to focus on the outcomes. The approach taken employs multiple-strategy learning with a post-processing step which automatically discovers hot spots in the discovered knowledge. This step uses additional data and background knowledge that may not have been used in the actual data mining. Traditional tools can be used with no or little modification to identify large collections of patterns, which can then be mined for the key discoveries. We present two case studies which illustrate this general approach to the discovery of hot spots in very large real world databases.

2 Motivation

In many businesses the identification of key client groups is a core concern. Targeting business development towards identified opportunities can give an organisation their competitive edge. Target groups are traditionally identified from market research, past experience, and business intuition. With the advent of very large databases and high performance computers, data mining provides an alternative approach to identifying key customer groups that may in the past have been overlooked, and that may continue to be overlooked by competitors.

Such mission critical information we refer to as hot spots. This is information that an organisation needs to know in order to improve its performance or its processes. In general it may not be information that describes large groups of the customer base but more importantly identifies key customer groups—groups which although small, have above average significance to the business of the organisation.

When dealing with extremely large and complex datasets, as we do in data mining, the descriptions and models built, whether from statistical tools or machine learning tools, are often extremely complex. The key knowledge or discoveries may be lost in this wealth of knowledge structures. For example, C4.5 (Quinlan 1993), when used on data with 30 or 40 attributes and many millions of records, will quite happily generate decision trees and rule sets with thousands of nodes or rules (ignoring the issue of over training). While the overall model may represent a good prediction tool it gives us little insight. New techniques are required to assist in the task of taking this model and extracting key “discoveries” from it.

3 The Hot Spot Methodology

We regard a *hot spot* as an identified set of entities which are of some particular, but crucial, importance to the domain of interest. Examples include loyal customer groups in a supermarket dataset or the group of regular high claiming insurance policy holders in a motor vehicle insurance portfolio. The discovery of these hot spots in the data can assist management in targeting their promotional efforts, for example. Simple techniques such as clustering or segmentation can help this task but are often computationally expensive and/or build groups that are not well described. An aim of data mining is to highlight areas of the data in such a way that makes sense to the domain experts—providing human understandable discoveries.

A heuristic approach to this segmentation task that we have empirically found to be effective in many real world problems involves the combination of a clustering tool and a decision tree induction tool. These are augmented with post analysis tools. This methodology we refer to as the hot spots methodology: cluster; rule induction; nugget evaluation.

Suppose we have a dataset D consisting of a set of real world entities such as a set of policy holders in an insurance company or a set of Medicare patients. We generally assume that D is relational with only one universal relation $\mathbf{R}(a_1, a_2, \ldots, a_m)$, where the a_i are the attributes of the entities, and are defined by the basic data types such as integers, reals, and strings. The dataset then consists of the set of entities $D = \{e_1, e_2, \ldots, e_n\}$, with each entity being a tuple $\langle v_1, v_2, \ldots, v_m \rangle$. For real world problems m and n are typically “large” (m may be anywhere from 20 to 1000 and n typically greater than 1 million).

Step 1 The first step of our approach is to develop a raw (and unsupervised) clustering of D. A particular clustering of D might be $C = \{C_1, C_2, \ldots, C_p\}$. The clusters are constructed to be complete and disjoint: $D = \bigcup C_i$ and $C_i \cap C_j = \emptyset, i \neq j$. A mixed data-type clustering algorithm has been used (Huang 1997). This efficient algorithm is based on a k-means clustering algorithm extended to handle categorical attributes. Other clustering algorithms tend to be computationally expensive, particularly in the context of data mining where the datasets are very large. Typically, anywhere between 10 and 1000 clusters may be constructed, depending on the size of the dataset.

Step 2 The second step records with each entity the cluster to which it now belongs. Thus, the entity becomes $\langle v_1, v_2, \ldots, v_m, c \rangle$, where $c \in \{1, 2, \ldots, p\}$. Supervised learning can then be used to build a symbolic description of the clusters. Using C4.5, for example, the resulting decision tree can be converted to a rule set and pruned. The end result is a set of rules, $R = \{r_1, r_2, \ldots, r_q\}$, usually with $q \geq p$, and usually much greater (since for each of the p clusters multiple rules will be induced). We will refer to a rule as a description of a nugget (or simply as a nugget). Each nugget now corresponds to a subset of the original dataset D. We use the notation r_i to represent both the nugget description and the nugget subset. R will be referred to more generally as the *nugget set*. Note that the nugget subsets are no longer disjoint: $r_i \cap r_j$ is not necessarily empty for $i \neq j$.

Step 3 The third and final step is to evaluate each nugget in the nugget set to find those of particular importance to the task at hand. We define the function $Eval(r)$ as a mapping from nuggets to a measure of the significance of the particular nugget r. Such a function is very much domain dependent, and is the key to effectively mining the knowledge mine. The nuggets may be evaluated in the context of all discovered nuggets or evaluated for their usefulness, novelty, and validity in the context of the application domain. Evaluation functions can be quite complex.

In the first instance domain experts often provide the most effective form of evaluation of discovered nuggets. Visualisation tools can be a critical component of this evaluation, providing effective presentations of the results of mining. However, as the nugget sets become large, such manual approaches become less effective.

An approach we have found empirically to be effective in evaluating nuggets is based on building statistical summaries of the entities associated with each nugget. Key variables that play an important role in the business problem at hand are characterised for each nugget and filters are developed to pick out those nuggets with profiles that are out of the ordinary. As the data mining exercise proceeds, the filters are refined and further developed.

The end result is a small (manageable) collection of nuggets that can be further investigated using human resources. The approach focuses attention on segments of the business portfolio where important discoveries may be made.

4 Hot Spots for Insurance Premium Setting

We now illustrate the approach with actual real world data in the context of two case studies. In this section we describe a case study involving insurance risk analysis. This is a joint project with NRMA Insurance Limited. The following section describes public fraud detection in the context of Medicare, a joint project with the Australian Health Insurance Commission.

A major task faced by any insurer is to ensure profitability. To oversimplify, the total sum of premiums charged for insurance must be sufficient to cover all claims made against the policies. However, the premiums must also be competitive against other insurers. The actuarial task is usually performed manually with the support of a variety of statistical tools resulting in only a small number of customer attributes being considered and very broad generalisations being made. With more powerful computing resources available many insurers are now looking to perform more detailed and complex analyses to assist them in developing and annually (or even more frequently) refining their premium setting formulae. An approach we have taken with this problem is to identify, describe, and explore customer groups that have significant impact on the insurance portfolio—using the hot spots methodology for insurance risk analysis.

The original data for our task was collected for business purposes other than for data mining. Consequently, considerable effort was required to transform the data for data mining—a common observation of data miners (Williams and Huang 1996). After preprocessing the dataset the three step hot spot methodology was used: clustering; rule induction; nugget evaluation.

For illustration and to protect confidentiality we present an example here consisting of a dataset of just some 72,000 records. This dataset was clustered into some 40 clusters, ranging in size from tens of records to thousands of records. Treating each cluster as a class we can build a decision tree to describe the clusters and then prune the tree through rule generation. This leads to some 60 rules (nugget descriptions). A sample rule is given in Figure 1. For each cluster there may be many such rules that together describe the data records that (mostly) belong to that cluster.

An evaluation function was developed to identify those nuggets that exhibit “peculiarities” or collections of customers that are important to the business. This is an ongoing task of refinement and exploration, and we present in Tables 1 and 2 an illustration of this process. For these examples the evaluation function identifies nuggets which exhibit characteristics significantly different from the “normal.”

If NCB < 60 and Age ≤ 24 and Address is Urban and
Vehicle $\in \{$Utility, Station Wagon$\}$
Then Cluster $= 1$

Figure 1: Sample nugget from the motor vehicle insurance dataset.

The first step was to derive for each nugget a collection of important indicators. For our example we use the number and proportion of claims lodged by the group of clients in the associated nugget subset, and the average and total cost of a claim for each nugget subset. This information is presented in Table 1 for some of the nuggets. For the whole dataset there were 3800 claims in all, representing a proportion of some 5%. The overall average claim cost is $3000, with a total of some $12 million. (Again, because of confidentiality these figures are indicative only.)

Nugget	*Claims*	*Total*	*Proportion*	*Average Cost*	*Total Cost*
2	148	1391	**11.90**	**3685**	545,000
3	141	2300	6.51	**3795**	535,000
19	3	25	**13.64**	**4338**	13,015
24	10	123	8.85	**7915**	79,150
34	22	344	6.83	**5293**	116,440
35	64	523	**13.94**	**4386**	280,728
36	3	3	100	**6771**	20,314
40	800	1400	5.9	**3500**	2,800,000
All	3800	72000	5.0	3000	12,000,000

Table 1: Number of claims, number of records, and summary nugget data.

From this “raw” data we define an evaluation function that specifies the conditions under which the nuggets are deemed to be of significant interest for the business problem at hand. The evaluation may, for example, highlight nugget subsets containing a very large number (or proportion) of claims (greater than 10%) is important. Alternatively, any nugget having significantly more than the average for any particular measure may be a candidate for further investigation. Table 2 identifies the nuggets that have higher than average characteristics.

Nugget	*By Claims*	*By Proportion*	*By Average Cost*
2	Y	Y	Y
3	Y	Y	Y
19		Y	Y
24		Y	Y
34	Y	Y	Y
35	Y	Y	Y
36		Y	Y
40	Y		Y

Table 2: Risk Areas by Various Criteria (Y indicates risk area).

By then investigating the associated descriptions of the nuggets, and the associated customers, a better understanding of these key groups can be obtained. This exploration can then be used as a key input to the process of defining the insurance premium setting formulae.

5 Hot Spots for Fraud Detection

Fraud is an area receiving significant attention from data mining. In very large collections of data (such as held by credit card companies, telephone companies, insurance companies, and the tax office), it is often expected that there is a small percentage of customers who are practising fraudulent behaviour. It is not surprising then that data mining has been a player in tackling this problem.

The Health Insurance Commission maintains a large database recording information relating to payments made to doctors and patients from the government's Medicare program. Their database is measured in terms of terabytes, and, if it could be analysed, would paint a very comprehensive picture of the health of Australia.

Like any large and complex payment system, Medicare is open to fraud. The HIC is very active in developing and refining its arsenal for identifying and procedurally eliminating fraud, committed both by doctors and by the public. Data mining is now being used to explore and to bring new tools to the problem of public fraud detection.

Once again we have used the hot spots methodology to identify areas within the very large datasets which may require further investigation. For reasons of confidentiality we will present artificially constructed, but indicative, results.

The dataset for this exercise consisted of some 40,000 Medicare numbers (a small subset of the many millions of Medicare numbers available), with over 30 raw attributes (e.g., age, sex, etc.) and some 20 derived attributes (e.g., number of times a patient visited a doctor over a year, number of different doctors visited, etc.). A variety of clusters have been developed for the exploratory analysis, but we will present results from just two: a clustering consisting of just 10 clusters and a clustering consisting of 100 clusters. For the 10 cluster, Table 3 lists the prototypes for each of the clusters over 5 of the variables, and the size of each cluster.

Cluster:	0	1	2	3	4	5	6	7	8	9
Age	30	30	65	40	30	45	50	10	50	55
Sex	F	F	F	F	M	F	F	M	F	F
Services	15	24	34	33	28	21	15	12	32	47
Benefits	430	841	1288	2233	1390	743	463	360	1125	1912
Weeks	2	4	2	2	1	1	2	2	10	7
Size	9000	150	3000	1000	80	5500	9000	7000	2000	800

Table 3: Prototypes for the 10 clusters.

Rules were generated for each of the clusters using C5.0 (the successor to C4.5). For the "10 cluster", some 280 nuggets (rules) were generated and for the "100 cluster" over 1000 nuggets were generated. Figure 2 gives an indicative example.

If	Age is between 18 and 25 and Weeks $\geq$ 10 and AnkSpc is either X or Y
Then	Cluster = 4

Figure 2: Sample nugget from the Medicare dataset.

Once the collection of nuggets gets to these sizes it is not longer feasible to manually scan through them looking for those that are interesting. For the 100 cluster, for example, we can generate a listing of over 1000 lines recording for each nugget the characteristics of that nugget. This might include the average number of claims made on those Medicare numbers in that cluster, the average size of those claims, and so on. Such a listing covers many pages and hides much interesting information. A hot spots evaluation function is used to mine this information to find those nuggets that, for example, have significantly higher rates of claim than the overall average. The evaluation function is tuned iteratively as we (both the data miners and the domain experts) gain more insights and understanding.

The approach has already identified interesting areas in the data that have required further detailed investigation by the Health Insurance Commission.

6 Conclusions

Data Mining has proved to be a useful tool in exploring and discovering interesting and useful knowledge in very large datasets. However, as the datasets get larger the tools being used generally produce significantly more complex models. In this paper we have presented a methodology that we have found useful in deploying data mining in real world problems. The hot spots methodology introduces the idea of mining the knowledge mine. The results of data mining are themselves being "mined" to find those interesting discoveries that are of particular importance to the business problem at hand.

The methodology we have presented employs clustering to provide a first cut segmentation of the data. Decision tree induction and then rule set pruning, using C4.5 and C5.0, deliver symbolic rule sets. Each rule can then be regarded as defining a segment of the dataset. The records from the original dataset associated with each rule are then analysed to find those nuggets that correspond to areas of the data that are significant to the domain problem.

Two case studies have been presented where the approach was used. For NRMA Insurance Limited, key areas of a dataset were those that had significantly greater than average impact on the motor vehicle insurance portfolio. For the Health Insurance Commission, the key areas of the dataset were those that had odd patterns of Medicare claim lodgment or claim amounts. For both case studies these segments of the data were identified for further human investigation, leading to potential refinements to their business.

The Evaluation function is of key importance to the hot spots methodology. Empirically we have found a simple filter approach to be effective so far, but more effort is needed to further develop our skills in defining evaluation functions. Future work will focus on using machine learning approaches, once again, but on the data associated with the nuggets, and incorporating domain knowledge, to derive evaluation formulae from the data.

Acknowledgements

The authors wish to acknowledge that this work was carried out within the Cooperative Research Centre for Advanced Computational Systems established under the Australian Government's Cooperative Research Centres Program.

References

Fayyad, U. M., Piatetsky-Shapiro, G. and Smyth, P.: 1996, From data mining to knowledge discovery: An overview, *in* U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy (eds), *Advances in Knowledge Discovery and Data Mining*, AAAI Press / The MIT Press, pp. 1–34.

Huang, Z.: 1997, Clustering large data sets with mixed numeric and categorical values, *in* H.-J. Lu, H. Liu and H. Motoda (eds), *Knowledge discovery and data mining: techniques and applications*, World Scientific.

Mallows, C. and Pregibon, D.: 1996, The analysis of call-detail data, *The Sydney International Statistical Congress*.

Quinlan, J. R.: 1993, *C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, CA, 1993.*, Morgan Kaufmann, San Mateo, CA.

Viveros, M. S., Nearhos, J. P. and Rothman, M. J.: 1996, Applying data mining techniques to a health insurance information system, *Proceedings of the 22nd VLDB Conference*, Mumbai (Bombay), India, pp. 286–293.

Williams, G. J. and Huang, Z.: 1996, A case study in knowledge acquisition for insurance risk assessment using a kdd methodology, *in* P. Compton, R. Mizoguchi, H. Motoda and T. Menzies (eds), *Pacific Knowledge Acquisition Workshop*, pp. 117–129.