Evolutionary Hot Spots Data Mining

An Architecture for Exploring for Interesting Discoveries*

Graham J Williams
CRC for Advanced Computational Systems,
CSIRO Mathematical and Information Sciences,
GPO Box 664, Canberra, ACT 2601, Australia,
Graham.Williams@cmis.csiro.au,
http://www.cmis.csiro.au/Graham.Williams

May 20, 1999

Abstract

Data Mining delivers novel and useful knowledge from very large collections of data. The task is often characterised as identifying key areas within a very large dataset which have some importance or are otherwise interesting to the data owners. We call this hot spots data mining. Data mining projects usually begin with ill-defined goals expressed vaguely in terms of making interesting discoveries. The actual goals are refined and clarified as the process proceeds. Data mining is an exploratory process where the goals may change and such changes may impact the data space being explored. In this paper we introduce an approach to data mining where the development of the goal itself is part of the problem solving process. We propose an evolutionary approach to hot spots data mining where both the measure of interestingness and the descriptions of groups in the data are evolved under the influence of a user guiding the system towards significant discoveries.

*Presented at the Pacific Asia Conference on Knowledge Discovery and Data Mining (PAKDD99) held in Beijing, China 26–28 April 1999, and published as *Methodologies for Knowledge Discovery and Data Mining*, Lecture Notes in Artificial Intelligence, Volume 1574, Springer-Verlag, 1999.

1 Introduction

Data mining is an inherently iterative and interactive process. A fundamental concern is the discovery of useful and actionable knowledge that might be contained in the vast collections of data that most organisations today collect but usually can not effectively analyse. In applying data mining techniques in a number of case studies with industrial collaborators (in health care, taxation, and insurance) we have developed the hot spots methodology for assisting in the task of identifying interesting discoveries (Williams and Huang 1997).

The hot spots methodology originally entailed the use of clustering and rule induction techniques to identify candidate groups of interesting entities in very large datasets. These groups were evaluated to assess their interestingness (i.e., whether they represented useful and actionable discoveries for the particular domain of application). In dealing with very large datasets the number of identified candidate groups becomes very large and is no longer amenable to simple nor manual evaluation. The groups, described by symbolic rules, serve as a reasonable starting point for the discovery of useful knowledge. However, it has been found empirically that an exploration of other but related areas of the data (e.g., nearby regions within the dataset), in concert with the domain user, leads to further interesting (and sometimes much more interesting) discoveries.

The focus of our data mining work is to *support* the domain user (fraud investigators, auditors, market analysts) in the task of focusing on *interesting* groups of a very large collection of data (many millions of records). The emphasis on support is important for data mining as our experience suggests that domain expertise will remain crucial for successful data mining. While the hot spots methodology established a useful starting point, it provided only little support to proceed further in identifying the most interesting discoveries from amongst the many thousands that were being made and others that lay nearby in the vast search space.

In this paper we develop an architecture for an evolutionary hot spots data mining system that is under development. The starting point is the many discoveries that the current hot spots methodology identifies. An evolutionary approach is employed to evolve nuggets using a fitness measure that captures aspects of interestingness. Since “interestingness” is itself hard to capture we closely couple with the nugget evolution a measure of interestingness that is itself evolved in concert with the domain user. We describe an architecture where a number of measures of interestingness compete to evolve alternative nugget sets from which the best nuggets are presented to the domain user for their ranking. This ranking is fed back into the system for further evolution of both the measure of interestingness and of the nuggets.

2 The Search for Interesting Nuggets

Padmanabhan and Tuzhilin (1998) demonstrate the need for a better grasp on the concept of interestingness for data mining with an example from marketing. Applying a traditional apriori association algorithm to the analysis of 87,437 records of consumer purchase data, over 40,000 association rules were generated, "many of which were irrelevant or obvious." Identifying the important and actionable discoveries from amongst these 40,000 "nuggets" is itself a key task for data mining.

The concept of *interestingness* is difficult to formalise and varies considerably across different domains. A growing literature in data mining is beginning to address the question. Early work attempted to identify objective measures of interestingness, and the confidence and support measures used in association algorithms are examples of objective measures. One of the earliest efforts to address the explosion of discoveries by identifying interestingness was through the use of rule templates with attribute hierarchies and visualisation (Klemettinen, Mannila, Ronkainen, Toivonen and Verkamo 1994). Silbershatz and Tuzhilin (1996) partition interestingness measures into objective and subjective measures, and further partition subjective measures into those that capture unexpectedness and those that capture actionability.

Many authors have focussed on capturing unexpectedness as a useful measure, particularly in the context of discovering associations (Silbershatz and Tuzhilin 1995) and classifications (Liu, Hsu and Chen 1997). Most recently Padmanabhan and Tuzhilin (1998) develop an unexpectedness algorithm based on logical contradiction in the context of expectations or beliefs formally expressed for an application domain.

Capturing actionability is a difficult and less studied proposition. Matheus, Piatetsky-Shapiro and McNeill (1996) discuss the concept of payoff as a measure of interestingness, where they attempt to capture the expected payoff from the actions that follow from their discoveries (deviations). There is little other work specifically addressing actionability.

3 The Hot Spots Methodology

We characterise our concept of interestingness in terms of attempting to identify areas within very large, multi-dimensional, datasets which exhibit surprising (and perhaps unexpected) characteristics that may lead to some actions being taken to modify the business processes of the data owners. At this stage, rather than specifically formalising how interestingness can be expressed we are exploring how we can facilitate the domain user in their search for interesting discoveries in their data using the hot spots methodology.

We now introduce some terminology to describe the hot spots methodology, following Williams and Huang (1997). A *dataset* D consists of a set of real world *entities* (such as a set of policy holders in an insurance company or a set of Medicare patients). Generally 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. The dataset consists of a set of entities: $D = \{e_1, e_2, \ldots, e_n\}$, where each entity is a tuple $\langle v_1, v_2, \ldots, v_m\rangle$ of values, one value for each attribute. For real world problems the number of attributes m and the number of tuples n are typically "large" (m may be anywhere from 20 to 1000 and n typically greater than 1,000,000).

The hot spots methodology uses a data-driven approach to generate a set of rules $R = \{r_1, r_2, \ldots, r_p\}$, where each rule describes a *group* or set of entities $g_i = \{e_j | r_i(e_j)\}$, $g_i \subset D$. (The Boolean function $r_i(e_j)$ is true when entity e_j is described by rule r_j.) We will find it convenient to refer to the set of groups described by R as $G = \{g_1, g_2, \ldots, g_p\}$ but regard R and G to be essentially synonymous and call each element of R (or as the purpose suits, each element of G) a *nugget*. The set of nuggets is synonymously $N = \{r_1, r_2, \ldots, r_p\}$ or $N = \{g_1, g_2, \ldots, g_p\}$. We note that p is generally much smaller than n but can still be substantial (perhaps one or two thousand for n in the millions). A rule consists of a conjunction of conditions, each condition being either: $A_i \in [v_1, v_2]$ for numeric attributes or $A_i \in \{v_1, v_2, \ldots, v_q\}$ for categorical attributes. While we have reduced the dimensionality of the problem (from n down to p) for real world applications p generally remains too large for manual consideration.

We identify a *hot spot* as a set of entities which are of some particular interest to the domain user (e.g., loyal customer groups or regular high insurance claimers). Simple techniques such as clustering or segmentation can help with the task of identifying nuggets that are candidate hot spots, but are often computationally expensive and/or build groups that are not well described. A heuristic approach to this segmentation task that we have empirically found to be effective in many real world problems involves the combination of clustering and rule induction, followed by an exploration of the discovered groups (Williams and Huang 1997), which we call the hot spots methodology.

The hot spots methodology is a three step process:

Step 1: Cluster D into p complete and disjoint clusters $C = \{C_1, C_2, \ldots, C_p\}$ where $D = \bigcup C_i$ and $C_i \cap C_j = \emptyset, i \neq j$. We generally use a mixed data-type k-means based clustering algorithm (Huang 1998).

Step 2: By associating with each record its cluster membership we use rule induction to build discriminatory descriptions of each cluster, leading to the rule set $R = \{r_1, r_2, \ldots, r_q\}$. Usually $q \geq p$ and usually much greater (for each clusters multiple rules may be induced). We will refer to a rule as a description of a nugget (or simply as a nugget). Each nugget describes a subset of the original dataset D

and r_i represents both the nugget description and the nugget subset. Note that $r_i \cap r_j$ is not necessarily empty.

Step 3: The third step is to evaluate each nugget in the nugget set to find those of particular interest. We define the function $Eval(r)$ as a mapping from nuggets to a measure of the interestingness of nugget r. Such a function is 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 actionability, unexpectedness, and validity in the context of the application domain. This is the heart of the problem of interestingness.

An empirically effective approach to evaluating nuggets is based on building statistical summaries of the nugget subsets. 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.

A visualisation of the summaries provides further and rapid insights to aid the identification of hot spots using a simple yet effective matrix-based graphic display of the data. This facilitates the task of working towards a small (manageable) collection of nuggets towards which further resources can be devoted.

Domain users provide the most effective form of evaluation of discovered nuggets. Visualisation tools are also effective. However, as the nugget sets become large, such manual approaches become less effective.

4 Hot Spots Applications

We illustrate the hot spots methodology in the context of two case studies involving data from commercial collaborators. These relate to actual data mining exercises carried out on very large collections of data. While the actual results and data remain confidential we present indicative results in the following sections.

4.1 Hot Spots for Insurance Premium Setting

NRMA Insurance Limited is one of Australia's largest general insurers. A major task faced by any insurer is to ensure profitability, which, to oversimplify, requires that the total sum of premiums charged for insurance must be sufficient to cover all claims made against the policies, while keeping the premiums competitive. Our approach has been to identify, describe, and explore customer groups that have significant impact on the insurance portfolio—using the hot spots methodology for risk assessment by better understanding and characterising customers. After preprocessing the dataset the three step hot spot methodology was used: clustering; rule induction; nugget evaluation.

We present an example here consisting of a dataset of just some 72,000 records with 20 attributes, 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 prune the tree through rule generation. This leads to some 60 nuggets. An example is:

$$\begin{array}{l}\text{No Claim Bonus} < 60 \;\text{ and }\; \text{Address is Urban and} \\ \text{Age} \leq 24 \;\text{ and }\; \text{Vehicle} \in \{\text{Utility, Station Wagon}\}\end{array}$$

An evaluation function was developed to identify interesting nuggets (groups of customers that exhibited some important characteristics in the context of the business problem). This began by deriving for each nugget a collection of indicators, such as the number and proportion of claims lodged by clients and the average and total cost of a claim for each nugget subset.

This summary information is presented in Table 1 for some nuggets. Over the whole dataset (the final row of the table) there were 3800 claims, representing a proportion of some 5% of all clients (a typical figure). The overall average claim cost is $3000 with a total of some $12 million of claims. Particular values that are out of the ordinary in the context of the whole dataset are italicised and nuggets that are above a certain threshold for interestingness based on these values are highlighted.

Table 1: Summary motor vehicle insurance nugget data.

Nugget	*Size*	*Claims*	*Proportion*	*Average Cost*	*Total Cost*
1	1400	150	*11*	3700	545,000
2	2300	140	6	3800	535,000
3	25	5	*20*	4400	13,000
4	120	10	8	*7900*	79,100
5	340	20	6	*5300*	116,000
6	520	65	*13*	4400	280,700
7	5	5	100	*6800*	20,300
...	...	...	...	...	...
60	800	1400	5.9	*3500*	2,800,000
All	3800	72000	5.0	3000	12,000,000

Our evaluation function identifies nuggets of reasonable size containing a high proportion of claims (greater than 10%) and having large average costs. This exploration and the refinement of the measure of interestingness is performed by the domain user.

4.2 Hot Spots for Fraud Detection in Health

The Australian Government's public health care system, Medicare, is managed by the Health Insurance Commission (HIC) who maintain one of the largest data holdings world wide recording information relating to all payments to doctors and patients made by Medicare since 1975. Like any large and complex payment system, Medicare is open to fraud. The hot spots methodology has been used to identify areas which may require investigation.

A subset of 40,000 of the many millions of patients is used for illustration here. The data consists of 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.).

Nuggets were generated from clusters leading to over 280 nuggets. An example is:

$$\begin{array}{l} \text{Age} \in [18, 25] \text{ and Weeks Claimed} \geq 10 \text{ and} \\ \text{Hoarding Index} \geq 15 \text{ and Benefit} > \$1000 \end{array}$$

Table 2 lists some nuggets, with cells of particular interest italicised and rows above a threshold for interestingness highlighted.

Table 2: Summary Medicare nugget data.

Nugget	*Size*	*Age*	*Gender*	*Services*	*Benefits*	*Weeks*	*Hoard*	*Regular*
1	9000	30	F	10	30	2	1	1
2	150	30	F	*24*	*841*	4	2	4
3	1200	65	M	7	220	20	1	1
4	80	45	F	*30*	*750*	10	1	1
5	90	10	M	12	*1125*	10	*5*	2
6	800	55	M	8	550	7	1	*9*
...								
280	30	25	F	15	450	15	2	6
All	40,000	45		8	30	3	1	1

With 280 nuggets it becomes difficult to manually scan for those that are interesting. For larger Medicare datasets several thousand nuggets are identified. The evaluation function takes account of the average number of services, the average total benefit paid to patients, etc.

The approach has successfully identified interesting groups in the data that were investigated and found to be fraudulent. A pattern of behaviour identified by this process was claim hoarding where patients tended to collect together many services and lodge a single claim. While not itself indicative of fraud a particularly regular subgroup was found to be fraudulent.

5 Evolving Interesting Groups

5.1 Background

The hot spots methodology was found to be a useful starting point for an exploration for interesting areas within very large datasets. It provides some summary information and visualisations of that information. However, as the datasets become larger (typically we deal with many millions of entities) the number of nuggets becomes too large to manually explore in this way. The simple expression for interestingness based on comparisons of summary data to dataset averages is of limited use.

To address this we propose an evolutionary approach that builds on the framework provided by the hot spots methodology. The aim is to allow domain users to explore nuggets and to allow the nuggets themselves to evolve according to some measure of interestingness. At a high level the two significant problems to be addressed are:

1. How to *construct* nuggets?

2. How to define the *interestingness* of a nugget?

In a perfect world where we could define interestingness precisely there would be no problem. The definition would be used directly to identify relevant nuggets. The nature of the domains and problems we consider in data mining though is such that both the data and the goals are very dynamic and usually ill-defined. It is very much an exploratory process requiring the sophisticated interaction of domain users with the data to refine the goals. Tools which can better facilitate this sophisticated interaction are needed.

We employ evolutionary ideas to evolve nuggets (described using rules). Previous work on *classifier systems* in Evolutionary Computation has also considered the evolution of rules, and there is a limited literature on using evolutionary ideas in data mining (Freitas 1997; Radcliffe and Surry 1994; Teller and Velosa 1995; Turney 1995; Venturini, Slimane, Morin and de Beauville 1997).

The hot spots methodology begins to address both of the high level problems identified above: constructing nuggets and measuring interestingness. For constructing nuggets a data driven approach is used: employing clustering and then rule induction. For measuring interestingness we define a measure based initially on the simple statistics of a group (e.g., a group is interesting if the average value of attribute A in the group is greater than 2 standard deviations from the average value of the attribute over the whole dataset). This may be augmented with tests on the size of the group (we generally don't want too large groups as they tend to exhibit "expected" behaviour) and meta conditions that limit the number of hot spots to 10 or fewer.

Assessing true interestingness, relying on human resources to carry out actual investigations, can be a time consuming task. As we proceed, working closely with the domain user, our ideas of what is interesting amongst the nuggets being discovered becomes refined and often more complex. This requires constant refinement of the measure and further loops through the nugget construction process to further explore the search space. An evolutionary approach which attempts to tackle both the construction of nuggets and the measure of interestingness is developed.

5.2 Proposed Architecture

We describe an evolutionary architecture to refine the set of nuggets N derived through either a hot spots analysis, random generation, or in some other manner. A small subset of N is presented to the domain user who provides insights into their interestingness.

A *measure of interestingness* of a nugget is to be determined. The aim is to develop an explicit function $\mathbf{I}(g)$ (or $Eval(r)$) that captures this. An initial random measure of interestingness will set the process going, or else we can employ the simple measures used in the current hot spots methodology.

The measure of interestingness can then be used as a fitness measure in an evolutionary process to construct a collection of nuggets. By using an evolutionary process we can explore more of the search space. Having evolved a fit population of nuggets we present some small subset of these to the domain user for their evaluation (to express whether they believe these to be interesting nuggets). We are working towards capturing the user's view on the interestingness of the nuggets and to feed this directly into the data mining process. An interesting twist is that the population of measures of interestingness is also evolved, based on the user feedback. The top level architecture of the evolutionary hot spots data mining system is presented in Fig. 1.

At any time in the process we will have some number, q, of measures of interestingness: $I = \{\mathbf{I}_1, \mathbf{I}_2, \ldots, \mathbf{I}_q\}$. The rules in the ruleset R will be evolved independently in each cycle through the process, leading to q independent rule sets, each being fit as measured by one of the measures of interestingness $\mathbf{I}_k$. Typically, a population of rules will consist of some thousand rules and it is not practical to present all rules in a "fit" population to the user. Instead a small subset of s rules is chosen from each population. Thus, for each cycle through the process, q (typ-

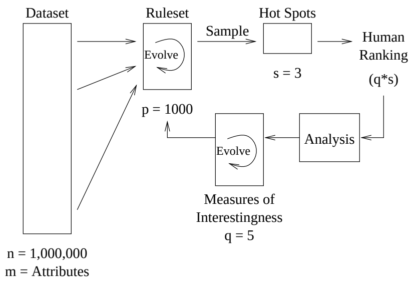

Figure 1: The basic model for an evolutionary hot spots data miner with indicative sizes.

ically 5) independent sets of s (typically 3) rules will be presented to the user for ranking. The user ranks all of these $q \times s$ rules in terms of their assessment of the interestingness of the rules and the entities associated with the rules. This ranking is then used to evolve the interestingness measures through another generation to obtain a new population of q measures.

The analysis of the ranking might, for example, find in the top rules some commonality in conditions. These could be identified as atoms to genetically engineer the next generation of the rule set $\mathcal{R}$ (retaining these atoms in any rules that are modified). Alternatively we increase the fitness of other rules that contain the commonality. Indeed, under this scheme we can identify conditions that either increase or decrease the fitness of other rules in $\mathcal{R}$. A tuning parameter is used to indicate the degree to which fitness can be modified. The evolutionary process then takes over to generate the next population $\mathcal{R}'$ using these new fitness measures.

At each stage the domain users are involved in the process, and as interesting discoveries are brought to their attention, they assess whether further investigation is required.

6 Summary

We present an architecture that is being implemented and tested in actual data mining exercises. At this stage we can not make any claims about its usefulness although early feedback indicates that it does provide a useful expansion of the search space under the control of a domain user, allowing a focus on relevant discoveries to be established. The expression of the formulae capturing interestingness still requires considerable research. We expect to adopt various approaches to the actual expression of interestingness as developed by our experience and that of others. With a formal and well developed language the evolution of the measures of interestingness will be better able to capture the user's insights, as well as providing some direction setting suggestions for exploration.

Acknowledgements

The author acknowledges the support provided by the Cooperative Research Centre for Advanced Computational Systems (ACSys) established under the Australian Government's Cooperative Research Centres Program. Dr. Xin Yao (Australian Defence Forces Academy) has contributed tremendously in discussion of the evolutionary aspects and Reehaz Soobhany (Australian National University) has implemented much of the rule evolution code.

References

Freitas, A. A.: 1997, A genetic programming framework for two data mining tasks: classification and generalized rule induction, *in* J. R. Koza, K. Deb, M. Dorigo, D. B. Fogel, M. Garzon, H. Iba and R. L. Riolo (eds), *Proceedings of the Second Annual Conference on Genetic Programming*, Morgan Kaufmann, San Francisco, CA, pp. 96–101.

Huang, Z.: 1998, Extensions to the k-means algorithm for clustering large data sets with categorical values, *Data Mining and Knowledge Discovery* **2**(3), 283–304.

Klemettinen, M., Mannila, H., Ronkainen, P., Toivonen, H. and Verkamo, A. I.: 1994, Finding interesting rules from large sets of discovered association rules, *in* N. R. Adam, B. K. Bhargava and Y. Yesha (eds), *Proceedings of the Third International Conference on Information and Knowledge Management*, ACM Press, pp. 401–407.

Liu, B., Hsu, W. and Chen, S.: 1997, Using general impressions to analyse discovered classification rules, *KDD97: Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*, AAAI Press, pp. 31–36.

Matheus, C. J., Piatetsky-Shapiro, G. and McNeill, D.: 1996, Selecting and reporting what is interesting, *in* U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy (eds), *Advances in Knowledge Discovery and Data Mining*, AAAI Press, pp. 465–515.

Padmanabhan, B. and Tuzhilin, A.: 1998, A belief-driven method for discovering unexpected patterns, *KDD98: Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, AAAI Press.

Radcliffe, N. J. and Surry, P. D.: 1994, Cooperation through hierarchical competition in genetic data mining, *Technical Report EPCC-TR94-09*, University of Edinburgh, Edinburgh Parallel Computing Centre, King's Buildings, University of Edinburgh, Scotland, EH9 3JZ.

Silbershatz, A. and Tuzhilin, A.: 1995, On subjective measures of interestingness in knowledge discovery, *in* U. M. Fayyad and R. Uthurusamy (eds), *KDD95: Proceedings of the First International Conference on Knowledge Discovery and Data Mining*, AAAI Press, pp. 275–281.

Silbershatz, A. and Tuzhilin, A.: 1996, What makes patterns interesting in knowledge discovery systems, *IEEE Transactions on Knowledge and Data Engineering* **8**(6), 970–974.

Teller, A. and Velosa, M.: 1995, Program evolution for data mining, *The International Journal of Expert Systems* **8**(3), 216–236.

Turney, P. D.: 1995, Cost-sensitive classification: Empirical evaluation of a hybrid genetic decision tree induction algorithm, *Journal of Artificial Intelligence Research* **2**, 369–409.

Venturini, G., Slimane, M., Morin, F. and de Beauville, J.-P. A.: 1997, On using interactive genetic algorithms for knowledge discovery in databases, *in* T. Bäck (ed.), *Proceedings of the Seventh International Conference on Genetic Algorithms*, Morgan Kaufmann, pp. 696–703.

Williams, G. J. and Huang, Z.: 1997, Mining the knowledge mine: The Hot Spots methodology for mining large, real world databases, *in* A. Sattar (ed.), *Advanced Topics in Artificial Intelligence*, Vol. 1342 of *Lecture Notes in Computer Science*, Springer-Verlag, pp. 340–348.