Exploratory Hot Spot Profile Analysis Using Interactive Visual Drill-Down Self-Organizing Maps

Denny[1,2], Graham J. Williams[3,1], and Peter Christen[1]

[1] Department of Computer Science, The Australian National University, Australia
denny@cs.anu.edu.au, peter.christen@anu.edu.au
[2] Faculty of Computer Science, University of Indonesia, Indonesia
[3] The Australian Taxation Office
graham.williams@ato.gov.au

Abstract. Real-life datasets often contain small clusters of unusual sub-populations. These clusters, or 'hot spots', are usually sparse and of special interest to an analyst. We present a methodology for identifying hot spots and ranking attributes that distinguish them interactively, using visual drill-down Self-Organizing Maps. The methodology is particularly useful for understanding hot spots in high dimensional datasets. Our approach is demonstrated using a large real life taxation dataset.

Keywords: self-organizing maps, hot spot analysis, attribute ranking, imbalanced data, interactive drill-down visualization.

1 Introduction

The complexity of knowledge contained in large datasets is often easier to explore by grouping similar entities together, which is known as cluster analysis. For example, clustering of customers sharing similar characteristics generally makes it easier to devise marketing strategies. Self-Organizing Maps (SOMs) [1] are popularly used in cluster analysis for several reasons. First, SOMs topologically map high-dimensional data into a two-dimensional map with similar entities being placed close to each other. Second, SOMs produce a smaller but representative dataset that exhibits the distribution of the original dataset. Third, SOMs offer various map visualizations that allow non-technical users to explore a dataset.

In real datasets cluster sizes are normally not equal and clusters do not have the same level of interest for a user. The cluster distribution is often very skewed with interesting clusters being a small fraction of the full dataset. Also, variance of the items at the tail/margin of the normal distribution of a population is also larger compared to the center of the distribution. Thus it is common to find large dense clusters for common sub-populations, and small sparse clusters that might be of interest. In a taxation context, for example, this could be a small

T. Washio et al. (Eds.): PAKDD 2008, LNAI 5012, pp. 536–543, 2008.

group of tax entities who have unusual tax debts, while in an insurance context this may be a small group of high claiming clients.

Hot Spots aims to identify important or interesting groups in very large datasets [2] using a combination of clustering and rule induction. By understanding attributes that distinguish these small and interesting clusters (hot spots), businesses can improve their processes, such as the choice of treatment strategies for ensuring tax compliance. We advance the hot spots methodology using attribute selection measurement and visualization. With our methodology, analysts can identify and understand distinguishing characteristics of hot spots through interactive visualizations and by performing drill-down exploration.

2 Hot Spots Analysis

Hot Spots data mining identifies key areas in very large datasets that are interesting to an analyst [2]. A dataset is clustered to identify between 10 and 1,000 clusters. Each entity is then labelled with the cluster it is assigned to. Supervised learning (e.g., tree induction) is used to generate distinguishing descriptions for each cluster. The resulting tree is pruned and transformed into a rule set. Finally, the interestingness of the clusters are evaluated. As it is difficult to formalize interestingness, this is domain dependent and therefore, such an analysis is often exploratory and evolutionary [3].

There are several drawbacks with the Hot Spots methodology. When correlated attributes exist in a dataset only one of them will be used in the rule set to describe a cluster, reducing the description of the clusters. Also, the supervised learning step is highly dependent on the results of the previous clustering step, and also on the clustering technique employed (usually k-means). When a large number of clusters is chosen some clusters might have quite similar characteristics, yet a small number of clusters would reduce the required detail extracted from the dataset. Exploring for the right number is difficult.

3 Self-Organizing Maps

A SOM is an artificial neural network that performs unsupervised competitive learning [1]. Importantly, SOMs can be visualized and be used to explore high-dimensional data spaces through a non-linear projection onto a lower-dimensional manifold, most commonly a 2-D plane [4]. Artificial neurons are arranged on a low-dimensional grid, with each neuron represented by an n-dimensional prototype vector (with n the dimension of the input data) and connected to its neighbouring neurons.

Exploring for Hot Spots we find that interesting clusters are usually located at the border of the map because of the topological ordering property. However, SOMs have a *border effect* problem [4] where the neighbourhood definition is not symmetric at the borders of the map—the number of neighbours per unit on the borders and corners of the map is not equal to the number of neighbours in the

middle of the map. As the density estimation for the border units is different to the units in the middle of the map, the tails of the marginal distributions of variables (normally located at border units) are less well represented than their centers [4]. A visual drill-down approach using a SOM can alleviate this [5]. Here, several nodes of a region can be selected by an analyst for interactive drill down to target regions of interest.

Furthermore, SOMs tend to merge small sparse clusters. This further reduces the detail in the analysis. Increasing the map size of a SOM gives a better resolution map but with significant additional computational cost.

4 SOM Hot Spot Profile Analysis Methodology

The contribution of this paper is the development of a methodology to perform profile analysis of hot spots. We present this as data pre-processing, map training, hot spots identification, profile analysis, drill-down, and sub-map analysis.

4.1 Data Pre-processing and Map Training

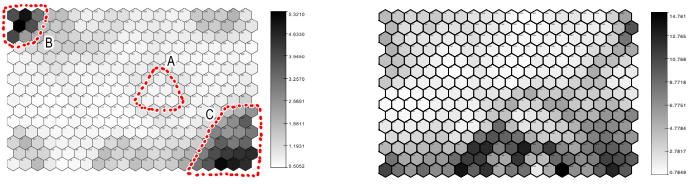
Data pre-processing is important prior to training any maps [5]. SOMs only handle numeric attributes—each non-numeric (categorical) attribute is transformed into a set of numeric attributes, encoding each categorical value into a binary indicator (1 or 0). Normalization of the numeric attributes ensures that attributes with larger ranges won't have an unduly larger influence on the distance calculations [6].

Linear initialization is recommended for initialising a SOM, resulting in an order of magnitude improvement in time taken for learning compared to random initialization [4]. Also, we train a SOM in two phases using batch training [4]. This combined linear initialization and batch training produces the same map each time the learning process is repeated (random initialization might produce different orientations of the map). Batch training can also utilize multi-processor environments to speed up the training process. The map size, training length, initial and final radius are chosen by considering a best practice approach [7].

4.2 Identifying Hot Spots in Self-Organizing Maps

Hot spots in SOMs can be identified by two approaches: first by using the distance matrix visualizations and second by analysts' feedback based on component plane visualizations. Noting that entities in hot spots are usually less homogeneous because they are often located at the tail of distributions, these regions can be identified using the distance matrix. Distance-matrix based visualizations, such as u-matrix visualization [8], show distances between neighbouring nodes using a colour scale representation on a map grid. As shown in Fig. 1, white indicates a small distance between a node and its neighbouring nodes while black indicates a large distance between a node and its neighbours[1].

[1] SOM graphics are best in colour but printing requirements necessitated gray scale.



(a) Whole dataset (b) The sub-map of region 'C'

Fig. 1. Distance matrix (median of a node to its neighbours [5]) visualization

The distance matrix visualization can be used to identify borders between clusters. Large distances that show highly dissimilar features between neighbouring nodes divide clusters, i.e. the dense parts of the map with similar features (white regions) [8].

Distance-matrix visualizations can be used to acquire the initial cluster structure of the dataset. By using this visualization, an analyst can see the cluster structure of the dense part of a map. An example is the cluster in the center of the map (marked 'A') in Fig. 1(a). However, it is difficult to see the cluster structure of the sparse regions of the lower-right and the upper-left corners of the map (marked 'B' and 'C').

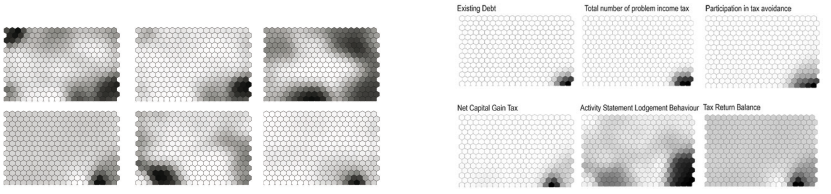
The distance matrix visualizations in Fig. 1 show homogeneous (low variation) groups with smaller neighbour distances (white regions) and high variation groups (dark regions). Regions with larger neighbour distances can be further investigated through component plane visualizations. In Fig. 1(a) two hot spots are identified according to the above criteria (the regions marked 'B' and 'C').

4.3 Profile Analysis of Hot Spot

Descriptive statistics (e.g., average values) of entities mapped to a hot spot provides a simple characterization. However, this approach does not provide an analyst with insight, as it is difficult to find the average value with respect to the spread of the values of the whole dataset.

Component plane visualizations can be used to show the spread of values of a certain component of all prototype vectors in a SOM [9]. The value of a component in a node is the 'average' value of entities in the node and its neighbours according to the neighbourhood function. The colour coding of the map is created based on the minimum (white) and the maximum values (black) of the component of the map. When analyzing the characteristics of hot spots in high dimensional datasets, it is difficult to identify components which distinguish hot spots from the remaining population by visualizing all component planes, except by ranking their importance, as shown in Fig. 2(a).

An analyst is supported in our methodology by sorting the component planes by the importance of the attributes that distinguish a hot spot from the rest of the population. This ranking can be done using an attribute selection measure [6], such as information gain or gain ratio. As attributes in a SOM are



(a) Unsorted component planes

(b) Top six component planes sorted based on region 'C'

Fig. 2. Component planes. Six of 90 attributes are shown.

numeric, a supervised discretization measure [6], such as entropy-based discretization, should be applied to the numeric attributes before ranking. To rank attributes by their importance, the nodes of the selected region are labeled as 'hot spot' and the rest as 'non-hot spot'. An analyst can then choose an attribute selection measure for attribute importance based on the prototype vectors. The component planes are then ordered by this rank. Fig. 2(b) shows the sorted component planes of the hot spot of region 'C' in Fig. 1(a) using the Gain Ratio. With this ordering, an analyst is able to identify the attributes that distinguish a hot spot from the rest of the population.

As a SOM produces a smaller but representative dataset, the prototype vectors can be used as an approximation of the whole dataset. Efficient computation allows an analyst to explore the profile of any region of the map interactively.

4.4 Drill Down and Visualizing Hot Spots

The analyst has chosen the region of the top level map of interest, allowing a sub-map to be trained to gain more detail for these sparse regions. In training the sub-map, consistency of interpretation of the visualization of the sub-map needs to be preserved while maintaining the sub-map quality with respect to the sub-population [5].

For consistent interpretation of the visualization of the sub-map, the orientation of the map is preserved and the colour coding is made consistent [5]. A drawback of using linear initialization for the sub-map based on the entities in the sub-map is that its orientation might be different to the orientation of the top level map. For example, entities located at the bottom-right corner of the top level map might be located at the top-left corner as we drill down, particularly when the two largest principal components of the whole population and the sub-population are different.

We propose that the top level map be used as the initial map of the sub-map [5]. The radius of the rough phase training must be wide enough to avoid subregions of the map becoming empty. We find that setting the initial radius of the rough phase to be half of the longest side and the initial radius of the fine tune phase to be a quarter of the longest side works well.

4.5 Visualization and Analysis of the Sub-map

Sub-maps are also visualized using the distance matrix and component plane visualizations introduced above. To display the distribution of values of the sub-map with respect to the whole population, we use the colour map for the whole population to visualize the component planes of the sub-map. In other words, black in the sub-map visualizations is used for the maximum value of the component of the top level map, not necessarily the maximum value of the component of the sub-map. As the sub-map has better quality in terms of quantization error (more homogeneous within a node), the component value in the sub-map might exceed the maximum value of the top-level map. The colour for such values are also black and this needs to be kept in mind in reviewing the visualization.

With sub-regions consisting of considerably fewer data vectors the training of the sub-map is considerably faster. An analyst is thus able to interactively explore hot spots once the top level map has been trained. The sub-map can be further explored using the methods introduced in Sects. 4.2 and 4.3.

5 Results and Discussion

Our new visual SOM drill-down approach has been applied to the task of exploring taxpayer compliance for the Australian Taxation Office (ATO), using a de-identified taxpayer dataset. Here, we provide aggregate indicative results that demonstrate the effectiveness of our methodology, without breaching the confidentiality of the data or the discoveries made.

The analysis is motivated by the need to understand the logic and structures that drive taxpayers' compliance behaviour (behavioural archetypes). The idea is to construct 'psychographic groups' [10] by using data mining. Understanding the difference between low and high risk taxpayers is important.

The dataset consists of 6.5 million entities with 90 attributes that reflect taxpayer behaviour. The attributes can be categorized into: income profile (details of income sources), propensity to lodge correctly and on time (lodgement profile), propensity to pay (debt profile), market segments, demographics, socio-economic indicators for areas (SEIFA) [11], and participation in tax avoidance schemes. These attributes were selected by domain specialists. The dataset was normalized and categorical attributes were transformed into numerical attributes.

A map size of 15x20 units with a hexagonal lattice structure [4] was chosen. The initial radius of the rough phase was 8 and for the fine tuning phase it was 4. The training length for the rough phase was 6 iterations and for the fine tuning phase 10 iterations. The training of the top-level map took about 5 hours under Debian GNU/Linux with two AMD64 dual-core 3GHz processors and 16 GB memory using our Java-based SOM Toolbox.

In interpreting multiple visualizations it must be understood that the visualizations are linked by position or by colour. A visualization of the same map is linked by position so that the position of each entity remains the same in each visualization. Figs. 1(a), 3(a), and 3(b) are linked by position. The visualization

(a) 'Number of debt cases' (b) 'Number of debt cases paid'

Fig. 3. Component plane of the whole population [5]

of the top-level map is linked by colour to the sub-map so that the colours of the top level map are directly used for the sub-maps.

The visualization of the dataset distance matrix can be seen in Fig. 1(a). The 'common' population in real life datasets is usually located in the center of a map. In Fig. 1(a), the entities in the center of the map of the whole population are relatively homogeneous. According to the criteria presented in Sect. 4.2, there are two hot spots, located in the top-left corner ('B') and in the bottom-right corner ('C'). Based on the ranking of the component planes (Sect. 4.3) using gain ratio as the attribute selection measure, hot spot 'C' can be distinguished by the following attributes in decreasing importance: existing debt, total number of problem income tax returns, participation in tax avoidance schemes, net capital gain tax, activity statement lodgement behaviour, and the balance of the tax return (Fig. 2(b)). Hot spot 'B' can be distinguished by the attributes: allowances, dividends, and total income. Based on these rankings, 'C' is more interesting, and further explored.

The entities in 'C' have highly dissimilar characteristics (Fig. 1(a)). However, at this level, it is difficult to differentiate the debt behaviour, as shown in Figs. 3(a) and 3(b). Therefore, to see the debt behaviour in detail, we drill down into the lower-right corner of the top level map (Sect. 4.4).

At this level we can also use the distance matrix visualization (Fig. 1(b)) to highlight the hot spots in this sub-map, which are located along the bottom of the map. It is also interesting to note that the hot spot of the sub-map consists of entities that are involved in tax avoidance activities. Furthermore, this group has characteristics of longer debt age, higher levels of compliance enforcement, and lower percentage of cases paid.

6 Conclusion and Future Work

We have introduced a methodology for understanding characteristics of hot spots in large real world datasets, such as from the taxation domain. Based on our experiments, the methodology is effective for hot spots exploration, offering interactive visualizations that are easy to understand. An analyst is able to identify discriminating characteristics of hot spots. As a SOM produces a considerably smaller-sized set of prototype vectors, it allows an efficient use of attribute selection measurements. In using the methodology introduced here analysts have

the flexibility to explore regions or clusters based on map visualization, and are able to drill-down into sparse regions or clusters. Analysts are now able to select regions or clusters based on their business needs.

This work is part of a larger research project where we are interested in observing the dynamics of hot spots over time, such as to find entities who are moving in or out of hot spots. Such knowledge will be valuable as an analyst can derive strategies to encourage or deter moves in or out of the hot spots (which might be regions of non-compliance or of high compliance). It can also be used to evaluate the effectiveness of such business strategies over time.

Acknowledgement

This research has been supported by the Australian Taxation Office. The authors express their gratitude to Grant Brodie, Georgina Breen, Nicole Wade, and Warwick Graco for providing data and domain expertise.

References

1. Kohonen, T.: Self-organized formation of topologically correct feature maps. Biological Cybernetics 43, 59–69 (1982)
2. Williams, G.J., Huang, Z.: Mining the knowledge mine: The hot spots methodology for mining large real world databases. In: Sattar, A. (ed.) Canadian AI 1997. LNCS, vol. 1342, pp. 340–348. Springer, Heidelberg (1997)
3. Williams, G.J.: Evolutionary hot spots data mining - an architecture for exploring for interesting discoveries. In: Zhong, N., Zhou, L. (eds.) PAKDD 1999. LNCS (LNAI), vol. 1574, pp. 184–193. Springer, Heidelberg (1999)
4. Kohonen, T.: Self-Organizing Maps, 3rd edn. Springer Series in Information Sciences, vol. 30. Springer, Heidelberg (2001)
5. Denny, Williams, G.J., Christen, P.: Exploratory multilevel hot spot analysis: Australian Taxation Office case study. In: AusDM 2007, Gold Coast, Australia, ACS. CRPIT, vol. 70, pp. 73–80 (2007)
6. Han, J., Kamber, M.: Data Mining: Concepts and Techniques, 2nd edn. Morgan Kaufmann, San Francisco (2006)
7. Vesanto, J., Himberg, J., Alhoniemi, E., Parhankangas, J.: SOM toolbox for Matlab 5. Report A57, Helsinki University of Technology, Neural Networks Research Centre, Espoo, Finland (April 2000)
8. Iivarinen, J., Kohonen, T., Kangas, J., Kaski, S.: Visualizing the clusters on the Self-Organizing Map. In: Proceedings of the Conference on AI Research in Finland, vol. 12, pp. 122–126, Helsinki, Finland, Finnish AI Society (1994)
9. Tryba, V., Metzen, S., Goser, K.: Designing basic integrated circuits by Self-Organizing Feature Maps. In: International Workshop on Neural Networks and their Applications, Nanterre, France, ARC, SEE, EC2, November 1989, pp. 225–235 (1989)
10. Wells, W.D.: Psychographics: A critical review. Journal of Marketing Research (JMR) 12(2), 196–213 (1975)
11. Trewin, D.: Socio-economic indexes for areas: Australia 2001. Technical Report 2039, Australian Bureau of Statistics (2003)