Visual Tools for Analysing Evolution, Emergence, and Error in Data Streams

Sol Hart, John Yearwood, Adil M. Bagirov
University of Ballarat
School of Information Technology and Mathematical Sciences
Mt Helen, Victoria, Australia
{shart, j.yearwood, a.bagirov}@ballarat.edu.au

Abstract

The relatively new field of stream mining has necessitated the development of robust drift-aware algorithms that provide accurate, real time, data handling capabilities. Tools are needed to assess and diagnose important trends and investigate drift evolution parameters. In this paper, we present two new and novel visulisation techniques, Pixie and Luna graphs, which incorporate salient group statistics coupled with intuitive visual representations of multidimensional groupings over time. Through the novel representations presented here, spatial interactions between temporal divisions can be diagnosed and overall distribution patterns identified. It provides a means of evaluating in non-constrained capacity, commonly constrained evolutionary problems.

1 Introduction

Analysing trends and developing improved data mining and machine learning techniques in stream environments has never been as crucial as in the current environment. When considering streams there is a marked departure from traditional machine learning and data mining methods. In such a dynamic environment there is a need to reconsider strategies for both the processing as well as the pre-processing techniques. The most prominent of these departures can be identified as 'concept drift', emergence detection and error.

The is a considerable amount of work on determining similarities or deviations in evolving data [6, 5], but few formalising the characteristics of evolution faced in stream environments. Our contribution here is to specifically look at exploratory and inferential tools for diagnosing stream evolution. A key aspect is the harnessing of unsupervised structural approaches that do not require shared points or expert knowledge. Clustering is well suited for studying streams in an unsupervised manner. One characteristic of clusters is their difficultly in determining the best clustering [3]. Clusterings are constructed on the underlying attribute distributions and rarely correlate perfectly with subjective notions of class or membership. We endeavor to demonstrate that unsupervised tracking of stream evolution is highly adept at providing intuitive insights into different clustering models and the trends they portray.

2 Motivations

Dimensionality of most datasets prohibits the direct interpretation of [12] the clusters formed in different time intervals. Not only does dimensionality reduction need to take place but also high fidelity data representations that capture the most salient and interesting aspect of the data quickly and intuitively. This paper puts forward two approaches to visualises such transitions, Pixie and Luna graphs, as a proof of concept. These techniques can be seen as appropriate precursors to improved machine learning methods.

Many stream techniques derived from traditional static mining methods, such as sliding windows and forgetting, together with novel approaches like ensembles, aim to minimise to effect of data evolution on accuracy without endeavouring to understand it. The graphing techniques described here explicitly depict the evolution, error, and emergence present in the datasets which has a number of advantages:

- Aid in the development of advanced mining techniques.
- Capacity to display high-dimensional data in low dimensional form so that inference can be made on important trends within the data.

3 Inherent complexities of changing data

Unless strict constraints are placed on data gathering, or production (an assumption often violated) all robust stream

6th IEEE/ACIS International Conference on Computer and Information Science (ICIS 2007)
0-7695-2841-4/07 $25.00 © 2007 IEEE

mining/learning techniques will have to account for introduced error, concept drift, and emergence.

3.1 Error and emergence

Although these factors constitute non-trivial complexity when trying to determine change in group representations they are easy to describe. Error, is by definition, any data that misrepresents a given class or cluster through human error or otherwise. Error becomes more interesting because in isolation it is quite easy to mediate through thresholding or statistical quality control. This is what we call 'actual' error. Which is different from 'perceived' error that results from change in the underlying distribution where drift occurs.

Emergence, similar to error and drift, initially presents itself as an outlier. An emergent group differentiates itself by exhibiting a unique distribution and drift from neighboring clusters. Through a thresholding measure, emergence can asserts itself as a new group. There are inherent difficulties with implementation when handling these factors. They all present themselves in very similar ways, yet need to be accurately handled in various manners.

3.2 Concept drift

Determining concept drift is of paramount importance in non-finite stream mining. In these environments, 'The underlying processes generating these concepts change over months and years' [8]. To accurately describe or predict an item's class effectively, you need to have a current and appropriate model of the concept at all times. This, as it turns out, is a non-trivial problem.

4 Diagnostic techniques

In a recent paper by C. Aggarwal [2] he describes methods for detecting and visualising changing distributions in streams. The technique employs velocity density estimation to indicate concept drift within the evolving streams. Velocity density is the rate of change of data in a give spatial region. This is then broken down and represented in traditional two and three-dimensional plots with axis of spatial and temporal densities. While this is a novel indicator for representing data evolution it fails to directly represent individual class or cluster distribution change. Traditional graphing techniques are still employed, where novel indicators are developed to fit typical representaions.

Although not incorparating novel visualisations nor aimed at streams, a diagnostic technique, ADCO [4], incorperates many of the same design objectives as the Pixie and Luna graphs. By exploiting a hyper-grid representation of clusterings, a similarity metric is derived that takes into account structural properties of the data, whilst operating without the need for all points to be shared. It utilises a density metric to give a final score that underpins an unsupervised evaluation technique but requires an equal number of clusterings for meaningful analysis. The application of ADCO highlighhts two relevant notions. Firstly, by including structual information in a diagnostic tool you are better able to evaluate changes there-in. Secondly, utilising methods that allow indirect comparisons to be made between clusterings without shared points allows for apraisal of newly incorparated data which is the case with streams. Pixie and Luna graphs have an advantage over ADCO techniques as theye can indicate emergent behaviour. This is demonstrated in the results section.

Data evolution has been widely investigated using supervised techniques [6, 5], however, this paper will be investigating purely unsupervised approaches.

5 Approach

5.1 Tools

For our experiments we make use of the K-means and Hierarchical clustering algorithms developed by the University of Tokyo in their micro-array clustering package Pycluster [7].

5.1.1 Hierarchical clustering algorithm

This algorithm employs an agglomerative styled approach to organise elements into clusters. The algorithm begins with as many clusters as there are elements. Elements are systematically joined via a linkage scheme. Unlike K-Means, if the same linkage scheme is used them all clusters will be the same for the same dataset. Time complexity is a little greater than K-Means $0(n^2 logn)$ and space complexity of $O(nlogn)$.

5.1.2 Global K-means

The Global K-means [1] implements a non-smooth, non-convex, optimised K-means. It derives this by minimising the sum-of-square by iterating toward the global minima. This process continues until a conditional threshold is met. It utilises a stopping function to avoid insignificant subclusters. Global K-means is well suited for high-dimensional datasets due to an objective function that reduces the number of data points needed to achieve a given accuracy. It employs a novel method to avoid local minima with the effectiveness of the algorithm proving stable over disjoint temporal data divisions. Global K-means has a similar time order of complexity to that of K-means of $O(nk)$, with k being the number of centers. It has a linear space complexity of $O(n)$.

5.2 Graph development

The newly designed Pixie and Luna graphs allow for quick inference of key elements of the interaction in an intuitive format.

5.2.1 Pixie graph

The Pixie graph was developed to depict cluster interactions over time intervals. Lines fan out in a many-to-many relationship. Stream intervals are represented on opposing sides of the graph by Bi (B^i), where i is the number of intervals. Similarly, clusters are represented by Cn(C_n), where n is the number of clusters for that interval. A complete line from one interval to the other represents the smallest centroid distance between two clusters in an inverse proportional relationship. These lines are normalised on the leftmost clusters interactions. The width of the line is proportional to the number of elements in overlapping regions(reffered to here as point inclusion). A strong overlap is described as an agreement. Such an agreement can be seen between B^1C_0 and B^2C_0 in figure 4.

5.2.2 Luna graph

The Luna graph has a similar layout as the agreement table with intervals represented on the left and top. The purpose of the graph is to show regions of overlap for agreements. Light grey regions depict no interaction and dark grey indicates overlap. Whereas the Pixie graph dealt with numbers of elements, Luna looks at geometric overlap. Complimenting the visual interpretations is the Hausdoff distance. An interaction of two clusters is represented by two radii and a center distance normalised over the largest value. The normalisation has the effect of reducing the size of the cluster to show distance even when there is no interaction present. The same strong agreement seen previously between B^1C_0 and B^2c_0 can be seen in figure 5.

5.2.3 Agreement tables

We developed agreement tables to facilitate a representation of the data that was as rich and intuitive as possible. Each cluster interaction has nine features attributed to it (As seen in figure 1). Information concerning the individual clusters is kept in the margins. Average cluster radius and cluster volume are displayed here. In the individual interactions we see forward and backward Hausdorff deviations. These deviations are followed by forward and backward point overlap (or inclusion) and under the dividing line is the distance between clusters. Cluster distances are characterised by differing font styles to indicate relative closeness. Bold font indicates that the distance between the two clusters is the closest for row clusters and, similarly, italics indicates least distance for the column clusters.

This table is specifically designed to allow the user to intuitively follow clusters over batches and time.

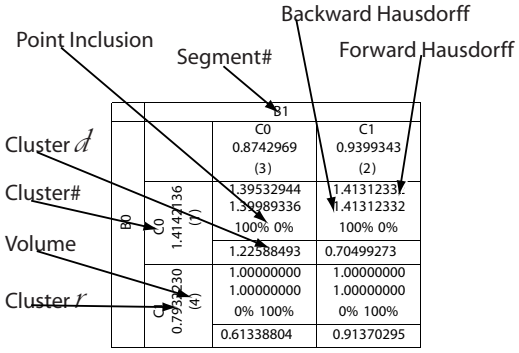

Figure 1. Agreement table field statistics

5.2.4 Hausdorf distance

Hausdorf distance is employed to represent cluster deviation in subsequent intervals. Hausdorff is directional having a forward and backward component indicating geometric deviation between clusters. As seen in equation 1 Hausdorf distance calculates the minimum distance from each element in one set to all elements in the other set in a many-to-many relationship. The maximum of these distances is the Hausdorff distance (in one direction). In figure 2, $|a_3 - b_1|$ is the Hausdorf distance in this instance.

$$h(A, B) = \max_{a \in A}\{\min_{b \in B}\{d(a, b)\}\} \quad (1)$$

Forward Hausdorff distance between point sets A and B.

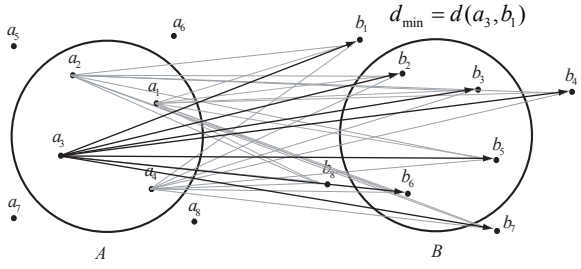

Figure 2. This is a graphical representation of the real Hausdorff distance

5.2.5 Unsupervised adjusted Rand

The Rand algorithm [11] was originally developed to serve as an objective measure of accuracy when testing different clustering algorithms. It has since been adapted to create a more meaningful coefficient in an adjusted implementation (adjusted Rand [9]) but the essential process remains unchanged. A confusion matrix is built by recording the intersection frequencies between the observed values and expected values.

As the graphing techniques are loosely based on matching matrices it was an intuitive step to incorporate the rand index that also expolits its funcionality and simplicity. As we are investigating similarities beteewn disjoint sections of the stream, shared point techniques, set matching and variation of information were unavaliable to us. Pair-counting measures, such as Rand and Jaccard indicies, could be easily adapted to be used without shared points and proved versatile enough to provide a meaningful metric over disjoint divisions.

Class \ Cluster	v_1	v_2	...	v_3	Sums
u_1	n_{11}	n_{12}	...	n_{1C}	$n_{1.}$
u_1	n_{21}	n_{22}	...	n_{2C}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
u_1	n_{R1}	n_{R2}	...	n_{RC}	$n_{R.}$
Sums	$n_{.1}$	$n_{.2}$	...	$n_{.C}$	$n_{..} = n$

Table 1. Notation for the contingency table comparing the two partitions

The dataset employed here is the flow control chart dataset provided by the UCI KDD repository [10]. In figure 3 Global K-means proves best at consistently determining similar structures across intervals. Becuase of its underlying randomness, K-means demonstrates a distinct disadvantage.

6 Results

6.1 Cluster identification and visualisation using synthetic data

To properly validate the interpretation techniques, the synthetic control flow chart was employed again to serve as a baseline. The objective of this experiment is to identify agreements over segments. Global K-means was utilsed as it demonstrated the most stability accross disjoint divisions by the adjusted rand index. Given this assuption, the stability should be easily infered from the proposed representations.

The synthetic control chart time series data [10] was selected from the UCI Machine Learning Repository for

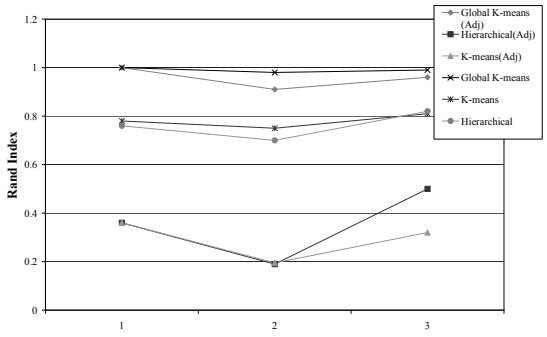

Figure 3. Rand and adjusted Rand index for three algorithms across three intervals of the flow control chart dataset

this experiment. This data represents six different classes of control charts containing one hundred points in each. Four equal segments were derived by equal spacing (non-random) stratification to simulate a stream environment. Immediately it can be seen that the agreements in the table (See table 2) match up seamlessly indicating perfect cluster alignment. Point inclusion further serves to highlight closeness of clusters.

It can immediately be seen that our six clusters and their counterparts in the second segment agree perfectly. Perfect agreements are seen as the unambiguous one-to-one relationships between clusters. Even though eight clusters are specified when there are actually six, the strength of the association, i.e. point association can be seen as the line thickness, indicating the correct number of clusters. From this visualisation it can be seen that the Pixie graph highlights the expected correspondence between like clusters in subsequent groups. Line thickness here may well indicate a drift in the target concept.

The Luna graph describes a high level view of the radius overlap and a high-grained view of corresponding clusters in n-dimensional space. Here we immediately see which clusters are in agreement and which are in doubt. In the third interaction it is also possible to discern subset point inclusions. In low point inclusion interactions we can see just how different the cluster spaces are away from each other.

Through the novel representations presented here, it has been demonstrated that spatial interations between temporal divisions can be diagnosed and overall distribution patterns identified.

		B1						
		C0	C1	C2	C3	C4	C5	C6	C7
B0		31.5492554 (30)	36.9097142 (34)	36.2855123 (30)	31.0599788 (7)	44.9993257 (9)	36.9472730 (5)	28.9388714 (11)	28.4944296 (15)
	C0 (29)	41.69117135	66.64790816	75.93278291	54.18730501	43.0586382	45.6630838	101.18855016	103.28260499
	32.2287196	48.70960201	72.54466869	81.17924911	78.3024315	75.5604975	65.2934671	101.1845216	107.2894300
		0% 0%	0% 0%	0% 0%	0% 0%	0% 0%	0% 7%	0% 0%	0% 0%
		10.3793504	*6.6736917*	*8.6196548*	98.0335219	94.173078	99.25800062	48.83496922	183.37665542
		83% 83%	0% 0%	36% 52%	0% 0%	0% 0%	0% 0%	0% 0%	0% 0%
	C1 (27)	77.5390424	134.3385857	18.8186523	81.78035135	67.53379399	75.03355797	44.66891012	176.3020177
	35.2900928	63.09106301	131.4440907	12.1938795	87.43512764	84.27463937	87.46004102	44.32932926	169.6587283
		0% 0%	0% 0%	0% 0%	0% 0%	0% 0%	0% 0%	0% 0%	0% 0%
		70.46651541	142.1061002	*8.6196548*	98.0335219	94.173078	99.25800062	48.83496922	183.37665542
	C2 (37)	70.46138285	14.99326558	138.1277179	82.71861185	72.68380006	71.62089473	163.46879304	45.11589705
	35.7903501	71.94729498	16.2513699	138.1277179	89.28034162	93.56463284	89.21045401	178.3377006	46.50277996
		0% 0%	53% 62%	0% 0%	0% 0%	0% 0%	0% 0%	0% 0%	0% 0%
		68.4413153	*8.6736917*	138.480822	95.3250465	90.8875994	92.9722995	181.2869428	49.6355321
	C3 (10)	70.5137065	87.1803418	94.31374155	75.7749233	44.6288149	71.4072539	116.1968565	125.3465435
	39.8516958	45.86837969	73.62075872	75.33765088	77.60437469	47.6046106	79.3477359	106.3939508	104.0503882
		3% 0%	0% 0%	0% 0%	0% 0%	33% 30%	0% 0%	0% 0%	0% 0%
		61.84148826	96.03255236	94.12898214	92.2390465	*41.9695426*	81.13064632	129.1206232	131.1928680
	C4 (9)	52.38124010	91.67423536	88.67531951	51.31848852	61.0279257	72.1271102	109.2834830	126.2464318
	34.8681187	49.89771904	81.71530063	76.35583963	46.6600232	69.9511484	80.2402365	106.9931965	111.2779672
		3% 0%	0% 0%	0% 0%	0% 11%	0% 0%	0% 0%	0% 0%	0% 0%
		60.25510624	100.6767188	92.42275362	*41.9690716*	87.1940692	98.005283	127.8860578	134.2083838
	C5 (12)	103.5164646	43.6656949	179.4805137	117.0792616	117.079265	106.179026	206.7448604	15.5713606
	29.1045522	110.515454	43.3304717	179.6176271	124.3877574	122.708379	121.3108753	223.4196148	6.6023011
		0% 0%	0% 8%	0% 0%	0% 0%	0% 0%	0% 0%	0% 0%	33% 17%
		112.834442	44.27003599	182.6875908	130.1375190	127.784213	127.3545008	226.6715215	*12.5834294*
	C6 (3)	65.91353908	96.41292630	100.6315767	81.5322963	80.7564350	63.4968030	118.9775944	129.2335259
	23.9615178	53.63254877	85.88451342	85.28070817	85.1334522	81.4706389	57.9887256	117.7303743	117.3206829
		0% 0%	0% 0%	0% 0%	0% 0%	0% 0%	0% 0%	0% 0%	0% 0%
		60.96370152	105.5418981	102.257103	101.153419	99.4934765	74.2038264	134.9015763	138.2036602
	C7 (23)	113.8266396	169.4419218	169.4419218	33.8286906	97.5639580	104.0706695	18.8473369	215.8054376
	31.8152678	98.46673249	168.1335965	168.1335965	40.41823191	114.209844	117.0341602	120.4568563	210.5131646
		0% 0%	0% 0%	0% 0%	0% 0%	0% 0%	0% 0%	73% 39%	0% 0%
		108.9713763	180.2189521	180.2189521	40.4549385	127.853464	125.2028197	130.8623677	*12.2966992*

Table 2. Agreement tables for the synth interaction.

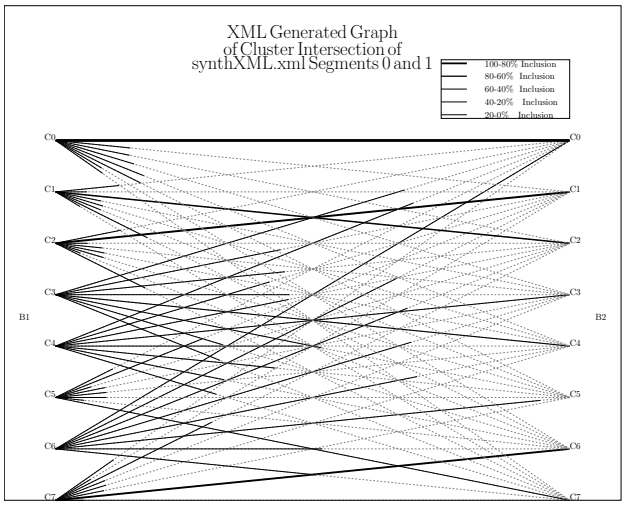

Figure 4. This figure represents the Pixie graph of the first interaction of the synthetic dataset

6.2 Cluster Emergence with Introduced Data

Here we will objectively evaluate how the Global K-means algorithm identifies emergence of an introduced cluster in the final interaction. The entire additional cluster is withheld until the final interaction. For this experiment the synthetic dataset has its first cluster removed and then added to the final interaction. We want to introduce a concept artificially and determine whether we can observe its emergence. With the combination of the Luna and Pixie graph concept emergence should be relatively straightforward to detect.

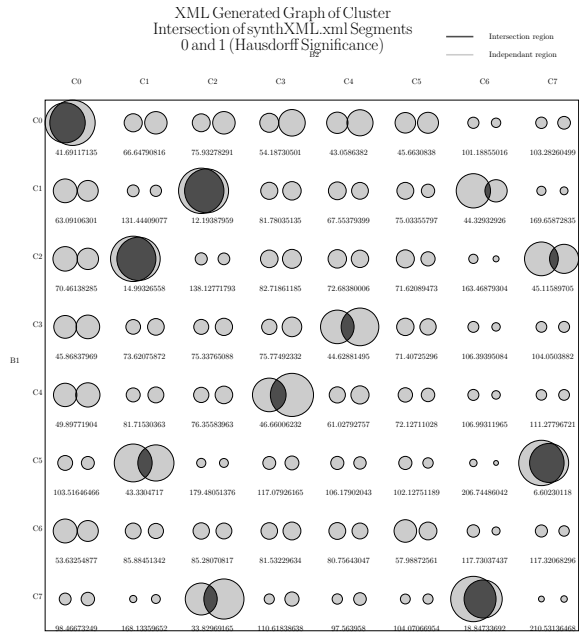

Figure 5. This figure represents the Luna graph of the first interaction of the synthetic dataset

The Pixie and Luna graph (See figures 6 and 7) show persuasive evidence of emergence. It can be seen in final interaction representations (The emergence interval) that cluster B^3C_0 has absorbed the retained group. In the Luna graph we first note that the first two clusters are very stable with the last two clusters without contentions.

Of B^3C_3, B^3C_5 we know that B^3C_3 has the largest radius intersection, is mutually closest, and has the lowest deviation score gained from the Luna graph. B^3C_5 is excluded due to its low point inclusion score and competing clusters in the column. The Pixie graph shows us the clearest of pictures that B^3C_0 has no counterpart in the preceding segment. Any ambiguity present in the Luna graph, as described above, is intuitively addressed and an accurate assessment of the cluster is immediately attainable. From the Luna representation B^2C_0 can align with only B^3C_3.

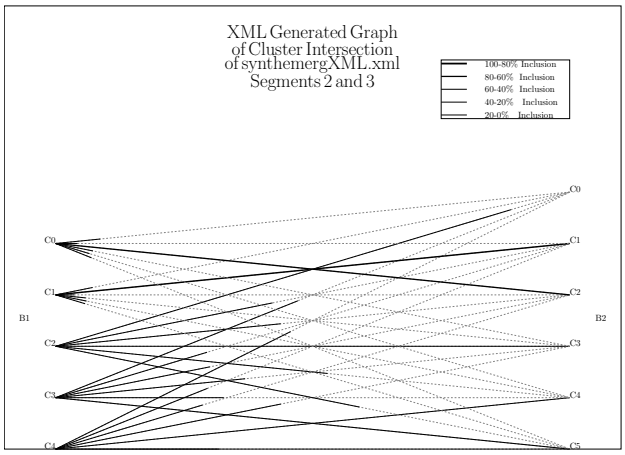

Figure 6. This figure represents the Pixie graph of the last interaction of the emerging synthetic dataset

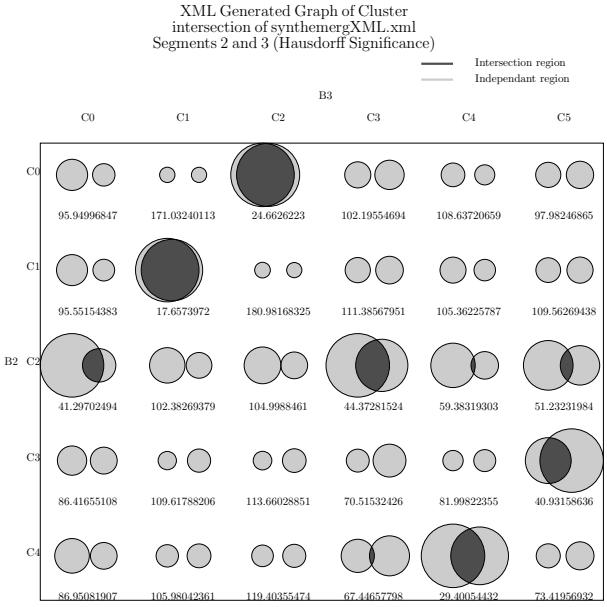

Figure 7. This figure represents the Luna graph of the last interaction of the emerging synthetic dataset

7 Conclusion

In this report we have investigated the important role of visualisation and metrics in determining data evolution. We have highlighted the increased complexities associated with the subdomain of stream mining. We have asserted that error, drift, and emergence are the main causes of change in streams. This being the case, there has been very little in the way of novel graphing techniques to explore these complexities.

Three main areas representing change are investigated. As stream techniques are only in the early days of development, crucial tools, such as evaluation measures and visualisations need to be specifically designed to accommodate streams. Pixie and Luna graphs have demonstrated the ability of such techniques at investigating evolution in data streams. Through the use of visualisation styles that are developed to fit the stream mining field, concepts can be quickly modelled and new approaches built upon intuitive understandings of the data.

References

[1] J. Y. Adil M. Bagirov, Alexander M. Rubinov. A global optimization approach to classification. *Optimization and Engineering*, 3(2):129 – 155, 2002.

[2] C. C. Aggarwal. A framework for diagnosing changes in evolving data streams. In *ACM SIGMOD Conference*, 2003.

[3] E. Bae and J. Bailey. Coala: A novel approach for the extraction of an alternate clustering of high quality and high dissimilarity. In *ICDM '06: Proceedings of the Sixth International Conference on Data Mining*, pages 53–62, Washington, DC, USA, 2006. IEEE Computer Society.

[4] E. Bae, J. Bailey, and G. Dong. Clustering similarity comparison using density profiles. In A. Sattar and B. H. Kang, editors, *Australian Conference on Artificial Intelligence*, volume 4304 of *Lecture Notes in Computer Science*, pages 342–351. Springer, 2006.

[5] V. Ganti, J. Gehrke, and R. Ramakrishnan. A framework for measuring changes in data characteristics. pages 126–137, 1999.

[6] V. Ganti, J. Gehrke, and R. Ramakrishnan. Mining data streams under block evolution. *SIGKDD Explor. Newsl.*, 3(2):1–10, 2002.

[7] M. D. Hoon, S. Imoto, J. Nolan, and S. Miyano. Open source clustering software. *Bioinformatics*, 20(9):1453–1454, 2004.

[8] G. Hulten, L. Spencer, and P. Domingos. Mining time-changing data streams. In *KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 97–106, New York, NY, USA, 2001. ACM Press.

[9] W. L. R. K. Y. Yeung. Details of the adjusted rand index and clustering algorithms. *Bioinformatics*, pages 1–6, 2001.

[10] D. Pham and A. Chan. Control chart pattern recognition using a new type of self organizing neural networ. *Proc. Instn, Mech, Engrs.*, 212(1):115–127, 1998.

[11] W. M. Rand. Objective criteria for the evaluation of clustering methods. *American Statistical Association*, 66:846850, 1971.

[12] L. G. Valiant. A theory of the learnable. In *STOC '84: Proceedings of the sixteenth annual ACM symposium on Theory of computing*, pages 436–445, New York, NY, USA, 1984. ACM Press.