

An Analog Online Clustering Circuit in 130nm CMOS

Junjie Lu, Steven Young, Itamar Arel, and Jeremy Holleman

The University of Tennessee, Knoxville

Abstract—An analog clustering circuit is presented. It is capable of inferring the underlying pattern and extracting the statistical parameters from the input vectors, as well as providing measures of similarity based on both mean and variance. A floating-gate analog memory provides non-volatile storage. A current-mode distance computation, a time-domain loser-take-all and a memory adaptation circuit implement efficient and robust learning algorithm. We show that our analog computation element can achieve more than 10× higher energy efficiency than its digital counterpart. An 8-dimension 4-centroid prototype was fabricated in a 130 nm standard CMOS process. Measurement results demonstrate vector classification at 16 kHz, and unsupervised online clustering at 4 kHz with a power consumption of 15 μW.

I. INTRODUCTION

Energy efficient signal processing is essential for ultra-low-power mobile systems. Compared to digital signal processing, analog signal processing (ASP) shows potential of much higher energy efficiency [1]-[4]. In addition, by performing signal processing and data compression at the front-end, the accuracy and bandwidth requirement of subsequent blocks can be relaxed, increasing the overall system efficiency.

This paper describes the implementation of an ASP system realizing an online k-means clustering algorithm, widely-used in feature extraction, pattern recognition, data compression, and other applications. It infers the underlying data pattern by capturing the regularity of it [5]. A vector quantizer (VQ) searches a set of stored centroids (templates) for the one nearest to the input vector. The proposed system enhances VQ with online construction and adaptation of templates to yield optimal performance under changing input statistics. While this algorithm is expensive in digital domain, it can be realized in ASP with relatively low cost in terms of power and area by exploiting the inherent computational primitives [6]. Analog or mixed mode VQ processors have been developed in [1], [2]. The lack of learning capability requires explicit programming. VQs with learning capability are presented in [3], [4], the centroids are stored in volatile capacitors or digital memories. The non-volatile memory used in this work enables intermittent powering, and the fully analog operation avoids the power and area overhead of internal A/D/A conversion.

We present a novel analog online k-means clustering circuit which performs unsupervised learning in real time. Parameters are stored in non-volatile analog memories compatible with standard digital CMOS. Confidence scores are constructed and

This work was partially supported by the Intelligence Advanced Research Projects Activity (IARPA) via Army Research Office (ARO) agreement number W911NF-12-1-0017, and by NSF grant #CCF-1218492. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, the Department of the Army, the NSF, or the U.S. Government.

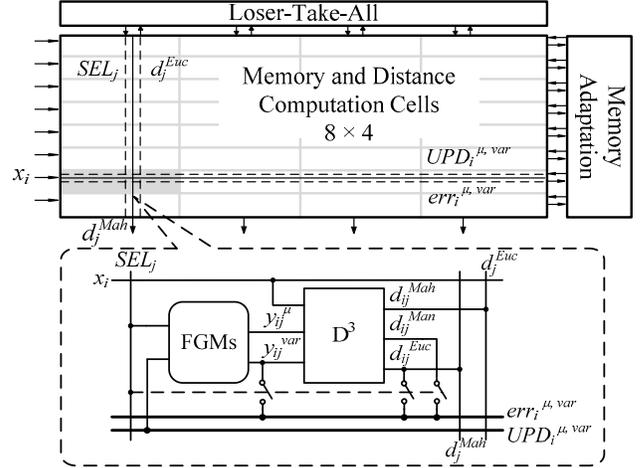


Fig. 1. The architecture of the proposed analog online clustering circuit, including an 8 × 4 array of memory and distance computation cells (one of them is shown in detail), memory adaptation circuits common to each row, and loser-take-all circuits common to each column.

passed to the higher hierarchical layers in a deep machine learning architecture [7]. The architecture and circuit design are optimized for scalable low-power fully-autonomous computing applications.

II. ARCHITECTURE AND ALGORITHM

The architecture of the clustering circuit is shown in Fig. 1. The signal processing is implemented in current mode to allow efficient arithmetic operations and wide linear range. The core of the prototype is an array of memory and distance computation cells (MDCs). The 4 columns form 4 centroids, each with 8 dimensions. The MDC consists of two analog memories (FGMs) and a distance computation block (D^3). The FGM stores the centroid mean y_{ij}^μ and variance y_{ij}^{var} , and is accessible and programmable from off-chip in test mode. The D^3 block provides 3 distance metrics between the input vector and the local centroid, necessary for different operation modes.

First, the i -th element x_i of input vector \mathbf{x} is broadcasted to each cell along row i , and the D^3 's compute the 1-D Euclidean distances between x_i and their local centroid means y_{ij}^μ . The Euclidean distances between \mathbf{x} and the centroid \mathbf{y}_j is obtained by wire-summing the D^3 's output currents along the columns,

$$d_j^{Euc} = \sum_i (x_i - y_{ij}^\mu)^2. \quad (1)$$

A time-domain loser-take-all (LTA) network common to the columns searches for the single centroid k with minimum distance to \mathbf{x} and sets SEL_k to high. The selected column of MDCs is then connected to a memory adaptation (MA) circuit.

In the adaptation phase, the D^3 block outputs both 1-D element of Manhattan distance $d_{ik}^{Man} = |y_{ik}^\mu - x_i|$, and Euclidean distance d_{ik}^{Euc} . The mean and variance error signals between the centroid and input vector are propagated to the MA circuit,

$$\begin{aligned} err_{ik}^\mu &= x_i - y_{ik}^\mu, \\ err_{ik}^{\text{var}} &= d_{ik}^{\text{Euc}} - y_{ik}^{\text{var}}. \end{aligned} \quad (2)$$

Note that the Euclidean distance is reused to construct the error signal for variance because of its simple quadratic form. Then the MA circuit generates update signals UPD_i for the centroid mean and variance memories. Although the UPD signals are broadcasted, logics in the MDCs ensure that only the ones selected by the LTA are updated. The amount of update is proportional to the errors in (2),

$$\begin{aligned} y_{ik}^\mu(n+1) &= y_{ik}^\mu(n) + \alpha \cdot err_{ik}^\mu, \\ y_{ik}^{\text{var}}(n+1) &= y_{ik}^{\text{var}}(n) + \beta \cdot err_{ik}^{\text{var}}, \end{aligned} \quad (3)$$

where α and β are the learning rates. From (3), the memory values follow the exponential moving averages and converge to the true mean and variance of the input data clusters. The memory adaptation is fully parallel.

Utilizing the diagonal elements of the covariance matrix, the proposed clustering circuit constructs a confidence estimate for each centroid based on the Mahalanobis distance to the input.

$$d_j^{\text{Mah}} = \sum_i \frac{(x_i - \mu_{ij})^2}{\sigma_{ij}^2}, \quad (4)$$

where μ_{ij} and σ_{ij}^2 are the estimated mean and variance of the input data cluster. By including variance, the Mahalanobis distance better estimates the probability that a given input came from a centroid. In classification stage, using Mahalanobis distance can lead to a larger "capture range" for a centroid with larger variance estimate, which in turn results in a further increased variance estimate. To avoid this positive feedback, Euclidean distance is used in classification.

III. CIRCUIT IMPLEMENTATION

A. Floating-Gate Analog Memory

The analog memories utilize floating gates for non-volatile analog storage. The update magnitude is pulse-width controlled, and negative feedback is employed to keep the floating gate voltage constant so as to get a smooth sigmoid update rule necessary for stable adaptation. The novel update scheme allows random-accessible control of both tunneling and injection without high-voltage switches, charge pumps or complex routing. The memory does not require special masks or double-poly, and is fully compatible with modern digital CMOS process. A detailed description is presented in [8].

B. Distance Computation (D^3) Block

The schematic of the D^3 block is depicted in Fig. 2. The two output currents from preceding stages (centroid mean y_{ij}^μ and input x_i) are summed with opposite polarities at its input node, and then rectified by M1-M4 and amplifier A1 to yield the unsigned output. The absolute value circuit is modified from [9]. Improvement is made by introducing a virtual ground using A1 to mitigate the error due to finite drain resistances, and improve the speed. A comparator provides the polarity of input current, Sgn , used in the adaptation phase.

In the translinear operator circuit M5-M10, M5 copies the current from the absolute value circuit to get the Manhattan distance $|y_{ij}^\mu - x_i|$, M9 forces the current into the drain of M8 by modulating its source voltage, and M6-8 and M10 form the

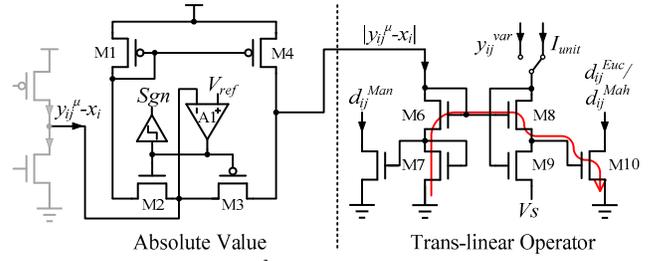


Fig. 2. The schematic of the D^3 block. The output devices of the preceding stages are shown in grey. A1 is built with a differential pair with current mirror load and the comparator with cascade of single-ended amplifiers. V_s biases M8,9 in saturation. The arrow indicates the translinear loop.

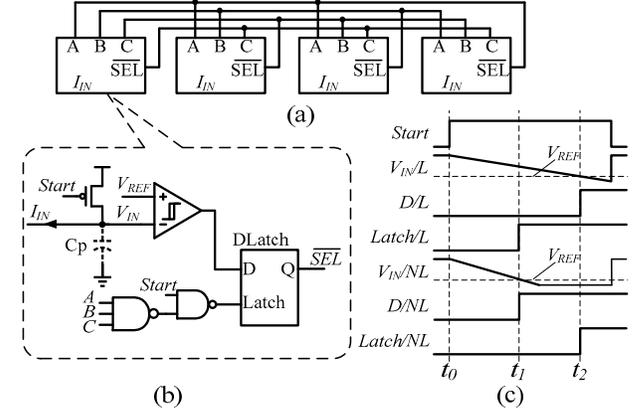


Fig. 3. The simplified schematic of (a) the LTA network, (b) one cell of the LTA circuit, (c) typical timing diagrams, the signals in the loser cell is suffixed with L , and non-loser cell with $/NL$.

translinear loop denoted by the arrow [10]. Assuming that M6-8 and M10 are matched and biased in subthreshold, the drain current of M10 is given by

$$I_{d,M10} = I_{d,M6}^2 / I_{d,M8} = (y_{ij}^\mu - x_i)^2 / I_{d,M8}. \quad (5)$$

The Euclidean distance d_{ij}^{Euc} can be obtained by supplying M8 with a unit current I_{unit} , and the Mahalanobis distance d_{ij}^{Mah} realized by connecting the variance memory output y_{ij}^{var} to M8.

C. Time-Domain Loser-Take-All (TD-LTA) Circuit

The LTA circuit receives the Euclidean distances d_{ij}^{Euc} , and searches for the centroid with smallest distance. It consists of 4 LTA cells interconnected as shown in Fig. 3(a). The LTA cell shown in Fig. 3(b) operates in time domain and exploits the dense digital blocks in modern process. The typical timing diagram of the "loser" and a "non-loser" cell is plotted in Fig. 3(c). The capacitor C_p is initially precharged to V_{dd} , and is discharged by the input current when $Start$ goes high (t_0). For the "loser" cell, the threshold crossing of the comparator (t_2) is the latest among the 4 cells, so the data input D of its D-latch is low when $Latch$ goes high. For the "non-loser" cell, D is high when $Latch$ goes high (t_1). Therefore, the output of the "loser" is latched to low while those of the others latched to high. At the end of LTA phase, $Start$ latches all the cells regardless of V_{IN} . Additional logic, omitted from Fig. 3 for clarity, prevents the selection of multiple losers.

Compared to a continuous time (CT) implementations similar to [11], the proposed TD-LTA can potentially yield lower power-delay product if C_p is realized with the parasitic

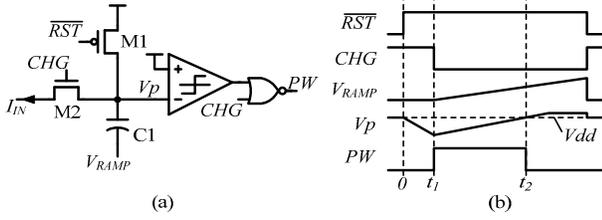


Fig. 4. (a) The simplified schematic and (b) timing diagram of the MA circuit.

capacitance at the input node. Since the input current is off before LTA phase, for CT-LTA, the input node has a voltage swing of about V_{th} , and a time constant of Cp/g_m , where V_{th} and g_m are the threshold voltage and transconductance of the input transistor. Neglecting the bias consumption and internal-node settling times, assuming $gm/Id \approx 20$ for transistors biased in weak inversion, and accounting for slewing and 3τ settling time, the current-delay product of the CT-LTA is

$$I_{IN} \cdot t_{settle,CT} = (V_{th} + 0.15) \cdot Cp, \quad (6)$$

while the current-delay product of the TD-LTA is

$$I_{IN} \cdot t_{d,TD} = (V_{dd} - V_{REF}) \cdot Cp. \quad (7)$$

Comparing (6) and (7), with same Cp , the TD-LTA has smaller current-delay product if the V_{REF} is set less than $V_{th} + 0.15$ below V_{dd} . In the prototype, V_{REF} is adjustable to balance between energy efficiency and accuracy. The accuracy also depends on the matching of Cp , and a good matching of wiring capacitances can be achieved with layout regularity.

To address unfavorable initial condition, a starvation trace mechanism is implemented [12]. Current is injected to decrease the apparent distance of the centroid which has not been selected by the LTA for a long time. As a result, centroids initialized too far away from populated regions of the input space will slowly move toward the input samples.

D. Memory Adaptation (MA) Circuit

The error currents between the input and the best-matching centroids' memory values are passed to the MA circuits. Each row of the MDC cells shares two MA circuits, for mean and variance memory respectively. The simplified schematic and timing diagram is shown in Fig. 4. The MA circuit utilizes the charging and discharging of capacitor to realize current-to-pulse-width conversion. The voltage V_p is first discharged from V_{dd} by the input current for a fixed period of t_1 , then ramped up by the external voltage V_{RAMP} at the bottom plate of $C1$, until V_p crosses V_{dd} at t_2 . The update pulse is defined by $t_2 - t_1$, and is proportional to the input error current, allowing the memory values to adapt to the moving averages. While the MA determines the magnitude of memory adaptation, the direction of adaptation for mean memory is determined by the Sgn output of the MDC. For variance memory, an absolute value circuit in Fig. 2 is added to the MA block, generating both the unsigned error current and the direction of adaptation.

IV. MEASUREMENT RESULTS

The proposed clustering circuit has been fabricated in a 130 nm CMOS process using IO FETs, occupying 0.18 mm^2 of active area including the programming registers and biasing circuits, shown in Fig. 5. The prototype has 8 input dimensions and 4 centroids, and consumes $15 \mu\text{W}$ with 3 V supply. Test

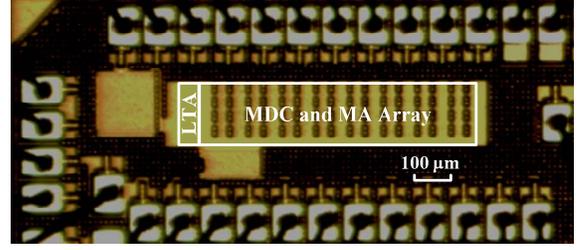


Fig. 5. The micrograph of the prototype clustering circuit.

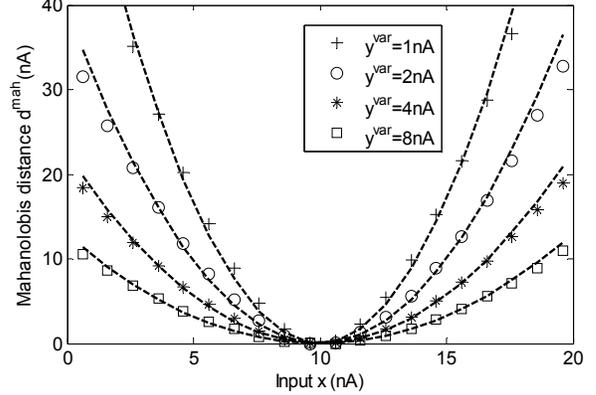


Fig. 6. Transfer characteristics of the D^3 block. The markers are the measured data and the dashed lines are curves fitted to the quadratic form $a(x-10)^2$, where a is the fitting parameter.

signals are provided to and acquired from the chip using PC-based data acquisition hardware and off-chip current-voltage converters on a custom test board.

To characterize the D^3 block, the stored centroid mean y^μ of a MDC cell is programmed to 10 nA, and its input x is swept from 0 to 20 nA. The Mahalanobis distance output with varying centroid variance y^{var} is measured and plotted in Fig. 6. The D^3 block shows good linearity over the entire input range, and a power law output corresponding to (4).

The classification test was performed by programming the centroids to fixed positions and disabling memory adaptation. The inputs are equally spaced and randomly presented to the circuit. To allow easier visual interpretation, only 2 out of 8 dimensions of input vectors are shown. The results are color-coded and the measured decision boundaries show good matching with the ideal boundaries, illustrated in Fig. 7. The prototype circuit runs classification at a speed of 16 kHz, limited by the settling time of the input current.

We demonstrated the full functionality of the prototype by solving a clustering problem. 40000 8-dimensional vectors were generated as the inputs to the circuit. The data set contains 4 underlying clusters, each drawn from Gaussian distributions with different means and variances. Initially the centroids were programmed to separated means and a same variance. The initial condition is not critical since the circuit will adaptively adjust to the inputs. During the test, the centroid means were read out every 0.5 s, plotted on top of the data scatter in Fig. 8, and shown together is the learned variance values at the end of test. Again 2 of 8 dimensions are plotted. The centroids adapt accurately to centers of the input data clusters despite their overlapping, and the extracted variances match with the true values, both confirming a robust

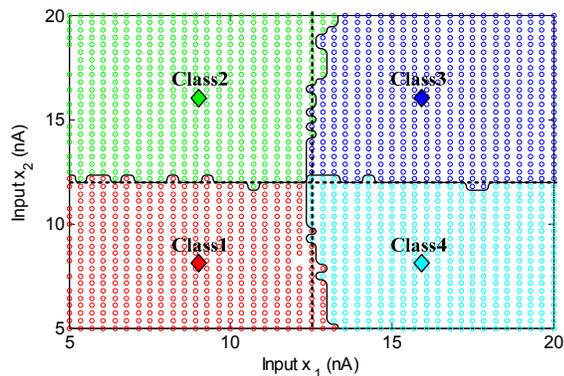


Fig. 7. Classification test results. The 4 centroids are shown in diamond shapes. The circuit assigns the input data to different centroids based on the Euclidean distances. The results are color coded and the measured decision boundaries are shown as solid lines and ideal boundaries as dashed lines.

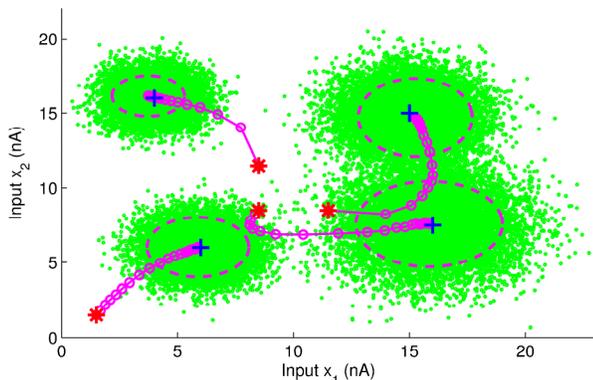


Fig. 8. Clustering test result. The input data are shown in green circles. The initial position of centroid means are marked with red stars, their positions for every 0.5 second during the test are plotted with magenta circles connect by lines. The extracted variances of the clusters are plotted with dashed ellipses in magenta, and the true means of data clusters are marked with blue crosses.

learning performance. The task takes 10 s at 4 kHz; higher speed is possible at the cost of lower learning rate.

The measured performance is summarized in Table. I. The bottleneck of the operation speed is the settling time of inputs, because small currents are delivered from off-chip. Simulation results indicate that the internal circuitry is operational at up to 50 kHz. Therefore, the energy efficiency of this proof-of-concept prototype can be largely improved by pre-distorting the input current to voltage [1] or using on-chip D/A converters [4]. To compare the proposed analog implementation to its digital counterpart, we focus on the translinear operator shown in Fig. 2, because it is the most computation-intensive block, and will dominate the power consumption when the system is scaled up. The translinear circuit has an SNR of 130 with typical bias current, and operates with an energy efficiency of 76.2 fJ/OP. A digital arithmetic block computing x^2/y with 7-bit resolution was synthesized in the same process and yields an energy efficiency of 1 pJ/OP, suggesting that the analog implementation in this work can have more than one magnitude higher energy efficiency. Further advantage over digital realization can be expected if the cost-free current-mode summing, efficient TD-LTA and memory reading/writing are included to the comparison.

TABLE. I. PERFORMANCE SUMMARY

Technology	130 nm digital CMOS
Total Area	$0.9 \times 0.2 \text{ mm}^2$ (8×4 array)
MDC Cell Area	$90 \times 30 \text{ }\mu\text{m}^2$
Power Consumption	$15 \text{ }\mu\text{W @ } 3\text{V}$
Classification Speed	16 kHz
Clustering Speed	4 kHz

V. CONCLUSIONS

An analog online clustering circuit has been presented. A floating-gate memory with random-accessible bi-directional update capability provides non-volatile storage. An analog computation block utilizes translinear principles to obtain 3 different distance metrics with significantly lower energy consumption than an equivalent digital implementation. A TD-LTA is proposed to improve energy efficiency, and an MA circuit implements a robust learning algorithm. The prototype circuit fabricated in a 130nm digital CMOS process demonstrates unsupervised real-time classification, statistical parameter extraction and clustering of the input vectors with a power consumption of 15 μW . The array-based architecture scales easily, the non-volatile floating-gate memory enables intermittent powering, and the fully analog operation avoids the power and area of internal A/D/A conversion.

REFERENCES

- [1] S. Chakrabarty and G. Cauwenberghs, "Sub-microwatt analog VLSI trainable pattern classifier," *IEEE J. Solid-State Circuits*, vol. 42, no. 5, pp. 1169-1179, May 2007.
- [2] R. Chawla, A. Bandyopadhyay, V. Srinivasan and P. Hasler, "A 531nW/MHz, 128x32 current-mode programmable analog vector-matrix multiplier with over two decades of linearity," in *Proc. IEEE Custom Integr. Circuits Conf. (CICC)*, Oct. 2004.
- [3] J. Lubkin and G. Cauwenberghs, "A micropower learning vector quantizer for parallel analog-to-digital data compression," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 1998, pp. 58-61.
- [4] K. Kang and T. Shibata, "An on-chip-trainable Gaussian-kernel analog support vector machine," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 57, no. 7, pp. 1513-1524, Jul. 2010.
- [5] D. J. C. MacKay, *Information Theory, Inference and Learning Algorithms*, New York, NY, USA: Cambridge University Press, 2003.
- [6] R. Sarpeshkar, "Analog versus digital: extrapolating from electronics to neurobiology," *Neural Comput.*, vol. 10, no. 7, pp. 1601-1638, Oct. 1998.
- [7] I. Arel, D. Rose and T. Karnowski, "A deep learning architecture comprising homogeneous cortical circuits for scalable spatiotemporal pattern inference," in *Proc. NIPS 2009 Workshop on Deep Learning for Speech Recognition and Related Applications*, Dec. 2009.
- [8] J. Lu and J. Holleman, "A floating-gate analog memory with bidirectional sigmoid updates in a standard digital process," in *Proc. IEEE Int. Symp. Circuit Syst. (ISCAS)*, May 2013.
- [9] Z. Wang, "Novel pseudo RMS current converter for sinusoidal signals using a CMOS precision current rectifier," *IEEE Trans. Instrum. Meas.*, vol. 39, no. 4, pp. 670-671, Aug. 1990.
- [10] B. Gilbert, "Translinear circuits: a proposed classification," *Electron. Lett.*, vol. 11, no. 1, pp. 14-16, 1975.
- [11] J. Lazzaro, S. Rytkebusch, M. A. Mahowald and C. Mead, "Winner-take-all networks of $O(n)$ complexity," in *Advances in Neural Information Processing Systems I*, D. S. Touretzky, Ed. San Mateo, CA: Morgan Kaufman, 1989, pp. 703-711.
- [12] S. Young, I. Arel, T. Karnowski and D. Rose, "A fast and stable incremental clustering algorithm," in *Proc. 7th International Conference on Information Technology*, Apr. 2010.