Role of Visual Assessment of Clusters for Big Data Analysis from Real-world Internet of Things

Marimuthu Palaniswami, *Fellow, IEEE,* Aravinda S. Rao, *Member, IEEE,* Dheeraj Kumar, Punit Rathore, *Member, IEEE,* and Sutharshan Rajasegarar

***Abstract*—Internet of Things (IoT) is playing a vital role in shaping today's technological world and our daily lives. By 2025, the number of connected devices because of IoT is estimated to surpass a whopping 75.44 billion. It is a challenging task to discover, integrate and interpret processed big data from such ubiquitously available heterogeneous and active nature of the resources and devices. Cluster analysis of IoT-generated big data is essential for meaningful interpretation of such complex data. However, often we have very limited knowledge of number of clusters really present in the given data. The problem of finding whether clusters are present even before applying clustering algorithms is termed as *assessing of clustering tendency*. In this article, we present a set of useful Visual Assessment of Cluster Tendency (VAT) tools and techniques developed with major contributions from James C. Bezdek. The article further highlights how these techniques are advancing the field of Internet of Things through large-scale IoT implementations.**

***Index Terms*—Internet of Things; Big Data; Clustering; Visual Assessment of Cluster Tendency (VAT).**

I. Introduction

Internet of Things (IoT) technology has significantly changed the way we live today. Physical objects (or devices) with the ability to sense, process and communicate the information to other devices has enabled IoT to empower "things" (or devices) to not only connect with other devices, but also to control the devices in far-off parts of the world [1]. This notion of sensing, processing, communicating and actuating provides a range of unprecedented opportunities to address many challenges in the world.

With the beginning of 21st century, the evolution of IoT and number of devices being used on the Internet is exponential—we see billions of devices from every corner of the world is now being connected to the Internet. For example, in 2018, there were about 23.14 billion Internet connected devices when compared with 15.41 billion in 2015. This is an addition of massive 7.73 billion devices in just under 3 years. By 2025, the number of IoT connected devices is predicted to surpass a whopping 75.44 billion, a five-fold increase in 10 years [2]. We see IoT penetration in almost all major industry sectors: agriculture and food, healthcare, energy and natural resources, water management, transportation and logistics, manufacturing, retail and advertisement, government, insurance and education. Specifically, we see high penetration in application areas of precision farming, wearable devices, smart homes, connected vehicles, industrial robots and smart cities. These IoT devices are producing a mind-boggling, 2.5 quintillion (*i.e.*, 2.5×10^{18}) bytes of data everyday (in 2018) [3]. The enormous amounts of data generated by devices poses several challenges to acquire, store, process, visualize and interpret the data, and use that knowledge to our own benefit.

In 1997, Michael Cox and David Ellsworth from National Aeronautics and Space Administration (NASA), first used the term "Big Data" for data sets that are large for visualization [4]. In 2001, Doug Laney identified 3Vs (Volume, Velocity and Variety) of data growth in a note for understanding and dealing with big data [5], [6]. Today, the term big data (or alternatively very large data sets [7]) is used to characterize the exponential increase of structured and unstructured data, which poses challenges to capture, store, manage and process using conventional data management and analysis techniques. It is a challenging task to discover, access, process, integrate and interpret data from such ubiquitously available heterogeneous and active nature of the resources and devices [8]. Since 2001, the number of "V"s to identify different characteristics of data has grown up to 7 (7Vs: **Volume**, **Velocity**, **Variety**, **Validity**, **Volatility**, **Variability** and **Visualization**) [9] of big data.

II. Visual Assessment of Clusters

Clustering or cluster analysis plays a major role in identifying clusters or patterns of subsets data, and relates to the problem of separating a set of data objects $O = \{o_1, o_2, \ldots, o_n\}$ into c self-similar subsets. These subsets are formed depending on the available data and some explicit measure of similarity of clusters [10]. Depending on the data, geometric descriptions of the clusters are also sought. While the clustering is conceived as an act of segregating the objects into convenient groups, cluster analysis aims to answer the following questions [11]: i) (tendency of clusters) how many clusters are there? ii) (data partition) to which cluster which objects belong and to what degree they are associated? iii) (validity of cluster) whether the partitions of data are good? Clustering algorithms require the number of clusters as input; however, many a times we may not know this beforehand and sometimes also not possible to determine by looking at the data. This is where the Visual

M. Palaniswami and A. S. Rao are with the Department of Electrical and Electronic Engineering, The University of Melbourne, Parkville, VIC 3010, Australia (e-mail: palani@unimelb.edu.au, aravinda.rao@unimelb.edu.au).

D. Kumar is with the Department of Electronics & Communication Engineering, Indian Institute of Technology (IIT) Roorkee, India (email: dheerajfec@iitr.ac.in).

P. Rathore is with Institute of Data Science, National University of Singapore, 117602, Singapore (email: idspr@nus.edu.sg).

S. Rajasegarar is with School of Information Technology, Deakin University, Burwood, VIC 3125, Australia (email: sutharshan.rajasegarar@deakin.edu.au).

This article is accpted for publication. The final published version is available at https://doi.org/10.1109/MSMC.2019.2961160.

Assessment of Cluster Tendency (VAT) algorithm [10] comes into picture. It is important to note that even if no "actual" clusters exist in the data, all clustering algorithms will be able to find c number of clusters, where $1 \leq c \leq n, n \in \mathbb{N}$. As a consequence, it is fundamentally important to ask oneself whether there are any "actual" clusters before applying any clustering algorithms [10].

Visual Assessment of Cluster Tendency (VAT) technique

The problem of finding whether clusters are present even before applying clustering algorithms is termed as *assessment of clustering tendency*. Several techniques, both formal (based on statistics) and informal (other approaches), have been proposed [12], [13], but they are not completely effective. On the other hand, visual approaches [14], [15] for analyzing data have been around for over four decades, forming the basis for many visual data analysis techniques. Bezdek's VAT tool [10] presents a pair wise dissimilarity knowledge about the set of objects $O = \{o_1, o_2, \ldots, o_n\}$. This is usually represented as a square digital image with n^2 pixels [10]. The advantage of VAT as opposed to other visual techniques is that it's ability to highlight the potential number of clusters in the data by suitably reordering the objects. This is achieved by reordering the dissimilarity matrix of the input data using modified Prim's algorithm, and visually estimating the number of clusters that appear as the dark blocks along the diagonal of reordered dissimilarity image (RDI). Fig. 1 shows how VAT tool helps to determine the number of clusters from dissimilarity matrix. The VAT tool is widely applicable to large real-world data sets, including big data. The VAT tool allows to display reordered dissimilarity data, and this can be accessed from the original data O. If O has any missing components, then any existing data imputation schemes can be used to fill in the missing parts of the data before applying VAT.

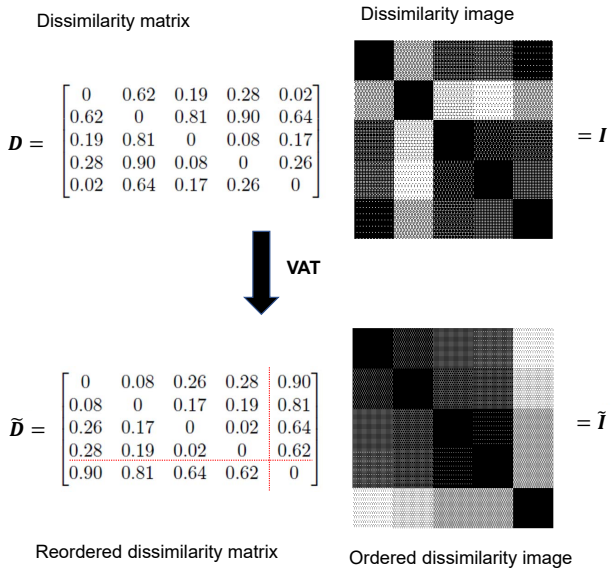


Fig. 1. Illustration of how VAT reorders the dissimilarity matrix. D is the dissimilarity matrix obtained from objects in O. From D, we notice that it is difficult to determine how many clusters are present. I represents the dissimilarity image. $\tilde{D}$ is the VAT reordered matrix obtained after applying VAT and $\tilde{I}$ is the reordered dissimilarity image. From the pair $(\tilde{D}, \tilde{I})$, we notice that there are two clusters of block sizes 4×4 and 1×1 in $\tilde{D}$ and is evident from the visual substructure suggested by $\tilde{I}$.

III. Extensions of VAT for handling Big Data

A. Improved Visual Assessment of Cluster Tendency (iVAT) for Improved Contrast

Suppose we have a pairwise dissimilarity matrix D of a set of n objects, VAT generally portrays D as an $n \times n$ image $I(\tilde{D})$, where the objects are reordered to reveal hidden cluster structure as "dark blocks" along the diagonal of the image [16]. However, VAT fails to clearly display the dark block if the data has complex structures. On the other hand, iVAT proposes to improve the performance of VAT by transforming the reordered dissimilarity matrix $\tilde{D}$ using the graph-theoretic geodesic distance. Evidently, iVAT significantly enhances the separation of the "dark blocks" in VAT images [11].

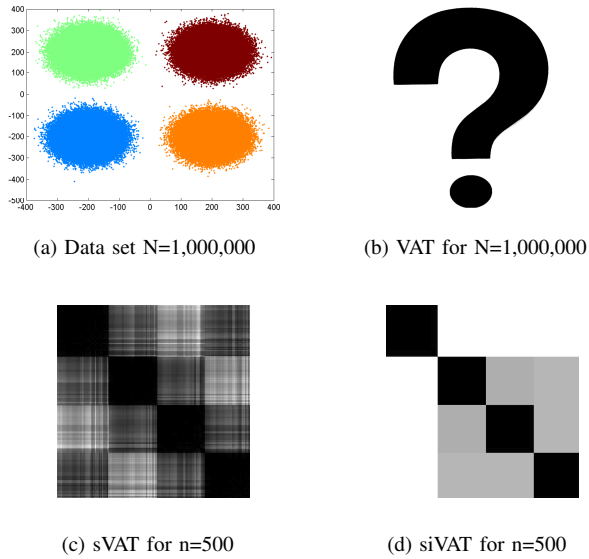
B. Scalable Visual Assessment of Cluster Tendency (sVAT) for Large Data Sets

Although the VAT tool finds its usefulness in many IoT applications, it can be computationally expensive as the size of the data set grows. An algorithm is said to be scalable if there is a linear increase in the run-time complexity with the increase in number of observations in the input data [17]. The VAT has a run-time complexity of $\mathcal{O}(N^2)$, which is not attractive for large data sets. On the other hand, sVAT algorithm is scalable, and uses sample-based version of VAT that can handle large data sets [18]. The sVAT chooses a sample of size n from the complete set of objects $O = \{o_1, o_2, ..., o_N\}$, and executes VAT on the distance matrix of n sample. The sample is selected such that it contains a cluster structure identical to the full set. This operation requires one to first pick a set of k' distinguished objects using the *maximin sampling* [19] to provide a depiction of each of the clusters. Subsequently, the remainder of the sample is produced by selecting additional data near each of the distinguished objects, leading to *maximin and random sampling* (MMRS). To assess the cluster tendency of large volume data set, we apply VAT algorithm on the MMRS samples. In Fig. 2(a) we see the scatterplot of $N = 1,000,000$ two-dimensional points drawn from four Gaussian components and $250,000$ points per cluster. However, we cannot create a VAT image—as shown Fig. 2(b) with question mark (?). On the other hand, sVAT and siVAT allow us to create images by sampling $n = 500$ points (0.05 % of the total data set) from O. In Fig. 2(c), the sVAT image indicates that there are four clusters and these clusters are enhanced in the siVAT image (Fig. 2(d)).

C. Scalable Single-Linkage Visual Assessment of Cluster Tendency (sVAT-SL) for High Volume Data Sets

While the sVAT is helpful in understanding the cluster structure of the big data and determining the optimal value of k—the number of clusters to seek based on the visual evidence, they on their own do not partition the data into k clusters. To tackle this issue, Scalable Single-Linkage Visual assessment of

This article is accpted for publication. The final published version is available at https://doi.org/10.1109/MSMC.2019.2961160.



(a) Data set N=1,000,000 (b) VAT for N=1,000,000

(c) sVAT for n=500 (d) siVAT for n=500

Fig. 2. Images of big data Gaussian clusters: (a) data scatterplot, (b) VAT, (c) sVAT, and (d) siVAT.

Cluster Tendency (sVAT-SL) [20] extends the sVAT algorithm to return single-linkage partitions of big data. The sVAT-SL works by calculating a single-linkage partition of the sVAT-sampled data and then extending this partition to the entire data set using a *nearest prototype rule* (NPR). It is shown that is a scalable instantiation of single-linkage clustering for data sets that contain k compact-separated clusters; and, the sVAT-SL produces a good approximation of single-linkage partitions on data sets not containing compact-separated clusters [20].

D. clusiVAT and Fast-clusiVAT for Faster Computations

The *clusiVAT* algorithm (renamed from sVAT-SL) is superior in regards to cluster quality and computation time over popular big data clustering algorithms, such as Minimum Spanning Trees (MSTs) constructed with the Filter-Kruskal, k-means, single pass k-means, online k-means, and clustering using representatives (CURE) [21], [22]. The clusiVAT is fast in cases where the objects are represented by their feature vectors and the Euclidean distance as a distance measure between objects. This superior computational speed is because the clusiVAT assumes that the distance function computation is quick and clustering can be executed as a batch *i.e.*, using matrix operations, we can compute the Euclidean distance of a datapoint from $M >> 1$ datapoints as a single operation. However, this fundamental assumption does not hold for many distance measures applicable to graphs and time series; there are many distance measures applicable for problems in different domains that are computationally expensive and can only be computed in a pair-wise manner. Fast-clusiVAT addresses this time-consuming distance measure issue by adopting maximin sampling and NPR while maintaining the accuracy [23].

E. Clustering of Streaming Data

Traditional clustering approaches, which provide a fixed outline of each datapoint's pledge to every group, may not be suitable for handling streaming data with new clusters or clubbing of existing concentrated data regions for streaming data sets. Incremental algorithms, such as *incremental VAT* (inc-VAT), *incremental iVAT* (inc-iVAT), *decremental VAT* (dec-VAT), and *decremental iVAT* (dec-iVAT) [24] provide efficient mechanisms to update the VAT and iVAT RDI in the case a new point is added to or an existing point is removed from the current data set.

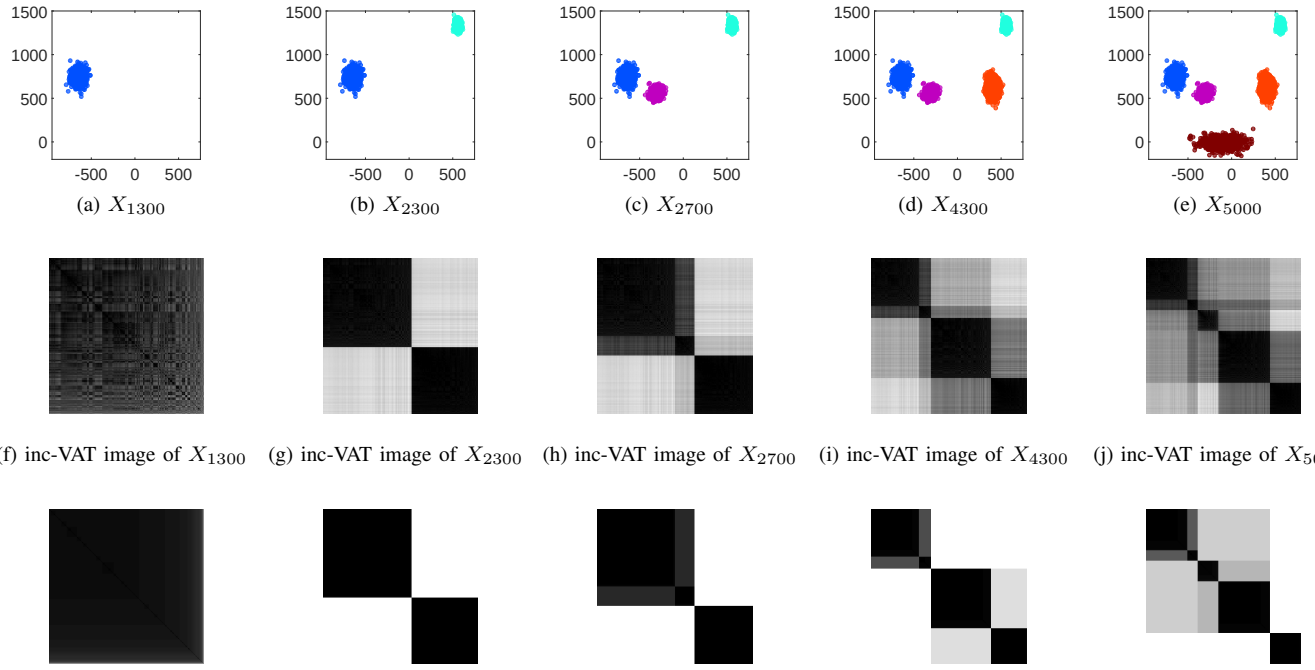
The time complexity of inc-VAT, dec-VAT, inc-iVAT, and dec-iVAT matches fairly with that of VAT and iVAT, respectively. These set of algorithms find applications in detecting anomalies as well as in sliding-window based visual cluster assessments for detecting clusters in streaming data in an online fashion. To illustrate the effectiveness of these algorithms in visualizing the evolving nature of cluster structures and computational efficiency compared to VAT/iVAT, an experiment conducted on a 2-dimensional Gaussian mixture (X) of 5 clusters are shown in Figs. 3 and 4. Each cluster in Fig. 3 has 1300, 1000, 400, 1600, and 700 datapoints, respectively. The datapoints in X are ordered based on cluster membership. As a result, the first 1,300 rows of X are x and y coordinates correspond to the datapoints of first cluster; the subsequent 1,000 rows correspond to the second cluster and so forth. Different columns of Fig. 3 show a subset of X coming from the first cluster, first two clusters, and the rest.

To emphasize the difference in time complexities of VAT/iVAT and inc-VAT/inc-iVAT, we shuffle the rows of X such that the datapoints of the same cluster are apart. This is initiated with two datapoints and then by adding a datapoint at each time step. At each time step we measure the time required by each algorithm (VAT, iVAT, inc-VAT, and inc-iVAT) to compute the reordered dissimilarity matrices. From Fig. 4(a) we we see that, as n increases, VAT+iVAT requires more time to update when compared to inc-VAT+inc-iVAT. Likewise, to reveal the time complexity between dec-VAT/dec-iVAT and VAT/iVAT, we perform the experiment on the aforesaid 2-dimensional data (X). As we are comparing the decremental nature of the algorithms, we initiate the process with $n = 5000$ datapoints and eliminate a single arbitrarily selected datapoint at each time step. From Fig. 4(b) we see that, as n decreases, dec-VAT+dec-iVAT requires much less time VAT+iVAT ($\mathcal{O}(n^2)$).

F. Clustering Large Volumes of High-dimensional Data

Majority of the clustering algorithms are designed to handle data sets either: (1) with a very large sample size, or (2) with a very high number of dimensions. However, they are usually impractical when the data set (generated especially from IoT devices) is large (both in sample size and dimensions). From the earlier sections, we see that both sVAT-SL and clusiVAT algorithms have the ability to handle the data cardinality with sampling schemes; however, they cannot deal with high-dimensional data. To address this critical issue, *FensiVAT* [25] is proposed. FensiVAT is a fast ensemble-based scalable iVAT algorithm. It integrates a new random projection-based distance matrix with MMRS sampling and iVAT to cluster large volumes of high dimensional data. FensiVAT is also several

This article is accpted for publication. The final published version is available at https://doi.org/10.1109/MSMC.2019.2961160.

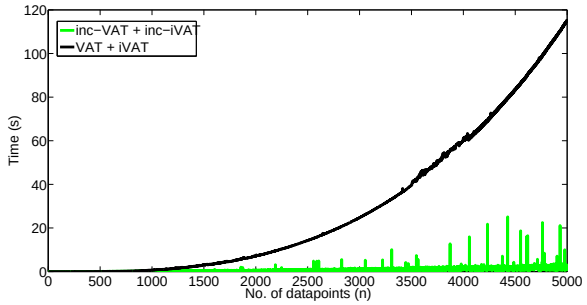


(a) X_{1300} (b) X_{2300} (c) X_{2700} (d) X_{4300} (e) X_{5000}

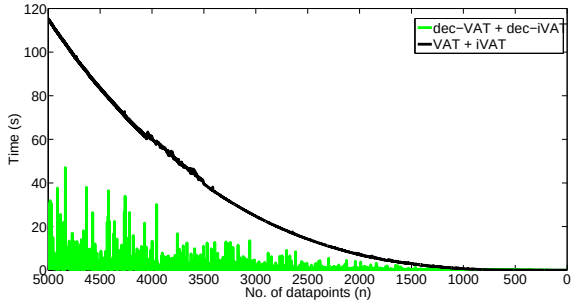
(f) inc-VAT image of X_{1300} (g) inc-VAT image of X_{2300} (h) inc-VAT image of X_{2700} (i) inc-VAT image of X_{4300} (j) inc-VAT image of X_{5000}

(k) inc-iVAT image of X_{1300} (l) inc-iVAT image of X_{2300} (m) inc-iVAT image of X_{2700} (n) inc-iVAT image of X_{4300} (o) inc-iVAT image of X_{5000}

Fig. 3. The 2-D data scatterplot, inc-VAT and inc-iVAT images of X at $n = 1300$, 2300, 2700, 4300, and 5000.



(a) VAT + iVAT vs. inc-VAT + inc-iVAT



(b) VAT + iVAT vs. dec-VAT + dec-iVAT

Fig. 4. Time required for a combination of VAT, iVAT, inc-VAT, inc-iVAT, dec-VAT, and dec-iVAT algorithms for the 5000-point 2D data set.

orders of magnitudes faster than the alternative clustering techniques including clusiVAT, without sacrificing accuracy.

IV. Real-world Applications of VAT families to Internet of Things

A. Monitoring the Great Barrier Reef of Australia

The Great Barrier Reef (GBR) of Australia comprises of 3200 coral reefs spanning over 280,000 square kilometers [26]. GBR is both economically and ecologically sensitive, however, the burning of fossil fuels has led to absorption of carbon dioxide (CO_2) in oceans, resulting in acidification of ocean. This process prevents corals from secreting calcium carbonate exoskeletons, diminishing the reef-building mechanism and linked organisms. Human-induced activities are increasing the stress on coral reefs, leading to coral bleaching, wherein the symbiotic relationship between the coral and algae breaks down during rapid changes in sea-water temperature (hot or cold) [26].

The Great Barrier Reef Ocean Observing System (GBROOS) Project aims to provide observational data to figure out the long-term effects of Coral Sea on the ecosystems and the impact on the GBR. To monitor the reef ecosystem, we collect temperature profiles and weather data from the Heron Island of the GBR. The iVAT algorithm detects the passage of Tropical Cyclone Hamish in March 2009 (see Fig. 5) [27]. We considered one month of data (21-Feb-2009 to 22-Mar-2009, 9:00am to 3:00pm with 10 minutes of sampling frequency) as a case study. Fig. 6 shows the Cyclone Hamish event as two anomalous clusters.

This article is accpted for publication. The final published version is available at https://doi.org/10.1109/MSMC.2019.2961160.

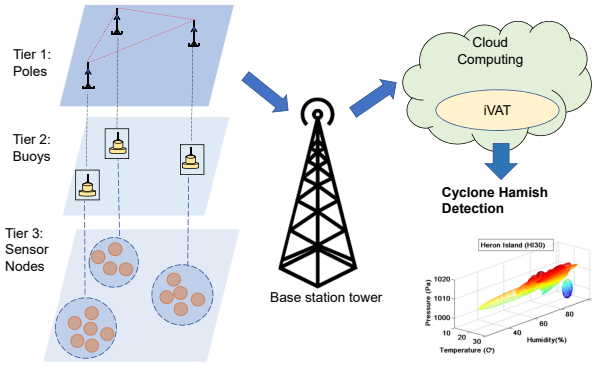


Fig. 5. Two-tiered (tree) hierarchical network architecture of wireless sensor nodes deployed on the Heron Island of GBR for continuous monitoring of the Reef. Using iVAT, we detected the passage of Cyclone Hamish (2009).

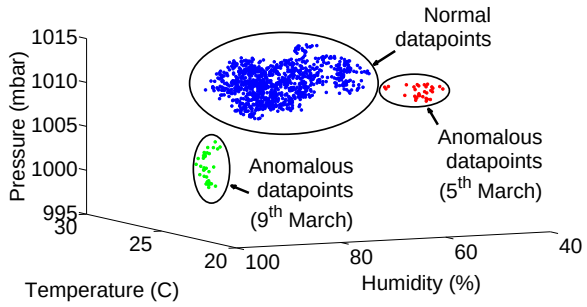


Fig. 6. Cyclone Hamish passed the Queensland state of Australia during 4 March–14 March 2009. The iVAT algorithm accurately clusters these events as two anomalous clusters, correspond to 5 Mar 2009 & 9 Mar 2009, respectively.

B. Urban Forest Monitoring in the City of Melbourne

The *Internet of Things* (IoT) infrastructure for the creation of smart cities consists of internet connected sensors, devices and citizens. This IoT infrastructure generates an enormous amount of data in the form of city-scale physical measurements and public opinions, constituting big data. Smart cities aim to efficiently use this wealth of data to manage and solve the problems faced by modern cities for better decision making. However, interpretation of the massive amount of smart city generated big data to create actionable knowledge is a challenging task. Environmental sensors measuring luminosity, humidity and temperature were deployed at Fitzroy Gardens and Docklands Library to study the effects of canopy cover and the impact of canopy cover on extrapolating the effects to enhance citizen interactions with the city infrastructures.

To this effect, we cluster sensor data (luminosity, humidity and temperature) from four sensors at the Docklands Library utilizing sliding windows. We re-sample the data so that there is one measurement from all four sensors in every 30 minutes time interval, resulting in a 12-dimensional feature vector. In total, we have measurements spanning 72 days. With a window size of 2 days, we have 96 samples. In each time step, the inc-VAT/inc-iVAT includes a new datapoint while the dec-VAT/dec-iVAT removes the oldest datapoint. In the first time step, the inc-VAT/inc-iVAT appends 96 samples to the MST, whereas in the last step dec-VAT/dec-iVAT terminates when sample size equals to 2. Fig. 7 shows the application of inc-iVAT/dec-iVAT to Docklands Library data to detect anomalous samples.

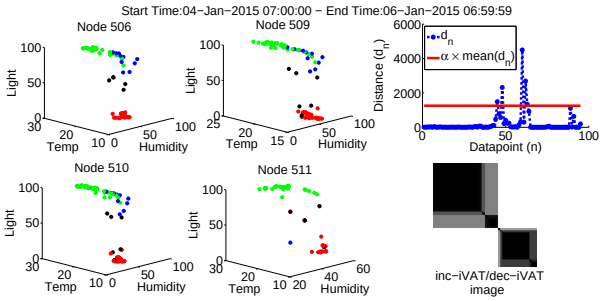


Fig. 7. The sensor data of temperature, humidity, and light from each sensor node (510 and 511). In addition, the figure shows the MST cut magnitude (d_n) and inc-iVAT/dec-iVAT shows the visual estimate of the number of clusters at each time step. For this graph, parameters used were $\alpha = 6$ and $\beta = 0.1$. Anomalous points are shown using dark dots, whereas clusters are depicted in red, green and blue colors.

C. Urban Mobility Patterns of Taxis in Singapore

Analyzing clusters is a fundamental challenge in trajectory mining; however, existing trajectory clustering algorithms are not appropriate for large numbers of trajectories in a city road network because of inadequate distance measures between two trajectories. We utilize GPS traces of 15,061 taxis within Singapore (equaling 3.28 million trajectories) gathered over one-month period. To cluster the origin and destination pairs of taxi rides, we use clusiVAT sampling and density-based spatial clustering of applications with noise (DBSCAN) [28] to provide useful insight into urban hot-spots, usage of road networks and crowd movements. For large numbers of overlapping trajectories, we use Dynamic Time Warping (DTW) coupled with Dijkstra distance measure (trajDTW) [23], [29]. For predicting trajectories, we utilize *Traj-clusiVAT* [30] that combines scalable clustering and Markov chain for predicting both short- and long-term trajectories. In addition, Traj-clusiVAT can determine the clusters representing diverse behaviors.

D. Detecting Anomalies in Mobility Patterns of Vehicles and Pedestrians

Knowing templates of pedestrian movement has many useful applications in managing pedestrian flows, maintaining public security and safety. We use iVAT+ and clusiVAT+ [31] for detecting anomalous pedestrian trajectories. These trajectories are classified as normal or abnormal depending on the number of trajectories in the clusters. Experiments on the vehicle and pedestrian trajectories from a parking space data set[1] showcases the ability of VAT-based approach in producing natural and informative trajectory clusters and finding representative anomalies.

E. Future Work

IoT is going to drive the increased use of connected devices for many applications. This results in generation of big data for

[1]http://www.ee.cuhk.edu.hk/ xgwang/MITtrajsingle.html

This article is accpted for publication. The final published version is available at https://doi.org/10.1109/MSMC.2019.2961160.

diverse applications and requires analyzing the data streams in real-time. From the literature and the work presented in this article, we can take cognizance of existing algorithms for clustering big data, assessing tendency of clusters and detecting anomalies from big data. However, with high-velocity streaming data generated by IoT devices, there are very limited algorithms that are: (1) suitable for extracting structure from streaming data, and (2) infer the exact number of clusters from the streaming data. Our future work includes designing algorithms that could not only cluster streaming big data, but also provide validation of clusters to handle streaming big data.

V. Conclusion

In this article, we presented an overview of how VAT-family techniques can be elegantly used to analyze the number of clusters present in the big data generated by IoT devices, even before we apply clustering algorithms. The article explores how James C. Bezdek's pioneered algorithms are effective in analyzing the IoT-generated big data. We presented four real-world IoT case studies (monitoring the impact on GBR, monitoring urban forest, understanding urban mobility patterns, and detecting anomalies in vehicles and pedestrians), wherein VAT techniques and their extensions were applied for solving key issues. These techniques primarily developed with James C. Bezdek are advancing the field of IoT for effective practical implementations.

References

[1] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of Things (IoT): A vision, architectural elements, and future directions," *Future generation computer systems*, vol. 29, no. 7, pp. 1645–1660, 2013.

[2] IHS, "IoT: number of connected devices worldwide 2012-2025," Nov 2016. [Online]. Available: https://www.statista.com/statistics/471264/iot-number-of-connected-devices-worldwide/

[3] B. Marr, "How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read," www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#7e687cf60ba9, 2018, [Online]; accessed June 5, 2019.

[4] M. Cox and D. Ellsworth, "Application-controlled demand paging for out-of-core visualization," in *Proceedings. Visualization'97 (Cat. No. 97CB36155)*. IEEE, 1997, pp. 235–244.

[5] D. Laney, "3D data management: Controlling data volume, velocity, and variety," META Group, Tech. Rep., February 2001.

[6] ——, "Deja VVVu: Others claiming gartner's construct for big data," https://blogs.gartner.com/doug-laney/deja-vvvue-others-claiming-gartners-volume-velocity-variety-construct-for-big-data/, January 2012.

[7] T. C. Havens, J. C. Bezdek, C. Leckie, L. O. Hall, and M. Palaniswami, "Fuzzy c-means algorithms for very large data," *IEEE Transactions on Fuzzy Systems*, vol. 20, no. 6, pp. 1130–1146, 2012.

[8] P. Barnaghi, A. Sheth, and C. Henson, "From data to actionable knowledge: big data challenges in the web of things," *IEEE Intelligent Systems*, no. 6, pp. 6–11, 2013.

[9] M. F. Uddin, N. Gupta *et al.*, "Seven v's of big data understanding big data to extract value," in *Proceedings of the 2014 Zone 1 Conference of the American Society for Engineering Education*. IEEE, 2014, pp. 1–5.

[10] J. C. Bezdek and R. J. Hathaway, "VAT: A tool for visual assessment of (cluster) tendency," in *Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No. 02CH37290)*, vol. 3. IEEE, 2002, pp. 2225–2230.

[11] T. C. Havens and J. C. Bezdek, "An efficient formulation of the improved visual assessment of cluster tendency (iVAT) algorithm," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 5, pp. 813–822, 2011.

[12] A. K. Jain, R. C. Dubes *et al.*, *Algorithms for clustering data*. Prentice hall Englewood Cliffs, 1988, vol. 6.

[13] B. S. Everitt, *Graphical techniques for multivariate data*. North-Holland, 1978.

[14] J. W. Tukey, *Exploratory Data Analysis: Limited Preliminary Ed*. Addison-Wesley Publishing Company, 1970.

[15] W. S. Cleveland, *Visualizing data*. Hobart Press, 1993.

[16] L. Wang, U. T. Nguyen, J. C. Bezdek, C. A. Leckie, and K. Ramamohanarao, "iVAT and aVAT: enhanced visual analysis for cluster tendency assessment," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2010, pp. 16–27.

[17] J. C. Dunn, "A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters," *Journal of Cybernetics*, vol. 3, no. 3, pp. 32–57, 1973.

[18] R. J. Hathaway, J. C. Bezdek, and J. M. Huband, "Scalable visual assessment of cluster tendency for large data sets," *Pattern Recognition*, vol. 39, no. 7, pp. 1315–1324, 2006.

[19] M. E. Johnson, L. M. Moore, and D. Ylvisaker, "Minimax and maximin distance designs," *Journal of statistical planning and inference*, vol. 26, no. 2, pp. 131–148, 1990.

[20] T. C. Havens, J. C. Bezdek, and M. Palaniswami, "Scalable single linkage hierarchical clustering for big data," in *2013 IEEE eighth international conference on intelligent sensors, sensor networks and information processing*. IEEE, 2013, pp. 396–401.

[21] D. Kumar, J. C. Bezdek, M. Palaniswami, S. Rajasegarar, C. Leckie, and T. C. Havens, "A hybrid approach to clustering in big data," *IEEE transactions on cybernetics*, vol. 46, no. 10, pp. 2372–2385, 2015.

[22] D. Kumar, M. Palaniswami, S. Rajasegarar, C. Leckie, J. C. Bezdek, and T. C. Havens, "clusiVAT: A mixed visual/numerical clustering algorithm for big data," in *2013 IEEE International Conference on Big Data*. IEEE, 2013, pp. 112–117.

[23] D. Kumar, H. Wu, S. Rajasegarar, C. Leckie, S. Krishnaswamy, and M. Palaniswami, "Fast and scalable big data trajectory clustering for understanding urban mobility," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 11, pp. 3709–3722, 2018.

[24] D. Kumar, J. C. Bezdek, S. Rajasegarar, M. Palaniswami, C. Leckie, J. Chan, and J. Gubbi, "Adaptive cluster tendency visualization and anomaly detection for streaming data," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 11, no. 2, p. 24, 2016.

[25] P. Rathore, D. Kumar, J. C. Bezdek, S. Rajasegarar, and M. Palaniswami, "A rapid hybrid clustering algorithm for large volumes of high dimensional data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 4, pp. 641–654, 2018.

[26] M. Palaniswami, A. S. Rao, and S. Bainbridge, "Real-time monitoring of the great barrier reef using internet of things with big data analytics," *ITU J.: ICT Discoveries*, vol. 1, no. 13, pp. 1–10, 2017.

[27] J. C. Bezdek, S. Rajasegarar, M. Moshtaghi, C. Leckie, M. Palaniswami, and T. C. Havens, "Anomaly detection in environmental monitoring networks [application notes]," *IEEE Computational Intelligence Magazine*, vol. 6, no. 2, pp. 52–58, 2011.

[28] D. Kumar, H. Wu, Y. Lu, S. Krishnaswamy, and M. Palaniswami, "Understanding urban mobility via taxi trip clustering," in *2016 17th IEEE International Conference on Mobile Data Management (MDM)*, vol. 1. IEEE, 2016, pp. 318–324.

[29] D. Kumar, S. Rajasegarar, M. Palaniswami, X. Wang, and C. Leckie, "A scalable framework for clustering vehicle trajectories in a dense road network," in *The 4th International Workshop on Urban Computing (UrbComp), Held in conjunction with the 21th ACM SIGKDD*, 2015, pp. 1–9.

[30] P. Rathore, D. Kumar, S. Rajasegarar, M. Palaniswami, and J. C. Bezdek, "A scalable framework for trajectory prediction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 10, pp. 3860–3874, 2019.

[31] D. Kumar, J. C. Bezdek, S. Rajasegarar, C. Leckie, and M. Palaniswami, "A visual-numeric approach to clustering and anomaly detection for trajectory data," *The Visual Computer*, vol. 33, no. 3, pp. 265–281, 2017.

This article is accpted for publication. The final published version is available at https://doi.org/10.1109/MSMC.2019.2961160.