2013 12th International Conference on Machine Learning and Applications

Zebrafish Larva Locomotor Activity Analysis Using Machine Learning Techniques

Hao Zhang*, Scott C. Lenaghan‡, Michelle H. Connolly§, and Lynne E. Parker*
* *Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville, TN 37996*
Email: {haozhang, leparker}@utk.edu
‡ *Department of Mechanical, Aerospace and Biomedical Engineering, University of Tennessee, Knoxville, TN 37996*
Email: slenagha@utk.edu
§ *Center for Environmental Biotechnology, University of Tennessee, Knoxville, TN 37996*
Email: mconnol3@utk.edu

***Abstract*—Zebrafish larvae have become a popular model organism to investigate genetic and environmental factors affecting behavior. However, difficulties exist in the analysis of complex behaviors from a large array of larvae. In this paper, we present the new application of machine learning techniques in bioinformatics to automatically detect and investigate the locomotor activities of zebrafish larvae. To achieve this, twelve features were defined and seven unsupervised learning methods were implemented. Next, seven performance measures were applied to evaluate and compare these methods. In order to empirically evaluate the machine learning algorithms, a large dataset was collected that contained 6847 valid instances. Using this dataset, the characteristics of the features were analyzed and the most appropriate unsupervised learning algorithm, i.e., Unweighted Pair Group Method with Arithmetic mean (UPGMA), for locomotor activity analysis was identified. In addition, UPGMA's ability to reveal underlying patterns of zebrafish locomotor activities was demonstrated. In general, this study shows that machine learning techniques have the potential to construct effective, high-throughput systems to automate the process of identifying zebrafish behaviors influenced by genetic manipulation, pharmaceuticals, and environmental toxins.**

***Keywords*-Bioinformatics, zebrafish larva, locomotor behavior analysis, machine learning**

I. Introduction

Bioinformatics is an interdisciplinary field of science in which biology, computer science, and information technology merge to develop computational methods of retrieving, organizing and analyzing biological data [1]. Machine learning techniques have been successfully applied in bioinformatics because of their ability to deal with the randomness and uncertainty of noisy biological data [1]. The objective of this work is to demonstrate the effectiveness of unsupervised machine learning techniques to automatically analyze the locomotor activities of zebrafish larvae. Through validation of the machine learning techniques, the ultimate goal will be to assist biomedical researchers in the analysis of complex behavioral changes induced by pharmaceutical chemicals, toxins, or genetic modification.

Zebrafish (*Danio rerio*) are small cyprinid fish that are native to the streams of southeast Asia. Zebrafish are among the most commonly-used vertebrate model organisms, and have been extensively used for drug discovery research and developmental genetic studies [2]. Recently, the use of zebrafish as a model for human diseases, such as cancer, Parkinsons disease, diabetes, and amyotrophic lateral sclerosis has become increasingly common [3]. The numerous reasons that zebrafish have emerged as a highly popular and attractive model organism have been extensively reviewed in [4]. Perhaps the most important justifications for the use of zebrafish as a model organism are the low cost, small size, and the rich repertoire of natural behaviors exhibited during larval development [5].

Of particular importance to this study, the behavior of zebrafish larvae is reflected by their locomotor activities. In the laboratory environment, zebrafish larvae are typically placed into multiwell plates for high throughput screening, as depicted in Figure 1. To analyze the locomotor activities of zebrafish larvae, a traditional approach adopted by biomedical researchers has been to directly watch hours of recorded experimental videos, manually label and track each zebrafish larva, and manually compute features that can represent the larva's locomotor activities [2], [3], [5]. However, this manual analysis is inherently inefficient and largely ineffective, and usually introduces observational biases. For example, hours or even days of videos are typically recorded during a single experiment, which generally contains hundreds of thousands of frames with each frame containing multiple larvae. Consequently, manual analysis is time-consuming, and usually impractical. In addition, manual approaches are very likely to miss zebrafish larva's locomotor activities and behaviors, given such a large amount of information.

Therefore, the development of an automated system will contribute to the improved analysis of locomotor behaviors in this model organism. Although several imaging systems have been implemented to detect and track zebrafish larvae [6], [7], to our knowledge, no other systems are capable of automated pattern detection. Our work is the first automated pattern detection system, which is currently being demonstrated as a proof of concept with zebrafish larvae but is also applicable to other model systems and applications.

To address this issue, several unsupervised machine learning techniques were implemented. Then, internal and stability evaluation metrics were used to select the best unsupervised learning approach, and demonstrate the effectiveness of this approach for the construction of an intelligent system

978-0-7695-5144-9/13 $31.00 © 2013 IEEE
DOI 10.1109/ICMLA.2013.35

IEEE computer society

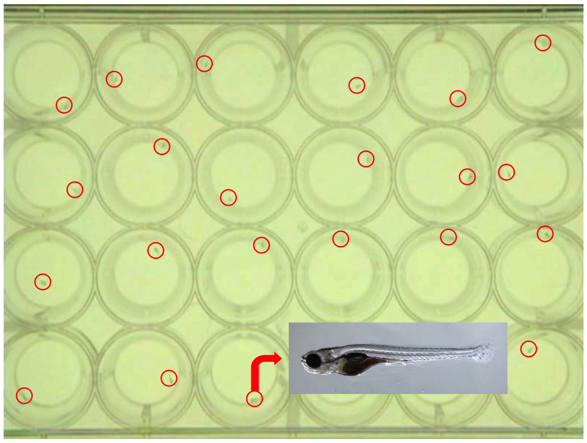


Figure 1: Multiwell plate and zebrafish larvae. Red circles represent positions of zebrafish larvae in the plate.

capable of automatically discovering underlying patterns in the locomotor activity of zebrafish larvae. The ultimate objective of this work is to use zebrafish larvae as a model organism, combined with machine learning techniques for data analysis, to investigate behavioral changes induced by pharmaceutical chemicals, toxins, or genetic manipulation.

II. Zebrafish Larva Perception and Feature Generation

In this section, the imaging system used in this work for detection and tracking of zebrafish larvae is described. Based on the visual data, such as recorded experimental videos or a sequence of online frames, the tracking results were obtained and a variety of features were generated. The hardware setup is demonstrated in Figure 2, which includes a video camera (e.g., a firewire camera or a webcam) that is installed on a tripod, a bottom lighting system (e.g., an LCD screen with a white background), a multiwell plate to contain zebrafish larvae, and a computer. The data processing procedure used to detect and track zebrafish larvae and generate features is described as follows.

A. Zebrafish Larva Perception

The input to the zebrafish perception system, i.e., detection and tracking, was a sequence of frames in the red-green-blue color space acquired using the hardware setup depicted in Figure 2. For example, Figure 1 shows an input frame, in which the positions of zebrafish larvae are marked with red circles. The output of the perception system is the trajectory of each larva, which is then used to generate features. Figure 3 provides an example of our perception system's output[1].

The zebrafish larva perception system contains two modules: detection and tracking. The detection module maintains the background model and uses a background subtraction technique to identify potential zebrafish larva candidates within a frame. The tracking module continuously tracks zebrafish larvae across the frame sequences using a global-local search. The global search employs color-based search techniques to locate regions that are most likely to contain zebrafish larva candidates. Then, the local search refines the tracking result of each candidate by removing noise and reflections of the larvae on the well wall in the local region.

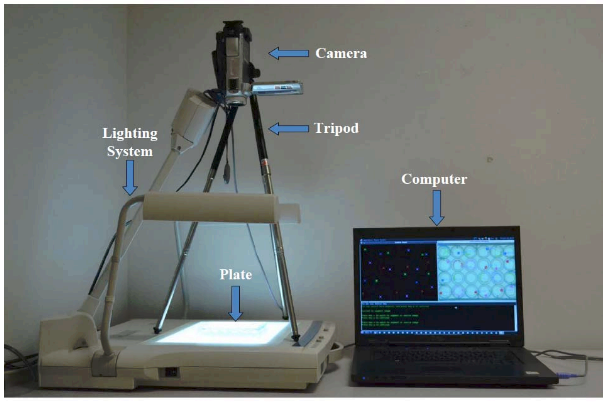


Figure 2: Imaging system's hardware setup.

B. Feature Generation

Twelve features with well-defined physical meaning have been identified for use by the machine learning techniques to automatically analyze the locomotor activity of the zebrafish larvae. The features used in this work are listed below:

- Total distance (pixels): the total distance that a zebrafish larva moves;
- Burst distance (pixels): the average distance that a larva moves within one burst;
- Deviation of burst distance (pixels): the standard deviation of the burst distance;
- Total absolute turning angle (°): the total absolute angle that a zebrafish larva turns;
- Burst turning angle (°): the average turning angle that a zebrafish larva turns within one burst;
- Deviation of burst turning angle (°): standard deviation of burst turning angle;
- Absolute burst turning angle (°): the average absolute turning angle that a larva turns within one burst;
- Deviation of absolute burst turning angle (°): standard deviation of absolute burst turning angle;
- Time in inner area (seconds): the amount of time that a zebrafish larva stays in the inner area;
- Moving distance in inner area (pixels): the total distance that a larva moves in the inner area;
- Freeze time (seconds): the amount of time that a larva stays stationary;
- Meandering (°/pixel): the degree of turning per pixel;

where a burst swim is a movement with a distance greater than $R/5$; the inner area is the circular area centered at the well center with a radius of $R/\sqrt{2}$; and R is the well radius.

[1] An illustrative video of our perception system's detection and tracking result is available at: http://www.youtube.com/watch?v=Gi7FiAXWTdY.

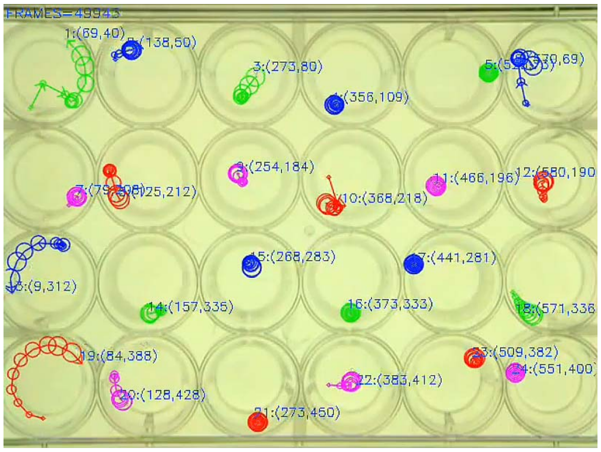


Figure 3: Trajectories of zebrafish larvae over ten seconds, where a larger circle represents a more recent position of a zebrafish larva, and its identity and position are displayed around its most recent position.

III. Unsupervised Machine Learning for Locomotor Activity Analysis

Unsupervised machine learning, also called clustering, is typically used to group together instances that are similar to one another in a multi-dimensional feature space. The goal of unsupervised learning is often to discover the underlying structure of the instance space. In order to analyze locomotor activity patterns of zebrafish larvae that move within circular wells of a multiwell plate, we discuss and implement seven unsupervised learning techniques. In addition, we use seven performance evaluation metrics to evaluate and compare the unsupervised learning techniques, and identify the algorithm that is the most appropriate for the bioinformatic application of recognizing the locomotor activity patterns of zebrafish larvae.

A. *Unsupervised Learning Algorithms*

Seven unsupervised learning algorithms from four different categories are implemented in this work:

- Partitioning unsupervised learning: K-means and Partitioning Around Medoids (PAM) are implemented;
- Hierarchical unsupervised learning: Unweighted Pair Group Method with Arithmetic mean (UPGMA), which is an agglomerative hierarchical method, and DIvisive ANAlysis (Diana) that is a divisive hierarchical method are implemented;
- Artificial neural network (ANN) based unsupervised learning: Self-Organizing Tree Algorithm (SOTA) and Self-Organizing Map (SOM) are implemented;
- Model-based unsupervised learning: Mixture Of Gaussian (MOG) is implemented.

These unsupervised learning algorithms are briefly described as follows[2]:

1) K-means is an iterative method which minimizes the within-class distances for a given number of clusters. The algorithm starts with an initial guess for the cluster centers, and each instance is placed in its nearest cluster. Then, the cluster centers are updated, and the entire process is repeated until the cluster centers no longer move.

2) PAM searches for K representative medoids among the instances. PAM is initialized by randomly selecting K instances as the medoids. Then, PAM iteratively assigns the instances to the nearest medoid, and updates the medoids to minimize the sum of the distance of the instances.

3) UPGMA initially treats each instance as a cluster. Then, instances are joined together to form new clusters according to their closeness that is determined by similarity measures. UPGMA yields a dendogram, which can be cut at a chosen height to produce the desired number of clusters.

4) Diana is a hierarchical unsupervised learning algorithm that initially starts with all instances in a single cluster. Then, it iteratively divides the clusters until each cluster contains a single instance.

5) SOM is based on the competitive unsupervised neural network. SOM initially populates nodes by randomly sampling instances. Then, it changes the weights in a systematic way that captures the distribution of the instances' variability. Each output node represents the average pattern of the instances that map into it.

6) SOTA is an unsupervised neural network with a divisive hierarchical binary tree structure. It offers a criterion to stop the growing of the tree based on the approximate distribution of the probability obtained by randomization of the original dataset, and therefore provides a statistical support for the cluster definition.

7) MOG is a probabilistic model that consists of a finite mixture of Gaussian distributions. Each mixture component represents a cluster, and the mixture components and group memberships are estimated using maximum likelihood.

B. *Performance Evaluation Metrics*

To select the most appropriate algorithm among the implemented unsupervised learning techniques for our bioinformatics application, we implement both internal and stability validation metrics [9]. Internal validation metrics take the instances and their partitions as input, and use intrinsic information contained in the instances to evaluate the clustering performance. On the other hand, stability validation metrics evaluate the consistency of clustering results by comparing the results with the partitions obtained using feature vectors with some attributes removed. In this paper, three internal validation metrics, i.e., Connectivity, Silhouette, and Dunn Index, and four stability validation metrics, i.e., Average Proportion of Non-overlap (APN), Average Distance (AD), Average Distance between Means (ADM), and Figure Of Merit (FOM), are implemented and described as follows[3]:

[2] Refer to [8] for detailed explanations of these techniques.

[3] Refer to [9] for detailed explanations of these validation metrics.

1) Connectivity measures the connectedness, which indicates to what extent instances are placed in the same cluster as their nearest neighbors. The value of connectivity is in the range $[0, 1]$, and a lower value indicates better performance.

2) Silhouette Width is the average value of all instances' silhouette values. The silhouette value measures the degree of confidence in the clustering assignment of an instance. The value of silhouette width falls in the range $[-1, 1]$ with a greater value indicating better performance.

3) Dunn Index is the ratio of the smallest distance of the instances not in the same cluster to the largest intra-cluster distance. Dunn Index lies in $[0, \infty]$. A greater value indicates better performance.

4) APN measures the average proportion of the instances not placed in the same cluster, by comparing the clustering results using the features with all attributes and features with one attribute removed. The value of APN lies in $[0, 1]$, and a smaller value indicates a higher clustering consistency.

5) AD measures the average distance of the instances that are not placed in the same cluster, by comparing the results using the features with all attributes and the features with one attribute removed. The value of AD lies in $[0, \infty)$, and a smaller value indicates a stronger consistency.

6) ADM measures the average distance of the means of all instances placed in the same cluster, by comparing the results using the features with all attributes and features with one attribute removed. ADM lies in $[0, \infty)$. A smaller value indicates a stronger consistency.

7) FOM evaluates the average intra-cluster variance of the instances using the deleted attributes. The clusters are obtained using the features with undeleted attributes. FOM takes values in the range $[0, \infty)$ and a smaller value indicates a stronger consistency.

IV. Empirical Study

The hardware setup in Figure 2 was used to collect data, where a 24-well plate with a radius of $R \doteq 8$ millimeters (around 26 pixels in a frame) was used to house the larvae, and a computer with 2GB memory and a 2.4 GHz dual core CPU was used to process the data. Approximately 100 hours of video was collected over the course of the study, with a resolution of 640×480 and a frame rate of 10 frames-per-second. Three-hour datasets, i.e., one-hour video each day from the 4 to 6 days post zebrafish fertilization, were used to analyze the locomotor activities of the zebrafish larvae. To obtain testing instances, the videos were temporarily segmented into short segments, each of which contained 52 frames and was 5.2 seconds in length. Since a 24-well plate was used, as depicted in Figure 1, a total of 49,824 instances were obtained. However, since the larvae remained still for the majority of the instances, these instances were removed, which leads to the final dataset that contains 6847 instances.

For each instance, the imaging system generates a feature vector that contains twelve features as described in Section II-B. Using the generated dataset, the statistical characteristics were analyzed, i.e., the histogram and distribution, of each feature, as graphically presented in Figure 4. Several phenomenon were observed from this figure. First, Figures 4a, 4b and 4k indicate that most zebrafish larvae travel a very short distance within the 5.2 second time period. Second, it was rare for a zebrafish larva to turn with a large angle, even when performing a burst swim, as indicated by Figures 4e, 4f, 4g, 4h and 4l. Third, zebrafish larvae tend to stay closer to the well walls instead of staying in the open area in the middle of the well, as presented by Figures 4i and 4j. Fourth, most of the feature's distribution has a single peak, indicating most of the zebrafish larva follow a certain pattern that represents their normal locomotor activity. On the other hand, all distributions have a long tail, indicating there exist different locomotor activity patterns.

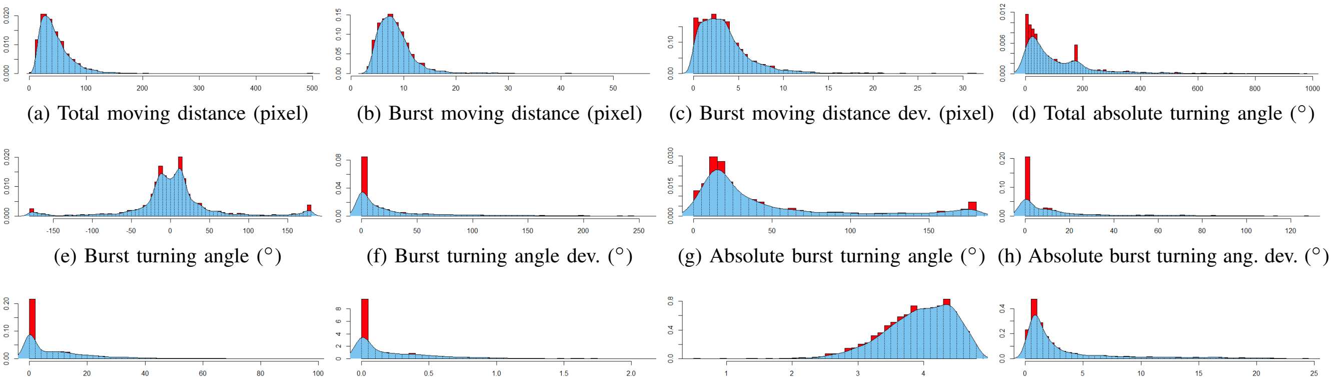
Table I: Top ranked unsupervised learning algorithms. The results are denoted using the format: *algorithm-#clusters*.

Performance metric	Rank 1	Rank 2	Rank 3
Connectivity	UPGMA-2	UPGMA-3	UPGMA-4
Dunn Index	UPGMA-4	UPGMA-5	UPGMA-6
Silhouette Width	UPGMA-2	UPGMA-3	Diana-2
APN	UPGMA-2	UPGMA-3	Diana-2
AD	PAM-8	SOM-8	PAM-7
ADM	MOG-2	UPGMA-2	UPGMA-3
FOM	MOG-8	PAM-8	PAM-7

After re-scaling for comparability, the features were used as the input to our machine learning algorithms for locomotor activity analysis. In the experiments, the Euclidean distance was used as the similarity measure, and the average-linkage was adopted as the linkage rule, which defines the cluster distance as the average distance between all pairs of instances belonging to different clusters.

The machine learning algorithms' performances based on internal validation measures are depicted in Figure 5. It is observed that the ranks of these algorithms are consistent across different internal validation measures, and UPGMA performs consistently better than the other algorithms. In addition, using three to six clusters, UPGMA obtains good performance. The experimental results using stability validation measures are illustrated in Figure 6. It is observed that different stability measures generally suggest conflicting conclusions. APN and ADM show that increasing the number of clusters generally reduces the stability. On the other hand, AD and FOM show an opposite trend that increasing the cluster number improves the stability performance. The top three machine learning algorithms are listed in Table I. No single algorithm achieves the best results for all measures and the rank varies significantly across different metrics.

In order to reconcile different ranks to produce the final rank, rank aggregation [10] was applied. Rank aggregation is capable of ranking the unsupervised learning algorithms through simultaneously incorporating evaluation results produced by all performance metrics. In our experiments, rank


(a) Total moving distance (pixel) (b) Burst moving distance (pixel) (c) Burst moving distance dev. (pixel) (d) Total absolute turning angle (°)

(e) Burst turning angle (°) (f) Burst turning angle dev. (°) (g) Absolute burst turning angle (°) (h) Absolute burst turning ang. dev. (°)

(i) Moving distance in inner area (pixel) (j) Time in inner area (second) (k) Freezing time (second) (l) Meandering (°/pixel)

Figure 4: Histograms and distributions of the features obtained by our imaging system. Histograms are denoted with red rectangles and probability distributions are depicted by the upper boundary of the blue area.

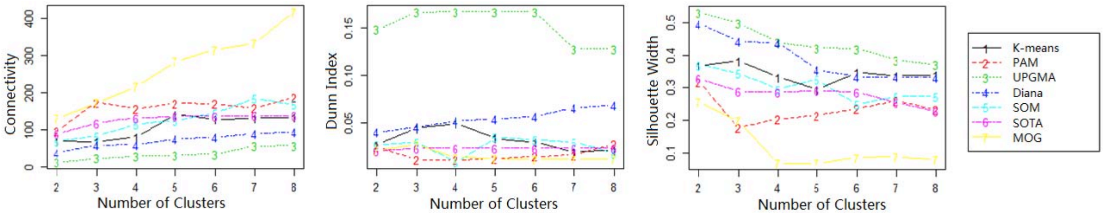


Figure 5: Internal evaluation results of the unsupervised learning algorithms over our zebrafish larva locomotor dataset.

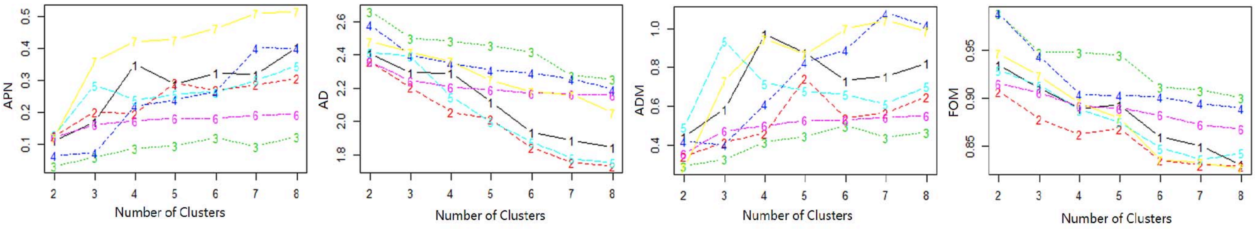


Figure 6: Stability evaluation results of the unsupervised learning algorithms. The legend is the same as in Figure 5.

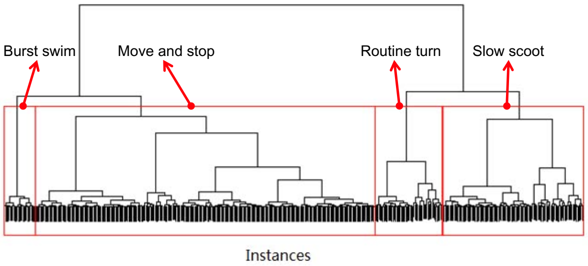
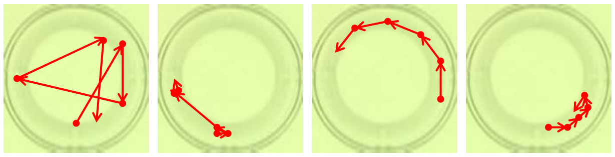


Figure 7: Dendrogram obtained by UPGMA. The instances in each red rectangle belong to the same locomotor activity.

aggregation was performed using the cross-entropy method with the weighted Spearman's footrule [10] to generate a list of the top three algorithms containing more than three clusters, as follows: UPGMA-4, UPGMA-3, and PAM-8. Given this rank, we select UPGMA with four clusters as our model to analyze zebrafish larva's locomotor activities. The dendrogram created by the algorithm is shown in Figure 7. Through checking the instance videos, each cluster was associated with its physical locomotor activity, which is marked in Figure 7. In order to intuitively visualize the identified zebrafish larva's locomotor activities, an illustrative example of each activity is depicted in Figure 8.

From Figure 7, it is observed that the identified groups are imbalanced, i.e., more than half of the instances are assigned to "move and stop", indicating that this activity is a common locomotor behavior. In addition, it is observed that the UPGMA algorithm is capable of recognizing small clusters (e.g., "burst swim"), which is helpful for the discovery of rare or abnormal locomotor activities. We believe that UPGMA is the most appropriate algorithm for this bioinformatic application also due to the following reasons. First, UPGMA outputs a hierarchy; as a result, users can obtain a desired number of clusters by controlling the cutting



(a) Burst swim (b) Move and stop (c) Routine turn (d) Slow scoot

Figure 8: Examples of zebrafish larva's locomotor activities.

level, which is very helpful for analyzing hierarchical activities with different granularity. Second, UPGMA is a nonparametric unsupervised learning technique that requires minimal manual tuning. Third, intuitive result visualizations can be integrated into UPGMA, as shown in Figure 7.

V. Related Work

In recent years, a very large number of studies have been conducted to investigate the behavior of zebrafish larvae, including spontaneous locomotor activities [11] and locomotor activity alterations induced by pharmaceutical chemicals [2], lighting changes, and other factors [12]. However, since these works were based on manually tracking zebrafish larva and analyzing their locomotor activities, they can only be applied to small datasets, and it is generally impractical to use these methods to process large datasets.

To automate the process of zebrafish larva detection and tracking, several imaging systems have recently been developed, including commercial systems, such as DanioVision [6] and ZebraLab [7], and some systems for research purposes [13]. However, the major objective, especially for the commercial systems, is to track zebrafish larvae, and these systems only provide distance or velocity as features, which are not informative enough to analyze complex behaviors.

Although some basic statistical analysis can be automated by using Microsoft Excel [13], no previous study has been conducted to investigate the effectiveness and efficiency of the use of machine learning techniques, especially unsupervised learning, to automatically analyze underlying patterns of the locomotor activity of zebrafish larvae.

VI. Summary and Conclusion

In this paper, we defined twelve features that can encode the characteristics of zebrafish larva's locomotor activities, based on the tracking results obtained from our imaging system. In addition, using these features as inputs, we implemented seven unsupervised machine learning techniques, and discussed their capability for investigation of the locomotor activity of zebrafish larvae. Seven evaluation metrics were then used to identify the most appropriate technique, i.e., the UPGMA algorithm, for locomotor activity analysis.

In order to empirically validate our findings, we collected a large dataset consisting of 6847 valid instances that contain discrete zebrafish larvae movements. In our experiment, after statistically analyzing the characteristics of the features, we validated the effectiveness of the UPGMA algorithm to automatically discover patterns of locomotor activities directly from the data. This study demonstrates that machine learning techniques have significant potential for the construction of automated, high-throughput systems to analyze biological datasets for applications in bioinformatics.

References

[1] P. Larranaga, B. Calvo, R. Santana, C. Bielza, J. Galdiano, I. Inza, J. A. Lozano, R. Armananzas, G. Santafe, A. Perez, and V. Robles, "Machine learning in bioinformatics," *Briefings in Bioinformatics*, vol. 7, no. 1, pp. 86–112, Mar. 2006.

[2] D. Kokel and R. T. Peterson, "Chemobehavioural phenomics and behaviour-based psychiatric drug discovery in the zebrafish," *Briefings in functional genomics proteomics*, vol. 7, no. 6, pp. 483–490, 2008.

[3] T. Ramesh, A. Lyon, R. Pineda, C. Wang, P. Janssen, B. Canan, A. Burghes, and C. Beattie, "A genetic model of amyotrophic lateral sclerosis in zebrafish displays phenotypic hallmarks of motoneuron disease," *Disease Models and Mechanisms*, vol. 3, no. 9–10, pp. 652–662, Sept. 2010.

[4] M. J. Winter, W. S. Redfern, A. J. Hayfield, S. F. Owen, J.-P. Valentin, and T. H. Hutchinson, "Validation of a larval zebrafish locomotor assay for assessing the seizure liability of early-stage development drugs," *Journal of Pharmacological and Toxicological Methods*, vol. 57, no. 3, pp. 176–187, 2008.

[5] R. Spence, G. Gerlach, C. Lawrence, and C. Smith, "The behaviour and ecology of the zebrafish, Danio rerio," *Biological Reviews*, vol. 83, no. 1, pp. 13–34, 2008.

[6] Noldus Information Technology, "DanioVision," www.noldus.com/animal-behavior-research/products/daniovision, accessed: 2013-06-07.

[7] Viewpoint Lifesciences, "ZebraLab," www.vplsi.com/content.php?content.72, accessed: 2013-06-07.

[8] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM Computing Surveys*, vol. 31, no. 3, pp. 264–323, Sept. 1999.

[9] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "On clustering validation techniques," *Journal of Intelligent Information Systems*, vol. 17, no. 2–3, pp. 107–145, Dec. 2001.

[10] V. Pihur, S. Datta, and S. Datta, "Weighted rank aggregation of cluster validation measures: a Monte Carlo cross-entropy approach," *Bioinformatics*, vol. 23, no. 13, pp. 1607–1615, Jul. 2007.

[11] R. M. Colwill and R. Creton, "Locomotor behaviors in zebrafish (danio rerio) larvae," *Behavioural Processes*, vol. 86, no. 2, pp. 222–229, 2011.

[12] S. Padilla, D. Hunter, B. Padnos, S. Frady, and R. MacPhail, "Assessing locomotor activity in larval zebrafish: Influence of extrinsic and intrinsic variables," *Neurotoxicology and Teratology*, vol. 33, no. 6, pp. 624–630, 2011.

[13] S. Xia, Y. Zhu, X. Xu, and W. Xia, "Computational techniques in zebrafish image processing and analysis," *Journal of Neuroscience Methods*, vol. 213, no. 1, pp. 6–13, 2013.