Journal of Applied Sciences

ISSN 1812-5654

Journal of Applied Sciences 7 (18): 2656-2660, 2007
ISSN 1812-5654

Application of Vertical Cluster Analysis Method to the Analysis of Time Dependent Biological Data Sets

[1]Sathees B.C. Chandra and [2]Yao Wang
[1]Department of Biological, Chemical and Physical Sciences,
[2]Department of Mathematics and Actuarial Sciences, Roosevelt University, Chicago,
430 S Michigan Avenue, IL 60605, USA

Abstract: The purpose of this study is to apply vertical cluster analysis method to interpret and analyze habituation of the leg movement response, to different odors, in fruit flies. In most cases cluster analysis methods are used to analyze data sets, which can be classified into categories. We define this type of method as horizontal cluster analysis method. In this study, instead of dividing the data into categories, we divide the data based on different periods of time. We define this method as a vertical cluster analysis method. Here we apply vertical cluster analysis method to evaluate the habituation of leg movement responses of fruit fly, *Drosophila melanogaster*. The vertical cluster analyses helped us to identify hidden features of fruit fly behavior.

Key words: Time dependent variables, vertical cluster analysis method, habituation, fruit flies

INTRODUCTION

Animals have evolved with specialized sense organs to recognize sensory information in the environment for the evaluation of food, to detect predators and mates. Some of the most valuable insights into chemosensory coding have been derived from studies from the olfactory reception in mammals and insects (Hildebrand and Shepherd, 1997; Strausfeld and Hildebrand, 1999; Chandra and Singh, 2005). An important step in understanding this coding scheme is to behaviorally quantify the degree to which sets of odors are discriminable. We recently evaluated the odor discriminability in the fruit fly, *Drosophila melanogaster*, by first conditioning individual flies to not respond to any of several odorants using a nonassociative conditioning protocol (habituation) (Chandra and Singh, 2005). In that study, we analyzed the habitation response of flies to different odors using a variety of statistical analysis. However, those methods did not elucidate important features of insect behaviors. In this paper we re-analyze a part of data sets using cluster analysis to bring out hidden and valuable features of the fruit fly behaviors.

Cluster analysis is a class of statistical techniques, encompasses a number of algorithms, which can be applied to data that exhibit natural groupings. Cluster analysis sorts through the raw data and groups them into clusters (Kaufman and Rousseeuw, 1990). A cluster is a group of relatively homogeneous cases or observations. Objects in a cluster are similar to each other. They are also dissimilar to objects outside the cluster, particularly objects in other clusters. All clustering algorithms use standard procedures for searching through the set of all the possible clustering to find one that fits the data reasonably well.

On the other hand, the categorical variables can be classified as either nominal or ordinal variable. Categorical variables having unordered scales are called nominal variables (Lioyd, 1999; Agresti, 1996). Categorical variables having ordered scales are called ordinal variables (Agresti, 1994; Clogg and Shihadeh, 1994). However, there are some data sets that vary along with time; here we refer to these types of data sets as time dependent variables or time dependent data sets. The time dependent variable is a special case of ordinal variable. In this research we apply the vertical cluster analysis method to time dependent data sets, to classify the data into different time periods, to further understand the behavior of fruit flies.

Here, we re-analyze, interpret and make an effort to bring out hidden features of habituation of the leg movement response to odor data sets for fruit flies by using vertical cluster analysis method. This kind of analysis has never been done in this organism and it will

Corresponding Author: Sathees B.C. Chandra, Department of Biological, Chemical and Physical Sciences, Roosevelt University, 430 S Michigan Avenue, Chicago, IL 60605 Tel: 847-619-7968 Fax: 847-619-8555

give us the comprehensive effect of odor at three different time periods in the data set. Further more, it is critical to understand the effect of odor at each of these periods from the viewpoint of behavioral ecology and neurobiology.

MATERIALS AND METHODS

This study was conducted at the Ohio State University, Columbus, OH and USA. The subjects used were male and female *Drosophila melanogaster*. All flies were cultured on a standard cornmeal medium in a half-pint milk bottles and maintained on a 16 h: 8 h light/dark cycle at 25°C and 50% relative humidity. Groups of 20-25 flies were transferred everyday to an empty vial, where they were anaesthetized by cooling in a freezer (-20°C) for approximately 1 min. Each fly was mounted in a 200 μL micropipette tip in such way that its head and forelegs protruded through the small hole at one end, which was widened to approximately of 0.1 mm inner diameter (Vargo and Hirsch, 1982). The wide end of the micropipette tip was closed with modeling clay, which prevented the fly from escaping and provided flexibility for positioning the flies in the training arena. Mounting of flies was done between 10.00 am and 1.00 pm. Experiments were started 30 min later.

Each fly was conditioned for movement of its first pair of legs using odor as the Conditioned Stimulus (CS). We used geraniol, 1-hexanol or blank as the CS depending on the experiment. A conditioning trial began after we placed a subject into a conditioning arena through which air was constantly drawn over the subject's antennae and into an exhaust vent. The exhaust ensured that odor was removed quickly from the arena. Odor delivery began approximately 30 sec after a subject was placed into the arena and it lasted 4 sec. Odorant delivery was accomplished by activation of a valve that was controlled by a parallel port on a computer. When it was activated, the valve carried air through a 1 cc glass syringe that contained 3 μL of pure odorant applied to small strip of filter paper. Each trial lasted for 60 sec, after which the next subject was placed into the conditioning arena and the procedure was repeated.

Response measures: The response of flies to odor was categorized into three levels based on their movement of their first pair of legs. Flies normally move their legs intermittently. Generally, the CS was presented when the fly was motionless. Most flies either did not respond or paused a response during the 4 sec presentation of CS. Flies typically initiated leg movement after termination of the CS. We recorded the response as 0 if a fly failed to start moving its legs within 3 sec after termination of the CS. We recorded the response as 1 if a fly started moving its legs within 3 seconds after termination of the CS and continued to move its legs up to but not exceeding 9 sec. We recorded the response as 2 if the fly continued to move its legs for more than 9 sec after the presentation of CS. The latter response is the most vigorous response, which is on average the response from flies when first presented with a novel odor.

Habituation to odors: The goal of this experiment was to demonstrate habituation of the leg movement response to odor. In this experiment, flies were exposed to 25 trials of odor presentation as described above. Three different treatment groups were used and each group corresponded to different odor treatments: 1-hexanol, geraniol and air alone. Odor treatments were presented in an air background. By comparison of the response to odor+air to air alone we could evaluate whether any observed habituation was to the odor, to the mechanosenosry stimulation provided by air, or to a combination of both. The response of each fly was recorded on each trial as described above. We conditioned five flies per day and therefore maintained an inter-trial interval of 5 min. Twenty animals were conditioned for each treatment group.

Cluster analysis method: The statistical derivation for the above experiment is delineated below. For nominal variables, the order of listing of the categories is irrelevant and the statistical analysis should not depend on that ordering. Methods designed for ordinal variables utilize the category ordering. In order to group the objects, the term distance was first defined mathematically. Then according to the distances between every two objects with a certain criterion, the two most similar objects were classified into one category.

A valid definition of distance should satisfy following conditions: Let Ω denote the set containing all elements. The domain and range of the distances are non-negative.

- $d(x,y) \geq 0, \forall x,y \in \Omega$;
- $d(x,y) = 0$, if and only if $w = y$;
- $d(x,y) = d(y,x), \forall x,y \in \Omega$;
- $d(x,y) \leq d(x,z)+d(y,z), \forall x,y,z \in \Omega$;

The definition of the distance between two objects reflects their overall similarity. For example, if there are n observations, the algorithm is as follows (Fisher, 1969):

- At first, each observation combines one category. Group the two closest observations into one category through comparing the distances, thus we have (n-1) categories.
- Get the gravity of the newly created category and regarded as a new observation. At the same time, ignore all observations in the category.
- Repeat step 1 and 2, until appropriate categories are left.

For our data sets we introduced a new criterion, $d^*_{i,j}$. The definition is as follows:
Let $\{(x_{11}, x_{12}, ..., x_{1p}), (x_{21}, x_{22}, ...x_{2p}), ...(x_{n1}, x_{n2}, ..., x_{np})\}$ denote the sample points. In this case, we have n observations, each of which has p measured values. For our data set, p is 3 and n is 25.

We defined the square of the distance between objects i and j:

$$d^*_{i,j} = \sum_{\beta=i}^{j}\sum_{\alpha=1}^{p}(x_{\beta,\alpha} - \bar{x}_\alpha(i,j))^2$$

where:

$$\bar{x}_\alpha(i,j) = \frac{1}{j-i+1}\sum_{\beta=i}^{j} x_{\beta,\alpha}, 1 \le i \le n, i \le j$$

Here, $x_{\beta,\alpha}$ is the value of the α-th variable of the β-th observation. The order of i and the order of j cannot be changed.

The objects should be sorted. The smaller $d^*_{i,j}$, the closer two observations are within one category. The best partition should minimize the sum of $d^*_{i,j}$. Therefore, we use the sum of as the measurement. We need to get values of all possible pairs of i and j. After we get all $d^*_{i,j}$, it is possible to compare the sum of the possible $d^*_{i,j}$.

For example, we need to group n observations into 3 categories. The first category is from O_1 to O_{m1}, the second from O_{m1+1} to O_{m2} and the last from O_{m2+1} to O_n. Then our solution must satisfy $(d^*_{1,m1} + d^*_{m1+1,m2} + d^*_{m2+1,n})$.

This is the only way we can calculate all $d^*_{i,j}$ and possible sums and then compare the sums to get the minimum one. A few books and papers mention the cluster analysis of ordinal variables. Cluster analysis of ordinal variables is not included in most of the popular statistics packages. In our case, a self-developed package is used to calculate the distances and automatically compare them and find the minimum sum of them.

To determine the most appropriate number of categories, the algorithm below was followed:

- Decide the possible numbers of categories, k.
- Get all $d^*_{i,j}$ and all-possible sums of all appropriate combinations of $d^*_{i,j}$. We should always take care of some consecutive numbers together and keep their orders for the data sets of time dependent variables.
- Compare all the sums and get the minimum one as the solution.
- Compare the results of different k; determine the most appropriate number of categories.

In this data set, we had three groups and each group had 25 test results following the order of the time, i.e., Present data is time dependent data. This goal is not to divide data sets into categories, but to divide data sets into different time periods. We define the method to divide data set into categories as the horizontal cluster analysis method and we define the method to divide data sets into different time periods as the vertical cluster analysis method.

RESULTS AND DISCUSSION

Each fly had a response of 0 or 1 or 2 in each trial. Each point at each trial number (from 1 to 25) shown in the Fig. 1 represents an average response of 14 flies to one of the three stimuli Hexanal or Geraniol or air. There was a significant effect of stimulus ($df = 2$; $F = 22.24$; $p<0.001$). Flies showed more habituation to an air stimulus than to air that contained either of the two odor stimuli. This effect is most evident after 17, when response levels of flies tested with air alone were consistently lower than either group exposed to air that contained odor. There was also a significant effect of trial ($df = 24$, $F = 9.01$, $p<0.001$), which reveal habituation. In the first trials flies showed on average significantly higher levels of response to all stimuli than in later trials (Fig. 1). This habituation across trials was evident to all three stimuli, because the interaction between stimulus and trial was not significant ($df = 48$; $F = 0.65$; ns). The complete description of our results can be found in our previous article (Chandra and Singh, 2005).

By using the vertical cluster analysis method, we tried different numbers of categories. In the end, we found that three categories fit the data better than any other number of categories. We divided all test results of three groups into three periods. The boundary points of these three periods were 4 and 13, i.e., the first period is from 1 to 4, the second period was from 5 to 13 and the third periods was from 14 to 25.

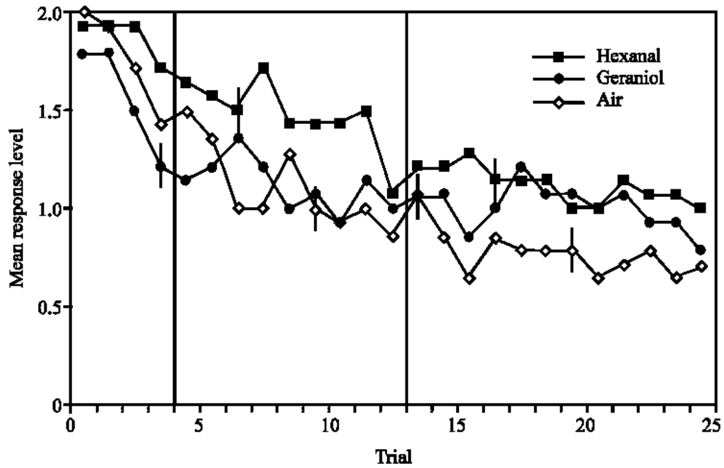

Fig. 1: Habituation of fruit flies to air or to air containing geraniol or hexanal. Three groups of restrained flies were exposed to 25 trials (x-axis) during which they were exposed to a stimulus for 4 sec every 5 min (inter-stimulus interval). Leg movement behavior was scored as 0, 1 or 2 based on criteria in the text and the mean response level (y-axis) is shown as a function of trial. Vertical bars at trials 10 and 20 represent standard errors for the group exposed to air (N = 14). Vertical bars at trials 4 and 14 represent standard errors for the group exposed to Geraniol (N = 14). Vertical bars at the trials 7 and 17 represent standard errors for the group exposed to Hexanal (N= 14)

We have demonstrated generalization of habituation to odor in the fruit fly. We found that fruit flies show strong unconditioned responses to stimulation either with air or with air that carries odor. Although the overall conclusion didn't differ much from our previous analyses (Chandra and Singh, 2005), the vertical cluster analysis method certainly delineated the intricate details of each time period to a greater extent. This kind of details, in most cases, will significantly help us better understand the biological significance of a particular animal behavior.

To apply cluster analysis to this kind of data set, we had to arrive at the appropriate number of categories. For our data set, we started with the following numbers of categories: 2, 3, 4 and 5. The boundary point of the two periods was 13. The boundary points of the three periods were 4 and 13. The boundary points of the four periods were 4, 7 and 13. The boundary points of the five periods are 4, 7, 10 and 13. We compared the behavior of flies during these different time periods. We then compared the results of all our categories to find the appropriate number of categories in order to find simple and valuable structures of the data set. Based on this preliminary analysis, we decided that three periods would fit the data better than any other periods.

The clustering is not useful until the ways of similarity and difference are made clear. Let us summarize the similarity and difference among these three periods. In the first period, fruit flies responded less to air and Geraniol. The response to air and Geraniol went down by a factor of 8, however Hexanal drops down only by 3. The response to the Geraniol was the lowest in the first class period. In the second period, fruit flies responses to air and Geraniol intersects four times, suggesting pretty much equal response to both odors. As a result the response to both these stimuli was not significantly different in this period. However, in this period the response to Hexanal reduced considerably compare to the first period. There was a slight reduction in the response to both Geraniol and air stimulus compared to the first period. It means that fruit flies are getting habituated to air and Geraniol stimulus much faster than to the Hexanal. In the third period, there was not much difference in the slope of response to all three stimuli. However, the response to air stimulus was lower than that of geraniol and hexanal. In addition, the response to hexanal was slighter higher than that of Genraniol, although it's not significant. This indicates that fruit flies are able to discriminate between air versus other important stimulus.

The next obvious question is why chose three clusters solution rather than any other number of categories. To arrive at this conclusion, we first compared the three clusters solution with the two clusters solution. For the two clusters solution, the boundary point of the two periods was 13. The first period was from 1 to 13 and the second period was from 14 to 25. For the two solutions, we realized that one can not get the information that indicates different character between the part from 1 to 4 and the part from 5 to 13. In the part from 5 to 13, the response to air and Geraniol intersects four times. However, we failed to find any intersect between the response to air and Geraniol in the part from 1 to 4. In the part from 5 to 13 the response to Hexanal reduces considerably compare to the part from 1 to 4. There is a slight reduction in the response to both Geraniol and air stimulus in the part from 5 to 13 compare to the part from 1 to 4. It was fairly apparent that we would lose lots of information by applying the two clusters solution. However, the three clusters solution rectifies some of these drawbacks. More over, when you compare the three clusters solution with the four clusters solution, for the four clusters solution, the boundary points are 4, 7 and 13. However, here again we failed to find the significant difference between the second period (from 5 to 8) and the third period (from 9 to 13). This indicates that we cannot get significantly more information from the four clusters solution comparing it with the three clusters solution. Similarly, we arrived at the same conclusion by comparing the three clusters solution with the five clusters solution. The simple three-cluster solution with enough information is better than the complicated four clusters solution or the five clusters solution.

In conclusion, the vertical cluster analyses clearly help us to understand similarity and variation of habituation responses of fruit flies with more depth than any other methods that has been used so far. This method not only gives us the over view of the all the three period of responses, but also provides us with detailed intricate activities of fruit flies within any one of three periods. Vertical cluster analysis can be an excellent tool in understanding and analyzing biologically important time dependent data sets.

ACKNOWLEDGMENTS

Authors would like to thank, Chun Xiao, graduate student at Roosevelt University, for writing up a cluster analysis program. We would also like to thank Sandeep Singh for helping us in collecting the fruit fly data.

REFERENCES

Agresti, A., 1994. Analysis of Ordinal Categorical Data. John Wiley and Sons.

Agresti, A., 1996. Introduction to Categorical Data Analysis. John Wiley and Sons.

Chandra, S.B.C. and S. Singh, 2005. Chemosensory processing in the fruit fly, *Drosophila melanogaster*: Generalization of feeding response reveals overlapping odor representations. J. Biosci., 30: 679-688.

Clog, C.C. and E.S. Shihadeh, 1994. Statistical Models for Ordinal Variables. Sage Publications.

Fisher, W.D., 1969. Clustering and Aggregation in Economics. The Johns Hopkins Press.

Hildebrand, J.G. and G.M. Shepherd, 1997. Mechanisms of olfactory discrimination: Converging evidence for common principles across phyla. Annu. Rev. Neurosci., 20: 595-631.

Lioyd, C.J., 1999. Statistical Analysis of Categorical Data. John Wiley and Sons.

Kaufman, L. and P.J. Rousseeuw, 1990. Finding Groups in Data: An Introduction to Cluster Analysis. Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics, New York.

Strausfeld, N.J. and J.G. Hildebrand, 1999. Olfactory systems: Common design, uncommon origins. Curr. Opin. Neurobiol., 9: 643-639.

Vargo, M. and J. Hirsch, 1982. Central excitation in the fruit fly (*Drosophila melanogaster*). J. Comp. Physiol. Psychol., 96: 452-459.