Building Concepts for AI Agents using Information Theoretic Co-clustering

Jason R Chen
Dept. of Engineering, College of Engineering and Computer Science
Australian National University, Canberra, ACT, Australia 0200
Email: jason.chen@anu.edu.au

***Abstract*—High level conceptual thought seems to be at the basis of the impressive human cognitive ability, and AI researchers aim to replicate this ability in artificial agents. Classical top-down (Logic based) and bottom-up (Connectionist) approaches to the problem have had limited success to date. We review a small body of work that represents a different approach to AI. We call this work the Bottom Up Symbolic (BUS) approach and present a new BUS method to concept construction. While valid concepts have been constructed using previous methods under this approach, we show in this paper that the one-sided clustering methods generally used there may fail to uncover valid concepts even when they clearly exist. We show that by using a Co-clustering algorithm that searches for an optimal partitioning based on the Mutual Information between the category and consequent components of a concept, the concept formation outcome is improved. We test our approach on data from experiments using a real mobile robot operating in the real world, and show that our Co-clustering based approach leads to significant performance improvement compared to previous approaches.**

I. Introduction

AI researchers have long recognised how impressive human cognitive abilities stem from a symbolic representation of the world, ie. in a human's ability to form and manipulate *concepts*. This recognition is reflected by the two main approaches to AI to date. First, the classical Top-Down approach based in Mathematical Logic, which attempts to directly mimic the symbol formation and processing of humans. Concepts are preinstalled in an agent, and logic is used as the basis for manipulating these concepts in order for intelligent behaviour to occur. The second approach to AI has been called the Bottom Up approach. Artificial neuronal elements are linked together into networks (neural networks) with the aim that a higher level symbolic representation of the world will emerge. Both approaches have had limited success to date. Indeed it has been conjectured [1] that the symbols preinstalled into the agent in the Top-down approach can never mean anything to the agent, and hence does not form a valid pathway to AI.

A large amount of research has gone into discovering what form concepts take in human thought processes [2] and many competing models have been proposed. Here we take a common view that a Concept is made up of (a) a collection of instances, called a Category, and (b) the consequences of category membership, which we call here the Consequent [1]. For example, the human concept for "Dog" consists of a category made up of the different dogs in the world, and the consequents of identifying an instance of this category is that it will have four legs, it will bark, etc.

An alternative to the existing Logic (top down) and Neural Net (bottom up) approaches to AI is what we call the Bottom Up Symbolic (BUS) approach. The idea here is to grow a symbolic representation of the world out of low level data streams from the agent's sensors and actuators. Note that it is distinct from the the Bottom Up Non-symbolic (BUNS) approach of Neural Nets since it does not require the substrate of the approach to consist of (non-symbolic) neurons. A small number of BUS works exist [3], [4], [5], [6], [7], including by the author of this paper [8]. One of the main findings in [8] was the importance of a probabilistic approach to searching for concepts in the underlying sensor/actuator data. This is because building concepts is about finding a deterministic relation between categories and consequents in the seemingly random data streams obtained from real world agents operating in real world environments.

In this paper we continue with a probabilistic approach to concept formation for AI agents, however we propose a novel new way to construct the category and consequent components of concepts on two fronts. First, we root our search for determinism in the important measure of Mutual Information from the field of Information Theory. Second, we show that a fundamental part of discovering such determinism is to link the formation processes of categories and consequents together. Specifically, we propose co-clustering (rather than traditional one-sided clustering) as the means to form categories and consequents, with a significant improvement in the final concept formation outcome.

II. Problem Formulation

Our framework for growing concepts is envisaged for a wide range of applications, from a simple valve controller, to a complex humanoid robot. Such agents will have sensors, which are read by the agent, and which reflect the state of the environment. They will also have actuators, which are written to by the agent, and which operate on the environment. The idea is to deploy the agent in its environment with a basic (maybe random) set of behaviours such that exploration results. The agent will experience a series of sensory inputs, and the consequences of actuatory outputs, through time. The framework we detail below aims to construct concepts from

[1]Note that in previous work we called this the Entailment

this experience. In essence, the framework takes a set of low level sensor and actuator readings, and derives from them a higher level discretised (symbolic) representation of the world in the form of set of meaningful concepts, ie. categories of experience and their consequents.

Imagine an agent with a set $\mathbf{s}$ of n_s distinct sensors and a set $\mathbf{a}$ of n_a distinct actuators. We are considering here sensors in $\mathbf{s}$ and actuators in $\mathbf{a}$ that are scalar in their output at each point in time (those that deal in vector quantities at each time step, like laser range finders, can be viewed as sets of scalar returning apparatus). Call Sensor Space $\mathcal{S} = \mathcal{S}_1 \times \ldots \times \mathcal{S}_i \times \ldots \times \mathcal{S}_{n_s}$, where $\mathcal{S}_i$ is the set of possible outputs from the ith sensor in $\mathbf{s}$. Similarly, call actuator space $\mathcal{A} = \mathcal{A}_1 \times \ldots \times \mathcal{A}_j \times \ldots \times \mathcal{A}_{n_a}$. The space $\mathcal{S} \times \mathcal{A}$ then represents the full possibility of an agent's sensory motor experience, and the agent's experience at each point in time will correspond to a single point in $\mathcal{S} \times \mathcal{A}$. Given that categories are defined as collections of sensorimotor experience, they then necessarily correspond to subsets in $\mathcal{S} \times \mathcal{A}$.

Our formulation of the consequent part of a concept differs to that for categories. The consequent infers predictive ability to the agent for some category of sensorimotor experience. The agent will be interested in predicting the category's effect on the state of the environment, and its view of the environment is encompassed by its sensory regime alone. Hence, consequents will correspond to subsets in $\mathcal{S}$ only, not $\mathcal{S} \times \mathcal{A}$.

The question of how subsets in $\mathcal{S} \times \mathcal{A}$ ($\mathcal{S}$) should be formed to represent categories (consequents) of sensorimotor experience is the key research question in this area. Assume the Euclidean metric on $\mathcal{S} \times \mathcal{A}$ and on $\mathcal{S}$. In previous work in the literature, the approach has been to form categories (consequents) as sets of closely spaced points in $\mathcal{S} \times \mathcal{A}$ ($\mathcal{S}$). The motivation for this approach has been that humans seem to form categories and consequents out of objects/instances that have *similar* characteristics. For example, the category "dog" used by humans contains instances that have fur, pant, have four legs, etc. If a Euclidean Space is formed so that each aspect of an object/instance's characteristics spans one dimension, then closely spaced point clouds should correspond to categories of the objects/instances. While anecdotally this approach makes sense, it ignores one important aspect of the category-consequent formation process in humans; a category is formed to provide predictive ability on its consequents. In this paper we propose the alternative approach to category (consequent) formation by partitioning $\mathcal{S} \times \mathcal{A}$ ($\mathcal{S}$) in such a way that predictive ability from categories to consequents is maximised.

III. Mutual Information as the metric for Predictive ability

We have talked about the "predictive ability" of a category on a consequent, but lets now use Information Theory to quantify this term. Let X be a discrete random variable (RV) that takes a distinct real value for each element in the set $\mathcal{S} \times \mathcal{A}$. Similarly, let Y be a discrete random variable that takes a distinct real value for each element in the set $\mathcal{S}$. To form categories (consequents) we have said we need to partition $\mathcal{S} \times \mathcal{A}$ ($\mathcal{S}$) into clusters. Let our search be for K clusters in category space ($\mathcal{S} \times \mathcal{A}$) and L clusters in consequent space ($\mathcal{S}$). That is, we need to form membership maps C_X and C_Y such that

$$C_X : \mathcal{S} \times \mathcal{A} \rightarrow \{\hat{x}_1, \ldots, \hat{x}_K\} \tag{1}$$

$$C_Y : \mathcal{S} \rightarrow \{\hat{y}_1, \ldots, \hat{y}_L\} \tag{2}$$

So C_X maps points in category space into one of K classes $\hat{x}_1, \ldots, \hat{x}_K$, and C_Y maps points in consequent space into one of L classes $\hat{y}_1, \ldots, \hat{y}_L$. We can then write $\hat{X} = C_X(X)$, where $\hat{X}$ is a RV that can take any of the values in the set $\{\hat{x}_1, \ldots, \hat{x}_K\}$, and which is a deterministic function of the RV X. Similarly, we can write $\hat{Y} = C_Y(Y)$, where $\hat{Y}$ is a RV that can take any of the values $\hat{y}_1, \ldots, \hat{y}_L$, and which is a deterministic function of the RV Y.

We can now quantify what we mean by "predictive ability". We measure the predictive ability of a particular set of categories in category space (ie. a partitioning of category space) on a particular set of consequents in consequent space (a partitioning of consequent space) as the mutual information $I(\hat{X}; \hat{Y})$ existing between the the RV's $\hat{X}$ and $\hat{Y}$. In information theory, the Mutual Information between two random variables X and Y is written as $I(X; Y)$ [9] and represents the amount of uncertainty removed from the outcome of Y given we know the outcome of X (and visa versa). Hence, by "predictive ability" here we mean the amount of uncertainty removed about what consequent the agent will next experience give the category that was experienced at this time step. Clearly our aim is to find categories and consequents (ie. partitions of category and consequent space) that lead to high values of mutual information.

IV. Co-clustering

Given we want to partition $\mathcal{S} \times \mathcal{A}$ and $\mathcal{S}$ so that $I(\hat{X}; \hat{Y})$ is maximised, let us now explore how this partitioning should occur. First, we could use some standard partitioning tools like standard (eg. K-means) clustering algorithms to partition $\mathcal{S} \times \mathcal{A}$, and then search for the partitioning of $\mathcal{S}$ that maximises $I(\hat{X}; \hat{Y})$. However, what if the chosen partition of $\mathcal{S} \times \mathcal{A}$ precludes any partitioning of $\mathcal{S}$ from achieving a reasonable maximisation? The solution is neither to first partition $\mathcal{S}$ and then find the partitioning in $\mathcal{S} \times \mathcal{A}$ that maximises mutual information, since this will suffer the same drawback. What is required is a search for a maximum $I(\hat{X}; \hat{Y})$ which allows the optimisation to search for possible partitions in both $\mathcal{S} \times \mathcal{A}$ and $\mathcal{S}$ simultaneously. Co-clustering is the name of just such an optimisation process.

A number of Co-clustering (also called bi-clustering or block clustering) algorithms exist the literature [10], [11], [12], [13]. Usually the problem is formulated as clustering the rows and columns of a 2D contingency table (matrix). All have the property that they look for partitions in a joint wise fashion between the 2 dimensions of the table. The algorithms differ

in the way they search for an optimum partition (some adopt a heuristic approach, others optimise on a particular metric). In this paper we use the approach proposed by Dhillion et al [13] because it uses mutual information as the metric on which the optimisation process is based. Details of co-clustering can be found in [13], however we now provide details of how we embed the co-clustering process into our application.

V. Implementation Details

The value of each read (write) from (to) the agents sensors (actuators) is recorded during the environment exploration phase of our regime. While the number of distinct read and write values for an agent's sensors/actuator apparatus is finite (given the limited range and precision of sensors/actuators), in practice the number can be large. This leads to large storage and computational requirements when the co-clustering algorithm is run at a later step. These requirements can be significantly reduced, without (in our practical experience) loss in performance, by dividing each sensor and actuator range into a reasonable number (say 20) of discrete bins. For example, a sensor returning readings between 0 and 200 can have its data preprocessed in this way by relabelling any data between 1 and 10 as "5", 10 and 20 as "15", etc.

Recall that we are looking for determinism between Category and Consequent space, ie between $\mathcal{S}\times\mathcal{A}$ and $\mathcal{S}$. The next step in the process is then to form the datasets corresponding to each of these spaces. Practically, this means 2 arrays, one n (rows) by $n_s + n_a$ (columns) for categories and the other n by n_s for consequents (where n is the number of time steps that occurred in the exploration phase). Integral to this step is the selection of a temporal offset. That is, at what time lag after a "category" has occurred are we looking for determinism in the "consequent" outcome. Call the lag δ. In practice we achieve a particular lag by matching the first $n-\delta$ rows of the category dataset (call this D_c) up with the the last $n-\delta$ rows of the consequent dataset (call this D_e) to form the final dataset D. The temporal offset, in seconds, can be calculated by multiplying δ by the reciprocal of the recording frequency (in Hz) of the sensors/actuators during the exploration phase.

Recall from Section III the RV's X (associated with Category Space) and Y (associated with Consequent Space). The next step in the process is to construct a statistical estimate of the joint distribution $p(X,Y)$ as a 2-dimensional contingency table (a 2D array). Let each row in this contingency table represent one possible outcome for X (call it x), and each column represent one possible outcome for Y (call it y). Each entry of the contingency table (ie. $p(X=x, Y=y)$) can then be calculated by (i) counting the number of rows in the final dataset D where the entry in the D_c part of the row corresponds to x and the entry in the D_e part of the row corresponds to y, and (ii) dividing this by the number of rows in D.

The contingency table just derived forms one of the inputs into the co-clustering algorithm proposed in [13]. The other inputs are K (the number of clusters to be formed in Category Space) and L (number of clusters in consequent space). For experiments in this paper we manually selected the K and L according to what made sense for the experiment. Automatic selection of these parameters is outside the scope of our present work in the area. Also of note is that the algorithm in [13] finds local optima dependant on the initialisation seeds provided. To promote the selection of global optima, we repeatedly ran the algorithm 50 times with random initial seeds, and selected the optimum solution of the 50 runs. The final outcomes are the membership maps C_X and C_Y, which partition Category and Consequent spaces respectively. Given these maps, and our previously derived contingency table $p(X,Y)$, we could then calculate $p(\hat{X},\hat{Y})$, and consequently $I(\hat{X};\hat{Y})$.

Fig. 1. Experiments: Pioneer DX3 Robot roaming in part of our laboratory

	Co-clustering			1-sided clustering		
	$\hat{y}_1$	$\hat{y}_2$	$\hat{y}_3$	$\hat{y}_1$	$\hat{y}_2$	$\hat{y}_3$
$\hat{x}_1$	0.2663	0.0880	0.0013	0.1823	0.1455	0.0001
$\hat{x}_2$	0.0143	0.2829	0.0678	0.0235	0.3421	0.0100
$\hat{x}_3$	0.0003	0.0243	0.2549	0.0006	0.1970	0.0990

TABLE I
Expt. 1: Clustered variable joint probability matrix

VI. Results

Experiments were conducted on a Pioneer DX3 mobile robot operating in an office type environment. The robot was allowed to roam randomly in a partitioned-off area of our lab (Figure 1), programmed only with some very low level behaviours that allowed it to bounce off obstacles and not get stuck in corners, etc. The roaming experiment lasted almost 3 hours and resulted in 46077 points of recorded data for each robot sensor (compass (C), gyro (G)) and actuator (wheels: Left (Lw), right (Rw)). Note that the Pioneer robot had more sensors than just the Compass and Gyro, however we have hand picked these sensors for forming two specific category and consequent space combinations (we call them Experiment 1 and 2), since these allowed us to demonstrate the contributions of this paper in the most straight forward way.

Experiment 1 took category space as $Lw \times Rw$ and consequent space as G. One would expect to find concepts (ie. determinism) in the relation between the current left and right wheel actuation and the angular velocity of the robot about its vertical axis measured by the gyro. According to our claims from previous sections, we would like to show

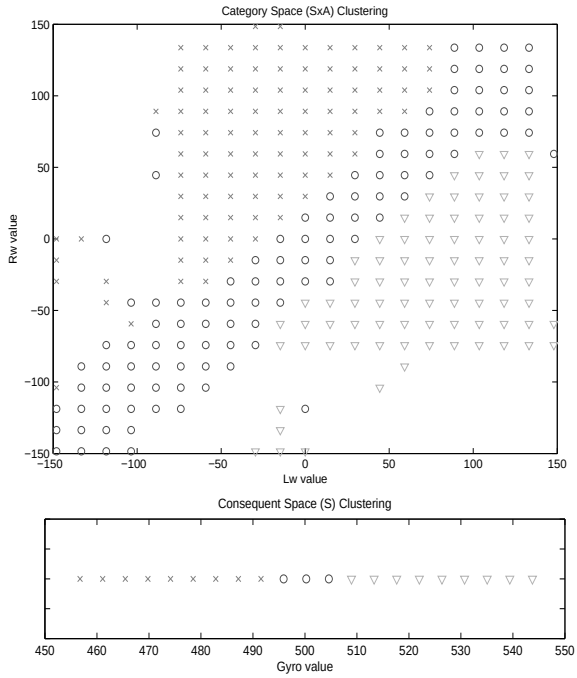

Fig. 2. Experiment 1: Co-clustering results

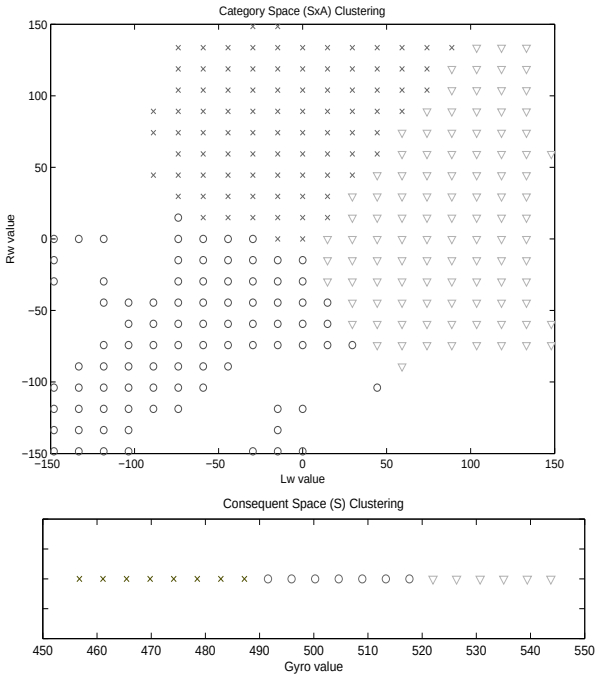

Fig. 3. Experiment 1: 1-sided clustering results

that the concepts formed from this data by our information theoretic, co-clustering approach better capture and reflect the form of the determinism that exists here compared to previous approaches. To this end, we ran both (i) our co-clustering approach, and (ii) previous approaches as represented by the traditional K-means 1-sided clustering algorithm, on the same data. In both cases we searched for 3 categories and 3 consequents, and used a temporal offset of one time step (we talk more about this below).

Table I shows the resulting joint probability distribution $p(\hat{X}, \hat{Y})$ for each approach. Recall from Section III that $\hat{X}$ and $\hat{Y}$ are random variables whose mutual information $I(\hat{X}; \hat{Y})$ is maximised when good concepts and consequents are found. Using the values in Table I we can calculate $I(\hat{X}; \hat{Y})$ for co-clustering as 0.5641, and for 1-sided clustering as 0.2745. On this basis, the co-clustering partitioning has better captured the determinism exiting between Category and Consequent Space in the original data. If we study the values in Table I we can see why the mutual information numbers come out as they do. For co-clustering, the categories $\hat{x}_1$, $\hat{x}_2$ and $\hat{x}_3$ very much predict the occurrence of the consequents $\hat{y}_1$, $\hat{y}_2$ and $\hat{y}_3$ respectively. In contrast, there is not this "one-to-one' type prediction in the 1-sided clustering results.

If we now look at what the clustering outcomes actually looked like in category and consequent space, we can see why co-clustering gave the better outcome. Figure 2 shows the co-clustering outcome for both category space (top) and consequent space (bottom). The cross markers show points belonging to $\hat{x}_1$ in the top plot and $\hat{y}_1$ in the bottom plot. The circle markers show points belonging to $\hat{x}_2$ and $\hat{y}_2$, while the triangle markers correspond to $\hat{x}_3$ and $\hat{y}_3$. Noting the "diagonal" nature of the co-clustering part of Table I, each marker type in the category (top) plot then predicts the same marker type in the consequent (bottom) plot. That is, a positive (forward) command on the left wheel, and a backward (negative) command on the right wheel, predicts a clockwise (> 502) reading on the gyro. Conversely backward left wheel and forward right wheel commands predict anti-clockwise gyro output. Finally, when wheel commands are more or less equal, a no rotation consequent is predicted. Note how the co-clustering algorithm has correctly identified that category space should be broken down into "diagonal strips" if a high level of mutual information between categories and consequents is to be maintained. We shall now see how this is not the case for the 1-sided clustering outcome.

Figure 3 shows the 1-sided clustering outcome for both category space (top) and consequent space (bottom). Each of the three marker types in the plot were assigned to the same cluster variables as in the co-clustering case. We can see straight away why significant entries lie off the diagonal in the 1-sided algorithm part of Table I: both Category and Consequent spaces have simply been partitioned into three regions of closely spaced points without consideration to what outcome leads to better prediction of consequents by categories. For example, the cross marked category now includes scenarios where (i) the left wheel command is negative and the right wheel command positive and (ii) the left and right wheel commands are more or less the same. It is clear then why it should provide the mixed prediction of (i) clockwise rotation, and (ii) no rotation, as reflected by the values (i) 0.1823 and (ii) 0.1455 in Table I. Similar observations can be made for the circle and triangle marker categories.

Experiment 2 took category space as $C \times G$ and consequent space as C. One would expect to find determinism in

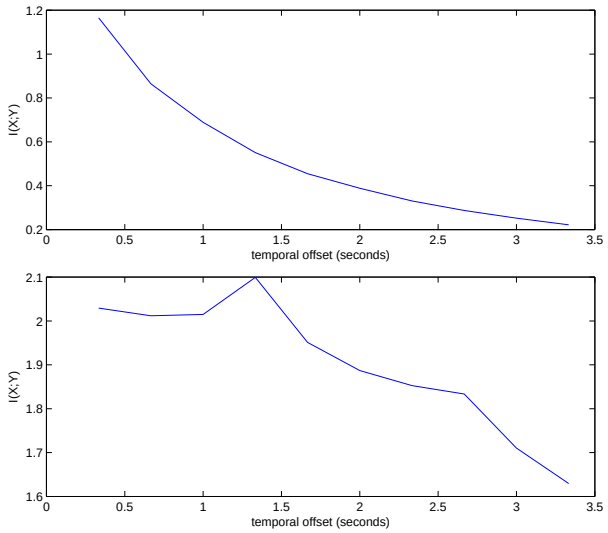

Fig. 4. Selecting a temporal offset using $I(X;Y)$

the relation between the current heading and turn rate with future heading. While (as we shall see) co-clustering more appropriately uncovers determinism in this case also, a key difference here compared to Experiment 1 is the temporal offset at which the determinism occurs. As left and right wheel commands change, the change experienced on the gyro is more or less instantaneous (for example, a "forward" left wheel and "backward" right wheel command will instantaneously change the gyro output to a "clockwise" reading). Hence the temporal offset of one time step used there. Here, the heading may take some time to change given any current heading and gyro command (eg. a clockwise gyro command and a current North heading will take some time to reach East as the robot gradually changes heading). How should we determine the best temporal offset to use here?

Two possibilities exist. We could plot $I(\hat{X};\hat{Y})$ versus temporal offset to find the offset that produces maximum determinism between the clustered category and consequent random variables. However, this has the computational disadvantage that the co-clustering algorithm needs to be run for each possible offset. While such a scheme would be easily tractable for our current simple examples, we have one eye on future work where data from larger combinations of sensor and actuators will form Category and Consequent spaces. An alternative, less computationally expensive approach is to plot $I(X;Y)$ versus temporal offset. The beauty here is that co-clustering does not need to occur, and that (by Lemma 2.1 in [13]) $I(\hat{X};\hat{Y}) \leq I(X;Y)$. In other words, we search for the temporal offset where the most mutual information exists between the raw data RV's X and Y, and then we use the co-clustering algorithm to find the partitioning that gives the maximum $I(\hat{X};\hat{Y})$ at that temporal offset. Figure 4 shows the result. The top plot shows that the optimal offset for the wheel/gyro experiment is indeed 1 time step, or approximately 0.4 seconds (with higher offsets producing a drop off in mutual information as the relationship between current wheel

	Co-clustering				1-sided clustering			
	$\hat{y}_1$	$\hat{y}_2$	$\hat{y}_3$	$\hat{y}_4$	$\hat{y}_1$	$\hat{y}_2$	$\hat{y}_3$	$\hat{y}_4$
$\hat{x}_1$	0.158	0.006	0	0.009	0.094	0.108	0.062	0.014
$\hat{x}_2$	0.008	0.254	0.011	0	0.029	0.065	0.112	0.040
$\hat{x}_3$	0	0.008	0.289	0.009	0.009	0	0.150	0.254
$\hat{x}_4$	0.007	0	0.008	0.235	0.004	0.012	0.038	0.006

TABLE II
EXPT. 2: CLUSTERED VARIABLE JOINT PROBABILITY MATRIX

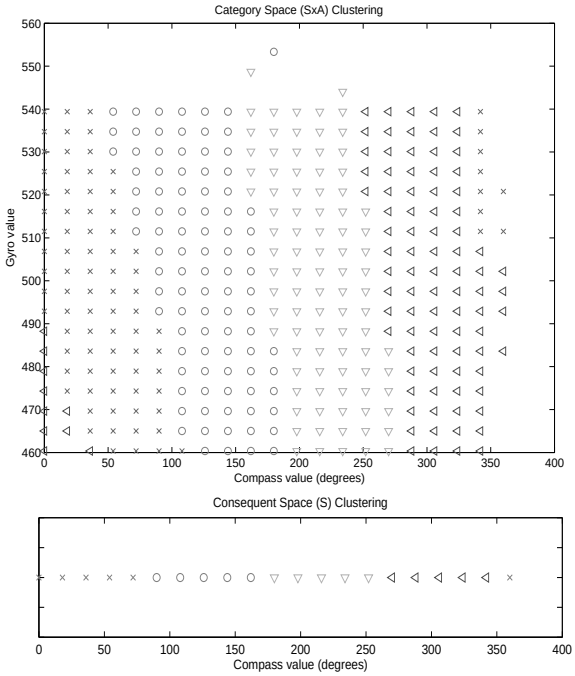

Fig. 5. Experiment 2: Co-clustering results

commands and future turn rate weakens). In contrast, the result (bottom plot) for the compass/gyro experiment shows an optimal offset of approximately 1.5 seconds, or 3 time steps. Again temporal offsets larger than this result in a drop off in mutual information.

Given the result in Figure 4, we formed our data set in Experiment 2 with a temporal offset of 3. We then ran both the co-clustering and k-means algorithms on the data, this time searching for 4 categories and 4 consequents [2]. The resulting mutual information value $I(\hat{X};\hat{Y})$ obtained for co-clustering was 1.0846, and for k-means was 0.2898. On the basis of our desire for categories to predict consequents, co-clustering has again produced the better partitioning. Table II shows the joint probability matrix $p(\hat{X},\hat{Y})$ for both the co-clustering (left 4 columns) and k-means (right 4 columns) outcomes. One can again clearly see the 1 to 1 relationship between the 4 categories and 4 consequents for the co-clustering distribution . This is in contrast to the outcome for k-means, where such a one-to-one correspondence is missing, and categories generally predict a number of different consequents.

In Figures 5 and 6 we show the results in Category and Consequent Space for co-clustering and K-means respectively. In both figures, the top plot shows the Category space partitioning, while the bottom plot shows the Consequent

[2]We chose the value 4 given the 4 headings on the compass rose.

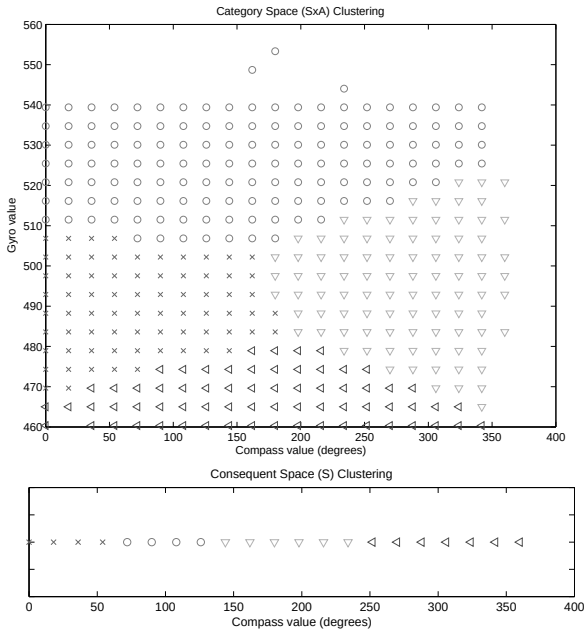

Fig. 6. Experiment 2: 1-sided clustering results

space partitioning. For co-clustering, the interesting feature in Category space is that partitions have been organised into four vertical strips that are inclined from right to left and that are centred over the partitions in Consequent space of the same marker type. What this means is that, for gyro values around 500 (ie. no rotation) a particular compass reading predicts the same compass reading to occur as the consequent 3 time steps later (ie. the strips are centred). However, if the gyro is measuring a clockwise rotation (say 540), then for a particular compass reading, a consequent in the clockwise direction on the compass rose is predicted, (ie. category partitions at these gyro readings are to the left of the predicted same-marker-type partitions in consequent space). Conversely, category partitions at low (say 450) gyro readings (representing anticlockwise rotation) are moved to the right of the same partitions in consequent space). This is exactly the structure of dynamics present in the view of the environment provided by these sensors/actuators, and such structure is not altogether obvious from the data. Indeed the naive partitioning of the 1-sided k-means clustering approach in Figure 6 has not discovered it. [3] K-means has simply partitioned Category and Consequent Spaces into 4 groups of closely spaced points. No account has been taken of how these partitions should be formed in terms of the predictive ability of categories on consequents, and hence it is obvious why the poor mutual information value and mixed prediction scenarios reflected in Table II occur. Given the results from both Experiments 1 and 2, it would seem that co-clustering should be a part of any framework for concept formation in AI agents.

[3]Note that the Euclidean metric does not make sense in a Category or Consequent space where the units along each dimension are distinct (eg. Gyro units on the Y-axis versus Compass units on the X-axis). The solution was to scale the readings along each dimension to obtain a mean of zero and standard deviation of 1, and then perform the K-means clustering. To make the plots comparable again to those produced by co-clustering, we rescaled the clustered data back into the original units for plotting purposes.

VII. Conclusion

We have presented a new approach to concept formation for AI agents. The approach was distinct from previous work on two fronts. First, we promoted the idea that the two components of a concept (ie. categories and consequents) need to be formed in a joint wise fashion so that they "match up" in their description of the dynamics in the environment. Co-clustering was used as the means to achieve this end. Previous work has looked at independently forming the category and consequent components of concepts using 1-sided clustering algorithms like K-means, which, as we have seen, produce substantially poorer outcomes than our co-clustering approach. The second novelty of our work was the use of Mutual information as the basis for forming categories and consequents. Previous work in the field constructed these entities by forming partitions of closely spaced points in category and consequent space. While it has been argued that this is what humans do, we took the different, more direct approach of basing the formation of partitions directly on the predictive ability of categories on consequents. The result was the formation of concepts with greater predictive ability, which is what is important given the role concepts must play for agents in their interaction with the world. One valid criticism of our work is that the category and consequents space combinations tested here were quite simple. However, it seems clear to us that the principle behind the success of the co-clustering examples presented here will be relevant to more complicated scenarios. This will be the focus of our future work in the area.

References

[1] S.Harnad, "The symbol grounding problem," *Physica D*, pp. 335–346, 1990.
[2] G.L.Murphy, *The Big Book of Concepts*. MIT Press, 2002.
[3] P.Vogt, "Bootstrapping grounded symbols by minimal autonomous robots," *Evolution of Communication*, vol. 4, pp. 89–118, 2000.
[4] R.Sun and T.Peterson, "Some experiments with a hybrid model for learning sequential decision making," *Information Science*, vol. 111, pp. 83–107, 1998.
[5] A.Billiard and K.Dautenhahn, "Experiments in learning by imitation - grounding and use of communication in robotic agents," *Adaptive Behaviour*, vol. 7, pp. 411–434, 1999.
[6] D.Pierce and B.J.Kuipers, "Map learning with uninterpreted sensors and effectors," *Artificial Intelligence*, vol. 92, pp. 169–227, 1997.
[7] M.Rosenstein and P.R.Cohen, "Continuous categories for a mobile robot," in *Proceedings of the Sixteenth National Conference on Artificial Intelligence*, 1999, pp. 634–640.
[8] J.R.Chen, "Symbol statistics for concept formation in ai agents," in *Proceedings of the Intelligent Agent Technology Conference*, Milan, Italy, September 2009.
[9] T.Cover and J.Thomas, *Elements of Information Theory*. Wiley, 1991.
[10] J.A.Hartigan, "Direct clustering of a data matrix," *Journal of the American Statistical Association*, vol. 67, no. 337, pp. 123–129, 1972.
[11] Y.Cheng and G.Church, "Biclustering of expression data," in *Proceedings of International Conference on Intelligent Systems for Molecular Biology*, 2000, pp. 93–103.
[12] I.S.Dhillon, "Co-clustering documents and words using bipartite spectral graph partitioning," in *Proceedings of SIGKDD Conference on Knowledge Discovery and Data Mining*, 2001, pp. 269–274.
[13] I.S.Dhillon, S.Mallela, and D.S.Modha, "Information theoretic co-clustering," in *Proceedings of SIGKDD Conference on Knowledge Discovery and Data Mining*, Washington DC, USA, 2003.