УДК 681.3

ЭМОЦИИ, АПРИОРНЫЕ ЗНАНИЯ И ДРУЖЕСВЕННОЕ ПОВЕДЕНИЕ РОБОТА

А.В. Гаврилов, НГТУ[1]

В работе описывается архитектура интеллектуальной системы, основанная на выработке базовых эмоций (положительной и отрицательно) с использованием априорного знания. Эти эмоции используются в качестве обратной связи для обучения нижележащих нейронных сетей, обеспечивающих восприятие окружающей среды и поведение в ней. Предлагается простой язык для представления априорных знаний и его использование для формулирования законов робототехники А.Азимова, что обеспечивает обучение интеллектуальной системы дружественному поведению.

Введение

В настоящее время можно выделить следующие тенденции в развитии искусственного интеллекта:

1. Проникновение ИИ во все аспекты информатизации жизни – от разработки программного обеспечения до приготовления пищи и уборки на кухне. Проявлением этой тенденции является появление и разработка таких парадигм как “ubiquitous computing”, “pervasive computing” и “smart environment”[Tarik-Ul Islam, A.V.Gavrilov et al, 2008].
2. Появление парадигмы «настоящего искусственного интеллекта» или универсального ИИ (Artificial General Intelligence) [Goertzel, Pennachin, 2007], способного обучаться и решать любые задачи (а не определенный тип задач) подобно тому, как это делает человек.
3. Создание искусственных интеллектуальных систем (в частности, интеллектуальных роботов) с человеко-подобным поведением, характеризующимся моделированием коллективного поведения, эмоций, мотивации, самообучения [Добрынин, 2006], .
4. Появление реальных попыток создания самовоспроизводящихся роботов.
5. Появление интереса к этическим вопросам, связанным с появлением такого человеко-подобного искусственного интеллекта.

Суммируя эти тенденции можно сформулировать главное направление развития ИИ – создание универсального обучаемого искусственного разума, подобного человеческому и ориентированного на восприятие и действие.

Для того, чтобы создать такой искусственный разум, необходимо ответить на следующие основные нерешенные в большой степени вопросы:

1) роль и механизмы эмоций в процессе принятия решений и реализации поведения;

[1] *630092, Новосибирск, Пр. Карла Маркса, 20, andr_gavrilov@yahoo.com*

2) необходимо или нет использовать априорное знание, если да, то в какой степени, и как его предсталять и использовать;
3) как реализовать рассуждения о будущем и связать их с планированием и действием;
4) как реализовать сознание и рассуждения о себе;
5) как обеспечить дружественные по отношению к человеку мысли и поведение этого разума.

В последнее время было опубликовано много гипотез и теорий о создании такого искусственного разума [Прибрам, 1971], [Арбиб, 1976], [Wermter and Sun, 2000],], [Hawkins, Blakeslee, 2004], [Hoya, 2005], [P. Wang, 2006], [Yingxu Wang, Ying Wang, 2006], [Minsky, 2006], [Hecht-Nielsen, 2006]. Но все они в основном уделяют внимание одной проблеме из выше перечисленных. В этой статье автор пытается связать 1-ый, 2-ой и 5-ый вопросы. А именно, в статье предлагается гибридная модель искусственного разума [Gavrilov, 2008], основанная на следующих постулатах

1. Все многообразие эмоций сводится к двум базовым эмоциям – положительной и отрицательной.

2. Базовые эмоции вырабатываются на самом верхнем уровне обработки информации (восприятия) и посредством обратных связей сверху вниз управляют процессом обучения и, как следствие, поведения интеллектуальной системы.

3. Выработка базовых эмоций является достаточно жестко запрограммированной (не считая того, что имеют дело с нечеткой информацией снизу), а нижележащие уровни обладают большей гибкостью при обучении в контексте, задаваемом верхнем уровнем.

Предлагается закладывать в такую систему априорное знание о выработке базовых эмоций в виде набора нечетких правил с лингвистическими переменными, в основе которых лежат три закона робототехники А. Азимова [Азимов, 2004]. А нижележащие уровни могут быть реализованы на основе нейронных сетей. Таким образом, эта модель является разновидностью гибридных интеллектуальных систем [Medsker, 1995], [Wermter, Sun, 2000], [Колесников, 2001], [Гаврилов, 2003], [Ярушкина, 2004], [Колесников, Кириков, 2007]. Предложенная архитектура обеспечивает выработку в процессе обучения дружественного поведения искусственной интеллектуальной системы.

Роль эмоций

Эмоции в нашем разуме играют очень важные и разные роли. В настоящее время нет достаточно четкого и однозначного научного представления об эмоциях. Это объясняется, вероятно, большим разнообразием точек зрения на эмоции в различных отраслях знаний. Например, психологи в основном интересуются внешним проявлением эмоций в процессе общения между индивидами. С этой точки зрения они используют свою классификацию

эмоций. Обычно в качестве основных (базовых) эмоций в психологии рассматривают следующие эмоции: интерес, удовлетворение, надежда, радость, пренебрежение, гнев, страх и горе.

Но разработчиков ИИ интересуют еще и внутренние состояния, связанные с продуцированием эмоций и их использованием в обучении и поведении. Следует иметь в виду, что выраженные (мимикой, жестами) эмоции далеко не всегда совпадают с внутренними («настоящими») эмоциями. Таким образом, можно рассматривать эмоции с разных точек зрения, например, с точки зрения их влияния на внимание, ускорения или замедления принятия решений. Можно говорить о связи эмоций с метаболизмом, использования эмоций в коммуникативной деятельности, в мотивации и планировании поведения и т.п. В этой статье мы рассматриваем эмоции как основной механизм для мотивации, обучения и планирования поведения, делая упор на обучении, как основе для организации поведения. Что касается мотивации, то надо иметь в виду, что мотивация базируется отчасти на инстинктивных процессах, а отчасти, является результатом обучения.

Эмоции и мотивация очень связанные между собой процессы. Более того, если мы будем ограничиваться базовыми эмоциями (положительными и отрицательными), то можно сказать, что выработка мотивации основана на базовых эмоциях. Это одно из основных свойств нашего (и братьев наших меньших) мозга. Живое существо стремится получать положительные и избегать отрицательные эмоции, и на этом строится обучение поведению всего живого. Что касается других видов эмоций, таких как гнев, радость, горе, жалость, страх, ужас, нерешительность, счастье, удивление, сомнение, презрение, надежда и т.д., то их можно рассматривать, как эмоции, производные от базовых с учетом особенностей их проявления, интенсивности и целей, преследуемых индивидом при общении с другими особями.

Автор исходит из того, что “любая классификация эмоций может считаться «истинной» или «естественной», коль скоро она удовлетворяет своему назначению» [Джеймс, 1911]. С этой точки зрения использование только двух базовых эмоций для целей управления обучением и, следовательно, поведением кажется достаточно обоснованным, учитывая опыт, накопленный в практической психологии, педагогике и дрессуре животных (использование поощрения и наказания), а также, применение этого подхода в ИИ в рамках разновидности обучения, известной как “reinforcement learning” [Sutton, Barto, 1998].

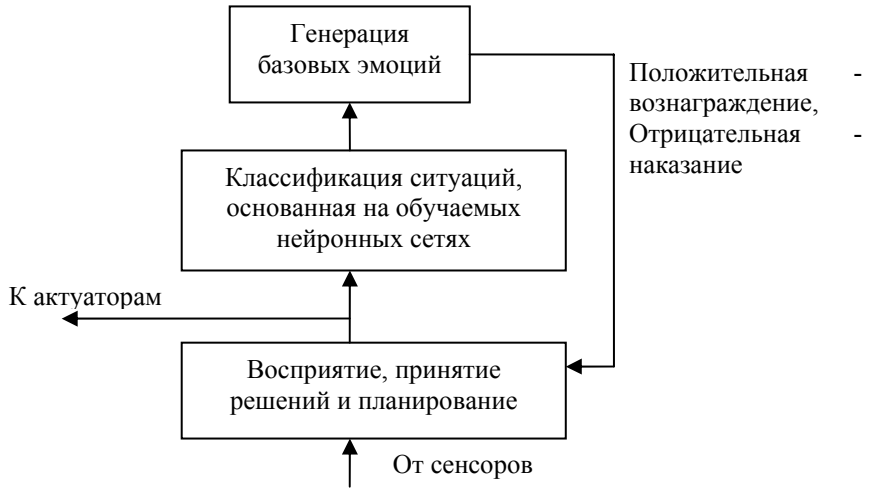

Рис. 1. Общая схема предлагаемой модели

Следует заметить, что в последнее время во многих лабораториях мира ведутся работы по моделированию эмоций в робототехнике [Picard, 1997], [Goerke, 2006]. При этом упор делается в основном на применении эмоций для целей обеспечения комфортного общения с роботом.

Представление априорного знания

Как связать между собой априорные знания и способность обучения интеллектуальной системы? В начале истории искусственного интеллекта такой вопрос не возникал, т.к. исследователи имели дело с моделированием символического (вербального, логического) мышления на основе априорного знания и использование этого знания для решения задач. Кульминацией такого подхода явились классические экспертные системы [Хейес-Рот, Уотерман, Ленат, 1987]. Но позже, когда многим исследователям в области ИИ стало ясно, что обучение посредством взаимодействия с окружающей средой является главным свойством настоящего интеллекта, иногда можно услышать другое крайнее мнение, что априорного знания не существует, а разум есть продукт только прежизненного обучения. Представляется более перспективным наряду с разработкой механизмов обучения, продуцирующих категории, концепции, правила, закодированные в и извлекаемые из обученных нейронных сетей (или, например, байесовских сетей) использовать и априорные знания, позволяющие управлять роботом в заранее определенных ситуациях. Так, например, в наших экспериментах с моделью мобильного робота при непосредственном столкновении его с препятствием робот выбирает направления поворота или отхода в соответствии с простыми правилами в отличие от поведения, управляемого нейронной сетью относительно далеко от препятствия, когда у робота есть свобода выбора с учетом множества факторов (входов нейронной сети) [Gavrilov, Lee, 2007, 1]. Такие априорные знания являются аналогом инстинктов. Конечно, такое априорное знание

является слишком примитивным, и для построения искусственного разума необходимо предусмотреть возможность закладывать в него более сложное априорное знание с учетом нечеткостей.

Мы предлагаем простой язык для представления априорных знаний, описанный ниже в БНС нотации.

<Rule> ::= If <Antecedent> then < Consequent >
<Antecedent> ::= <Fuzzy symbolic value> | <Function> (<Fuzzy symbolic value>) | <Condition>
| not <Antecedent> | <Antecedent> and <Antecedent>
<Condition> ::= <Fuzzy symbolic value> <Relation> <Operand>
<Relation> ::= = | < | > | ≤ | ≥ | ≠
<Operand> ::= <Fuzzy symbolic value> | <Constant>
<Consequent> ::= <Action>
<Function> ::= Increase | Decrease

Здесь <Fuzzy symbolic value> означает значение лингвистической переменной [Zadeh, 1975]. В этой грамматике не описаны терминальные символы < Fuzzy symbolic value > и < Action >, т.к. их возможные значения зависят от конкретной реализации интеллектуальной системы. Ниже они будут описаны для случая генерации базовых эмоций на основе трех законов робототехники. Функции «Increase» и “Decrease” применяются к значениям нечеткой переменной и означают, что оно увеличилось или уменьшилось, соответственно, по сравнению с предыдущим значением.

Семантика этой грамматики определяется в основном нечеткой логикой. Это означает, что каждая нечеткая переменная <Fuzzy symbolic value> определяется ее коэффициентом достоверности или функцией принадлежности. При интерпретации правил <Consequent> наследует свой коэффициент достоверности от <Antecedent >. Коэффициент достоверности конъюнкции вычисляется с использованием сигмоидной функции от суммы коэффициентов достоверности членов конъюнкции. Такая модель правила подобна модели нейрона в нейронных сетях прямого распространения. Так что нет различий в представлении знаний в нейронной сети и правилах, как в некоторых гибридных моделях [Ярушкина, 2004]. Таким образом, можно легко переводить знания, сформированные в обученной нейронной сети в разряд априорных (неизменных).

Три закона робототехники

Разработка человеко-подобных (с точки зрения поведения и мышления) роботов чревата проблемами при их взаимоотношениях с человеком, особенно если по своим психо- и

физическим качествам они будут нас в какой-то степени превосходить. Об этой проблеме писал еще Карел Чапек – родоначальник слова «робот». Хорошо известны три закона робототехники, предложенные А. Азимовым [Азимов, 2004] для обеспечения дружественного поведения роботов:

1. Робот не может причинить вреда человеку или своим бездействием допустить, чтобы человеку был причинён вред.
2. Робот должен выполнять приказы человека в той мере, в которой это не противоречит Первому Закону.
3. Робот должен заботиться о своей безопасности в той мере, в которой это не противоречит Первому и Второму Законам.

Предположим, что наш классификатор обеспечивает распознавание следующих трех классов ситуаций:

1) *Danger_for_man* – опасность для человека,
2) *Danger_for_myself* – опасность для себя (робота),
3) *Command_executed* – успешное выполнение текущей команды (процесс достижения текущей цели).

Каждый из этих трех распознаваемых классов может рассматриваться как нечеткая переменная или лингвистическая переменная с соответствующими двумя символьными значениями, означающими истинность или ложность, описанные функцией принадлежности. В обоих случаях можно их представить переменной, принимающей значения из диапазона (0, 1). Это означает, что это могут быть три выхода многослойной нейронной сети прямого распространения, и их значения могут интерпретироваться как значения коэффициента достоверности. При этом можно предположить, что истинность распознавания класса *Command_executed*, т.е. значения данного выхода должно возрастать по мере выполнения плана (приближения к решению поставленной человеком задачи).

Предположим, что действие (<Action>) в грамматике может порождать положительные и отрицательные эмоции, т.е. мы имеем дело с действиями, соответственно “Positive_emotion” и “Negative_emotion”, соответственно.

Тогда законы робототехники можно записать как априорное знание в следующем виде:

If *Danger_for_man* then Negative_emotion;

If *Danger_for_myself* and not *Danger_for_man* then Negative_emotion;

If not *Command_executed* and not *Danger_for_man* and not *Danger_for_myself* then
Negative_emotion;

If *Command_executed* and not *Danger_for_man* and not *Danger_for_myself* then

Positive_emotion;

Следует иметь в виду, что одинаковые решения на выходах нескольких правил могут объединяться по правилам нечеткой логики. Кроме того, надо иметь в виду, что в отличие от оригинальных законов А.Азимова наши законы не непосредственно управляют поведением, а опосредованно через эмоции.

Если мы хотим усилить эти законы, можно добавить правила, отражающие динамику в изменении исходных лингвистических переменных с использованием функций Increase и Decrease.

If Decrease(*Danger_for_man*) then Positive_emotion;
If Decrease(*Danger_for_myself*) and not *Danger_for_man* then Positive_emotion;
If Increase(*Command_executed*) and not *Danger_for_man* and not *Danger_for_myself*
then Positive_emotion;

Конечно, этот механизм реализации законов робототехники не гарантирует дружественное поведение, т.к. все зависит от качества обучения классификации «хороших» и «плохих» ситуаций, так же как и в случае с людьми. Ведь никакие моральные или религиозные принципы не избавляют человечество от насилия и других проявлений «нехорошего» поведения. Это еще связано и с относительностью таких понятий как «хорошо» и «плохо». Сам автор законов робототехники во многих своих произведениях анализирует случаи их неработоспособности.

Для большей надежности можно на верхнем уровне экспертом задать правила классификации, расширив априорные знания вниз. В этом случае можно говорить о двух уровнях представления знаний в интеллектуальной системе – логическом (в виде правил) и ассоциативном (в виде нейронных сетей), т.е. мы имеем дело с «двухполушарной экспертной системой» [Гаврилов, 2003]. При этом можно предусмотреть перевод твердо укоренившихся качественных (с точки зрения дружественного поведения) ассоциаций в разряд правил, извлекая знания из обученной нейронной сети, используя известные методы.

Обучение нейронных сетей прямого распространения с использованием обратной связи в виде базовых эмоций можно осуществлять с использованием предложенного автором модифицированного алгоритма обратного распространения ошибки, который позволяет обучать многослойный персептрон как на положительных так и на отрицательных примерах [Gavrilov, Lee, 2007, 2]. При этом сигнал обратной связи (базовая эмоция) запускает процесс

перестройки весов, рассматривая выходной вектор как положительный или отрицательный пример в зависимости от знака эмоции.

Заключение

В статье рассматривается взаимосвязанное представление о роли и моделировании эмоций, о представлении априорных знаний и обеспечении дружественного поведения искусственного интеллекта. Предлагается простой язык для представления априорных знаний в виде правил и нечетких переменных, легко согласующийся с использованием нейронных сетей для предварительной обработки информации с целью формирования этих нечетких переменных. Предлагается обеспечивать дружественное поведение искусственного интеллекта, запрограммировав как априорное знание три закона робототехники А.Азимова, продуцирующих базовые эмоции (отрицательную или положительную), используемые в качестве обратной связи для обучения робота воспринимать окружающий мир и планировать свое поведение.

Список литературы

[Азимов, 2004] А. Азимов. Мечты роботов. М.: Эксмо, 2004.

[Арбиб, 1976] М. Арбиб М. Метафорический мозг. - М.: Мир, 1976.

[Гаврилов, 2003] Гаврилов А.В. Гибридные интеллектуальные системы. - Новосибирск: НГТУ, 2003. - 162с.

[Гаврилов и др., 2004] А.В. Гаврилов, В.В. Губарев, К.-Х. Джо, Х.Х. Ли. Архитектура гибридной системы управления мобильного робота. – Мехатроника, автоматизация, управление, 2004, №8. – С. 30-37.

[Джеймс, 1911] В. Джеймс. Психология. Часть II, СПб: Изд-во К.Л. Риккера, 1911.

[Добрынин, 2006] В.А. Добрынин. Интеллектуальные роботы вчера, сегодня и завтра. Труды конференции КИИ-2006, Дубна, 2006.

[Колесников, 2001] Колесников А.В. Гибридные интеллектуальные системы: Теория и технология разработки. – СПБ: Изд-во СПбГТУ, 2001. – 711с.

[Колесников, Кириков, 2007] Колесников А.В., Кириков И.А. Методология и технология решения сложных задач методами функциональных гибридных интеллектуальных системы. – М.: ИПИ РАН, 2007. – 387 с.

[Прибрам, 1971] K.H. Pribram Языки мозга. Prentice Hall/Brandon House, N.Y., 1971.

[Рейковский, 1979] Я. Рейковский. Экспериментальная психология эмоций. М., 1979.

[Хейес-Рот, Уотерман, Ленат, 1987] Построение экспертных систем. Под ред. Ф. Хейес-Рота, Д. Уотермена, Д. Лената. - М.: Мир, 1987.

[Ярушкина, 2004] Ярушкина Н.Г. Основы теории нечетких и гибридных систем: Учеб. пособие.- М.: Финансы и статистика, 2004.- 320 с.

[Gavrilov, 2008] A.V. Gavrilov. Emotions and a priori Knowledge Representation in Artificial General Intelligence. Int. Conf. on Intelligent Information and Engineering Systems. Bulgaria, Varna, June 23-July 03, 2008.

[Gavrilov, Lee, 2007, 1] A.V. Gavrilov, S,-Y. Lee. Usage of Hybrid Neural Network Model MLP-ART for Navigation of Mobile Robot. Proceedings of International Conference on Intelligent Computing ICIC'07, China, August, 2007, LNAI 4682. Springer-Verlag, Berlin, Heiderberg, 2007, 182-191.

[Gavrilov, Lee, 2007, 2] Andrey Gavrilov, Sungyoung Lee. *Unsupervised hybrid learning model (UHLM) as combination of supervised and unsupervised models*. Proc. of IEEE Int. Conf. on Cybernetic Systems SMC UK&RI , Dublin, 6-7 September, 2007.

[Goerke, 2006] N. Goerke. EMOBOT: A robot Control Architecture based on Emotion-like Internal Values. In book: Mobile Robots, Moving Intelligence (eds. J. Buchli). ARS/pIV, Germany, 75-94.

[Goertzel, 2006] Ben Goertzel. Patterns, hypergraphs and embodied general intelligence. Proceedings of International Joint Conference on Neural Networks (IJCNN 2006), July, 16-21, 2006, Vancouver, BC, Canada., (2006), 451-458.

[Goertzel, Pennachin, 2007] Ben Goertzel and Cassio Pennachin. Artificial General Intelligence. Springer-Verlag, Berlin, Heidelberg, 2007.

[Hawkins and Blakeslee, 2004] J. Hawkins and S. Blakeslee, On Intelligence, Times Books, 2004.

[Hecht-Nielsen, 2006] R. Hecht-Nielsen. The mechanism of thought. Proceedings of International Joint Conference on Neural Networks (IJCNN 2006), July, 16-21, 2006, Vancouver, BC, Canada., (2006), 1146-1153.

[Hoya, 2005] Tetsuya Hoya. Artificial Mind System. Kernel Memory Approach. Springer, Berlin, 2005.

[Medsker, 1995] L. R. Medsker, Hybrid Intelligent Systems, Kluwer Academic Publisher, 1995.

[Minsky, 2006] M. Minsky. The emotion machine. Simon&Shuster, New York, 2006.

[Picard, 1997] R. Picard. Affective Computing. MIT Press, 1997.

[Sutton, Barto, 1998] Sutton R.S., Barto A.G. Reinforcement learning: An introduction. – MIT Press, Cambridge, MA, 1998.

[Tarik-Ul Islam, A.V.Gavrilov et al, 2008] Kh. Tarik-Ul Islam, Jehad Sarkar, Kamrul Hasan, M. Rezwanul Huq, **Andrey V. Gavrilov**, Young-Koo Lee, Sungyoung Lee. A Framework of Smart Objects and their Collaboration in Smart Environment. The 10th Int. Conf. on Advanced Communication Technology ICACT-08, Phoenix Park, Korea, Febrary, 2008. – Pp. 852-855.

[P. Wang, 2006] P. Wang, Rigid Flexibility: The Logic of Intelligence, Springer, 2006.

[Yingxu Wang, Ying Wang, 2006] Yingxu Wang, Ying Wang. Cognitive Informatics Models of the Brain. IEEE Trans. on Systems, Man, and Cybernetics — Part C: Applications and Reviews, 36(2), 2006, 203-207.

[Wermter, Sun, 2000] Stefan Wermter, Ron Sun, Hybrid Neural Systems. Springer-Verlag, Heidelberg, Germany. 2000.

[Zadeh, 1975] L.A. Zadeh, The concept of linguistic variable and its application to approximate reasoning, Inform. Sci. I: 8, (1975), 199-249.