ИСПОЛЬЗОВАНИЕ АЛГОРИТМОВ ПРЕДСКАЗАТЕЛЬНОГО МОДЕЛИРОВАНИЯ ДЛЯ ПРОГНОЗИРОВАНИЯ УРОВНЯ ПОДГОТОВКИ УЧИТЕЛЯ ИНФОРМАТИКИ

Симченко Н. Н., канд. пед. наук, доцент
Оренбургский государственный университет

В современном обществе накоплено большое количество данных, которые подвергаются анализу. Кроме того, в огромной объеме информации, которую человек не в силах исследовать самостоятельно, содержатся знания. Для обнаружения «скрытых» знаний применяется специальные методы автоматического анализа, при помощи которых приходиться практически добывать знания из «завалов» информации. За этим направлением закрепился термин «добыча данных (Data Mining)» или «интеллектуальный анализ данных», который применяется для реализации масштабных аналитических проектов в бизнесе, маркетинге, интернете, телекоммуникациях, промышленности, геологии, медицине, фармацевтике и других областях.

Предсказательное моделирование основано на подходе Data Mining, что делает его наиболее полезным в ситуациях, когда:

- пользователь имеет дело с многомерной проблемой: есть множество факторов, оказывающих влияние на объект анализа;
- в данных имеются пропуски или неверно заполнены поля;
- не совсем понятно, подходят ли имеющиеся данные для анализа (первичная оценка данных);
- требуется быстрый наглядный результат, поскольку пользователь не владеет навыками настройки модели и ее интерпретации;
- решение нужно «день-в-день»;
- желательно проанализировать все имеющиеся данные (без лимита на число переменных).

Довольно часто смоделировать ту или иную ситуацию с использованием обычных вычислительных средств крайне сложно, например, если необходимо сопоставлять слишком много критериев или данных между собой. Тут на помощь и приходят методы компьютерного моделирования. Процесс моделирования начинается с «загрузки» в компьютер необходимых данных. На основе этих сведений создается и настраивается компьютерная модель – виртуальный образ ситуации. Затем в дело вступают эксперты, которые формулируют к образу всевозможные сценарии его развития. Это не значит, что они предсказывают саму ситуацию, – они лишь определяют условия, в которых ситуация будет предположительно развиваться.

Для того, чтобы подчеркнуть, что целью компьютерного моделирования является предсказание (прогноз, оценка) характеристик проектируемого объекта, в последнее время часто используется термин «предсказательное моделирование». Компьютерные системы предсказательного моделирования

(называемые также системами поддержки принятия инженерных решений) вместе с компьютерными системами проектирования давно используются для автоматизации труда инженера-проектировщика и повышения качества принимаемых решений. В рамках данного исследования была поставлена задача проанализировать зависимость уровня подготовки учителя от места проживания, категории, стажа работы и образования.

Для этого была спроектирована структура данных, основанная на результатах тестирования, которое было проведено для определения уровня предметной составляющей профессиональной компетенции учителя информатики. Для контроля знаний в тест были включены вопросы двух уровней сложности: низкий и высокий.

В качестве выходного определен атрибут –Уровень подготовки учителя, показывающий уровень подготовки учителя, принимающий значение «высокий», если учитель, верно ответил на вопросы высокого уровня сложности и значение «низкий», если учитель ответил верно на вопросы на распознавание, которые представлены тестами с выборочными ответами. Содержание тестовых вопросов отражало все содержательные линии школьного предмета «Информатика» в полном соответствии с ФГОС полного (среднего) образования. 60% заданий аналогичны заданиям части «В» Единого государственного экзамена по информатике, 40% заданий предусматривали проверку теоретических знаний учителя по темам, которые недостаточно полно представлены в ЕГЭ, например, компьютерным сетям. Для обширного контроля знаний в тест были включены вопросы двух уровней сложности: низкий и высокий.

Низкий уровень – вопросы на распознавание, которые представлены тестами с выборочными ответами. Они представлены заданиями с кратким ответом

Высокий уровень – вопросы на воспроизведение знаний и на применение при решении нетиповой или измененной задачи. Они представлены заданиями с кратким ответом (задания на вычисление определенной величины; задания на установление правильной последовательности, представленной в виде строки символов по определенному алгоритму) и в виде заданий с открытым (развернутым) вариантом ответа.

Для анализа зависимости уровня подготовки учителя от места проживания, категории, стажа работы и образования, были определены атрибуты:

1) Район/Город – место проживания учителя, возможные значения: областной центр, город, районный центр, село.

2) Категория – уровень квалификации педагога. Возможные значения: нет, вторая, первая, высшая.

3) Образование – уровень подготовки. Возможные значения: профильное (педагогическое высшее (учитель информатики), непрофильное-педагогическое высшее (учитель другой дисциплины, педагог-психолог и другое).

4) Стаж работы. Возможные значения6: >20, <5, 5-10, 11-20.

В качестве выходного определен атрибут –Уровень, показывающий уровень подготовки учителя, он будет принимать значение «высокий», если учитель, верно ответил на вопросы на воспроизведение знаний и на применение при решении нетиповой или измененной задачи и принимать значение «низкий», если учитель ответил верно на вопросы на распознавание, которые представлены тестами с выборочными ответами. Возможные значения: низкий, высокий.

Для построения дерева решений был рассмотрен следующий фрагмент хранилища данных, полученный экспериментально (таблица 1).

Таблица 2.1 – Хранилище данных

Образование	Район/Город	Категория	Стаж	Уровень
Профильное	Город	Высшая	>20	высокий
Профильное	Город	Высшая	>20	высокий
Профильное	Город	Высшая	<5	высокий
Профильное	Город	Первая	5-10	высокий
Профильное	Город	Нет	5-10	высокий
Непрофильное	Город	Первая	<5	низкий
Профильное	Город	Первая	11-20	высокий
Непрофильное	Город	Первая	<5	низкий
Профильное	Город	Первая	5-10	высокий
Непрофильное	Город	Первая	<5	низкий
Непрофильное	Город	Нет	<5	низкий
Непрофильное	Город	Нет	<5	низкий
Непрофильное	Город	Вторая	<5	низкий
Профильное	Областной_центр	Высшая	>20	высокий
Непрофильное	Областной_центр	Первая	<5	низкий
Непрофильное	Районный_центр	Нет	<5	низкий
Профильное	Районный_центр	Первая	5-10	низкий
Непрофильное	Районный_центр	Нет	<5	низкий
Профильное	Село	Первая	5-10	высокий

Для проведения предсказательного моделирования был выбран алгоритм ID3. Алгоритм строит такое решающее дерево, в котором с каждым узлом ассоциирован атрибут, являющийся наиболее информативным среди всех атрибутов, еще не рассмотренных на пути от корня дерева. В качестве меры информативности обычно используется теоретико-информационное понятие энтропии. Для того, чтобы получить более оптимальное дерево принятия решений, нужно на каждом шаге выбирать атрибуты, которые «лучше всего» характеризуют целевую функцию [1].

Предположим, что имеется множество A из n элементов, m из которых обладают некоторым свойством S. Тогда энтропия множества A по отношению к свойству S вычисляется по формуле (1):

$$H(A,S) = -\frac{m}{n} log_2 \frac{m}{n} - \frac{n-m}{n} log_2 \frac{n-m}{n} \quad (1)$$

То есть энтропия зависит от пропорции, в которой разделяется множество. По мере возрастания этой пропорции от 0 до ½ энтропия тоже возрастает, а после ½ – симметрично убывает.

Если свойство S – не бинарное, а может принимать s различных значений, каждое из которых реализуется в m_i случаях, то энтропия обобщается естественным образом, формула (2):

$$H(A,S) = -\sum_{i=1}^{s} \frac{m_i}{n} \log \frac{m_i}{n} \quad (2)$$

Понятие энтропии тесно связано с теорией информации. Грубо говоря, энтропия – это среднее количество битов, которые требуются, чтобы закодировать атрибут S у элемента множества A. Если вероятность появления S равна ½, то энтропия равна 1, и нужен полноценный бит; а если S появляется не равновероятно, то можно закодировать последовательность элементов множества A более эффективно.

При выборе атрибута для классификации нужно выбрать его так, чтобы после классификации энтропия стала как можно меньше (свойство S в данном случае – значение целевой булевской функции). Энтропия при этом будет разной в разных потомках, и общую сумму нужно считать с учетом того, сколько исходов осталось в рассмотрении в каждом из потомков. Общепринятое в теории деревьев принятия решений определение выглядит следующим образом:

Предположим, что множество A элементов, некоторые из которых обладают свойством S, классифицировано посредством атрибута Q, имеющего q возможных значений. Тогда прирост информации (information gain) определяется формулой (3).

$$Gain(A,Q) = H(A,S) - \sum_{i=1}^{q} \frac{|A_i|}{|A|} H(A_i,S) \quad (3)$$

где A_i – подмножество элементов множества A, на которых атрибут Q имеет значение i.

Практическое применение классической реализации ID3 сталкивается с рядом проблем, характерных для моделей, основанных на обучении вообще и деревьев решений в частности. Основными из них являются переобучение и наличие пропусков в данных.

Например, если данные содержат шум, то число уникальных значений атрибутов увеличится, а в крайнем случае для каждого примера обучающего множества значения атрибутов окажутся уникальными.

Следуя логике ID3, можно предположить, что при разбиении по такому атрибуту будет создано количество узлов, равное числу примеров, так как в каждом узле окажется по одному примеру. После этого каждый узел будет объявлен листом, и дерево даст число правил, равное числу примеров обучающего набора.

Энтропия при этом будет разной в разных потомках, и общую сумму нужно считать с учетом того, сколько исходов осталось в рассмотрении в каждом из потомков. Кроме энтропии применяется величина *Gain,* показывающая количество информации, которое получают благодаря некоторому атрибуту. Алгоритм ID3 использует эту величину для оценки информативности атрибута при построении решающих деревьев, что позволяет получать деревья минимальной высоты.

Далее была произведена программная реализация алгоритма ID3, выбран язык программирования C#, среда разработки Microsoft Visual Studio 2012. Для построения дерева принятия решений разработан рекурсивный статический метод create_tree.

```
class item_tree
{
   public attributes a;
   public int num_attr;
   public int obuch = 0;
   public string res;
   public item_tree[] children;
   public bool[] enter=new bool[count];
}
```

На рисунке 1 продемонстрирована работа программы: выбор атрибута «Образование» в качестве корневого и перебор ветвей полученного узла.

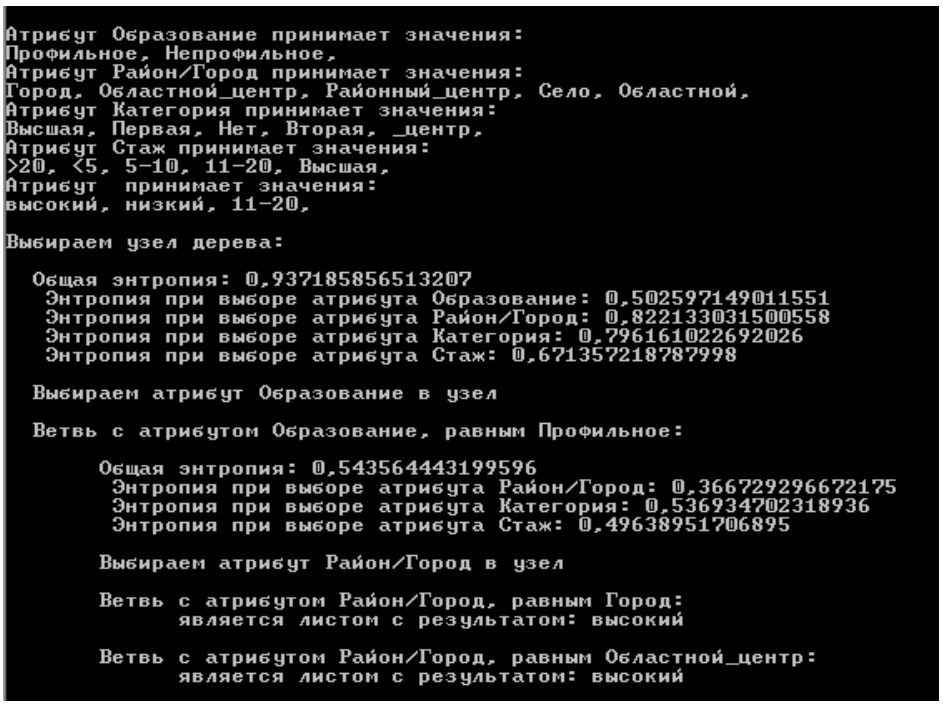

Рисунок 1 – Выбор атрибута, наиболее уменьшающего энтропию

По окончании построения дерева принятия решений, можно спрогнозировать уровень, определив район/город, категория, стаж. Результаты такого прогноза представлены на рисунке 2.

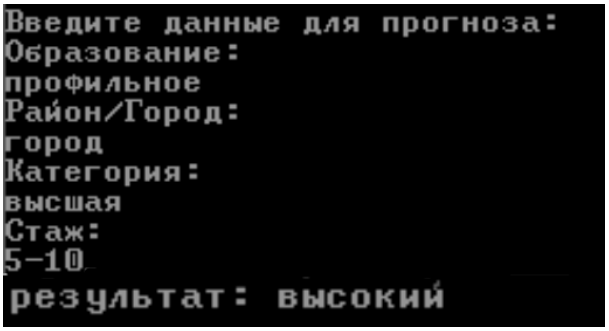

Рисунок 2– Результат прогноза

В результате работы программы на приведенном в таблице 1 фрагменте хранилища данных, было построено дерево принятия решений. Анализ построенного дерева принятия решений, показывает, что на уровень подготовки учителя наибольшее влияние оказывает образование (профильное/непрофильное). Проанализировав ветвь «Профильное», было выявлено, что наиболее информативным является атрибут «Стаж», если категория «Высшая», то «Уровень» – «Высокий», если «Вторая», то «Низкий».

Таким образом был сделан вывод, что для того, чтобы уровень подготовки учителя был высоким необходимы средства повышения квалификации, т.к. меняющаяся ситуация в системе общего образования

формирует новые образовательные потребности педагогов. Учитель постоянно находится между практикой и теорией, наращивая свой опыт преимущественно практическими умениями. Любая педагогическая работа – это практическая деятельность. Часто бывает так, что между теоретическими знаниями и практическими умениями продолжает сохраняться серьёзный разрыв. Преодолеть этот разрыв в современной школе можно средствами профессиональной переподготовки. Человека с современным мышлением, способного успешно самореализоваться в жизни, могут только педагоги, обладающие высоким профессионализмом. Повышение квалификации помогает учителю избавиться от устаревших взглядов, делает его более восприимчивым к внешним изменениям, что в конечном итоге повышает его конкурентоспособность.

Список литературы

1. *Вагин, В. Н. Достоверный и правдоподобный вывод в интеллектуальных системах / В. Н. Вагин., Е. Ю. Головина, А. А. Загорянская, М. В. Фомина // Москва.: ФИЗМАТЛИТ, 2004. – 704 с. -*
2. *Вьюгин, В.В. Математические основы теории машинного обучения и прогнозирования / В. В. Вьюгин. – Москва: МЦНМО, 2013. — 390 с.*
3. *Осипов, Г. С. , Методы искусственного интеллекта / Г .С. Осипов.– Москва: Физматлит, 2011. - 296 с.*