ПРЕДСКАЗАНИЕ РАЗМЕЩЕНИЯ НОВЫХ ФАЙЛОВ В ДРЕВОВИДНОЙ ФАЙЛОВОЙ СТРУКТУРЕ СРЕДСТАВМИ DATA MINING

Азеев А. А.,
научный руководитель старший преподаватель Бархатов А. В.
Сибирский федеральный университет
Институт математики

Целью данной работы было исследование возможности применения алгоритмов интеллектуального анализа данных (Data Mining) для предсказания размещения новых файлов в древовидной файловой структуре компьютера. Зачастую у пользователей компьютеров имеются свои принципы распределения файлов по структуре каталогов, поэтому эффективное предсказание размещения новых файлов может иметь не только теоретический, но и практический интерес. Потому, кроме собственно исследования эффективности алгоритмов Data Mining, было решено разработать программный инструмент для автоматизации процесса размещения файлов по каталогам пользователя.

Технология Data Mining — это процесс выделения из данных неявной и неструктурированной информации и представления ее в виде, пригодном для использования.

Классификация является одной из основных задач Data Mining. В результате решения задачи классификации обнаруживаются признаки, характеризующие группы объектов исследуемого набора данных — классы. По этим признакам новый объект можно отнести к тому или иному классу.

Поставленная задача в данной работе является фактически задачей классификации. В роли классов можно взять каталоги файловой структуры, объектом выступает файл директории. Признаками, характирзующими файлы каждой директории, являются такие свойства файлов, как расширение, размер, слова из содержимого и названия файла и т. д. Поэтому структура задачи позволяет использовать различные методы классификации.

Классификация происходит в два этапа: конструирование модели классификатора и её использование.

На первом этапе определяется набор данных (в нашем случае — существующие файлы и их параметры) и соответствующие им классы (в нашем случае — каталоги, в которых расположены файлы). Эти данные используются как обучающее множество, на нем происходит конструирование (обучение) модели. В данной работе используется три вида модели для классификации — классификационные правила, деревья решений и математические формулы.

На втором этапе на основе полученной обученной модели происходит классификация новых (неизвестных) данных.

Говоря о качестве классификации той или иной модели, используют термин «уровень точности», который вычисляется просто как процент правильно классифицированных данных. Уровень точности можно определить с помощью кросс-проверки. Для этого из исходного набора данных выделяют какую-то часть (например, 10% файлов), обучение выполняют на оставшихся данных (90%), а тестирование для определения точности проводят на выделенном наборе с известными значениями распределения по классам.

Если точность модели допустима, то возможно её использование для классификации данных, класс которых неизвестен.

Для построения конкретных классификаторов и разработки программы в данной работе была применена библиотека Weka на языке Java, разработанная в Университете Вайкато (Новая Зеландия). Были использованы следующие методы классификации этой библиотеки:

1) байесовская (наивная) классификация;

2) классификация с помощью деревьев решений (J4.8);

3) классификация при помощи метода ближайшего соседа (OneR).

Сама программа также реализована на языке Java и имеет графический интерфейс.

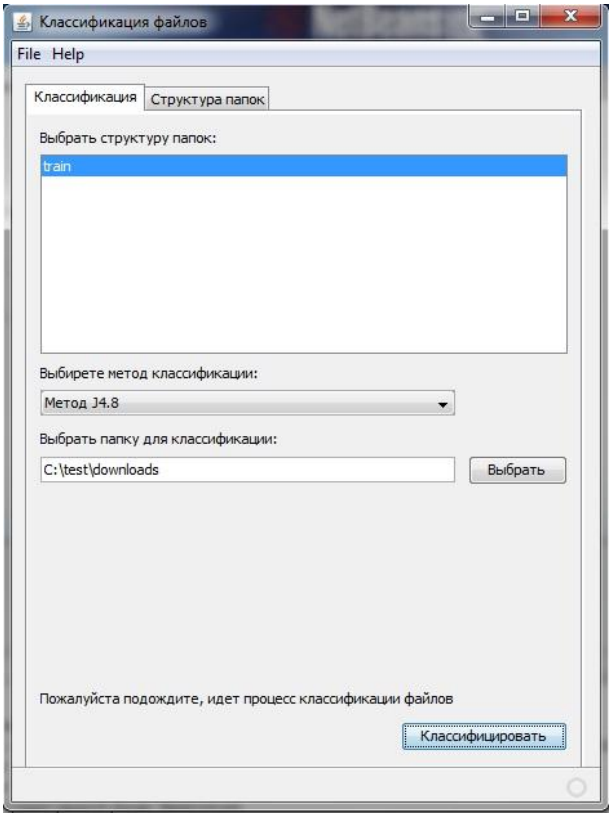

Рис. 1

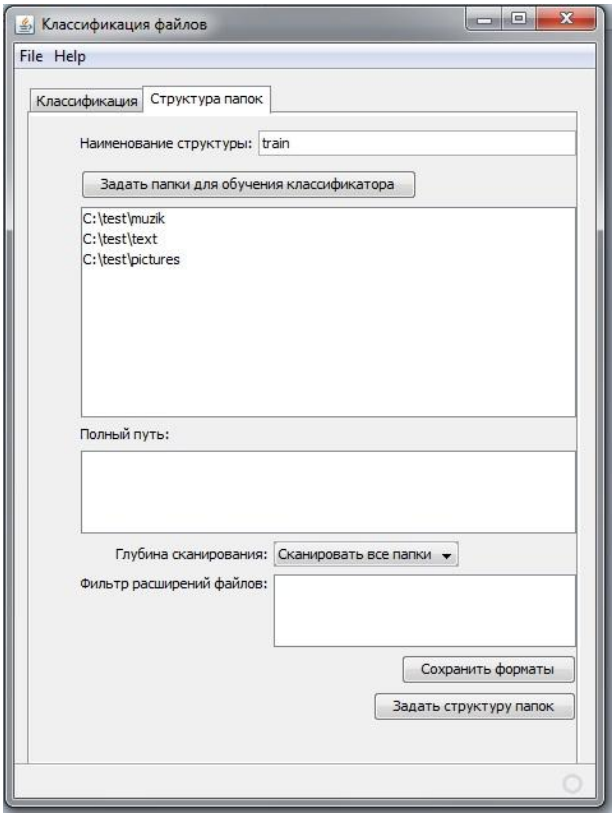

Рис. 2

На первой вкладке окна программы может выполнить классификацию всех файлов из указанного каталога (рис. 1). Настройка каталогов для обучения и сам процесс обучения выполняется с помощью второй вкладки (рис. 2).

После нажатия кнопки «Задать папки для обучения классификатора», появляется окно, в котором необходимо выбрать те каталоги, которые предназначены для обучения

классификатора. После выбора каталогов структура папок сохраняется и происходит обучение (кнопка «Задать структуру папок»).

После окончания процесса обучения на первой вкладке следует выбрать каталог, метод классификации, нажать кнопку «Классифицировать» и программа произведёт классификацию файлов по построенной (обученной) ранее модели. В результате программа выдаст предложения по распределению файлов по существующим каталогам (рис. 3).

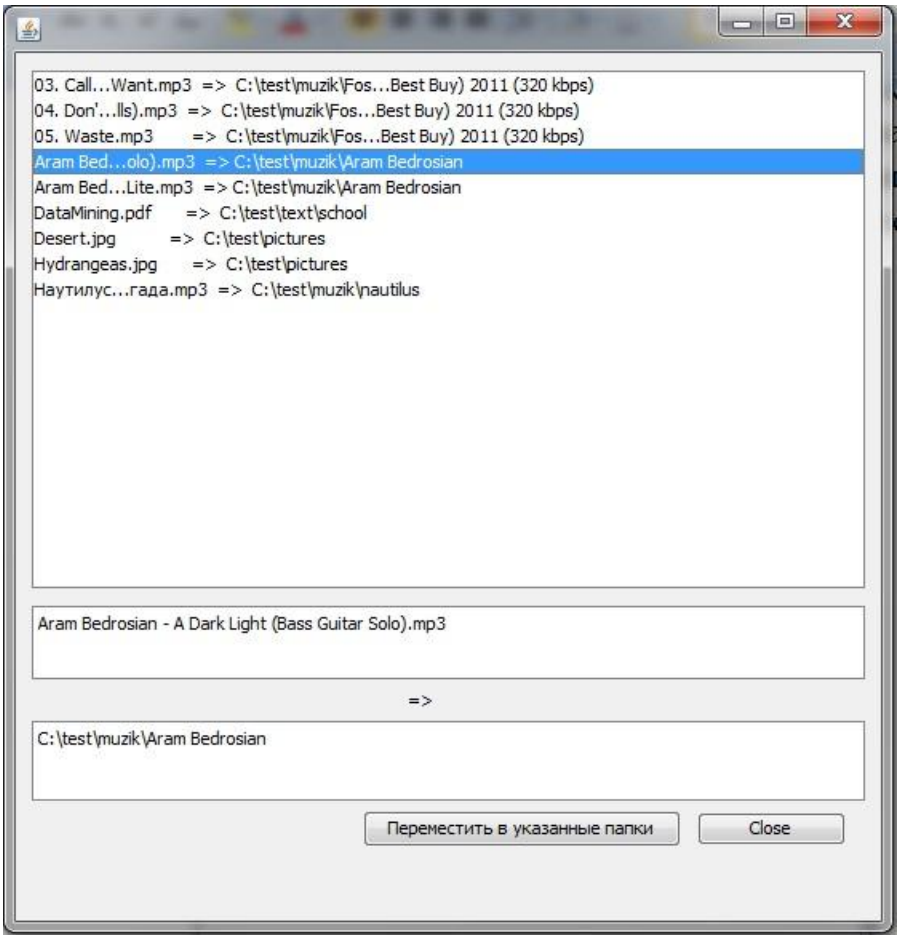

Рис. 3

Предварительные тесты показали, что выбранные методы классификации работают и могут использоваться на практике. Причём было замечено, что метод деревьев (J4.8) работает лучше остальных. Позже планируется провести более обширное тестирование и получить более конкретные показатели точности методов на различных наборах данных.