С.М. Гусакова, А.С. Комаров

Интеллектуальная система для решения идентификационной задачи в почерковедении

Аннотация. В статье рассматривается версия программы интеллектуального анализа почерковедческих данных, предназначенная для решения идентификационной задачи почерковедческой экспертизы. Интеллектуальный анализ данных опирается на ДСМ-метод автоматического порождения гипотез. Предлагается вариант ДСМ-метода, адаптированный для решения задач этого типа. Описывается экспериментальное исследование, проведенное при помощи предложенной модификации программы.

Ключевые слова: интеллектуальный анализ данных, ДСМ-метод, почерковедческая экспертиза, почерк, идентификация.

Введение

Задача криминалистического исследования почерка состоит в установлении исполнителя либо в установлении характеристик и особых состояний исполнителя или выполнения рукописи (пол, рост, возраст, профессиональные навыки и особенности, состояние алкогольного опьянения, выполнение рукописи в неудобной позе и т.п.). Задача установления исполнителя рукописи называется в почерковедении идентификационной задачей. Решение этой задачи – наиболее распространенный вид криминалистической экспертизы. Для идентификации исполнителя рукописи по его почерку наибольшее значение имеют особенности графического и технического навыков пишущего.

Существующие методики для решения идентификационной задачи основаны на выявлении идентификационнозначимых признаков. Выявлению этих признаков предшествует сбор обширного почерковедческого материала. Обработка этого материала существенно опирается на принятые в данный момент нормы прописи. В связи с этим созданные ранее методики в настоящее время не работают, а разработка новых методик, отвечающих современным условиям, не финансируется в силу своей дороговизны и организационной сложности. Поэтому актуальной является задача создания методов, позволяющих решать идентификационную задачу без разработки указанных методик. Это могут быть только компьютерные методы, берущие на себя сложную обработку материала. Основная идея состоит в том, что сложная обработка обширного почеркового материала для выделения немногих идентификационно значимых признаков для облегчения работы эксперта заменяется обработкой компьютером максимально полного множества признаков без различения идентификационной значимости. Возможность полного перебора доступная компьютеру позволяет автоматически выделить значимые признаки. Таким образом, компьютерный метод, который может оказаться полезным в решении данной задачи должен обладать комбинаторными возможностями и способностью воспроизводить рассуждения эксперта. Таким методом является ДСМ-метод - логико-комбинаторный метод, названный в честь Джона Стюарта Милля, некоторые рассуждения которого этот метод формализует [1].

Мы не будем давать подробное описание ДСМ-метода в общем виде. Такое описание применительно к атрибуционным задачам почерковедческой экспертизы можно найти в [2,3]. Отметим только несколько существенных моментов. ДСМ-метод автоматически выдвига-

ет гипотезы о наличии или отсутствии определенных свойств у объектов. Для того чтобы сформировать эти гипотезы, предварительно выдвигаются гипотезы о причинах обладания или не обладания объектами свойств.

Для того чтобы применение ДСМ-метода было возможно, необходимо наличие в базе фактов (БФ), содержащей объекты и их свойства как положительных, так и отрицательных примеров. (Этим, конечно, необходимые условия ДСМ-метода не исчерпываются). ДСМ-метод реализует синтез правдоподобных рассуждений – индукции, аналогии и абдукции. Индуктивное рассуждение заложено в проверке выполнения предиката сходства, для чего необходимо определение математически корректной операции существенного сходства [4] на объектах. Для этого объекты должны быть хорошо структурированы, т.е. в них должны быть выделены признаки, характеризующие объекты с точки зрения решаемой задачи. Рассуждение по аналогии проводится при выполнении правил правдоподобного вывода II–го рода, а абдуктивное рассуждение реализует проверку критерия достаточного основания. Конкретная реализация этих правдоподобных рассуждений будет рассмотрена для решения идентификационной задачи почерковедческой экспертизы.

1. Идентификационная задача

Идентификационная задача решается в следующей постановке. Имеется набор рукописных текстов, выполненных разными людьми. Для каждого человека один или несколько выполненных им текстов содержат все буквы алфавита, т.е. имеют в совокупности достаточный объем. Имеется также набор коротких рукописных текстов. Про исполнителей коротких текстов известно, что они входят в число исполнителей больших тестов, но кем и какой текст написан, неизвестно. Требуется определить авторов коротких текстов.

Прежде всего, необходимо выбрать язык представления данных и перевести тексты на этот язык. Для решения идентификационной задачи был выбран язык признаков почерка, принятых в криминалистике. С учетом сказанного выше, для представления данных использовалось как можно большее количество признаков, характеризующих почерк с различных сторон. В результате были использованы три группы признаков:

- общие признаки, характеризующие почерк в целом;
- частные признаки для строчных букв, характеризующие особенности их написания вплоть до элементов;
- частные признаки для прописных букв.

Признаки делятся на абсолютные и относительные. Абсолютные признаки определяются одним параметром, например, признак «размер почерка» определяется высотой букв. Значения относительных признаков зависит от нескольких параметров. Таков, например, признак «разгон почерка», который определяется как соотношение высоты и ширины букв. Кроме того, есть признаки объективные и субъективные. К объективным относятся признаки, значения которых строго определены, например, размер букв до 2 мм – малый размер почерка, от 2 мм до 4 мм – средний размер и более 4 мм – большой размер. Или разгон почерка – если ширина букв менее или равна половине высоты – малый, если ширина более высоты – большой, промежуточные значения дают значение - средний разгон почерка. Субъективные признаки определяются экспертом «на глаз». Такие признаки, как правило, не являются независимыми, и их значение определяют по значениям других признаков. Такова выработанность почерка, определяемая по темпу и координации. Темп в свою очередь, зависит от связности, а координация от форм выполнения линий. Впрочем, есть и независимые субъективные признаки, такие, например, как «преобладающая форма движения при выполнении письменных знаков». Некоторые общие и частные признаки могут тоже быть связаны между собой. Эта зависимость некоторых признаков от набора других дает возможность проверить правильность определения экспертом значений признаков. Причем сделать это можно тоже с помощью ДСМ-метода.

Идентификационная задача принадлежит к классу задач, в которых наличие определенного свойства у объектов объясняется не тем, что эти объекты имеют общие признаки, как в задачах решаемых классическим вариантом ДСМ-метода, а всей совокупностью признаков, которыми обладают объекты с данным свойством. В этом классе задач объекты содержат, как

правило, небольшое число признаков и разные объекты, обладающие общим свойством, могут не иметь ни одного общего признака. Для этого класса задач был разработан модифицированный вариант ДСМ-метода [5], который был успешно применен для решения задачи датировки берестяных грамот.

С целью проверки возможности решения идентификационной задачи почерковедческой экспертизы с помощью модифицированного варианта ДСМ-метода была проведена серия экспериментов. Для этого в Московском университете МВД России (МУ МВД России) были собраны образцы почерка слушателей.

При решении этой задачи эксперт имеет дело с рукописными текстами, как правило, небольшого размера, выполненными одним и тем же лицом. В каждом из этих текстов характерные особенности почерка встречаются не в полном объеме и только все вместе они дают достаточно точное представление о почерке данного лица. Поэтому для сбора образцов почерков был подобран специальный текст, в котором каждая буква алфавита встречалась не менее 10 раз (прописная буква – не менее 2-3раз). Текст выполнялся слушателями под диктовку в обычном для них темпе, в привычных условиях, с использованием привычных пишущих приборов – шариковых ручек. Объем текста, выполненного почерком среднего размера, составлял полный лист формата А4. Таким образом, объем текста и его буквенный состав обеспечили полноту представления признаков почерка, а способ его выполнения исключил возможность влияния посторонних факторов. Поэтому такую рукопись можно считать эквивалентной сумме нескольких коротких текстов, выполненных одним лицом. Экспертами-почерковедами МУ МВД России тексты были переведены на язык представления данных, т.е. в них были выделены все три группы признаков.

Таким образом, была сформирована база фактов, положительными примерами в которой являются образцы почерка (объекты) и лица (свойства), выполнившие данный образец, а отрицательными примерами являются образцы почерка и лица, этот образец не выполнявшие. Кроме того, в базу фактов вошли краткие образцы почерков, сведения об исполнителях, которых в БФ отсутствуют. Известно только, что исполнители кратких текстов выполняли и большой образец почерка. (На самом деле исполнители кратких текстов, конечно, известны, но эти сведения будут использованы при оценке результатов экспериментов).

Всего был собран 301 образец почерков – 293 больших и 8 кратких. Общие признаки были выделены у всех 301 образцов, у 177 образцов (169 большого объема и 8 малого объема) были выделены частные признаки почерка по строчным буквам и у 31 (23 большого объема и 8 малого объема) образцов почерка – частные признаки по прописным буквам.

2. Решение посредством ДСМ-метода

После того, как сформирована база фактов, можно начинать эксперименты с помощью ДСМ-метода. Была проведена серия экспериментов – определение авторов кратких записей по общим признакам, определение авторов кратких записей по частным признакам для строчных букв и определение авторов кратких записей по частным признакам для прописных букв.

Прежде чем применять ДСМ-метод, необходимо определить операцию сходства на объектах. Эта операция должна быть идемпотентной, коммутативной и ассоциативной (ассоциативность не является обязательным условием, но с точки зрения реализации, очень желательным). Кроме математической корректности, эта операция должна быть содержательна, т.е. выражать общность объектов, обладающих одним и тем же свойством в рамках модели предметной области и решаемой в этой области задачи.

Как уже говорилось, общность объектов (образцов почерка), обладающих одним свойством (принадлежащих одному и тому же лицу), выражается собирательно, т.е. всей совокупностью значений признаков, встретившихся в различных образцах данного почерка. Поэтому естественно операцию сходства определить как объединение всех значений этих признаков. Поскольку, как отмечалось выше, текст для большого образца почерка был подобран так, что он полно характеризовал почерк выполнившего его лица, мы можем приравнять его к сумме нескольких образцов, и результатом операции сходства будет полный набор всех значений признаков, встретившихся в данном образце.

Так формируется положительная гипотеза первого рода: $v \Rightarrow_2 w$, которая читается как:

«v – причина проявления свойства w». Здесь v – результат операции сходства. Конечно, содержательно говорить о том, что признаки почерка являются причиной того, что рукопись выполнена данным лицом, нельзя. Реально психофизиологические характеристики лица, средовые факторы формирования его почерка являются причиной того, что у него проявляются именно такие значения признаков. Точнее было бы сформулировать так: «данный набор значений признаков является причиной того, что мы приписываем данный почерк данному лицу». Для аккуратности будем называть v квазипричиной.

Следующим шагом ДСМ-метода является применение правил правдоподобного вывода, которые с помощью полученных квазипричин порождают по аналогии гипотезы II-го рода: $X \Rightarrow_1 w$ – объект X обладает свойством w.

Правдоподобное рассуждение по аналогии строится следующим образом. Если объекты, обладающие свойством w, содержат причину v, то и объект с неизвестным свойством, содержащий причину v, должен обладать свойством w при условии, что он не содержит никакой отрицательной причины (т.е. причины, полученной из отрицательных примеров). Поэтому естественно, что для проведения такого рассуждения определяется отношение вложения, связанное с операцией сходства соотношением: $v \subseteq x \rightarrow v \rho x = v$, где ρ - обозначение операции сходства.

Для операции сходства, определенной как объединение это соотношение принимает вид: $v \subseteq x \rightarrow v \cup x = x$, только реально вложение осуществляется не гипотезы в объект, а объекта в гипотезу.

Однако в задаче идентификации есть нюансы, которые требуют пересмотра рассуждения по аналогии. Дело в том, что в почерковедении существенную роль играют два понятия устойчивость признака и его вариативность. При проведении почерковедческой экспертизы особо выделяют значения признаков, встречающиеся неоднократно, т.е. проявляющие устойчивость. Вместе с тем, почерк обладает и вариативностью - один и тот же признак может иметь разные значения, т.е. буква или ее элемент выполняется по-разному. В кратких образцах может встретиться случайный вариативный элемент, который не проявился в большом образце, выполненным этим же лицом. Поэтому полного вложения значений признаков краткого образца в большой образец может не быть. С другой стороны, значения признаков почерка не являются абсолютно уникальными. Одни и те же значения встречаются в почерках разных лиц. В связи с учетом этих особенностей и того факта, что объект вкладывается в гипотезу, а не наоборот, рассуждение по аналогии формулируется следующим образом: «если объекты, обладающие свойством w, вкладываются в гипотезу v, то объект, имеющий максимальное по мощности пересечение по сравнению с пересечениями этого объекта с другими гипотезами, тоже обладает свойством w». Максимальность мощности пересечения объекта с гипотезой означает и квазивложение в положительную гипотезу и одновременно отсутствие такого квазивложения в отрицательные гипотезы.

Таким образом, правила правдоподобного вывода II-го рода имеют вид:

$|X \cap v| = \max [|X \cap v_1|, \ldots, |X \cap v_k|] \rightarrow J_{(1,n)} (X \Rightarrow_1 w)$;

$|X \cap v| \neq \max [|X \cap v_1|, \ldots, |X \cap v_k|] \;\&\; (\exists v_i (|(X \cap v_i| = \max [|X \cap v|, |X \cap v_1|, \ldots, |X \cap v_k|]) \rightarrow J_{(-1,n)} (X \Rightarrow_1 w)$;

$|X \cap v| = \max [|X \cap v_1|, \ldots, |X \cap v_k|] \;\&\; (\exists v_i (|X \cap v_i| = \max [|X \cap v|, |X \cap v_1|, \ldots, |X \cap v_k|]) \rightarrow J_{(0,n)} (X \Rightarrow_1 w)$.

Здесь v_1 , v_k - гипотезы для других свойств, J – оператор Россера-Тюркета [6], n – номер шага работы ДСМ-метода.

Следует обратить внимание на две особенности, возникающие при таком определении правил правдоподобного вывода II-го рода.

Во-первых, противоречивая гипотеза означает в данном случае не фактическое противоречие, т.е. коллизию положительной и отрицательной гипотез, а гипотезу вида: «данный краткий образец выполнен лицом с почерком v, либо лицом с почерком v_k ».

Вторая особенность заключается в том, что практически любой объект доопределится, даже если реально его исполнителя нет среди лиц, выполнявших большие тексты. Т.е. реально неопределенный объект скорее определится неверно, чем не определится совсем.

За основу системы для реализации модифицированного варианта ДСМ-метода была взята интеллектуальная система анализа почерковедческих данных для решения задач атрибуции в почерковедении [6]. Планируемые в рамках данного исследования компьютерные эксперименты потребовали от системы дополнительные функциональные возможности. Так что одной из целей стало создание новой версии

системы интеллектуального анализа криминалистических данных, удовлетворяющей всем поставленным требованиям.

При проведении экспериментов предполагалось использовать уже имеющиеся в системе образцы, но дополнить некоторые из них специально собранными краткими образцами почерка этих же людей. В системе уже реализована возможность вводить, просматривать и редактировать описания образцов почерка человека, но для одного человека допускалось хранить не более одного образца. Сейчас в системе возможна работа с несколькими образцами почерка для одного человека.

Расширение языка описания данных за счет добавления частных признаков почерка в прописных буквах также потребовало дополнительных модификаций уже существующего интерфейса ввода новых образцов. Во все интерфейсы, связанные с манипуляциями над образцами почерка, добавлена поддержка прописных букв.

Для проведения экспериментов по идентификации кратких документов был добавлен специальный интерфейс анализа данных, реализующий модифицированный вариант ДСМ-метода, рассмотренный в этой статье. Интерфейс позволяет провести за один раз идентификацию одного образца почерка.

Работа с данным интерфейсом происходит в несколько шагов. В первую очередь выбираются типы используемых при анализе признаков почерка. Далее на основе этого выбора формируется список образцов почерка, у которых описаны и введены в базу признаки всех указанных типов. На следующем этапе среди отобранных документов указывается краткая запись, чья идентификация будет производиться, и выбирается массив образцов, для обучающей выборки. Затем для указанных документов запускается процесс анализа данных и по его окончании выводится отчет о результате идентификации. В отчете содержится список всех документов, породивших гипотезы, сами гипотезы с указанием их мощности, а так же реализована возможность для каждой отдельной гипотезы посмотреть набор вошедших в нее значений признаков почерка. Интерфейс является достаточным инструментом для проведения предполагаемых экспериментов.

В рамках данного исследования были проведены три эксперимента по идентификации исполнителя краткой рукописи с использованием определенного типа признаков:

- общих признаков почерка,
- частных признаков почерка в строчных буквах,
- частных признаков почерка в прописных буквах.

Учитывая неполноту описания некоторых образцов почерка в базе данных, массивы пригодных для эксперимента документов отличаются в зависимости от выбранного типа признаков. Рассмотрим каждый эксперимент в отдельности.

Эксперимент по решению задачи идентификации исполнителей кратких записей на основе общих признаков почерка был проведен для 8 кратких образцов на массиве из 169 больших образцов почерка. В результате правильно был идентифицирован всего один краткий документ, для остальных 7 правильные парные документы оказались на низких позициях относительно мощности полученных гипотез об авторстве.

Для эксперимента по идентификации с использованием частных признаков почерка в *строчных* буквах оказались доступны 6 кратких образцов и 169 больших образцов почерка. В ходе эксперимента правильно идентифицирован был только один краткий документ (тот же, что и в первом эксперименте). Соответствующие парные образцы для оставшихся 5 документов, как и в первом случае, породили гипотезы с маленькой мощностью.

Эксперимент по идентификации с использованием частных признаков почерка в *прописных* буквах был проведен для 8 кратких образцов на массиве из 23 больших образцов почерка. В результате правильно идентифицированы оказались 7 документов, один документ был идентифицирован неверно. Подробное изучение этого документа и его парного образца с целью установить причину ошибки, выявило следующий факт. Для ошибочно идентифицированного документа его парный большой образец оказался выполненным некачественно. Некоторые прописные буквы, встретившиеся в кратком образце, фактически отсутствуют в большом образце почерка, поскольку они выполнены как строчные. Таким образом, большой образец оказался неполным, что привело к уменьшению потенциальной мощности гипотезы. При пополнении информации о почерке данного исполнителя проблема его ошибочной идентификации должна быть решена.

На основе результатов экспериментов, можно сделать вывод, что наибольшую информативность для решения задачи идентификации исполнителя краткой рукописи представляют частные признаки почерка в прописных буквах.

Подобные результаты объяснимы с позиций почерковедческой практики. Так как задача идентификации решалась для кратких документов, нужно учитывать тот факт, что в кратком образце могут встретиться варианты написания не всех букв, кроме того, буква может встретиться недостаточное количество раз для выделения устойчивого варианта ее написания. Количество вхождений строчной буквы в краткий образец может оказаться недостаточным для выделения устойчивого, характерного для почерка, варианта написания. Поэтому в кратком документе могут с большей вероятностью проявиться не свойственные почерку в целом варианты написания строчных букв, что в свою очередь может привести к его ошибочной идентификации. Этот факт и был подтвержден результатами эксперимента. Для идентификации образцов с большим объемом текста результат может оказаться лучше, но в почерковедческой практике в большинстве случаев решается задача именно идентификации кратких записей.

Объяснима и неинформативность общих признаков почерка для задачи идентификации. В рамках эксперимента мы анализировали краткий образец, сопоставляя его с полными образцами. Эти образцы относятся к разным типам документов и их общие признаки определяются в первую очередь типом документа, а уже потом особенностями почерка человека. Кроме того, на общие признаки в большей мере, чем на частные, оказывают влияние условия, в которых писался документ. Заявление или служебная записка всегда будут расположены на листе иначе, чем конспект лекции. Поэтому в различных образцах почерка одного человека, тем более образцах разного типа, могут проявляться разные общие признаки в зависимости от ситуации.

Прописные буквы из-за своей сравнительно высокой сложности написания считаются почерковедами наиболее индивидуальными для почерка человека и требуют для устойчивости малого количества вхождений, а значит, имеют большую информативность для идентификации автора рукописи в не зависимости от объема текста исследуемого образца.

Заключение

Подводя итог вышесказанному, можно утверждать, что решение задачи идентификации кратких записей с применением ДСМ-метода возможно. Для решения этой задачи целесообразно использовать частные признаки почерка в прописных буквах, так как они наименее всего зависят от ситуативных факторов, влияющих на почерк человека в конкретном документе.

Литература

1. Дж.С. Милль. Система логики силлогической и индуктивной. М.: Книжное дело, 1900 г.
2. Финн В. К. Синтез познавательных процедур и проблема индукции. / В кн. В.К. Финна «Интеллектуальные системы и общество». М.: URSS, 2006.
3. Аншаков О.М. "ДСМ - метод и модификационные исчисления" // Искусственный интеллект и принятие решений – М.: 2008 – выпуск 1 – с. 55-79
4. Гусакова С.М., Финн В.К. Сходство и правдоподобный вывод // Известия АН СССР. Сер. Тех. киб. – 1987 – выпуск 5 – с.42- 63.
5. Гусакова С.М. Логико-комбинаторный метод анализа исторических данных // Проблемы исторического познания. Сборник статей по материалам круглого стола. М.: ИВИ РАН, 2008 г.
6. Финн В.К. Правдоподобные рассуждения в интеллектуальных системах типа ДСМ // Итоги науки и техники. – Москва,1991.- Том 15.- с. 54-98
7. Комаров А.С. Интеллектуальный анализ данных в почерковедении: программная реализация // Вестник РГГУ. Серия Информатика. Защита информации. Математика. - М.: РГГУ, 2010 (в печати)

Гусакова Светлана Марковна. Старший научный сотрудник Всероссийского института научной и технической информации РАН. Окончила Московский государственный университет в 1968 году. Кандидат физико-математических наук. Автор 31 печатной работы. Область научных интересов: теория отношений, теория сходства, искусственный интеллект, ДСМ-метод, логико-комбинаторные методы, интеллектуальный анализ данных в почерковедении. E-mail: sgusakova@gmail.com.

Комаров Алексей Сергеевич. Программист в ИД «Первое сентября». Окончил Российский государственный гуманитарный университет в 2007 году. Автор 4 печатных работ. Область научных интересов: искусственный интеллект, ДСМ-метод, интеллектуальный анализ данных в почерковедении, логико-комбинаторные методы, нейронные сети, web-программирование. E-mail: alexskomarov@gmail.com.