УДК 004.622, 004.942

РАЗРАБОТКА МЕТОДА ВОССТАНОВЛЕНИЯ ДАННЫХ ГОРОДСКОГО КОНТЕКСТА С ИСПОЛЬЗОВАНИЕМ ГЕОПРОСТРАНСТВЕННОЙ ИНФОРМАЦИИ

Мишина М.Э. (Национальный исследовательский университет ИТМО),
Хрульков А.А. (Национальный исследовательский университет ИТМО)
Научный руководитель – к.т.н, директор Института дизайна и урбанистики
Митягин С.А. (Национальный исследовательский университет ИТМО)

Авторами предлагается метод восстановления данных городского контекста с учетом дополнительных сведений об окружении – соседних объектах. Анализ эффективности метода производится на данных, собранных для г. Санкт-Петербург.

Полная и точная информация о городских объектах является основой создаваемых интегрирующих систем управления городом и во многом определяет эффективность принимаемых решений. Однако параллельные процессы инвентаризации городского хозяйства и мониторинга параметров городской среды, осуществляемые различными ведомствами при ведении собственных бизнес-процессов, неизбежно приводят к ошибкам при объединении имеющихся разрозненных данных и порождают проблему неполноты и неточности используемой информации.

Наиболее часто проблема отсутствия полной и актуальной информации о городской среде освещается в разрезе социологических опросов. В других исследованиях вопрос обработки неполных данных рассматривается относительно определенных типов городских объектов или в отношении укрупненных показателей для всего города с учетом установленных паттернов. Результатов в данной области добились исследователи Международной лаборатории сетевого анализа Национального исследовательского университета «Высшая школа экономики», факультета компьютерных наук и телекоммуникаций Института транспорта и связи в Латвии и факультета инженерии и информационных технологий Технологического университета Сиднея.

Однако на данный момент разработанные научные методы восстановления недостающих данных не позволяют в полной мере учитывать специфику городских данных – наличие различных взаимосвязей между признаками, описывающими свойства городских объектов, в том числе их зависимость от геопространственного положения объекта.

Городские данные представляют информацию о объектах городской среды – элементах благоустройства, дорогах, зданиях, кварталах и др., включая сведения об их расположении и семантических свойствах, в которых часто наблюдаются различные взаимосвязи. Все объекты городской среды привязаны к местности через их географические координаты, агрегирующие в себе свойства объектов и формирующие городской контекст. Согласно «позиционному принципу», семантические свойства объектов зависят также от их положения в пространстве. Данный принцип отчетливо наблюдается в территориальном зонировании городов, определяющем, например, допустимую этажность и назначение здания в зависимости от его местоположения. Геопространственная информация, помимо явного представления в виде последовательности координат, после преобразований и расчетов выражается в новых производных семантических признаках, например, представляющих семантическую информацию об окружении объекта.

Учитывая существующую специфику предметной области, разработанный метод восстановления данных городского контекста основывается на выявлении оптимальной математической модели, наиболее точно описывающей взаимосвязи, существующие как в исходных, так и в производных признаках объектов. Расширение исходного пространства признаков осуществляется путем преобразования геопространственной информации. На данном этапе метод предполагает дополнение исходного набора признаков агрегированной информацией о соседних объектах.

Для каждого признака, содержащего неизвестные значения в исходном наборе данных, подбор оптимальной модели осуществляется при обучении по прецедентам различных алгоритмов (случайные леса, градиентный бустинг и др.). Настройка гиперпараметров применяемых алгоритмов для прогнозирования неизвестных значений признака производится по алгоритму решетчатого поиска с последовательным сокращением вдвое комбинаций гиперпараметров. Данный алгоритм основан на периодическом сокращении малоэффективных моделей и сохранении вычислительных ресурсов для более перспективных. Оценка моделей производится методом k-блочной перекрестной проверки. В качестве функций потерь для моделей регрессии используется среднеквадратическая ошибка, а для моделей классификации – логистическая функция ошибки.

Разработанный метод поддерживает восстановление многомерных недостающих данных, когда в исходном наборе значения отсутствуют более чем в одном признаке. В таком случае прогнозирование неизвестных значений производится итеративно для каждого признака, принимаемого за зависимую переменную. Поскольку не все машинные алгоритмы показывают корректные результаты при обучении и прогнозировании по неполным данным, предварительно, если пропуски в независимых переменных еще не были заполнены прогнозными значениями, на их местах инициализируются начальные значения – средневзвешенные по удаленности значениями других объектов из набора.

Для алгоритмов прогнозирования, имеющих стохастическую составляющую, прогнозирование неизвестных значений осуществляется несколько раз, а полученные результаты усредняются, Множественные вычисления позволяют учитывать неопределенность восстанавливаемых значений.

Тестирование разработанного метода осуществлялось на наборе данных, содержащем сведения о жилых зданиях г. Санкт-Петербург. Моделирование пропусков в собранных данных производилось путем случайного удаления значений в признаках. Степень поврежденности входного набора данных определялась процентом «посеянных» пропусков по каждому признаку и находилась в диапазоне от 10 до 90 процентов с шагом в 10 процентов.

Результаты проведенных экспериментов показали, что использование представленного метода позволяет восстанавливать значения с точностью выше 80 процентов при степени поврежденности входного набора данных до 30 процентов. При этом было установлено, что дополнение исходного набора данных производными признаками, полученными на основании геопространственной информации, повышает точность восстанавливаемых значений на 3-10 процентов. Следует отметить, что на данном этапе представленный метод не позволяет делать уверенных суждений относительно значений признаков для отдельно взятых объектов, однако дает возможность получать корректные агрегированные показатели по территориальным единицам и формулировать достоверные статистические выводы при выработке решений в области управления и оптимизации городских процессов.

В дальнейших исследованиях планируется провести эксперименты, направленные на повышение точности восстанавливаемых значений за счет выявления новых взаимосвязей, существующих в городских данных. Отдельным направлением работ будет являться анализ возможности использования разработанного метода для решения проблемы неточности информации о городских объектах путем пересчета всех имеющихся значений признаков.