УДК 004.021:004.051

МЕТОД ОЦЕНКИ ЭФФЕКТИВНОСТИ ФИЛЬТРАЦИИ ГЕОКООРДИНАТ

Головков А.А., Иванова Г.С.

Московский государственный технический университет имени Н.Э. Баумана, Москва, e-mail: alexander.golovkov1@gmail.com

Мобильные устройства в настоящее время активно используются в областях логистики, технического обслуживания и производства. Ключевые возможности геоинформационных систем (ГИС) в данных сферах позволяют выполнять мониторинг перемещений и координировать действия полевых сотрудников, что тесно связано со сбором, обработкой, хранением, передачей и визуализацией геоданных. Модели и методы фильтрации данных местоположения, ориентированные на мобильные операционные системы, зачастую основаны на обработке информации в реальном времени. Их проектирование и разработка, а также экспериментальная оценка эффективности представляют собой нетривиальную задачу. Данная статья посвящена разработке метода оценки эффективности фильтрации геокоординат. В работе рассмотрены траектории движения, построенные по геоданным перемещаемых мобильных устройств, и предложен метод оценки качества фильтрации, основанный на определении значения функции ошибки, которая сравнивает временную и пространственную идентичность реального маршрута и траектории движения после фильтрации. Одной из особенностей метода является линейное время вычисления функции ошибки, что позволяет применять ее в качестве целевой функции при тренировке моделей машинного обучения в задачах фильтрации геокоординат в том числе в переносимых мобильных устройствах.

Ключевые слова: геоданные, функция ошибки, фильтрация, ГИС, траектория движения, оценка эффективности

METHOD FOR EVALUATING EFFICIENCY OF GEODATA FILTERING

Golovkov A.A., Ivanova G.S.

Bauman Moscow State Technical University, Moscow, e-mail: alexander.golovkov1@gmail.com

Mobile devices are now actively used in the fields of logistics, maintenance and production. The key capabilities of geoinformation systems (GIS) in these spheres allow monitoring movements and coordinating the activity of field staff, which is closely related to the collection, processing, storage, transmission and visualization of geodata. Models and methods of filtering location data, oriented to mobile operating systems, are often based on real-time information processing. Their design and development, as well as the experimental evaluation of efficiency, is a non-trivial task. The article is devoted to the development of a method for evaluating the efficiency of geodata filtering. The motion trajectories, constructed from the geodata of moving mobile devices, are considered and a method for estimating the quality of filtering based on the determination of the error function value is proposed, which compares the temporal and spatial identity of the real route and the trajectory after filtration. One of the features of the method is the linear time for calculating the error function, which makes it possible to use it as an objective function in the training phase of machine learning models in geo-coordinate filtering challenges in portable mobile devices.

Keywords: geodata, error function, filtering, GIS, trajectory, evaluation of effectiveness

В настоящее время мобильные устройства широко используются в различных областях логистики, технического обслуживания и производства. Ключевые возможности геоинформационных систем (ГИС) в данных сферах позволяют выполнять мониторинг перемещений и координировать действия полевых сотрудников, что тесно связано со сбором, обработкой и хранением геоданных, поступающих с мобильных устройств [1–3]. Важность и актуальность задач обработки геолокационной информации обусловлена высокими требованиями к геоинформационным системам в различных областях применения [4–6].

Стратегия разделения ответственности по обработке геоинформации в ГИС, когда приоритет отдается мобильным приложениям, является предпочтительной. В этом случае существенно снижается нагрузка на серверную часть и уменьшается объем передаваемых данных, что в случае плохого соединения с сетью или полного отсутствия связи является ключевым фактором. Такая стратегия позволяет реализовать потоковую обработку данных в реальном времени на мобильном устройстве, что даст возможность получить высокоточные данные местоположения настолько быстро, насколько это возможно. Это, несомненно, важно для ГИС, которые предполагают немедленное реагирование на различные ситуации на основе данных местоположения.

Модели и методы фильтрации геоданных зачастую основаны на обработке информации в реальном времени [1, 2]. Их проектирование и разработка, а также экспериментальная оценка эффективности представляют собой нетривиальную задачу [7, 8]. Для оценки качества фильтрации обычно используют такую характеристику, как функцию ошибки, которую констру-

ируют исходя из оценки идентичности целевой функции (целевого, реального маршрута, Ground Truth) и предсказаний классификатора или траектории движения на выходе фильтрации, учитывая специфику предметной области, таким образом, чтобы ошибка была максимально достоверной для данной конкретной задачи [9–11]. Значение функции ошибки должно уменьшаться с увеличением качества работы модели. При этом решая задачу минимизации функции, возможно найти оптимальные параметры модели. В качестве функции ошибки наряду с евклидовой метрикой часто используют кросс-энтропию или функцию потерь Хьюбера.

Определение функции ошибки

Рассматривая задачу фильтрации геокоординат, целевой маршрут можно представить в виде ломаной линии, заданной вектором кортежей. Каждый кортеж состоит из набора параметров: времени, широты и долготы, которые определяют точку в трехмерном пространстве:

$$L_{groundTruth} = (\ldots, t_i, lat_i, lon_i, \ldots). \quad (1)$$

Траектория движения, полученная после фильтрации, также представляет собой ломаную:

$$L_{predicted} = (\ldots, t_j, lat_j, lon_j, \ldots). \quad (2)$$

Особенностью такого представления является то, что количество элементов в векторах $L_{groundTruth}$ и $L_{predicted}$ в общем случае может быть разным. Координаты широты $lat_i(lat_j)$ и долготы $lon_i(lon_j)$ могут быть также разными в обеих ломаных. Только по параметру $t_i(t_j)$ возможно однозначно определить две точки, соответствующие одному и тому же времени. По этому параметру существует некая неравномерная дискретизация координат: каждой единице времени t_k соответствует не менее одного кортежа t_k, lat_k, lon_k в векторах $L_{groundTruth}$ и *Lpredicted*. То есть возможны случаи, когда:

- $\exists t_k : \exists! t_k, lat_k, lon_k \in L_{groundTruth}, \nexists t_k, lat_{k2}, lon_{k2} \in L_{predicted}$;
- $\exists t_k : \nexists t_k, lat_{k1}, lon_{k1} \in L_{groundTruth}, \exists t_k, lat_k, lon_k \in L_{predicted}$;
- $\exists t_k : \exists! t_k, lat_{k1}, lon_{k1} \in L_{groundTruth}, \exists! t_k, lat_{k2}, lon_{k2} \in L_{predicted}$.

Задача оценки качества фильтрации сводится к сравнению идентичности двух ломаных $L_{groundTruth}$ и $L_{predicted}$ функцией ошибки $E_{loss} \in \mathbb{R}$ как по пространственным координатам lat_i и lon_i, так и по временным t_i. Чем больше различаются ломаные, тем больше должно быть значение ошибки E_{loss} и наоборот. В общем случае $E_{loss} \in [0;\infty)$. При этом известно, что ломаные изначально так расположены в пространстве, что $E_{loss} = \min$, так как всегда рассматривается одна и та же реальная траектория движения. Вектор $L_{predicted}$ в случае идеальной фильтрации должен совпадать с $L_{groundTruth}$, и $E_{loss} = 0$.

Идентичность ломаных в данном контексте следует понимать как степень соответствия каждой точки $L_{predicted}$ с соответствующей точкой в $L_{groundTruth}$. Степень подобия в отличие от идентичности не будет адекватной оценкой, поскольку ломаные могут быть схожи с высоким показателем подобия, однако расположены далеко друг от друга, при этом значение ошибки будет мало, что неверно.

Задача конструирования E_{loss} усложняется тем, что координаты точек на ломаных lat_i, lon_i и lat_j, lon_j – это геокоординаты в замкнутом пространстве, эллипсоиде, а не точки на плоскости. Расстояние от точки до прямой здесь определяется дугой. Перевод геокоординат в прямоугольную систему координат может добавить существенную погрешность определения значения функции ошибки.

Конструирование функции ошибки

Рассмотрим возможные решения задачи нахождения максимально достоверной функции E_{loss}.

1. Ковариационный анализ. Метод предполагает составление матрицы ковариации между распределенными величинами, которые являются координатами точек на двух ломаных $L_{predicted}$ и $L_{groundTruth}$. После составления матрицы анализируются коэффициенты корреляции. По их значениям возможно определить степень идентичности двух ломаных, однако метод применим только в случае одинаковой дискретизации точек на ломаных, что в данном случае неверно.

2. Анализ Фурье и сравнение спектров. Проблема здесь аналогична п. 1. Для ее разрешения возможно аппроксимировать дискретные функции $f_{t+dt} = (lat_i, lon_i)$ на интервалах времени $t + dt$ непрерывными функциями для каждой ломаной, выполнить передискретизацию по времени и сравнить непрерывные

функции ковариационным анализом, анализом Фурье или простым вычислением интеграла разницы функций на интервалах $t + dt$. Однако определение значения dt для максимально достоверной оценки представляет собой нетривиальную задачу. Дело в том, что ломаные могут описывать окружности на больших интервалах времени, и в таком случае аппроксимация не имеет смысла. На малых интервалах возможна большая погрешность вычисления ошибки E_{loss}.

Предложим три подхода, основанные на векторном представлении ломаных. Первый определяет E_{loss} как разницу сумм изменения углов направления по двум ломаным. Второй подход заключается в сдвиге всех векторов в начало координат и подсчете разницы общих площадей под двумя ломаными. Однако в обоих случаях оценка может быть недостоверна, так как возможно построить два различных трека с нулевой ошибкой E_{loss}. Третий подход предполагает попарное сравнение векторов, но здесь появляется проблема дискретизации.

Все рассмотренные подходы с высокой степенью недостоверности вычисляют ошибку E_{loss}. Для декомпозиции задачи попробуем выделить частные случаи анализа двух ломаных. Допустим, моменту времени t_k соответствуют две точки на обеих ломаных, то есть:

$$\exists t_k : \exists! t_k, lat_{k1}, lon_{k1} \in L_{groundTruth}, x_{k1} = \\ = \left(lat_{k1}, lon_{k1}\right)$$

и $\exists! t_k, lat_{k2}, lon_{k2} \in L_{predicted}, x_{k2} = \left(lat_{k2}, lon_{k2}\right)$.

Ошибка здесь очевидно пропорциональна расстоянию между двумя точками x_{k1} и x_{k2} в метрах:

$$E^1_{loss} \sim D\left(x_{k1}, x_{k2}\right), \qquad (3)$$

где D – функция расстояния между точками x_{k1} и x_{k2}.

Рассмотрим второй случай, когда

$$\exists t_k : \exists! t_k, lat_k, lon_k \in L_{groundTruth},$$
$$x_k = \left(lat_k, lon_k\right), \nexists t_k, lat_{k2}, lon_{k2} \in L_{predicted}$$

или

$$\exists t_k : \nexists t_k, lat_{k1}, lon_{k1} \in L_{groundTruth},$$

$$\exists t_k, lat_k, lon_k \in L_{predicted}, x_k = \left(lat_k, lon_k\right),$$

то есть для момента времени t_k существует только одна точка на одной из ломаных. Выполнение любого из условий сводится к одной и той же задаче, поэтому далее будем рассматривать точку t_k, lat_k, lon_k на одной ломаной ($L_{groundTruth}$ или $L_{predicted}$), для которой нет соответствующей точки на другой ломаной ($L_{predicted}$ или $L_{groundTruth}$). Ошибка здесь пропорциональна расстоянию от точки x_k на первой ломаной до точки на второй ломаной $\widetilde{x_l}$ в метрах, которая на самом деле отсутствует:

$$E^2_{loss} \sim D\left(x_k, \widetilde{x_l}\right), \qquad (4)$$

Эту точку целесообразно определить как местоположение на второй ломаной в момент времени t_k. За $\widetilde{x_l}$ можно взять существующую точку на второй ломаной, ближайшую к моменту времени t_k. Однако данный подход может существенно увеличить погрешность определения итоговой ошибки E_{loss}, так как ближайшая точка по времени может располагаться на значительном расстоянии от x_k. Для нивелирования погрешности следует учитывать разницу во времени, что представляет собой нетривиальную задачу.

Второй способ определения $\widetilde{x_l}$ заключается в нахождении на второй ломаной двух точек $x_l = \left(lat_l, lon_l\right)$ и $x_{l+1} = \left(lat_{l+1}, lon_{l+1}\right)$, предыдущей и следующей по отношению к моменту времени t_k. Такой подход привнесет в определение ошибки E_{loss} не только оценку пространственной идентичности ломаных, но и временной. Следует учитывать, что в начале и/или в конце трека на одной из ломанных может не быть точек x_l и/или x_{l+1}. В этом случае ошибка пропорциональна расстоянию до существующей точки x_l или x_{l+1}:

$$E^2_{loss} \sim D\left(x_k, x_l\right), \qquad (5)$$

$$E^2_{loss} \sim D\left(x_k, x_{l+1}\right). \qquad (6)$$

Если обе точки присутствуют на ломаной, имеем треугольник с геокоординатами x_l, x_{l+1} и x_k, и ошибка определяется как

$$E^2_{loss} \sim D\left(x_k, \widetilde{x_l}\right) = D'\left(x_k, x_l, x_{l+1}\right), \qquad (7)$$

где D' – функция расстояния между точкой x_k и дугой, составленной из точек x_l и x_{l+1}.

Рассмотрим подходы для определения функции D'.

1. Определение D' и $\widetilde{x_l}$ по скорости. В качестве $\widetilde{x_l}$ здесь рассматривается точка, которая находится на отрезке $\left(x_l, x_{l+1}\right)$, на расстоянии $d_{l,k}$ от x_l. При этом $d_{l,k}$ – расстояние, которое возможно преодолеть за время $t_{k-l} = t_k - t_l$ со скоростью spd_l. Время t_l здесь соответствует кортежу t_l, lat_l, lon_l, spd_l – моментальная скорость в точке x_l. Итоговая ошибка определяется расстоянием между x_k и $\widetilde{x_l}$:

$$E^2_{loss} \sim D\left(x_k, \widetilde{x_l}\right) = \\ = D'\left(x_k, x_l, x_{l+1}\right) = D\left(x_k, \widetilde{x_l}\right). \qquad (8)$$

Однако скорость spd_l может быть определена с большой погрешностью; движение в интервале времени от t_l до t_k может быть неравномерным, что предполагает учет ненулевого ускорения и/или градиентов ускорения, погрешность вычисления которых может лишь расти по отношению к погрешности определения spd_l. В результате накопленная ошибка для E^2_{loss} может быть значительна.

2. Определение D' как длины кратчайшего отрезка дуги от точки x_k до отрезка дуги (x_l, x_{l+1}). При этом необходимо рассмотреть случаи, когда:

● точка $\widetilde{x_l}$ на дуге (x_l, x_{l+1}), которая лежит на кратчайшем отрезке дуги, не лежит на отрезке дуги (x_l, x_{l+1}). В этом случае в качестве E^2_{loss} следует использовать расстояние до ближайшей точки:

$$E^2_{loss} \sim D\left(x_k, \widetilde{x_l}\right) = D'\left(x_k, x_l, x_{l+1}\right) = \min\left(D\left(x_k, x_l\right), D\left(x_k, x_{l+1}\right)\right). \quad (9)$$

● точка $\widetilde{x_l}$ на дуге (x_l, x_{l+1}), которая лежит на кратчайшем отрезке дуги, лежит на отрезке дуги (x_l, x_{l+1}). В этом случае ошибка пропорциональна кратчайшему расстоянию:

$$E^2_{loss} \sim D\left(x_k, \widetilde{x_l}\right) = D'\left(x_k, x_l, x_{l+1}\right) = D\left(x_k, \widetilde{x_l}\right). \quad (10)$$

3. Определение D' как разницы расстояний. Данный подход предполагает расчет D' как разницы суммы расстояний от x_l до x_k, от x_k до x_{l+1} и от x_l до x_{l+1}:

$$E^2_{loss} \sim D\left(x_k, \widetilde{x_l}\right) = D'\left(x_k, x_l, x_{l+1}\right) = D\left(x_l, x_k\right) + D\left(x_k, x_{l+1}\right) - D\left(x_l, x_{l+1}\right). \quad (11)$$

Иллюстрация приведена на рисунке.

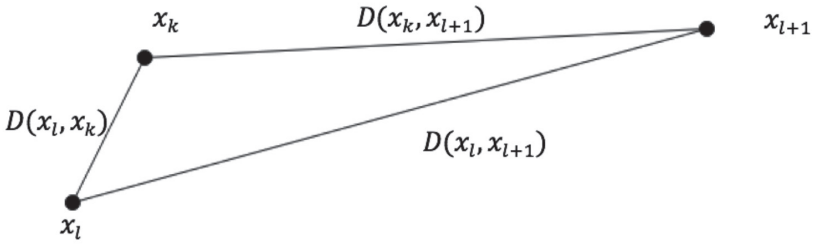

Иллюстрация расчета значения D' как разницы расстояний

Прямые между точками на рисунке эквивалентны дугам. Функция E^2_{loss} отражает оценку расстояния между x_k и дугой (x_l, x_{l+1}). Чем меньше E^2_{loss}, тем ближе точка x_k к дуге (x_l, x_{l+1}), чем больше – тем дальше. Если $E^2_{loss} = 0$, точка x_k лежит на дуге.

Исходя из минимизации погрешности и трудоемкости определения функции E_{loss} был выбран подход, учитывающий расстояние между точками в случае присутствия обеих точек в определенный момент времени на двух ломаных $L_{predicted}$ и $L_{groundTruth}$ и разницу расстояний между точками (см. п. 3 выше) в противном случае.

Поскольку ошибка может быть неабсолютной величиной и не должна в общем случае отражать какие-либо параметры траектории движения, например сумму отклонений расстояний и др., знаки пропорциональности в формулах (1) и (9) заменим на знаки равенства. Общая ошибка E_{loss} будет являться квадратным корнем из среднеквадратической ошибки модели – отношения суммы квадратов расстояний E^1_{loss} или E^2_{loss} в зависимости от наличия или отсутствия обеих точек на двух ломаных соответственно к количеству точек:

$$E_{loss} = \sqrt{\frac{\sum_{i=1}^{N_{\max}} E_i^{\,2}}{N_{\max}}}, \quad (12)$$

где $N_{max} = \max\left(N_{predicted}, N_{groundTruth}\right)$, так как вычисление ошибки должно быть выполнено для каждой точки ломаных $L_{predicted}$ и $L_{groundTruth}$;

$E_i = D\left(x_{i,1}, x_{i,2}\right), x_{i,1} = \left(lat_{i,1}, lon_{i,1}\right), x_{i,2} = \left(lat_{i,2}, lon_{i,2}\right)$, при условии, что $t_i, lat_{i1}, lon_{i1} \in L_{groundTruth}$ и $t_i, lat_{i2}, lon_{i2} \in L_{predicted}$. В противном случае, когда на одной из ломаных нет точки в момент t_i, т.е. $x_i = \left(lat_i, lon_i\right) = \left(lat_{i,1}, lon_{i,1}\right), t_i, lat_{i1}, lon_{i1} \in L_{exist} = L_{groundTruth}$ и $t_i, lat_{i2}, lon_{i2} \notin L_{notExist} = L_{predicted}$ или $t_i, lat_{i1}, lon_{i1} \notin L_{notExist} = L_{groundTruth}$ и $x_i = \left(lat_i, lon_i\right) = \left(lat_{i,2}, lon_{i,2}\right)$,

$t_i, lat_{i2}, lon_{i2} \in L_{exist} = L_{predicted}$:

- $E_i = D(x_i, x_l)$, при $x_l = (lat_l, lon_l), t_l, lat_l, lon_l \in L_{notExist}$, $t_l < t_i$, если $\forall t_{l+1} > t_i : \nexists t_{l+1}, lat, lon \in L_{notExist}$;
- $E_i = D(x_i, x_{l+1})$, при $x_{l+1} = (lat_{l+1}, lon_{l+1}), t_{l+1}, lat_{l+1}, lon_{l+1} \in L_{notExist}$, $t_{l+1} > t_i$, если $\forall t_{l-1} < t_i : \nexists t_{l-1}, lat, lon \in L_{notExist}$;
- $E_i = D(x_l, x_i) + D(x_i, x_{l+1}) - D(x_l, x_{l+1})$, при $x_l = (lat_l, lon_l), t_l, lat_l, lon_l \in L_{notExist}$ и $x_{l+1} = (lat_{l+1}, lon_{l+1}), t_{l+1}, lat_{l+1}, lon_{l+1} \in L_{notExist}$.

Стоит отметить, что алгоритм вычисления значения функции ошибки для одной траектории выполнится за линейное время, и его вычислительная сложность

$$Q = O(N_{\max}). \quad (11)$$

Заключение

В работе предложен метод расчета функции ошибки для оценки качества фильтрации геокоординат. Рассмотренный метод учитывает как оценку пространственной идентичности траекторий движения, так и временной, учитывая специфику предметной области задачи. Преимуществом метода является линейное время вычисления функции ошибки, что позволяет использовать его в качестве целевой функции при тренировке моделей машинного обучения в задачах фильтрации геокоординат, в том числе в переносимых мобильных устройствах. В качестве недостатка метода стоит отметить отсутствие учета метрики пространства в случае сравнения близости точки к отрезку дуги. Принимая во внимание скоростные и временные показатели движения, возможно сделать оценку более достоверной.

Список литературы

1. Головков А.А. Источники геоданных в мобильных устройствах / А.А. Головков // Динамика сложных систем – XXI век. – 2017. – № 4. – С. 94–101.

2. Головков А.А., Иванова Г.С. Адаптивная фильтрация потока геолокационных данных в реальном времени // Наука и Образование. МГТУ им. Н.Э. Баумана. Электрон. журн. – 2016. – № 4 [Электронный ресурс]. URL: http://technomag.edu.ru/jour/article/download/10/12 (дата обращения: 30.03.2018).

3. Макаренко Г.К. Исследование алгоритма фильтрации при определении координат объекта по сигналам спутниковых радионавигационных систем / Г.К. Макаренко, А.М. Алешечкин // Доклады Томского государственного университета систем управления и радиоэлектроники. – 2012. – № 2–2 (26). – С. 15–18.

4. Прохорцов А.В. Методы определения координат и скорости подвижных объектов с помощью спутниковых радионавигационных систем / В.В. Савельев, А.В. Прохорцов // Известия Тульского государственного университета. Технические науки. – 2011. – № 2. – С. 264–274.

5. Jaime Gomez-Gil, Ruben Ruiz-Gonzalez, Sergio Alonso-Garcia, Francisco Javier Gomez-Gil. A Kalman Filter Implementation for Precision Improvement in Low-Cost GPS Positioning of Tractors // Sensors. – 2013. – no. 13. – P. 15307–15323.

6. Салычев О.С. MEMS/GPS – малогабаритная интегрированная навигационная система / О.С. Салычев // Геопрофи. – 2013. – № 3. – С. 16–17.

7. Weiliang Zeng, Tomio Miwa, Takayuki Morikawa. Exploring Trip Fuel Consumption by machine Learning from GPS and CAN Bus Data // Journal of the Eastern Asia Society for Transportation Studies. – 2015. – vol. 11. – P. 906–921.

8. Huiqin Li, Gang Wu. Map Matching for Taxi GPS Data with Extreme Learning Machine // Advanced Data Mining and Applications: 10th International Conference proceedings. – 2014. – P. 447–460.

9. Katherine Ellis, Suneeta Godbole, Simon Marshall, Gert Lanckriet, John Staudenmayer, Jacqueline Kerr. Identifying active travel behaviours in challenging environments using GPS, accelerometers, and machine learning algorithms // Frontiers in Public Health. – 2014. – vol. 2. – P. 1–8.

10. Christian Manasseh, Raja Sengupta. Predicting driver destination using machine learning techniques // 16th International IEEE Conference on Intelligent Transportation Systems. – 2013. – P. 142–147.

11. Abhishek Goswami, Luis E. Ortiz, Samir R. Das. WiGEM: a learning-based approach for indoor localization // CoNEXT '11 Proceedings of the Seventh COnference on emerging Networking EXperiments and Technologies, 2011.