МЕТОД ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ ПРЕЦЕДЕНТОВ ДЛЯ ПРОГНОЗИРОВАНИЯ ВРЕМЕННЫХ РЯДОВ

Поминчук Е.В., Иващенко А.Б.
Донецкий национальный технический университет
jeka275@ukr.net

В данной статье рассмотрен метод интеллектуального анализа данных прецедентов, обсуждаются идеи алгоритма, представлена схема разрабатываемого программного обеспечения.

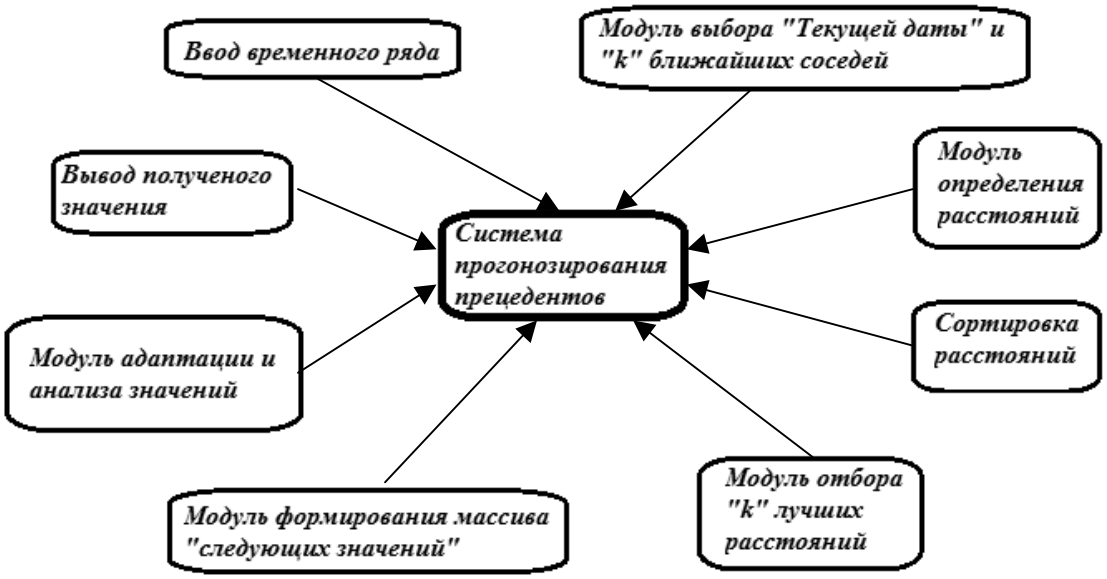
Актуальность. Необходимость предвидения вероятного развития событий на будущее в Украине, никогда ранее не была такой важной как сейчас. Решения, принимающиеся сегодня, опираются на признаки развития явлений. В свою очередь, они более или менее влияют на это в будущем. Именно поэтому, исследования моделей прогнозирования временных рядов в условиях недостаточной информации поможет избежать принципиальных ошибок при принятии каких либо решений. Изучение этой проблемы является актуальным как для теории, так и для практики.

Идея алгоритма.

Условно, алгоритм включает следующие этапы:

1. Ввод вр ряда для некоторой переменной;
2. Выбор «текущей даты» и числа “k”ближайших соседей .
3. Определение расстояний от значения текущей даты до значений предыдущих дат;
4. Сортировка расстояний по возрастанию;
5. Отбор k первых расстояний после сортировки;
6. Формирование массива «след значений»
7. Адаптация и анализ значений, например, поиск арифметическое среднего, минимального или максимального значения.
8. Вывод полученного значения как результат прогноза.

На рис. 1 представлена диаграмма модулей разрабатываемой системы.



Цель алгоритма.

Спрогнозировать значение вр ряда для будщего вр периода на основе закономерностей и связей выявленных в БД содержащей значения (измерения) данного ряда по прошедшему периоду.

Основные метрики

Евклидово пространство.

В изначальном смысле, пространство, свойства которого описываются аксиомами евклидовой геометрии. В этом случае предполагается, что пространство имеет размерность 3.

В современном понимании, в более общем смысле, может обозначать один из сходных и тесно связанных объектов, определённых ниже. Обычно n-мерное евклидово пространство обозначается E^n, хотя часто используется не вполне приемлемое обозначение R^n.

1. Конечномерное гильбертово пространство, то есть конечномерное вещественное векторное пространство R^n с введённым на нём (положительно определенным) скалярным произведением, порождающим норму:

$\|x\| = \sqrt{\langle x, x \rangle}$,

в простейшем случае (евклидова норма):

$$\|x\| = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2} = \sqrt{\sum_{k=1}^{n} x_k^2},$$

где $x = (x_1, x_2, \dots, x_n)$ (в евклидовом пространстве всегда можно выбрать базис, в котором верен именно этот простейший вариант).

2. Метрическое пространство, соответствующее пространству описанному выше. То есть R^n с метрикой, введённой по формуле:

$$p(x,y) = \|x-y\| = \sqrt{(x_1-y_1)^2 + (x_2-y_2)^2 + \dots + (x_n-y_n)^2} = \sqrt{\sum_{k=1}^{n}(x_k-y_k)^2}$$

.

,где $x = (x_1, x_2, \dots, x_n)$ и y= $(y_1, y_2, \dots, y) \in R^n$.

Расстояние Хэмминга.

Число позиций, в которых соответствующие символы двух слов одинаковой длины различны. В более общем случае расстояние Хэмминга применяется для строк одинаковой длины любых q-ичных алфавитов и служит метрикой различия (функцией, определяющей расстояние в метрическом пространстве) объектов одинаковой размерности.

Первоначально метрика была сформулирована Ричардом Хэммингом во время его работы в Bell Labs для определения меры различия между кодовыми комбинациями (двоичными векторами) в векторном пространстве кодовых последовательностей, в этом случае расстоянием Хэмминга $d(x,y)$ между двумя двоичными последовательностями (векторами) x и y длины n называется число позиций, в которых они различны — в такой формулировке расстояние Хэмминга вошло в словарь алгоритмов и структур данных национального института стандартов и технологий США (англ. NIST Dictionary of Algorithms and Data Structures).

Расстояние Махаланобиса.

Мера расстояния между векторами случайных величин, обобщающая понятие евклидова расстояния. Предложено индийским статистиком П.Ч. Махаланобисом в 1936 году. С помощью расстояния Махаланобиса можно определять сходство неизвестной и известной выборки. Оно отличается от расстояния Евклида тем, что учитывает корреляции между переменными и инвариантно к масштабу.

Формально, расстояние Махаланобиса от многомерного вектора $x = (x_1, x_2, \dots, x_n)^T$ до множества со средним значением $\mu = (\mu_1, \mu_2, \dots, \mu_n)^T$ и матрицей ковариации S определяется следующим образом:

$$D_M(x) = \sqrt{(x-\mu)^T * s^{-1} * (x-\mu)}.$$

Расстояние Махаланобиса также можно определить как меру несходства между двумя случайными векторами $\vec{x}$ и $\vec{y}$ из одного распределения вероятностей с матрицей ковариации S:

$$d(\vec{x},\vec{y}) = \sqrt{(\vec{x}-\vec{y})^T * S^{-1} * (\vec{x}-\vec{y})}.$$

Если матрица ковариации является единичной матрицей, то расстояние Махаланобиса становится равным расстоянию Евклида. Если матрица ковариации диагональная (но необязательно единичная), то получившаяся мера расстояния носит название нормализованное расстояние Евклида:

$$d(\vec{x},\vec{y}) = \sqrt{\sum_{i=1}^{n} \frac{(x_i-y_i)^2}{\sigma_i^2}}.$$

Здесь σ_i — среднеквадратичное отклонение x_i в выборке.

Расстояние Журавлева.

Впервые описано российским математиком Ю.И. Журавлевым.В сети к сожалению мною были найдены только ссылки на работы Журавлева, но ни в

одном из просмотренных мною материалов не описана в сколь-нибудь существенным образом. Поэтому постараемся проанализировать данную функцию для определения ее смысла.

Формула для определения расстояния:

$$d_{ik} = \sum_{j=1}^{N} I_{ik}^{j}.$$

где

$$I_{ik}^{j} = \begin{cases} 1, \text{если } |x_{ij} - x_{kj}| < s \\ 0, \text{иначе} \end{cases}.$$

Исходя, из отсутствия информации по этому вопросу попробуем проанализировать данную формулу.

i – номер строки в матрице исходных данных

k – номер столбца в той же матрице

j – последовательный номер столбца в обрабатываемых строках

x_{ij} – элемент строки i исходной матрицы

x_{kj} – элемент строки k исходной матрицы

ε – обычно означает положительное сколь угодно малое вещественное число. В данном случае расстояние между x_{ij} и x_{kj} в любом случае положительно. Следовательно, можно предположить, что ε это некоторый критерий минимального расстояния межу точками.

Функция $I(i,k,j)$ будет иметь значение 1 если расстояние между точками x_{ij} и x_{kj} меньше значения ε. Т.е. если точки находятся близко.

Соответственно значение функции будет равно 0 если точки находятся далеко друг от друга.

Таким образом, функция $I(i,k,j)$ является бинарным индикатором близости точек. Сумма в формуле расстояния Журавлева проходит по строке и попарно оценивает расстояние между эквивалентными элементами в двух строках i и k. Под эквивалентными элементами подразумеваются элементы, имеющие одинаковый порядковый номер.

Отсюда можно сделать вывод, что расстояние Журавлева равно количеству эквивалентных точек соответствующих критерию минимального расстояния ε. Следовательно, расстояние Журавлева является мерой сходства между двумя векторами данных. В контексте матрицы расстояний формируемой при кластерном анализе каждый элемент ее будет характеризовать расстояние Журавлева между каждыми парами строк в исходной матрице.

Таким образом, применение расстояния Журавлева при кластерном анализе дает меру сходства между всеми точками исходных данных.

Вывод

Преимуществом разрабатываемого алгоритма является возможность его использования для прогнозирования динамических показателей и факторов из любой области знаний и сферы деятельности человека. Например, для прогнозирования метеорологических параметров и состояния погоды, динамика курсов акций и валют, Прогнозирование потребительского спроса, прогнозирование обьемов кредитования на следующий отчетный период, прогнозирование урожайности и многое другое.