МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ, КОМПЬЮТЕРНЫЕ ТЕХНОЛОГИИ В ТЕОРИИ УПРАВЛЕНИЯ СЛОЖНЫХ ПРОЦЕССОВ

ПРИМЕНЕНИЕ ВИЗУАЛИЗАТОРОВ ДЛЯ ОЦЕНКИ КАЧЕСТВА МОДЕЛЕЙ ПРОГНОЗИРОВАНИЯ РАЗВИТИЯ ПОЖАРНОЙ ОБСТАНОВКИ

А.Ю. Иванов, доктор технических наук, профессор;
С.С. Чернов.
Санкт-Петербургский университет ГПС МЧС России

Рассмотрена одна из основных задач Data Mining – регрессия, ее применение для оценки качества модели прогнозирования развития пожарной обстановки, на основе диаграммы рассеяния. Также для определения корректности модели прогнозирования рассмотрена возможность применения частного случая задачи регрессии – прогнозирования временных рядов.

Ключевые слова: прогностическая модель, визуализатор, регрессия, прогноз, ретропрогноз, Data Mining, диаграмма рассеяния

APPLICATION VISUALIZATION FOR ASSESSING THE QUALITY PREDICTION MODEL DEVELOPMENT OF FIRE SITUATION

A.Y. Ivanov; S.S. Chernov.
Saint-Petersburg university of State fire service of EMERCOM of Russia.

Analyzed one of the main tasks of Data Mining - regression and its application to assess the quality of predictive models of fire situation, based on the scattering diagram. To determine the correctness of the prediction models consider the possibility of a special case of the problem of regression – time series prediction.

Key words: predictive model, visualizer, regression, forecast, retroprognosis, Data Mining, scatterplot

Современное общество вступило в новый этап своего развития, называемый информационным или постиндустриальным. Под этим обычно понимается, что знания и информация стали главной движущей и производительной общественной силой, определяющей как духовное, так и материальное состояние личности, общества в целом.

От умения производить, искать, анализировать, классифицировать, обобщать, распознавать, перерабатывать, представлять информацию и принимать решения сегодня напрямую зависит функционирование системы МЧС России, обеспечивающее качество жизни человека и общества, их информационную и общественную безопасность [1].

Это, безусловно, предъявляет новые требования к сотруднику МЧС России, к уровню его подготовки, к его информационной культуре. Умение работать с информацией для сотрудника МЧС России становится главным аспектом повышения эффективности принимаемых решений по управлению силами и средствами, а также позволяет решить проблемы, связанные с прогнозированием чрезвычайных ситуаций, в частности пожаров.

Это требует построения специализированной информационной системы, с помощью которой можно будет обнаружить в первичных данных ранее неизвестные, нетривиальные, практически полезные, доступные интерпретации знания, необходимые для принятия решений в сферах деятельности МЧС России.

Для решения задачи по повышению обоснованности принятия решений по предупреждению и ликвидации пожаров следует рассмотреть основные вопросы разработки модели системы интеллектуального анализа данных о пожарной обстановке.

Так как данная система предусматривает построение модели прогнозирования развития пожарной обстановки, то для определения качества таких моделей пользователем, необходимо применять специальные методы – визуализаторы.

Визуализаторы – тип программного обеспечения, предназначенный для преобразования различной информации в зрительные образы. Может являться либо отдельным приложением, либо плагином или частью другого приложения [2]. Независимо от вида построенной модели, прежде чем применять ее на практике, необходимо оценить ее качество, то есть определить, насколько правильно и точно она решает поставленную задачу. Для этого определим две составляющие качества модели: адекватность – точность описания моделью исследуемого объекта; корректность – насколько правильно модель может работать со всеми возможными входными данными.

Как показывают предварительные наработки в данной области, абсолютно точного описания моделью реального объекта или процесса достичь практически невозможно. Для разрабатываемой модели выберем разумный компромисс между точностью и сложностью. К примеру, зададим допустимую точность 10 %. Таким образом, модель из 50 примеров должна неправильно классифицировать не более 5.

Как показывает научная практика, для оценки качества модели прогнозирования развития пожарной обстановки, необходимо применить одну из важнейших задач Data Mining (интеллектуального анализа данных) [3] – регрессию, которая устанавливает взаимную связь между значениями входных и выходных переменных модели. Визуализатором, который применяется для оценки качества моделей в случае непрерывной выходной переменной, является диаграмма рассеяния, представляющая собой график, по одной оси которого откладываются целевые значения выходной переменной, то есть заданные в качестве эталонов для обучения, а по другой оси откладываются реальные значения, полученные на выходе модели. На таком графике можно построить линию идеальных значений: $y=y'$. На этой линии будет лежать любая точка, для которой реальное выходное значение y, сформированное моделью будет равно целевому значению y', при этом ошибка: $E= y-y'=0$.

Очевидно, что такая линия будет выходить из начала координат и пересекать координатную плоскость по диагонали. В то же время любая точка, для которой реальное значение будет отлично от эталонного целевого значения, отклонится от линии идеальных значений, при этом величина отклонения будет равна ошибке, допущенной моделью на данном примере (рис. 1).

Точка, лежащая на линии идеальных значений, соответствует случаю, когда реальный выход модели равен эталонному. На практике модель допускает ошибку E, которая приводит к отклонению точки, соответствующей реальному выходу модели, от эталонного и, следовательно, от линии идеальных значений. Таким образом, точка, соответствующая реальному выходу, будет иметь координаты: $y' = y+E$.

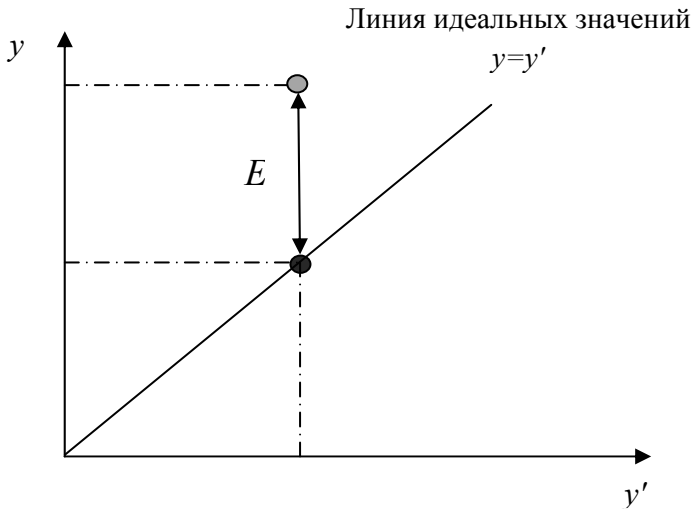

Рис. 1. **Принцип построения диаграммы рассеяния**

Множество точек на плоскости *(yy′)*, образованных парами целевых и реальных значений, будет представлено на диаграмме в виде облака, рассеянного вдоль линии идеальных значений. Степень отклонения точки от этой линии будет определяться ошибкой модели на соответствующем примере [4].

Пример диаграммы рассеяния представлен на рис. 2. Диагональная линия на рисунке – это линия идеальных значений. Точками, рассеянными вдоль линии идеальных значений, обозначены реальные выходные значения модели.

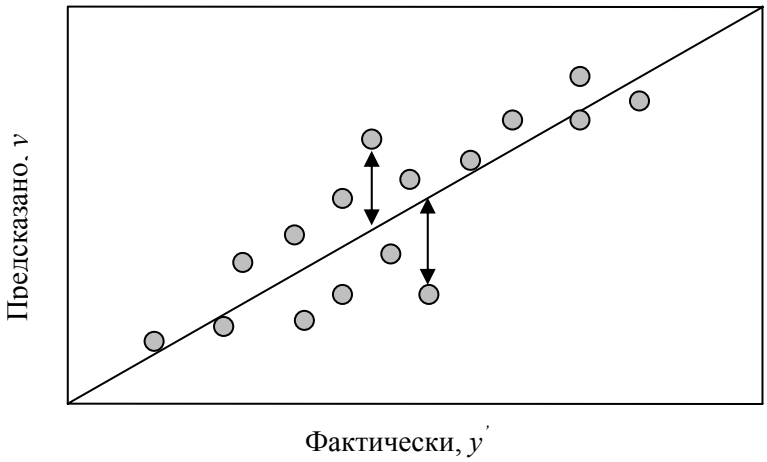

Рис. 2. **Диаграмма рассеяния**

Смысл диаграммы рассеяния достаточно прозрачен, если все точки, или основная их масса, представляющие реальные выходные значения модели, сосредоточены вблизи линии идеальных значений, то модель работает хорошо.

Если у облака, образуемого точками выходов модели, значительный разброс, то большинство выходных значений имеет большую ошибку и, в этом случае, качество модели является неудовлетворительным. Пример такой диаграммы представлен на рис. 3.

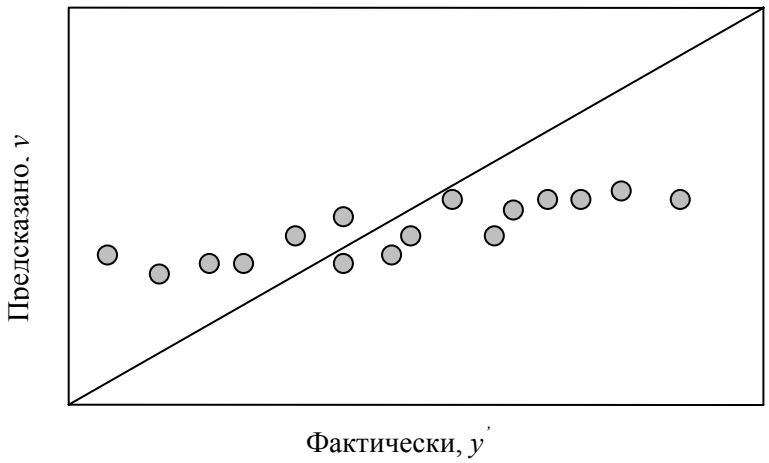

Рис. 3. **Диаграмма рассеяния для неудачной модели**

Также интерес могут представлять отдельные точки, наиболее далеко стоящие от линии идеальных значений, их появление обусловлено тем, что на соответствующих примерах модель дает очень высокую ошибку, поскольку в этих примерах содержатся аномальные значения, нехарактерные для поведения выборки в целом. Анализ отклонений позволит выявить редкие, но важные события, влияющие на исследуемый процесс. Если же аномальное значение просто следствие ошибки, то его следует исключить из рассмотрения или откорректировать.

На основе рассмотренных примеров, можно сделать однозначный вывод, что для определения адекватности и корректности моделируемых процессов прогнозирования, в частности развития пожарной обстановки, достаточно визуального определения значений, полученных с помощью диаграммы рассеяния.

Также для определения корректности модели прогнозирования можно применить частный случай задачи регрессии – прогнозирование временных рядов. В данном случае для прогнозирования строится регрессионная модель, которая на основе прошлых значений ряда рассчитывает прогнозируемые значения. Она имеет некоторый набор параметров, позволяющих получить прогноз с учетом поведения ряда в прошлом (рис. 4).

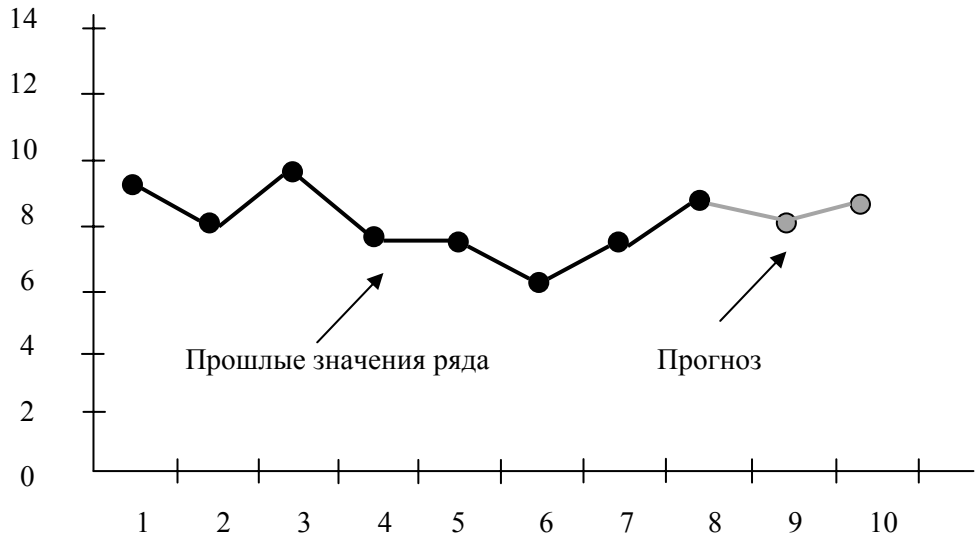

Рис. 4. **Диаграмма прогноза**

Для проверки качества прогностической модели применим специальный тип визуализатора – ретропрогноз, который буквально означает прогноз прошлых значений. Сущность данного метода заключается в том, чтобы применить построенную модель к подмножеству данных из прошлого и оценить, насколько полученный прогноз соответствует реальным значениям ряда, имевшим место в прошлом. Для построения ретропрогноза необходимо выбрать некоторое подмножество данных из прошлого, чтобы использовать их в качестве исходных. Далее к этому подмножеству нужно применить прогностическую модель с заданными параметрами. Модель формирует набор прогнозных значений, которые затем сравниваются с данными, реально имевшими место в прошлом, и если в результате такого сравнения обнаружится, что между значениями ретропрогноза и реальными данными существует большое расхождение, то это дает повод усомниться в корректности прогностической модели, а сам характер расхождения, его величина позволит выработать методику коррекции модели.

Принцип работы ретропрогноза показан при помощи диаграммы (рис. 5).

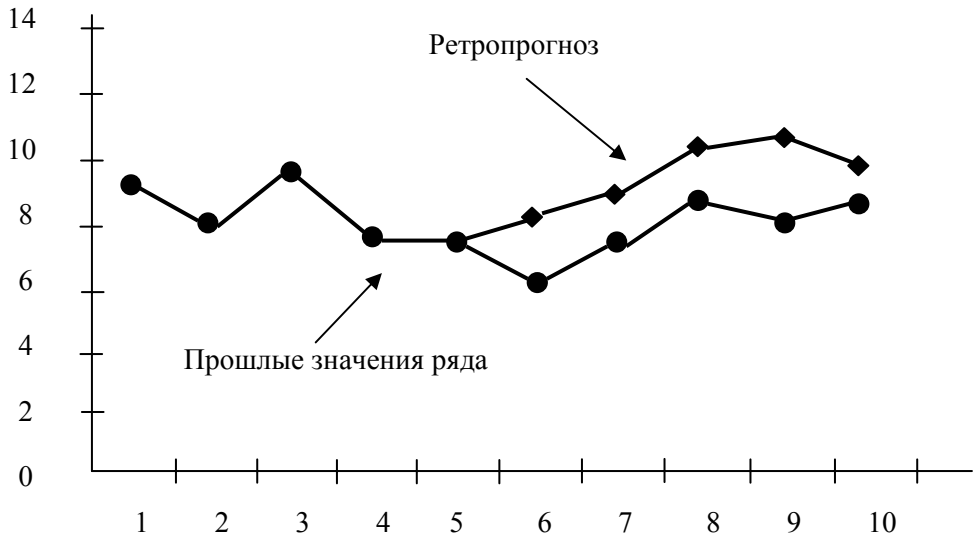

Рис. 5. **Диаграмма ретропрогноза**

Ретропрогноз направлен не в будущее ряда, а формируется параллельно его прошлым значениям и сравнивается с ними. На рис. 5 можно наблюдать достаточно хорошее согласование значений ретропрогноза и идеальных значений ряда, что позволяет предположить высокое качество прогностической модели, с помощью которой был построен ретропрогноз. На рис. 6 показан противоположный случай, когда ретропрогноз оказывается существенно завышенным или заниженным, что говорит о недостаточной точности модели.

Следует отметить, что при выборе множества значений временного ряда, которые будут использоваться в качестве исходных данных для ретропрогноза, необходимо учитывать его актуальность. Во временном ряду можно выделить два вида данных: актуальные и устаревшие. В первом случае данные сформированы под действием текущих условий, влияющих на процесс, описываемый временным рядом. Но со временем условия могут меняться, и тогда данные, сформированные под действием утративших силу условий, оказываются устаревшими, а данные, сформированные под действием новых условий, актуальными, другими словами самыми свежими данными. Следовательно, если модель прогноза была построена на основе актуальных данных, то и ретропрогноз должен строиться на основе свежих данных, если же ретропрогноз будет построен на основе устаревших

данных, то, вероятнее всего, его значения будут сильно расходиться с реальностью. Но это будет означать не низкое качество модели, а то, что она используется для данных, подчиняющихся другим закономерностям, которые в этой модели не учитываются. С учетом изложенного к интерпретации результатов ретропрогноза следует подходить с определенной долей осторожности.

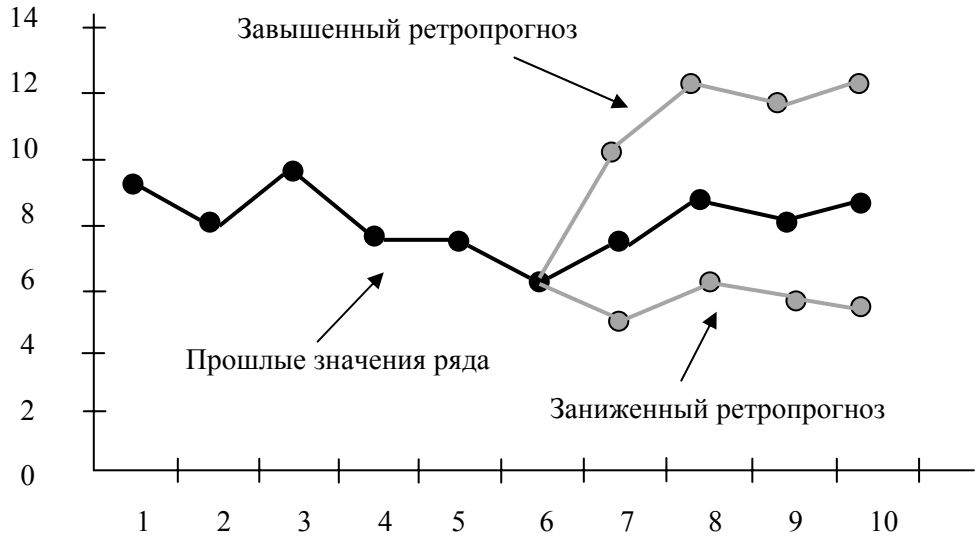

Рис. 6. **Ретропрогноз с большим отклонением прогнозируемых значений**

Литература

1. Материалы официального сайта МЧС России. [Электронный ресурс]. URL: http:// www.mchs.gov.ru (дата обращения: 24.06.2012).

2. Визуализатор. [Электронный ресурс]. URL: http://ru.wikipedia.org/wiki/ (дата обращения: 24.07.2012).

3. [Электронный ресурс]. URL: http://ru.wikipedia.org/wiki/Data_mining (дата обращения:11.07.2012).

4. Паклин Н.Б., Орешков В.И. Бизнес-аналитика: от данных к знаниям. СПб.: Питер, 2010. 704 с.