УДК 004.94(075.8)

Н.В. Богушевская, А.С. Плакса, Е.В. Ягодкина

ИСПОЛЬЗОВАНИЕ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ ДЛЯ ПРОГНОЗИРОВАНИЯ ДЛИТЕЛЬНОСТИ РЕВИЗИИ ДЛЯ ОРГАНОВ ГКРС

Аннотация: В статье рассмотрена задача определения зависимости длительности ревизии от суммы финансирования, выделяемой объекту контроля, математическая формулировка задачи представлена с помощью полинома Колмогорова-Габора. Показана эффективность одновременного применения метода группового учёта аргументов и метода кластеризации. Получены коэффициенты зависимости для частного случая, приведены графики зависимостей исходных данных и полученных при моделировании.

Ключевые слова: групповой учёта аргументов, кластеризация, прогнозирование, ревизия.

Введение

Двумя основными задачами интеллектуального анализа данных (ИАД) являются прогнозирование и описание. Задача *прогнозирования* состоит в создании общей модели всей системы, основываясь лишь на ограниченном наборе данных. Целью *описания* является обнаружение новой нетривиальной информации, основываясь на существующих данных.

Обе задачи решаются с помощью использования следующих шести основных методов [1]:

1. Классификация – заключается в поиске группы функций (или моделей), которые смогут определить принадлежность неизвестного объекта к одному из существующих классов.

2. Регрессионный анализ – заключается в поиске функций, с помощью которых было бы возможно предсказать значение целевой (входной) переменной.

3. Кластеризация – занимается поиском определенных сегментов (кластеров) по которым можно было бы распределить все исходные переменные (в отличие от классификации, где классы известны заранее). Кластеры формируются таким образом, чтобы объекты одного кластера были максимально схожи друг с другом и максимально отличались от объектов других кластеров.

4. Суммирование – включает в себя методы поиска краткого описания для определенного множества данных.

5. Моделирование зависимостей – заключается в поиске моделей, которые могли бы описать существенные зависимости между значениями переменных или значениями тех или иных свойств в массиве данных.

6. Выявление отклонений и изменений – поиск наиболее значительных изменений во множестве данных.

 ISSN 1562-9945

Использование методов ИАД является неотъемлемой частью любой системы поддержки принятия решений (СППР). Для СППР органов Государственной контрольно-ревизионной службы (ГКРС) [4] одной из важных задач является задача прогнозирования длительности проведения ревизии.

Длительность проведения ревизии на подконтрольном объекте зависит от различных бюджетных сумм, целевое использование которых проверяется в ходе контрольного мероприятия. Тогда задача прогнозирования длительности ревизии может быть сведена к решению задачи регрессионного анализа, где в качестве целевой переменной будет выступать длительность ревизии.

Постановка задачи

Для каждого подконтрольного объекта известна выделяемая сумма финансирования X, имеющая такие составляющие:

- финансирование на содержание X_1;
- выделение денег X_2;
- аккумулирование льгот налогообложения X_3;
- кредиты, полученные под гарантию правительства X_4;
- аккумулирование других государственных целевых денег X_5;
- деньги государственных целевых фондов X_6;
- внебюджетные деньги X_7.

Пусть t – длительность проведения ревизии, вычисляемая как количество человеко-дней, затраченных на её проведение.

Необходимо определить зависимость t от векторной переменной X, где $X = \{X_1, X_2, X_3, X_4, X_5, X_6\}$, т.е. определить коэффициенты полинома Колмогорова-Габора:

$$t = f(X_1, X_2, \ldots, X_6) = \\ = a_0 + \sum_{i=1}^{6} a_i X_i + \sum_{i=1}^{6}\sum_{j=1}^{6} a_{ij} X_i X_j + \sum_{i=1}^{6}\sum_{j=1}^{6}\sum_{k=1}^{6} a_{ijk} X_i X_j X_k + \ldots \quad (1)$$

Анализ методов решения

Для решения задачи воспользуемся методом группового учёта аргументов (МГУА). Эффективность метода многократно подтверждалась решением множества конкретных задач из областей экологии, экономики, гидрометеорологии и т.д.[2-3].

Для определения коэффициентов полинома (1) был использован программный продукт VariReg, предоставляемый бесплатно для некоммерческих целей (http://www.cs.rtu.lv/jekabsons/ regression.html). Визуализация и статистическая обработка данных выполнялась при помощи инструментального пакета Deductor (http://www.basegroup.ru/).

Анализ данных

В качестве первоначальных данных использовалась выборка на 1 800 элементов. На рисунке 1 представлен график зависимости длительности ревизии от суммы финансирования для первоначальной выборки.

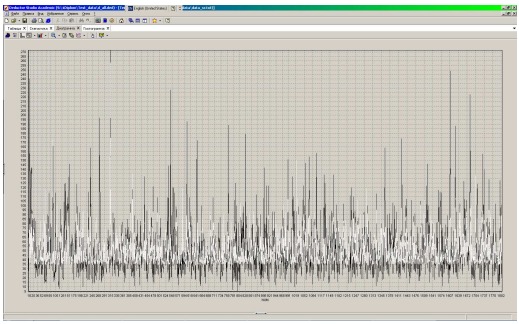

Рис. 1 – График зависимости длительности ревизии от суммы финансирования (чёрный цвет – исходные данные, белый – полученные при моделировании)

Диапазон отклонения исходных данных от полученных данных представлен на рисунке 2:

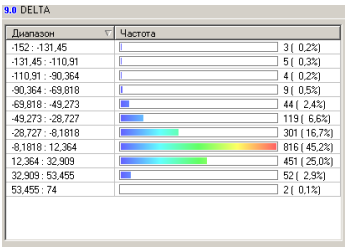

Рис. 2 – Отклонение исходных данных от полученных при моделировании

Значительный разброс данных может быть объяснён следующими факторами:

- ошибки в исходных данных – для решения этой проблемы может использоваться метод выявления отклонений и изменений;
- некорректное формирование выборки данных. Для решение может быть использована кластеризация исходных данных.

Выявление отклонений и изменений

Дополнительный анализ исходных данных проводился для тех ревизий, у которых отклонение фактических значений от значений, полученных при моделировании, было максимально. Анализ показал, что

 ISSN 1562-9945

в ряде случаев фактическая длительность ревизии была неверна, что объясняется внесением некорректных данных.

Наиболее распространёнными ошибками оказались:

1. неверное заполнение дат (к примеру, внесение даты “01.01.0208” вместо “01.01.2008”);
2. ошибка при внесении данных о привлечённых ревизорах (один и тот же ревизор указывался как проводящий ревизию и как привлечённый (на срок проведения ревизии), что приводило к неверному вычислению длительности ревизии).

Кластеризация данных

Для более детального анализа первичной выборки данных была проведена кластеризация данных с использованием инструментального пакета Deductor. В результате были выявлены 15 кластеров (рисунок 3).

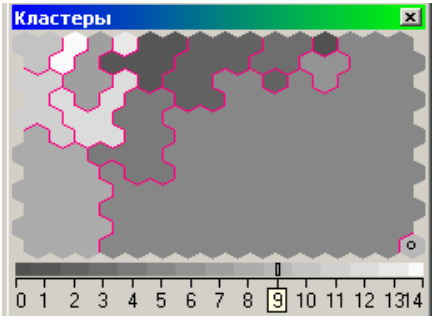

Рис. 3 – Карта Кохонена для исходных данных

Для нахождения наилучшего решения задачи прогнозирования необходимо определить коэффициенты зависимости (1) для каждого из кластеров.

Анализ данных в каждом из кластеров позволил выделить дополнительные критерии, которые не были учтены при первоначальном моделировании:

- организационно-правовая принадлежность подконтрольного объекта (общегосударственная, коммунальная и т.д.);
- порядок суммы финансирования X в тыс.грн.

Таким образом, первоначальная задача сводится к задаче нахождения коэффициентов полинома (1) для каждого из кластеров. К примеру, для общегосударственных подконтрольных объектов с выделенной суммой финансирования от 100 до 1000 тыс. грн., был получен следующий полином:

$$t = 3810472756499 - 0.00626047434711138 \cdot X_1 + 1.04719139472326E- \\ -6 \cdot X_1^2 - 0.013484445351341 \cdot X_2 + 1.5165563478114E - 5 \cdot X_1 \cdot X_2 + \\ +2.48959238118124E - 6 \cdot X_2^2 - 0.0122142250901004 \cdot X_7 + \\ +6.24197156080134E - 6 \cdot X_1 \cdot X_7 + 9.00229779504105E - 6 \cdot X_2 \cdot X_7 + \\ +2.18808623782868E - 6 \cdot X_7^2, \qquad (2)$$

зависящий только от переменных X_1, X_2 и X_7.

На рисунке 4 представлен график зависимости длительности ревизии от суммы финансирования для этого кластера.

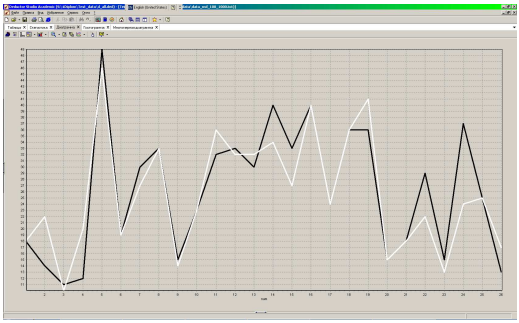

Рис. 4 – График зависимости длительности ревизии от суммы финансирования для кластера (чёрный – исходные данные, белый – полученные при моделировании)

Диапазон отклонения исходных данных от полученных данных для кластера представлен на рисунке 5:

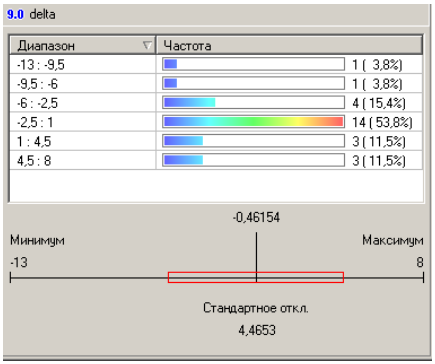

Рис. 5 – Отклонение исходных данных от полученных

Анализ рисунков (1) и (4) показывает, что использование полинома для каждого из кластеров даёт более точные результаты, чем использование единственного полинома.

Выводы

Применение методов ИАД для решения задачи прогнозирования длительности проведения ревизии, позволяет получить зависимость времени проведения ревизий от суммы финансирования, выделенной объекту контроля, и организационно-правовой принадлежности объекта.

Одновременное применение кластеризации по описательному критерию (организационно-правовая принадлежность объекта контроля) и регрессионного анализа с использованием МГУА по числовому (сумма финансирования) позволяет получить более точную зависимость, чем применение только регрессионного анализа.

Полученные зависимости могут быть использованы для составления годовых, полугодовых и квартальных планов деятельности ГКРС.

Литература

1. Kantardzic M. Data Mining: Concepts, Models, Methods, and Algorithm: John Wiley & Sons, 2003.
2. Ивахненко А.Г. Системы эвристической самоорганизации в технической кибернетике. - Киев: "Техніка", 1971. - 392 с.
3. Ивахненко А.Г. Долгосрочное прогнозирование и управление сложными системами. - Киев: "Техніка", 1975. - 311 с.
4. Богушевская Н.В., Плакса А.С., Ягодкина Е.В. Подсистема планирования в рамках системы поддержки принятия решений для органов Государственной контрольно-ревизионной службы//Науковий вісник Кременчуцького університету економіки, інформаційних технологій та управління "Нові технології". - 2009.- 1(23). - с.165

Отримано 07.12.2009 р.