УДК 004.732

МЕТОД ГЛАВНЫХ ФАКТОРОВ ДЛЯ ОЦЕНКИ ПРИОРИТЕТНОСТИ И ВЫБОРА ПАРАМЕТРОВ ПРИ ПРОГНОЗИРОВАНИИ ТЕХНИЧЕСКОГО СОСТОЯНИЯ КОМПЬЮТЕРНОЙ СЕТИ

Л.Н. ДУДНИК, Ю.Д. ШЕВЦОВ, Л.М. КРИЦКАЯ, В.В. БОГДАНОВ

Кубанский государственный технологический университет
350072, Российская Федерация, г. Краснодар, ул. Московская, 2
электронная почта: lududnik@gmail.com

В данной статье предлагается подход к решению задачи выбора информативной совокупности параметров для прогнозирования технического состояния компьютерной сети на основе использования методов факторного анализа с целью повышения точности прогноза и уменьшению затрат на обработку данных.
Ключевые слова: компьютерная сеть, прогнозирование, выбор информативных параметров, модель факторного анализа, корреляционная матрица, дисперсия факторов.

Для проведения эффективного контроля и прогнозирования технического состояния компьютерной сети необходимо осуществить выбор оптимального наиболее информативного набора параметров, для которого применяется приоритетный метод оценки, позволяющий построить упорядоченную последовательность параметров по мере возрастания или убывания их значимости по заданным критериям. В данной методике таким критерием оценки являются доли дисперсий каждого из параметров в общие факторы, которые достаточно точно воспроизводят взаимосвязи контролируемых параметров. Основной задачей является переход от описания системы, заданной большим набором измеряемых параметров к описанию меньшим числом максимально информативных признаков, отражающих наиболее существенные свойства системы.

Объектом исследования является некоторая система (компьютерная сеть КС), техническое состояние которой можно описать совокупностью параметров, представляющих собой вектор технического состояния $\vec{Y} = (y_1, y_2, ..., y_n)$, где n – число параметров системы.

Методология выбора параметров для прогнозирования основывается на использовании метода главных факторов.

Модель факторного анализа имеет матричный вид [1]:

$$R_h = A * A' \qquad (1)$$

или в развернутом виде:

$$\begin{pmatrix} h_1^2 & r_{12} & \dots & r_{1n} \\ r_{21} & h_2^2 & \dots & r_{2n} \\ \dots & \dots & \dots & \dots \\ r_{n1} & \dots. & \dots & h_n^2 \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1r} \\ a_{21} & a_{22} & \dots & a_{2r} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nr} \end{pmatrix} \times \begin{pmatrix} a_{11} & a_{21} & \dots & a_{n1} \\ a_{12} & a_{22} & \dots & a_{n2} \\ \dots & \dots & \dots & \dots \\ a_{1r} & a_{2r} & \dots & a_{nr} \end{pmatrix} (2)$$

где R_h - редуцированная корреляционная матрица;

h_i^2 - значение общности;

A - матрица факторного отображения;

A' - транспонированная матрица факторного отображения.

Метод главных факторов исходит из максимизации дисперсии в одном направлении при введении дополнительных условий.

Система уравнений (1) имеет однозначное решение при следующих дополнительных условиях:

- сумма квадратов нагрузок первого фактора должна составлять максимум от полной дисперсии,
- сумма квадратов нагрузок второго фактора должна составлять максимум от оставшейся дисперсии и т.д.,

т.е. максимизирует функцию: $D_1 = \sum_{i=1}^{n} a_{i1}^2 = \max$ при $\frac{n(n-1)}{2}$ независимых друг от друга условиях.

Для максимизирования функции D_1 используем метод множителей Лагранжа [1]. В результате получим систему из n однородных уравнений с n неизвестными a_{il}.

$$\begin{cases} (1-\lambda)a_{11} + & r_{12}a_{21} + & \dots + & r_{1n}a_{n1} = 0 \\ r_{21}a_{11} + & (1-\lambda)a_{21} + & \dots + & r_{2n}a_{n1} = 0 \\ . & . & \dots & . \\ r_{n1}a_{11} + & r_{n2}a_{21} + & \dots + & (1-\lambda)a_{n1} = 0 \end{cases} \qquad (3)$$

Для того, чтобы данная система уравнений имела нетривиальное решение необходимо и достаточно, чтобы определитель матрицы коэффициентов этих уравнений был равен нулю:

$$\begin{vmatrix} (1-\lambda) & r_{12} & \dots & r_{1n} \\ r_{21} & (1-\lambda) & \dots & r_{2n} \\ . & . & \dots & . \\ r_{n1} & r_{n2} & \dots & (1-\lambda) \end{vmatrix} = 0 \qquad (4)$$

Все n корней данного характеристического определителя действительны и являются возможными решениями, т.е. $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n \geq 0$.

Подставляя в (3) найденное значение λ_1, получим вектор решения α_{i1}=(α_{11}, α_{21}, $\dots, \alpha_{n1}$), который имеет максимум при $\sum_{i=1}^{n} \alpha_{i1}^2$ и удовлетворяет дополнительным условиям. Аналогично получаем для λ_2, в качестве решения вектор α_{i2}=(α_{12}, α_{22}, $\dots, \alpha_{n2}$), так чтобы при $\sum_{i=1}^{n} \alpha_{i2}^2$ обеспечивался максимум оставшейся дисперсии и т.д.

Получение результата уравнения (1) заключается в решении характеристического уравнения (3) относительно редуцированной корреляционной матрицы, которое предусматривает определение собственных значений λ_l и собственных векторов α_{il} матрицы корреляций. Факторные нагрузки a_{il} являются элементами собственных векторов матрицы корреляций, а дисперсии факторов равны собственным значениям этой матрицы. Путем нормирования получим значения a_{il} матрицы $A=(a_{il})$ по компонентам собственных векторов α_{il} матрицы R:

$$a_{il} = \alpha_{il}\sqrt{\lambda_l} / \sqrt{\alpha_{i1}^2 + \alpha_{i2}^2 + \dots + \alpha_{nl}^2}, \qquad (5)$$

где $i, l = 1, 2, \dots, n$

Метод главных факторов состоит в переходе от редуцированной корреляционной матрицы к матрице факторного отображения, которая позволит определить, сколько общих факторов необходимо для отражения всех корреляций между параметрами, а также определить нагрузки каждого фактора для всех параметров и нагрузки всех факторов для одной переменной.

Рассмотрим основные этапы алгоритма данной методики.

Этап 1. Построение корреляционной матрицы $\|R_{ij}\| = r_{ik}$ осуществляется на основе расчета коэффициентов парных корреляций параметров по алгоритму, описанному в [2].

Этап 2. Построение редуцированной корреляционной матрицы $\|R_h\| = r_{ik}^h$, где i =1,…, n, k = 1,…, n с общностями на главной диагонали.

Оценка общностей h_i^2 производится на основе вычисления квадрата множественной корреляции для каждой переменной с помощью обратной матрицы R^{-1} по формуле: $h_i^2 = 1 - \frac{1}{r^{ii}}$,где r^{ii} - диагональный элемент обратной матрицы.

Этап 3. Нахождение собственных векторов α_{il} и собственных значений λ_l редуцированной корреляционной матрицы R_h, где i, l = 1,…, n.

Используем итерационный метод определения собственных значений λ_l и собственных векторов α_{il} редуцированной корреляционной матрицы R_h. Итерационный процесс начинается с выбора вектора $\alpha^{(1)}$, элементы которого являются первыми приближениями значений элементов собственных векторов. Для этого, вычисляем суммы элементов строк матрицы R_h, делим каждую полученную сумму на максимальную из них и получаем элементы вектора $\alpha^{(1)}$.

Вектор $\alpha^{(1)}$ перемножаем с матрицей R по формуле:

$$\beta^{(1)} = R\alpha^{(1)} \qquad (6)$$

После деления результирующего вектора $\beta^{(1)}$ на наибольший по величине элемент этого вектора, получим вектор $\alpha^{(2)}$:

$$\alpha^{(2)} = \beta^{(1)} / \max(\beta_i^{(1)}) \qquad (7),$$

Повторяем процедуру (6) с вектором $\alpha^{(2)}$.

$$\alpha^{(k+1)} = \beta^{(k)} / \max(\beta_i^{(k)}) \qquad (8),$$

где верхние индексы в скобках означают шаг итерации.

Формула (8) является общей для k шагов итерации. Процесс повторяется до тех пор, пока не добиваются сходимости к первому собственному значению матрицы R: $\lambda_1 = \max(\beta_i^{(k)})$ и соответствующему первому собственному вектору $\alpha^{(k)}$. Итерационный процесс заканчивается, когда $\alpha^{(k)}$ и $\alpha^{(k-1)}$с достаточной точностью совпадают друг с другом, т.е. $\left|\alpha^{(k)} - \alpha^{(k-1)}\right| \le \varepsilon$. Если в качестве множителя для итерации в (6), вместо R брать R^2, R^4, или R^8, то скорость сходимости повышается, и число итераций сокращается. При этом R^2 означает двукратное перемножение матрицы $R*R$, а не возведения в степень и т.д.

Этап 4. Вычисление нагрузок первого фактора a_{i1} матрицы факторного отображения $A=(a_{il})$. Нагрузки матрицы факторного отображения $A=(a_{il})$, могут быть определены по компонентам собственных векторов матрицы R согласно формуле (5). Для этого находим сумму квадратов элементов первого собственного вектора α_{i1}по формуле: $T = \sum_{i=1}^{n} \alpha_{i1}^2$, а также обозначим и находим $t = \frac{\sqrt{\lambda_1}}{\sqrt{T}}$. Вычислим нагрузки первого фактора по формуле: $a_{i1} = \alpha_{i1} t$.

Этап 5. Определение матрицы воспроизведенных корреляций R^+ с учетом только первого фактора и остаточной матрицы корреляций R_1.

Матрица воспроизведенных корреляций R^+ определяется по формуле: $R^+ = a_1 \cdot a_1^{'}$, где a_1 - вектор столбец нагрузок первого фактора, $a_1^{'}$- вектор строка нагрузок первого фактора. Остаточную матрицу корреляций R_1, которая остается после выделения первого фактора, определяем по следующей формуле: $R_1 = R_h - R^+$.

Этап 6. Выделение второго фактора и определение второго собственного значения λ_2 и второго собственного вектора α_{i2} матрицы R_h. Расчет нагрузок второго фактора матрицы $A=(a_{il})$.

Повторяем процедуру с матрицей первых остаточных коэффициентов корреляции R_1 согласно этапам 3 – 5 данного алгоритма. В результате вычислений получаем второе собственное число, второй собственный вектор

α_{i2} матрицы R_h .

Этап 7. Определение следующего собственного вектора и собственного значения матрицы R_h .

Определение собственных векторов, а также соответствующих собственных значений осуществляется согласно этапам 3 – 5 данного алгоритма. Каждое собственное число характеризует конкретный выделенный фактор исследуемой системы.

Вес фактора в общей совокупности из *n* компонент определяется по формуле:

$$K_i = \frac{\lambda_i}{\sum_{i=1}^{n} \lambda_i} \qquad (9),$$

где λ_i - собственные числа, *n* – число компонент (переменных).

Этап 8. Построение матрицы факторного отображения $A = (a_{il})$, элементами которой являются факторные нагрузки, где *i* =1,..., *n* – параметры; *l* = 1, ..., *r* – факторы.

Матрица $A = (a_{il})$ позволяет выделить для каждого фактора группу параметров наиболее тесно с ним связанную, рассчитать коэффициенты приоритетности параметров с учетом их взаимосвязи и осуществить выбор наиболее значимых параметров для контроля и прогнозирования технического состояния основных блоков сети, обеспечив заданный критерий информативности.

Этап 9. Расчет квадратов факторных нагрузок a_{ik}^2 и дисперсий каждого фактора. Вычисление долей дисперсии факторов по матрице факторного отображения $A=(a_{il})$.

Нагрузки общих и характерных факторов связаны определенным соотношением через единичную дисперсию переменных [33]:

$$D_i^2 = 1 = (a_{i1}^2 + a_{i2}^2 + \ldots + a_{ir}^2) + u_i^2, \qquad (10)$$

где $h_i^2 = (a_{i1}^2 + a_{i2}^2 + \ldots + a_{ir}^2)$ (11)

h_i^2 - общность, которая представляет собой часть единичной дисперсии переменой и относится к общим факторам,

u_i^2 - характерность, т.е. часть единичной дисперсии не связанной с общими факторами: $u_i^2 = 1 - h_i^2$

Дисперсии факторов рассчитываются, как суммы квадратов факторных нагрузок a_{ik}^2 соответствующего фактора, равны собственным числам корреляционной матрицы и могут быть вычислены по формуле: $Di = \sum_{i=1}^{n} a_{ik}^2 = \lambda_i$.

Далее вычисляют доли дисперсии каждого фактора от полной дисперсии. В результате анализа оставляют в рассмотрении только те факторы, которые имеют максимальные дисперсии и обеспечивают 90-95% от полной дисперсии факторов. Факторы, имеющие дисперсию менее 5% от полной дисперсии, из рассмотрения исключаются.

Этап 10. Определение коэффициентов приоритетности параметров осуществляется на основе ранжирования квадратов факторных нагрузок в общий фактор, обеспечивающий максимум дисперсии от полной дисперсии факторов.

Этап 11. Расчет информативности *I* полученной совокупности факторов и потерь информации при исключении отдельных факторов.

Информативность рассчитывается согласно следующей формуле:

$$I = \frac{\lambda_1 + \lambda_2 + ... + \lambda_i}{\sum_{i=1}^{n} \lambda_i} \qquad (12),$$

где λ_i - собственные числа, $i = 1, 2, \ldots, n$.

Предложенная методика позволяет осуществить научно-обоснованный выбор информативной совокупности параметров для контроля и прогнозирования технического состояния компьютерной сети, минимизировать количество параметров для прогнозирования без потерь в информативности, уменьшить размерность массивов данных при обработке, а также сократить затраты на их контроль и прогноз.

Преимущества использования данной методики перед другими известными заключается в том, что большая совокупность параметров, описывающих техническое состояние КС, подвергается анализу, и выявляются группы взаимосвязанных параметров, относящиеся к определенным факторам (признакам). В результате анализа оставляют в рассмотрении только информативные факторы, которые имеют максимальные дисперсии от полной дисперсии факторов и отражают наиболее существенные свойства исследуемой системы.

ЛИТЕРАТУРА

1. Иберла К. Факторный анализ /Пер. с нем. В.М. Ивановой – М.: «Статистика» , 1980, – 398 с.

2. Выбор параметров контроля технического состояния для цифровых блоков корпоративной сети на основе использования методов факторного анализа. / Г.С. Петриченко, Л.Н. Дудник // Межотраслевой научно-технический журнал «Автоматизация и современные технологии» – М: Машиностроение – 2010 – №2, С.16-21.

REFERENCES

1. Iberla K. Factor analysis / M .: "Statistics", 1980. - 398 p.

2. Selection of parameters for monitoring the technical condition of digital blocks the corporate network through the use of factor analysis methods /Petrichenko G.S., Dudnik L.N. //Mezhotraslevoy nauchno-tekhnicheskiy zhurnal «Avtomatizatsiya i sovremennye tekhnologii. – M.: Mashinostroenie. – 2010. – №2, p.16–21.

PRINCIPAL-FACTOR ANALYSIS FOR PRIORITY ASSESSMENT AND PARAMETERS SELECTION FOR FORECASTING TECHNICAL STATE OF COMPUTER NETWORK

L.N. DUDNIK, Y.D. SHEVTSOV, L.M. KRITSKAYA, V.V. BOGDANOV

Kuban State Technological University
2, Moskovskaya st., Krasnodar, Russian Federation, 350072,
e-mail: lududnik@gmail.com

This article proposes the approach for selecting an informative set of parameters to forecast the technical state of the computer network using factor analysis methods. The objective of this approach is to increase the accuracy of forecast and reduce data processing costs.
Key words: computer network, forecasting, choice of informative parameters, the model of factor analysis, correlation matrix, dispersion factors.