СБОРНИК **НАУЧНЫХ ТРУДОВ НГТУ. – 2015. – № 4(82).** – 56–66

УДК 519.23

ГЛОБАЛЬНОЕ И ЛОКАЛЬНОЕ ОЦЕНИВАНИЕ ПАРАМЕТРОВ РЕГРЕССИОННЫХ МОДЕЛЕЙ ПРИ ИСПОЛЬЗОВАНИИ КОНЦЕПЦИИ НЕЧЕТКИХ СИСТЕМ*

А.А. ПОПОВ[1], А.А. ХОЛДОНОВ[2]

[1] *630073, РФ, г. Новосибирск, пр. Карла Маркса, 20, Новосибирский государственный технический университет, доктор технических наук, профессор кафедры теоретической и прикладной информатики. E-mail: a.popov@corp.nstu.ru*
[2] *630073, РФ, г. Новосибирск, пр. Карла Маркса, 20, Новосибирский государственный технический университет, аспирант кафедры теоретической и прикладной информатики. E-mail: firuz_530_11_29@mail.ru*

В работе рассматривается проблема построения регрессионных зависимостей в рамках концепции нечетких систем (Fuzzy Systems). Концепция нечетких систем является достаточно удобным инструментом моделирования при отсутствии априорных предположений о структуре модели. Основное внимание в работе уделено рассмотрению вопроса оценивания параметров результирующих моделей. В качестве метода оценивания неизвестных параметров используется метод наименьших квадратов в так называемом глобальном и локальном вариантах. При использовании локального метода наименьших квадратов параметры отдельных линейных моделей, входящих в систему правил, оцениваются независимо. В глобальном варианте наименьших квадратов осуществляется совместное оценивание всей совокупности неизвестных параметров. В качестве систем правил использовалась модель Такаги–Сугено. При разбиении области определения входных факторов использовались трапециевидные функции принадлежности. Для оценивания точности получаемых решений в работе использовалась среднеквадратичная ошибка (MSE). Для проведения вычислительного эксперимента было разработано соответствующее программное обеспечение. Вычислительный эксперимент проводился на модельных данных. В качестве модели, порождающей данные, использовалась нелинейная зависимость от входного фактора. Дисперсия помехи (уровень шума) определялась в процентах от мощности незашумленного сигнала. Результаты вычислительного эксперимента в работе отражены в табличной форме. В работе отмечаются преимущества и недостатки локального и глобального вариантов метода наименьших квадратов.

Ключевые слова: нечеткие системы (Fuzzy System), регрессионная модель, метод наименьших квадратов, система нормальных уравнений, функция принадлежности, метод центра масс, среднеквадратичная ошибка, оценивание параметров, модель Такаги–Сугено

DOI: 10.17212/2307-6879-2015-4-56-66

* Статья получена 05 октября 2015 г.

ВВЕДЕНИЕ

Технология построения регрессионных зависимостей в рамках концепции нечетких систем (FS) является достаточно удобным инструментом моделирования при отсутствии априорных предположений о структуре модели (составе регрессоров) [1–8]. В этом случае выбор модели происходит, как говорят, на основе самых данных. Универсальность данной методологии базируется на возможности управлять сложностью модели через выбор числа и формы нечетких партиций для входных факторов. Однако эта гибкость и универсальность создают и определенные трудности. Перечислим некоторые из них. Необходимо, например, контролировать полноту покрытия области определения факторов нечеткими партициями; число их достаточно быстро растет при попытке усложнить модель; высока вероятность получения переусложненной модели; есть сложности моделирования при наличии выбросов. В число факторов, действующих на объект, могут входить факторы, измеренные не только в абсолютной шкале или шкале отношений, но и в номинальной шкале. В этом случае необходимо учитывать условия идентифицируемости таких моделей [13, 16]. Некоторые вопросы идентифицируемости подобных моделей с лингвистическими переменными рассмотрены в [15], при этом использовалась методология, изложенная в [14].

В данной работе рассматриваются вопросы построения регрессионных моделей на основе размытых правил Такаги–Сугено с использованием схем глобального и локального оценивания параметров.

1. РЕГРЕССИОННАЯ МОДЕЛЬ. ОЦЕНИВАНИЕ ПАРАМЕТРОВ

Рассматривается классическая задача построения адекватной регрессионной зависимости отклика исследуемого объекта y от набора регрессоров $x = (x_1, \ldots, x_k)^T$. В силу сложности моделируемого объекта влиянию неучтенных факторов не всегда удается однозначно определить вид статистически устойчивой зависимости $y = f(x)$. Часто можно наблюдать, что в различных частях области определения регрессоров более адекватными могут оказываться различные модели. В этом случае можно пытаться получить единую адекватную на исходной выборке модель, значительно ее усложнив. Другой способ может состоять в построении кусочной регрессии. Недостаток первой (переусложненной) модели состоит в значительном риске ее использования для прогноза. Кусочные модели требуют достаточно точного определения области определения ее отдельных частей.

Альтернативным вариантом можно считать построение модели в виде системы размытых правил. Преимуществом таких моделей является то, что получаемое решение есть достаточно гладкая функция. Границы действия отдельных частей модели размыты, что снижает требования к их точному определению. Точность аппроксимации можно варьировать, увеличивая или уменьшая число используемых размытых правил.

Обычная TS (Takagi–Sugeno) система размытых правил может быть записана в следующем виде [1]:

$$IF\ \ x_1 \in A_{1i}\ \&\ x_2 \in A_{2i}\ \&\ \ldots\ \&\ x_k \in A_{ki}\ \ THEN\ \ y = \eta^i(x), \quad i = 1,\ \ldots,\ M, \tag{1}$$

где A_{ji} – нечеткое подмножество для переменной x_j с функцией принадлежности $\mu_{A_{ji}}(x_j)$; M − число правил; $\eta^i(x)$ − функция, определяющая локальную зависимость отклика y от набора регрессоров $x = (x_1, \ldots, x_k)^T$. Прогнозное значение для отклика y определяется обычно по методу центра масс:

$$\hat{y} = \frac{\sum_{i=1}^{M} \mu_i \eta^i}{\sum_{i=1}^{M} \mu_i}, \quad \mu_i = \prod_{j=1}^{k} \mu_{A_{ji}}(x_j). \tag{2}$$

Получить систему размытых правил вида (1), опираясь на имеющуюся выборку данных, − достаточно сложная и неоднозначная задача. На первых этапах ее решения может быть проведена кластеризация данных в пространстве $(x_1, \ldots, x_k, y)$ [9, 12]. Границы кластеров будут определять подмножества A_{ji}. Число кластеров и их границы можно определять достаточно грубо. Необходимое требование здесь – число объектов в *i*-м кластере должно быть достаточным для построения зависимости $\eta^i(x)$, например, линейной или квадратичной.

Модель в виде (1) и (2) будем называть FLR (Fuzzy Logic Regression) регрессионной моделью. Рассмотрим технику построения FLR регрессии для случая построения одномерной зависимости.

Для случая одной переменной x система правил (1) приобретает вид

$$IF \;\; x \in A_i \;\; THEN \;\; y = \eta^i(x), \;\; i = 1, \; \ldots, \; M, \tag{3}$$

где A_i имеют функцию принадлежности $\mu_{A_i}(x)$.

Необходимость нормировки в (2) отпадает, если считать, что функции принадлежности обладают тем свойством, что в любой точке x выполняется условие

$$\sum_{i=1}^{M} \mu_{A_i}(x) = 1. \tag{4}$$

В случае локальной линейной зависимости отклика от фактора функции $\eta^i(x)$ приобретают вид $\eta^i(x) = \theta_0^i + \theta_1^i x$, $i = 1, \; ..., \; M$. В итоге можно считать, что регрессия y по x подчиняется следующему уравнению наблюдения:

$$y_u = \sum_{i=1}^{M} (\theta_0^i + \theta_1^i x_u)\mu_{A_i}(x_u) + e_u, \quad u = 1, \; ..., \; N, \tag{5}$$

где e_u – случайная величина, центрированная и с конечной дисперсией. В случае использования метода наименьших квадратов в глобальном его варианте все неизвестные параметры, входящие в (5), оцениваются совместно. При этом в качестве регрессоров используются следующие:

$$\mu_{A_1}(x), ..., \; \mu_{A_M}(x), \; x\mu_{A_1}(x), ..., \; x\mu_{A_M}(x). \tag{6}$$

Одной из серьезных проблем построения нечетких TS (Takagi–Sugeno) моделей является быстрый рост числа правил вида (1) при увеличении числа нечетких партиций при разбиении области определения входных переменных, равно как и при увеличении числа входных факторов. Например, при двух факторах с разбиением областей их определения на три партиции и линейной правой частью вида $\eta^i(x) = \theta_0^i + \theta_1^i x_1 + \theta_2^i x_2$ нам придется формировать матрицу наблюдений с числом регрессоров, равным 27. При разбиении области определения на пять партиций число регрессоров будет равно 75. При такой размерности информационной матрицы не исключено появление вычислительных проблем, связанных с ее обращением.

В определенной степени снизить остроту данной проблемы возможно, если использовать метод раздельного (локального) оценивания зависи-

мостей $\eta^i(x),\ i=1,...,M$ по взвешенному МНК. Обозначим через $\mu_i = \mu_{A_{1i}}\mu_{A_{2j}},...,\ \mu_{A_{kl}}$ силу высказывания для i-го правила в (1). Введем в рассмотрение целевую функцию для взвешенного МНК следующего вида:

$$S(\theta^i) = (y - X\theta^i)^T W_i (y - X\theta^i) = y^T W_i y - 2\theta^{i\mathrm{T}} X^T W_i y + \theta^{i\mathrm{T}} X^T W_i X \theta^i ,$$

где $W_i = diag(\mu_{i1},\ \mu_{i2},...,\ \mu_{iN})$, μ_{ij} – значение μ_i в j-й точке.

Первые частные производные $S(\theta^i)$ по параметрам θ^i имеют вид

$$\frac{\partial S(\theta^i)}{\partial \theta^i} = -2X^T W_i y + 2X^T W_i X \theta^i .$$

Приравнивая их нулю, получаем систему нормальных уравнений

$$X^T W_i X \hat{\theta}^i = X^T W_i y$$

с решением

$$\hat{\theta}^i = (X^T W_i X)^{-1} X^T W_i y .$$

Видим, что параметры локальных моделей в этом случае оцениваются независимо. Следует ожидать, что качество получаемых таким способом решений будет несколько хуже, чем при использовании глобального МНК.

2. ВЫЧИСЛИТЕЛЬНЫЙ ЭКСПЕРИМЕНТ

Для проведения вычислительных экспериментов было разработано программное обеспечение, которое позволяет осуществлять построение кусочной регрессионной зависимости FLR типа для отклика исследуемого объекта от набора регрессоров с оцениванием параметров по глобальному и локальному МНК.

В качестве модели, порождающей данные в вычислительном эксперименте, использовалась следующая зависимость:

$$\begin{cases} if\ (x \le 0)\ then\ y = 2 + 8x^2, \\ if\ (x \ge 0)\ and\ (x \le 1)\ then\ y\ = 3 + \dfrac{x^2}{2}. \end{cases}$$

Интервал варьирования фактора входного фактора от –1 до +1. Количество наблюдений бралось равным 201, уровень зашумления варьировался от 5 % до 25 % от мощности полезного сигнала. Число подобластей, на которые разбивался весь интервал варьирования входного фактора, варьировался от 2 до 5. При этом использовались трапециевидные функции принадлежности.

Скриншот работы программы с визуализацией получаемого решения представлен на рис. 1.

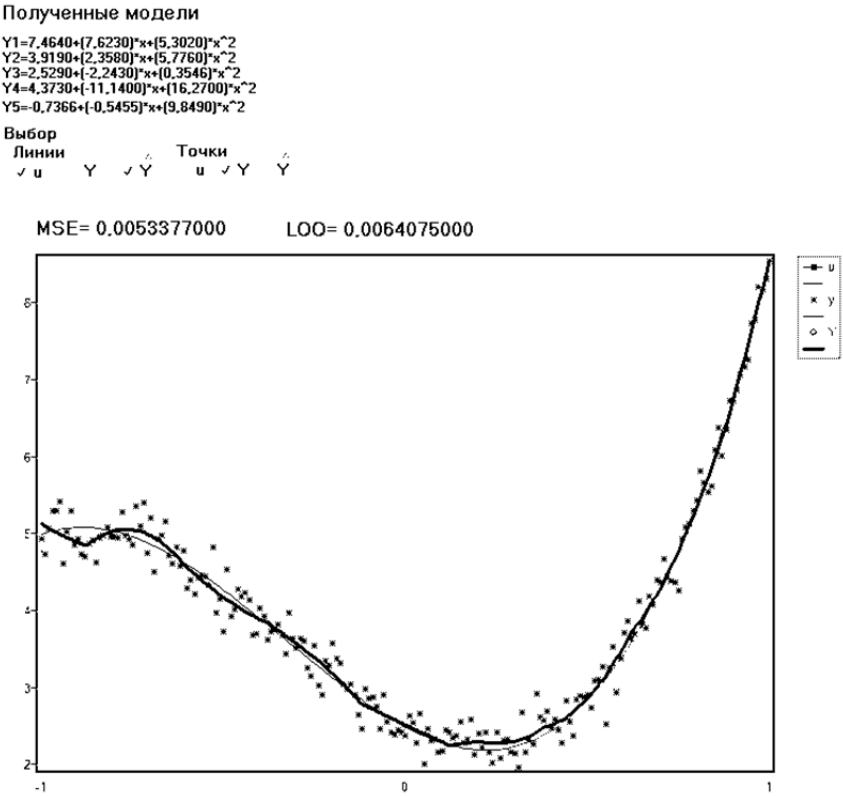

Рис. 1. Скриншот работы программы с графиками зависимостей:

u(−) – незашумленный отклик; Y(*) – зашумленный отклик; $\hat{Y}$(▬) – решение по построенной модели

В таблице отражены результаты проведенного вычислительного эксперимента. В качестве локальных моделей $\eta^i(x)$ использовались линейные и квадратичные. Анализ результатов вычислительного эксперимента позволяет сделать ряд заключений. Во-первых, точность получаемых решений существенно зависит как от числа партиций, на которые разбивается область определения действующего фактора, так и от выбираемой локальной модели (линейная или квадратичная). Во-вторых, излишнее усложнение модели приводит к эффекту переобучения. Это заметно, например, если использовать в качестве локальной квадратичную модель на четырех или пяти партициях.

Сравнение значений MSE для моделей с оценками параметров по локальному и глобальному МНК

Число партиций	Шум в (%)	Способ оценки параметров			
		Локальный МНК		Глобальный МНК	
		Модель		Модель	
		линейная	квадратичная	линейная	квадратичная
2	5	0,487966	0,024777	0,13024	0,016493
	10	0,492780	0,026029	0,13118	0,017481
	15	0,498059	0,027887	0,13273	0,019127
	20	0,503802	0,030352	0,1349	0,021433
	25	0,510009	0,033424	0,13769	0,024397
3	5	0,087776	0,010096	0,051116	0,002326
	10	0,090501	0,011731	0,052231	0,003557
	15	0,093847	0,014027	0,054089	0,005608
	20	0,097816	0,016981	0,056689	0,00848
	25	0,102407	0,020596	0,060033	0,012172
4	5	0,035654	0,001693	0,021307	0,000922
	10	0,037811	0,003048	0,022549	0,003188
	15	0,040585	0,005299	0,024619	0,006964
	20	0,043977	0,008444	0,027517	0,012251
	25	0,047986	0,012485	0,031243	0,019048
5	5	0,016513	0,000832	0,010585	0,001358
	10	0,018099	0,002685	0,01204	0,005338
	15	0,020427	0,005815	0,014464	0,01197

	20	0,023498	0,010222	0,017858	0,021256
	25	0,027311	0,015908	0,022222	0,033194

ЗАКЛЮЧЕНИЕ

Исследования показали, что реализованный метод построения регрессионной зависимости является достаточно гибким инструментом восстановления зависимостей по зашумленным данным в условиях структурной неопределенности. Оценивание параметров по локальному МНК несколько уступает по точности глобальному варианту. Однако использование локального метода наименьших квадратов (МНК) может оказаться востребованным в случае, когда число правил в нечеткой системе достаточно велико.

СПИСОК ЛИТЕРАТУРЫ

1. *Takagi T., Sugeno M*. Fuzzy identification of systems and its applications to modeling and control // IEEE Transactions on Systems, Man, and Cybernetics. – 1985. – Vol. 15, iss. 1. – P. 116–132.

2. *Kosko B*. Fuzzy systems as universal approximators // IEEE Transactions on Computers. – 1994. – Vol. 43, iss. 11. – P. 1329–1333.

3. *Hao Ying*. General SISO Takagi–Sugeno fuzzy systems with linear rule consequent are universal approximators // IEEE Transactions on Fuzzy Systems. – 1998. – Vol. 6, iss. 4. – P. 582–587.

4. *Круглов В.В.*, *Дли М.М.*, *Голунов Р.Ю*. Нечеткая логика и искусственные нейронные сети. – М.: Физматлит, 2001. – 224 с.

5. *Пегат А.* Нечеткое моделирование и управление: пер. с англ. – 2-е изд. – М.: Бином, 2013. – 798 с.

6. *Lilly J.H*. Fuzzy control and identification. – Hoboken: Wiley, 2010. – 231 p.

7. *Babuska R.* Fuzzy modelling for control. – London; Boston: Kluwer Academic Publishers, 1998. – 257 p.

8. *Попов А.А*. Регрессионное моделирование на основе нечетких правил // Сборник научных трудов НГТУ. – 2000. – № 2 (19). – С. 49–57.

9. *Babuska R*., *Verbruggen H.B*. Constructing fuzzy models by product space clustering // Fuzzy Model Identification: Selected Approaches / H. Hellendoorn, D. Driankov, eds. – Berlin: Springer, 1997. – P. 53–90.

10. *Abonyi J.*, *Szeifert F.*, *Babuska R*. Modified Gath–Geva fuzzy clustering for identification of Takagi–Sugeno fuzzy models // IEEE Transactions on Systems, Man, and Cybernetics. Pt. B. – 2002. – Vol. 32, iss. 5. – P. 612–621.

11. *Chen J.*, *Xi Y.*, *Zhang Z.* A clustering algorithm for fuzzy model identification // Fuzzy Sets and Systems. – 1998. – Vol. 98. – P. 319–329.

12. *Bezdek J.C.* Pattern recognition with fuzzy objective function algorithms. – New York: Plenum Press, 1981. – 272 p.

13. *Попов А.А.* Конструирование дискретных и непрерывно-дискретных моделей регрессионного типа // Сборник научных трудов НГТУ. – 1996. – № 1 (3). – С. 21–30.

14. *Лях К.Н.*, *Попов А.А.* Анализ линейных моделей мягкого дисперсионного анализа // Сборник научных трудов НГТУ. – 2003. – № 1 (31). – С. 85–90.

15. *Попов А.А.* Построение деревьев решений для прогнозирования количественного признака на классе логических функций от лингвистических переменных // Научный вестник НГТУ. – 2009. – № 3 (36). – С. 77–86.

16. *Попов А.А.* Оптимальное планирование эксперимента в задачах структурной и параметрической идентификации моделей многофакторных систем: монография. – Новосибирск: Изд-во НГТУ, 2013. – 296 с.

Попов Александр Александрович, доктор технических наук, профессор кафедры теоретической и прикладной информатики Новосибирского государственного технического университета. Основное направление научных исследований – методы анализа данных, оптимальное планирование экспериментов. Имеет более 150 научных работ, в том числе 2 монографии. E-mail: a.popov@corp.nstu.ru

Холдонов Абдурахмон Абдуллоевич, аспирант кафедры теоретической и прикладной информатики Новосибирского государственного технического университета. E-mail: firuz_530_11_29@mail.ru

Global and Local Parameter Estimation of Regression Models Using Fuzzy Systems Concept *

A.A. Popov[1], A.A. Holdonov[2]

[1] *Novosibirsk State Technical University, 20 K. Marx Prospekt, Novosibirsk, 630073, Russian Federation , Popov A.A., Doctor of Sciences (Engineering), Professor of the Department of Theoretical and Applied Informatics. E-mail: a.popov@corp.nstu.ru*

[2] *Novosibirsk State Technical University, 20 K. Marx Prospekt, Novosibirsk, 630073, Russian Federation, Postgraduate Student of the Department of Theoretical and Applied Informatics. E-mail: firuz_530_11_29@mail.ru*

* *Received 05 October 2015.*

This paper considers the problem of regression dependences construction within fuzzy systems (Fuzzy Systems) concept.. Such a concept is quite a handy simulation tool if there are no prior assumptions on the model structure. The main attention is paid to the estimation of the resulting models parameters. The least squares method in so-called global and local versions is used as the method of the unknown parameters estimation. When using the least squares local method , the parameters of some certain linear models, included in the system rules, are estimated independently. The least squares global method is carried out by joint estimation of unknown parameter part. Takagi-Sugeno model was used as a rule systems. While splitting the domain of the input factors was used trapezoidal membership functions. To assess the accuracy of the solutions calculated was used the mean square error (MSE). To carry out a computational experiment was developed appropriate software. Computer experiment was performed on simulated data. As a generating data model was used nonlinear dependence on the input factor. Noise dispersion (noise level) was determined as a percentage of not noise-contaminated signal power. Results of the computational experiment are given in a tabular form in this paper. The paper analyzes the advantages and the disadvantages of local and global versions of the least squares method estimation.

Keywords: fuzzy systems (Fuzzy System), regression model, least squares method, the system of normal equations, membership function, a method of center of mass, mean square error, parameter estimation, model of Takagi-Sugeno

DOI: 10.17212/2307-6879-2015-4-56-66

REFERENCES

1. Takagi T., Sugeno M. Fuzzy identification of systems and its applications to modeling and control. *IEEE Transactions on Systems, Man, and Cybernetics*, 1985, vol. 15, iss. 1, pp. 116–132.

2. Kosko B. Fuzzy systems as universal approximators. *IEEE Transactions on Computers*, 1994, vol. 43, iss. 11, pp. 1329–1333.

3. Hao Ying. General SISO Takagi–Sugeno fuzzy systems with linear rule consequent are universal approximators. *IEEE Transactions on Fuzzy Systems*, 1998, vol. 6, iss. 4, pp. 582–587.

4. Kruglov V.V., Dli M.M., Golunov R.Yu. *Nechetkaya logika i iskusstvennye neironnye seti* [Fuzzy logic and artificial neural networks]. Moscow, Fizmatlit Publ., 2001. 224 p.

5. Piegat A. *Fuzzy modeling and control*. Heidelberg, Physica-Verlag, 2001. 728 p. (Russ. ed.: Pegat A. *Nechetkoe modelirovanie i upravlenie*. 2nd ed. Translation from English. Moscow, Binom Publ., 2013. 798 p.).

6. Lilly J.H. *Fuzzy control and identification*. Hoboken, Wiley, 2010. 231 p.

7. Babuska R. *Fuzzy modelling for control*. London, Boston, Kluwer Academic Publishers, 1998. 257 p.

8. Popov A.A. Regressionnoe modelirovanie na osnove nechetkikh pravil [Regression modeling based on fuzzy rules]. *Sbornik nauchnykh trudov Novosibirskogo*

gosudarstvennogo tekhnicheskogo universiteta – Transaction of scientific papers of the Novosibirsk state technical university, 2000, no. 2 (19), pp. 49–57.

9. Babuska R., Verbruggen H.B. Constructing fuzzy models by product space clustering. *Fuzzy Model Identification: Selected Approaches*. Ed. by H. Hellendoorn, D. Driankov. Berlin, Springer, 1997, pp. 53–90.

10. Abonyi J., Szeifert F., Babuska R. Modified Gath–Geva fuzzy clustering for identification of Takagi–Sugeno fuzzy models. *IEEE Transactions on Systems, Man, and Cybernetics. Pt. B*, 2002, vol. 32, iss. 5, pp. 612–621.

11. Chen J., Xi Y., Zhang Z. A clustering algorithm for fuzzy model identification. *Fuzzy Sets and Systems*, 1998, vol. 98, pp. 319–329.

12. Bezdek J.C. *Pattern recognition with fuzzy objective function algorithms*. New York, Plenum Press, 1981. 272 p.

13. Popov A.A. Konstruirovanie diskretnykh i nepreryvno-diskretnykh modelei regressionnogo tipa [Construction of discrete and continuous-discrete models such as regression]. *Sbornik nauchnykh trudov Novosibirskogo gosudarstvennogo tekhnicheskogo universiteta – Transaction of scientific papers of the Novosibirsk state technical university*, 1996, no. 1 (3), pp. 21–30.

14. Lyakh K.N., Popov A.A. Analiz lineinykh modelei myagkogo dispersionnogo analiza [The analysis of linear models of soft-way analysis of variance]. *Sbornik nauchnykh trudov Novosibirskogo gosudarstvennogo tekhnicheskogo universiteta – Transaction of scientific papers of the Novosibirsk state technical university*, 2003, no. 1 (31), pp. 85–90.

15. Popov A.A. Postroenie derev'ev reshenii dlya prognozirovaniya kolichestvennogo priznaka na klasse logicheskikh funktsii ot lingvisticheskikh peremennykh [Construction of decision trees to predict the quantitative trait in the class of logic functions of linguistic variables]. *Nauchnyi vestnik Novosibirskogo gosudarstvennogo tekhnicheskogo universiteta – Science bulletin of the Novosibirsk state technical university*, 2009, no. 3 (36), pp. 77–86.

16. Popov A.A. *Optimal'noe planirovanie eksperimenta v zadachakh strukturnoi i parametricheskoi identifikatsii modelei mnogofaktornykh sistem* [The optimal design of experiments in the problems of structural and parametric identifycation of models of multivariate systems]. Novosibirsk, NSTU Publ., 2013. 296 p.