УДК 519.254

И.А. Карлов

ВОССТАНОВЛЕНИЕ ПРОПУЩЕННЫХ ДАННЫХ ПРИ ЧИСЛЕННОМ МОДЕЛИРОВАНИИ СЛОЖНЫХ ДИНАМИЧЕСКИХ СИСТЕМ

I.A. Karlov

THE MISSING VALUE ESTIMATION IN NUMERICAL MODELLING OF COMPLEX DYNAMIC SYSTEMS

Рассмотрена проблема пропущенных значений в массивах данных при моделировании сложных динамических систем. Изучены различные типы пропусков и общие подходы к работе с данными, содержащими пропуски. Приведен обзор самых распространенных методов восстановления пропущенных данных. Представлен оригинальный гибридный адаптивный метод восстановления с нейро-нечетким управлением. Приведена оценка эффективности различных методов восстановления применительно к массивам данных, содержащих информацию о процессах в сложных динамических системах. Отдельное внимание уделено вопросу влияния наличия пропусков в данных на эффективность численных моделей.

ПРОПУЩЕННЫЕ ДАННЫЕ; НЕЙРОННЫЕ СЕТИ; НЕЧЕТКАЯ ЛОГИКА; ЧИСЛЕННОЕ МОДЕЛИРОВАНИЕ; ГИБРИДНЫЕ СИСТЕМЫ.

This paper is deals with the problem of missing values in the modelling of complex dynamic systems. Considered different types of missing values and general approaches to dealing with missing data. This paper provides an overview of the most common methods of missing data estimation, and presents an original hybrid adaptive method of estimation with neuro-fuzzy control. Evaluate the effectiveness of various methods of missing data estimating applied to data sets that contain information about the processes in complex dynamic systems. Special attention is paid to the impact of missing data presence on the effectiveness of the models.

MISSING DATA; NEURAL NETWORKS; FUZZY LOGIC; NUMERICAL MODELLING; HYBRID SYSTEMS.

Одним из подходов, используемых при моделировании процессов в сложных динамических системах, является использование статистических массивов данных, содержащих информацию о процессах, происходящих в системе, и построение на их основе индуктивных численных моделей. При этом качество построенных моделей, их соответствие реальным процессам, происходящим в системе, напрямую зависит от качества массива данных.

На практике при построении численной модели на основе статистических данных исследователи часто сталкиваются с проблемой пропущенных значений в массивах данных. Наличие пропусков существенно осложняет применение различных методов обработки информации, снижает эффективность статистических моделей и сказывается на качестве принимаемых управленческих решений.

Данная проблема возникает при численном моделировании в различных областях и свойственна как для технических, так и для социальных систем. Так, впервые, с проблемой пропущенных значений мы столкнулись при моделировании нестабильности работы алюминиевого электролизера [1]. Число экземпляров, содержащих пропуски, достигало 10, а иногда и 15 %. Позже мы столкнулись с подобной проблемой при изучении зависимости детской смертности от показателей социально-экономического развития территорий [2, 3]. Здесь количество экземпляров данных с пропусками для различных массивов варьировалось в пределах от 3 до 25 %.

В представленной статье рассмотрены наиболее распространенные подходы к работе с данными, содержащими пропуски, приводится оригинальный метод восстановления пропусков в данных, дается оценка влияния наличия пропусков на эффективность численных моделей. Все результаты получены и проверены на реальных массивах данных, описывающих процессы в сложных динамических системах.

Основные подходы к работе с пропусками в массивах данных

На практике встречается несколько подходов к работе с массивами данных, содержащих пропущенные значения.

Первый подход, наиболее простой в реализации, – это удаление экземпляров, содержащих пропущенные значения, из массива и работа только с полными данными [4]. Использование данного подхода выглядит целесообразным, если пропуски данных носят единичный характер. Но даже в этом случае имеется серьезная опасность при удалении данных «потерять» важные закономерности. В том же случае, когда количество пропусков велико, удаление соответствующих экземпляров может привести к дефициту данных и даже невозможности дальнейшей обработки.

Вторым подходом является использование специальных модификаций методов обработки данных, допускающих наличие пропусков в массиве. В [5] приведен ряд модификаций методов классификации и кластеризации для работы с данными, содержащими пропущенные значения.

И, наконец, третьим подходом, наиболее распространенным, является использование методов оценки значений пропущенных элементов. Данные методы помогают заполнить пропуски в массивах, основываясь на некоторых предположениях о значении отсутствующих данных.

Принципиальная применимость и эффективность того или иного подхода зависит от количества пропусков в данных и причин, по которым они образовались. С точки зрения природы возникновения данных традиционно выделяют три категории пропусков [6].

1. Случайный пропуск. Факт появления пропуска в массиве не зависит ни от самого пропущенного значения, ни от значений других атрибутов. Пропуски случайны, носят единичный характер и связаны, как правило, с потерями при вводе, хранении или передаче данных. Очень важно понимать, что соответствующая информация на самом деле существует, но она по каким-то причинам отсутствует в массиве. В этом случае вполне эффективными оказываются как второй, так и третий подходы. В случае, когда число пропусков невелико, возможно использование и первого подхода.

2. Как бы случайный пропуск. Как и в первом случае, отсутствующие данные существуют. Основным отличием является то, что в появлении соответствующих пропусков наблюдаются закономерности. Факт появления такого пропуска, как правило, зависит от значений одного или нескольких других атрибутов. На практике такие пропуски возникают по причине сложности процессов измерения показателей и, как следствие, нерегулярности или меньшей частоты измерений, в случае систематических сбоев при вводе, хранении и передачи информации от какой-либо части изучаемой системы, например, удаленного производственного цеха или отдельной территории.

Поскольку пропуски в данных носят систематический характер, их удаление из массива может привести к потере важных зависимостей, поэтому в данном случае неприменим первый подход. Также открытым остается вопрос о применимости второго подхода. Наиболее эффективным является использование методов восстановления пропущенных данных.

3. Неслучайный пропуск. Значение пропущенного элемента не существует. Появление подобного пропуска в данных может зависеть от самого атрибута, от других пропусков в массиве, а также от значений других атрибутов. Факт появления может быть связан с наличием в выборке взаимоисключающих показателей либо показателей, которые измеряются только для отдельных групп экземпляров. Например, для разных типов объектов в технических и социальных

системах. Для работы с пропусками данного типа неприменимы методы восстановления, единственным допустимым подходом является проведение анализа предметной области и принятие решения на основе полученных выводов.

Методы восстановления пропусков в массивах данных

В настоящее время существует множество различных методов, отличающихся своей вычислительной сложностью, универсальностью и точностью работы.

Наиболее простыми являются метод подстановки среднего значения атрибута, вычисленного по всем известным значениям [4], и метод ближайших соседей [7]. Широкое распространение получили регрессионные методы [4, 8], локальные алгоритмы [9] и методы максимального правдоподобия [4].

Особого внимания заслуживают нелинейные методы, использующие искусственные нейронные сети, генетические алгоритмы и методы нечеткой логики. Среди них можно выделить алгоритмы, осуществляющие непосредственное предсказание пропущенного значения [10], и алгоритмы, основанные на минимизации ошибки работы системы [11].

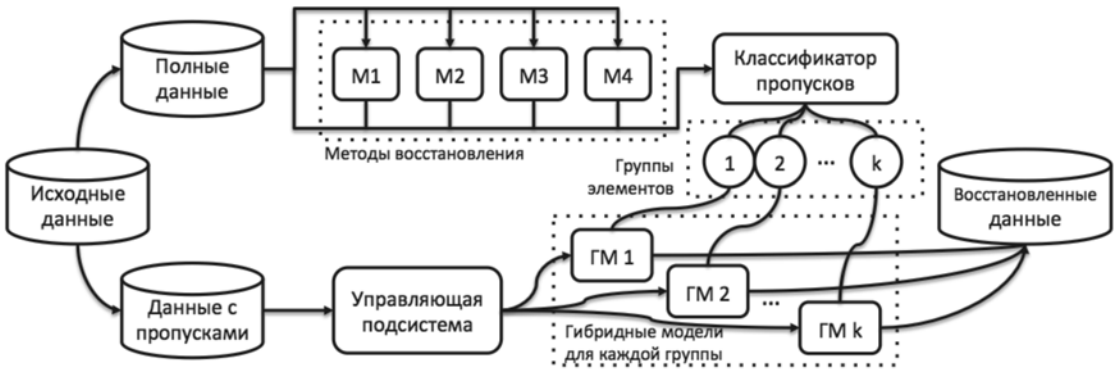
В [12] представлен сравнительный анализ эффективности ряда методов на различных массивах данных. В ходе проведенного анализа было замечено, что эффективность работы методов существенно варьируется не только на разных массивах, но и для одного массива на разных пропусках. В связи с этим возникла идея создания гибридного метода, который смог бы объединить сильные стороны каждого из методов.

Гибридный адаптивный метод с нейро-нечетким управлением. Предлагаемый метод включает в себя два этапа: построение гибридной системы и восстановление пропущенных значений. На рисунке представлена общая схема предлагаемого метода.

Как видно на рисунке, гибридная система состоит из набора гибридных моделей и управляющей подсистемы, которая принимает решение о выборе модели для каждого поступающего на вход вектора данных. Для создания гибридной модели были отобраны четыре метода: метод подстановки среднего значения [4] (М1), метод ближайших соседей [7] (М2), двухшаговый em-алгоритм [4] (М3), метод, основанный на использовании авто-ассоциативных систем [11] (М4).

При построении гибридной системы используется часть массива, содержащая только полные данные. Случайным образом оттуда удаляется некоторое количество элементов, и полученные пропуски восстанавливаются с применением всех методов, используемых в гибридной системе. Сравнивая полученные результаты с исходным значением, получаем погрешности работы каждого из методов на каждом исключенном значении.

Используя один из методов кластеризации на множестве векторов, содержащих ошибки, получаем разделение экземпляров с удаленными значениями на кластера –



Общая схема гибридного адаптивного метода с нейро-нечетким управлением

группы элементов. Наиболее перспективным на этом шаге видится использование метода нечеткой кластеризации [14].

Для каждой из групп значений строится собственная гибридная модель. Интерпретация результатов кластерного анализа позволяет получить оценку того, какие из методов целесообразно использовать в каждой конкретной модели. В зависимости от результатов анализа при построении модели мы можем использовать один из методов, несколько или весь набор методов. В случае использования более одного метода, в качестве инструмента гибридизации применяем линейные взвешенные модели, либо многослойные нейронные сети прямого распространения сигнала.

На последнем этапе мы настраиваем управляющую подсистему, которая, получая на вход вектор из массива данных, должна дать оценку того, какую из гибридных моделей необходимо использовать для данного вектора. В качестве инструмента для построения управляющей подсистемы предлагается использовать модифицированную систему нейро-нечеткого вывода, способную работать с векторами, содержащими пропуски [5].

Такой подход позволяет использовать преимущества отдельных методов, преодолевая некоторые недостатки других, и тем самым получить более мощный интеллектуальный анализатор.

Тестирование методов. Для тестирования методов использовалась часть данных, не используемая при построении гибридной модели, из которой случайным образом удалялось некоторое количество элементов. Затем с использованием различных методов восстановления получали оценку пропущенного значения. Полученные оценки сравнивались с исходным значением и вычислялись следующие параметры: точность – доля пропущенных значений, предсказанных с ошибкой, меньшей 20 %, и погрешность – средняя квадратичная ошибка предсказания.

Тестирование методов проводилось на двух массивах данных.

Массив 1. Выборка значений технологических показателей работы алюминиевой электролизной ванны. Число атрибутов – 10, число экземпляров – 1092.

Массив 2. Выборка значений показателей социально-экономического развития городов и муниципальных районов Красноярского края за период 2005–2009 гг. Число атрибутов – 26, число экземпляров – 250.

В табл. 1 представлены результаты тестирования гибридного метода и четырех методов, используемых в гибридной системе. Как мы видим, для обоих массивов гибридный метод дает более точную оценку пропущенных значений.

Влияние пропусков на эффективность численных моделей

Для оценки влияния пропусков на эффективность работы численных моделей использованы исходные данные и результаты исследований, представленных в [1–3]. Исходные массивы данных, содержащие пропущенные значения, обрабатывались с помощью гибридного адаптивного метода и затем использовались для обучения и настройки моделей, аналогичных построенным в оригинальных исследованиях. Сравнение показателей работы оригинальных

Таблица 1

Результаты тестирования гибридного и оригинальных методов

Методы	Массив 1, %		Массив 2, %	
	точность	погрешность	точность	погрешность
1	33,5	39,3	42,2	27,8
2	73,5	18,1	68,2	21,1
3	68,8	24,8	69,6	18,8
4	59,5	23,1	66,0	20,0
Гибридный метод	90,7	14,7	85,8	16,0

моделей и моделей, построенных на восстановленных данных, дает возможность оценить степень влияния пропусков на эффективность работы моделей.

Аппроксимационная модель технологического параметра в процессе производства алюминия. Промышленная электролизная ванна для производства алюминия (электролизер) состоит из угольных анодов, погружённых в расплавленный электролит, в котором растворен глинозем. Слой расплавленного электролита расположен над расплавом алюминия.

В процессе электролиза в ванне на поверхности раздела алюминия и электролита могут возникать волны. Образование таких волн приводит к переносу металла в область электролита и снижает экономические показатели работы электролизера. При некоторых условиях наблюдается рост амплитуд таких волн, что называется неустойчивостью работы электролизера. В результате колебаний раздела происходят изменения напряжения на ванне. Количественной характеристикой таких изменений является технологический параметр «уровень шума», измеряемый в вольтах.

В рамках проводимого исследования [1] решалась задача изучения зависимости параметра «уровень шума» от других технологических параметров. Были спроектированы и построены аппроксимационные модели на основе искусственных нейронных сетей прямого распространения сигнала, которые на основе значений 18 технологических параметров вычисляли значение параметра «уровень шума». Отметим, что при обучении моделей из массивов данных исключались экземпляры, содержащие одно или более пропущенных значений.

В данной работе мы использовали тот же подход к построению моделей и те же массивы данных, пропущенные элементы в которых были восстановлены с помощью гибридного адаптивного метода с нейронечетким управлением.

В табл. 2 приведено сравнение точности двух моделей, построенных ранее, с их эквивалентами, построенными сейчас на полных массивах данных. Отметим, что при использовании массивов с заполненными пропусками мы наблюдаем повышение числа примеров тестовой выборки, вычисленных с заданной точностью, а также заметное снижение величины максимальной погрешности как для первой, так и для второй модели. Что касается средней ошибки по всей тестовой выборке, то здесь также наблюдается снижение, но не настолько значительное.

Классификационная модель при изучении влияния различных факторов на показатели детской смертности. Детская смертность является важнейшей группой показателей, во многом определяющих демографическую ситуацию в стране. Правильный и своевременный анализ детской смертности позволяет разработать ряд конкретных мер по снижению заболеваемости и смертности, оценить эффективность проведенных ранее мероприятий, в значительной мере охарактеризовать работу местных органов здравоохранения по охране материнства и детства.

Таблица 2

Основные показатели работы аппроксимационных моделей, построенных на оригинальных и восстановленных данных

Ошибки	Модель 1, %		Модель 2, %	
	оригинальная	новая	оригинальная	новая
Средняя ошибка	8,24	7,96	4,23	4,19
Максимальная ошибка	45,46	21,63	41,32	29,59
Доля значений, вычисленных с ошибкой не более 10 %	88,75	93,17	95,46	97,52
Доля значений, вычисленных с ошибкой не более 5 %	74,31	80,37	83,20	87,73

Таблица 3

Основные показатели работы классификационных моделей, построенных на оригинальных и восстановленных данных

Возрастная категория	Количество правильно классифицированных объектов, %		
	исходные модели	модели на этапе 1	модели на этапе 2
До года	75,22	76,99	80,23
От 1 до 14 лет	81,31	81,51	83,48
От 15 до 19 лет	83,48	83,97	88,79

В рамках проводимых исследований [2, 3] был поставлен и решен ряд задач: анализ структуры показателей смертности детей в городских округах и муниципальных районах Красноярского края за 2006–2009 гг., выделение групп территорий со схожей структурой показателей; изучение зависимости показателей детской смертности от показателей социально-экономического развития территории.

Для проведения исследований использовались массив данных по показателям детской смертности, включающий данные более чем по 250 причинам смерти для каждой из групп возрастов (до года, 1–4 года, 5–9 лет, 10–14, 15–19 лет), а также массив данных, включающий результаты измерений 26 показателей социально-экономического развития, относящихся к различным сферам. Перечень показателей получен в результате экспертного опроса.

В ходе исследования были выделены три группы возрастов, различающиеся по структуре показателей смертности (до года, от 1 до 14 лет, от 14 до 19 лет), для каждой из возрастных групп определены наиболее существенные причины детской смертности (пять – для первой группы, семь – для второй и четыре – для третьей), все территории разбиты на группы со сходной картиной значений основных показателей детской смертности (шесть групп для возраста до 1 года, шесть для возраста от 1 до 14 лет, семь для возраста от 15 до 19 лет).

Для каждой из групп возрастов были построены классификационные модели, позволяющие по набору значений показателей социально-экономического развития территории определить, к какой группе относится данная территория. При построении оригинальных моделей все неполные экземпляры удалялись из массива и построение осуществлялось только на полных данных.

Оценка степени влияния пропусков в данных на точность работы этих моделей проходила в два этапа. На первом этапе были получены оценки и заполнены пропуски в первом массиве (показатели детской смертности), после чего снова выполнили все действия по построению и настройки модели. На втором этапе были получены оценки и заполнены еще и пропуски в массиве, содержащем значения показателей социально-экономического развития. Сравнение точности работы оригинальных моделей, моделей, построенных на этапах 1 и 2, приведены в табл. 3.

Заметим, что при использовании заполненного массива данных, содержащего показатели детской смертности, мы наблюдали лишь незначительное повышение качества классификационной модели (от 0,5 до 1,5 %). Это может быть связано как с незначительным количеством пропусков (около 8 % экземпляров), так и со случайным характером этих пропусков.

На втором этапе при использовании заполненного массива данных, содержащих показатели социально-экономического развития, мы наблюдали более серьезный рост точности работы модели. Это говорит о том, что данные пропуски являются существенными в контексте поставленной задачи.

Проведенное исследование показало применимость методов восстановления пропусков в массивах данных и, в частности, гибридного адаптационного метода

для заполнения пропущенных значений в массивах данных, содержащих информацию о процессах, происходящих в сложных динамических системах.

Кроме того, в ходе исследования получено экспериментальное подтверждение негативного влияния пропусков в данных на эффективность работы моделей, а также продемонстрированы возможности методов восстановления пропусков в массивах данных при численном моделировании сложных динамических систем.

СПИСОК ЛИТЕРАТУРЫ

1. **Карлов И.А., Проворова О.Г.** Новый подход к исследованию устойчивости алюминиевого электролизера // Вестник Красноярского гос. ун-та. Физико-математические науки. – 2002. – № 1. – С. 116–120.

2. **Карлов И.А.** Исследование структуры показателей детской смертности в городских округах и муниципальных районах Красноярского края // Нейроинформатика, ее приложения и анализ данных: Матер. XIX Всеросс. семинара. – 2011. – С. 66–72.

3. **Карлов И.А.** Анализ показателей детской смертности в городских округах и муниципальных районах Красноярского края с использованием искусственных нейронных сетей и методов нечеткой логики // Всеросс. конф. Математическое моделирование и информационно-вычислительные технологии в междисциплинарных научных исследованиях. – 2011. – С. 64–65.

4. **Литтл Р.Дж.А., Рубин Д.Б.** Статистический анализ данных с пропусками. – М.: Финансы и статистика, 1991. – 336 с.

5. **Garcia-Laencina P.J., Sanco-Gomez J.-L., Figueiras-Vidal A.R.** Pattern classification with missing data: a review. – London: Springer-Verlag Limited, 2009.

6. **Schafer J.L., Graham J.W.** Missing data: Our view to the state of the art // Psychological methods. – 2002. – Vol.7. – № 2. – С.147–177.

7. **Zloba E., Yatskiv I.** Statistical methods for estimating missing data // Computer Modelling and New Technologies. – 2002. – № 6(1). – P. 51–61.

8. **Россиев А.А.** Моделирование данных при помощи кривых для восстановления пробелов в таблицах // Методы нейроинформатики: сб. научн. тр.; под ред. Горбаня А.Н. – Красноярск: Изд-во КГТУ, 1998. – С. 6–22.

9. **Загоруйко Н.Г.** Прикладные методы анализа данных и знаний. – Новосибирск: Изд-во Ин-та математики, 1999.

10. **Gupta A., Lam M.S.** Estimation Missing Values using Neural Networks // J. of Operational Research Society. – 1996. – Vol. 47. – № 2. – С. 229–239.

11. **Nelwamondo F.V., Mohamed S., Marwala T.** Missing Data: A comparison of neural network and expectation maximization techniques // Current Science. – 2007. – Vol. 93. – № 11. – С.1467–1473.

12. **Карлов И.А.** Методы восстановления пропущенных значений с использованием инструментария Data Mining // Вестник Сибирского гос. аэрокосмического ун-та им. академика М.Ф. Решетнева. – 2011. – № 7(40) – С.29–33.

13. **Карлов И.А., Кошур В.Д.** Подходы к построению гибридной модели для оценки значений пропущенных элементов в массивах данных // Нейроинформатика, ее приложения и анализ данных: Матер. XX Всеросс. семинара. – 2012. – С. 174–179.

14. **Halkidi M., Batistakis Y., Vazirgiannis M.** On Clustering Validation Techniques // J. of Intelligent Information Systems. – 2003. – № 17:2/3. – С.107–145.

REFERENCES

1. **Karlov I.A., Provorova O.G.** Noviy podhod k issledovaniyu ustoychivosti aluminievogo electrolizera / Vestnik Krasnoyarskogo gos. un-ta. Fiziko-matematicheskie nauki. – 2002. – № 1. – S. 116–120. (rus)

2. **Karlov I.A.** Issledovanie struktury pokazateley detskoy smertnosti b gorodskih okrugah i municipalnyh raionov Krasnoyarskogo kraya / Neiroinformatika, ee prilogeniya i analiz dannyh: mater. XX Vseross. seminara. – 2011. – S. 66–72. (rus)

3. **Karlov I.A.** Analiz pokazatelej detskoj smertnosti v gorodskih okrugah i municipal'nyh rajonah Krasnojarskogo kraja s ispol'zovaniem iskusstvennyh nejronnyh setej i metodov nechetkoj logiki / Vseross. konf. Matematicheskoe modelirovanie i informacionno-vychislitel'nye tehnologii v mezhdisciplinarnyh nauchnyh issledovanijah. – 2011. – S. 64–65. (rus)

4. **Little R.J.A., Rubin D.B.** Statisticheskiy analiz dannyih s propuskami. – Moscow: Finansy i statistika, 1991. – 336 s. (rus)

5. **Garcia-Laencina P.J., Sanco-Gomez J.-L., Figueiras-Vidal A.R.** Pattern classification with missing data: a review. – London: Springer-Verlag

Limited, 2009.

6. **Schafer J.L., Graham J.W.** Missing data: Our view to the state of the art / Psychological methods. – 2002. –Vol. 7. – № 2. – P. 147–177.

7. **Zloba E., Yatskiv I.** Statistical methods for estimating missing data / Computer Modelling and New Technologies. – 2002. – № 6(1). – P. 51–61.

8. **Rossiev A.A.** Modelirovanie dannyh pri pomoshhi krivyh dlja vosctanovlenija probelov v tablicah / Metody nejroinformatiki: sb. nauchn. tr.; pod red. Gorbanja A.N. – Krasnojarsk: Izd-vo KGTU, 1998. – S. 6–22. (rus)

9. **Zagoruiko N.G.** Prikladnye metody analiza dannyh i znanij. – Novosibirsk: Izd-vo In-ta matematiki, 1999. (rus)

10. **Gupta A., Lam M.S.** Estimation Missing Values using Neural Networks / J. of Operational Research Society. – 1996. – Vol. 47. – № 2. – P. 229–239.

11. **Nelwamondo F.V., Mohamed S., Marwala T.** Missing Data: A comparison of neural network and expectation maximization techniques / Current Science. – 2007. –Vol. 93. – № 11. – P. 1467–1473.

12. **Karlov I.A.** Metody vosstanovlenija propushhennyh znachenij s ispol'zovaniem instrumentarija Data Mining. / Vestnik Sibirskogo gos. ajerokosmicheskogo un-ta im. akademika M.F. Reshetneva. – 2011. – № 7(40). – S. 29–33. (rus)

13. **Karlov I.A., Koshur V.D.** Podhody k postroeniju gibridnoj modeli dlja ocenki znachenij propushhennyh jelementov v massivah dannyh / Nejroinformatika, ee prilozhenija i analiz dannyh: Mater. XX Vseross. seminara. – 2012. –S. 174–179. (rus)

14. **Halkidi M., Batistakis Y., Vazirgiannis M.** On Clustering Validation Techniques / J. of Intelligent Information Systems. – 2003. – № 17:2/3. – P. 107–145.

КАРЛОВ Иван Александрович – *заместитель руководителя информационно-телекоммуникационного комплекса Сибирского федерального университета, докторант кафедры вычислительной техники Института космических и информационных технологий СФУ, кандидат технических наук.*

660041, Россия, г. Красноярск, пр. Свободный, д. 79.
E-mail: IAKarlov@sfu-kras.ru

KARLOV, Ivan A. *Siberian Federal University.*
660041, Svobodny Pr. 79, Krasnoyarsk, Russia.
E-mail: IAKarlov@sfu-kras.ru