К.Ю. Брестер, В.В. Становов, О.Э. Семенкина, Е.С. Семенкин

О применении эволюционных алгоритмов при анализе больших данных[1]

Аннотация. Статья носит обзорный характер: на ряде примеров демонстрируется целесообразность использования эволюционных алгоритмов для анализа больших данных. Эволюционные алгоритмы обладают очевидными преимуществами: высокая масштабируемость и гибкость, способность решать задачи глобальной оптимизации и оптимизировать несколько критериев одновременно являются существенными при отборе информативных признаков, селекции обучающих примеров и решении других задач снижения размерности данных. В частности, рассматривается применение эволюционных алгоритмов в сочетании с такими инструментами машинного обучения как нейронные сети и нечеткие системы. Приведенные примеры доказывают, что эволюционное машинное обучение становится все более и более востребованным и ожидается дальнейшее развитие данной области, в особенности в отношении анализа больших данных.

Ключевые слова: эволюционные алгоритмы, большие данные, отбор признаков, селекция обучающих примеров, заполнение пропусков.

Введение

Нередко для проектирования эффективных моделей машинного обучения стремятся использовать как можно больше выборочных данных. Однако большие объемы данных (*Big Data*) не обязательно означают высокое качество их обработки. Во-первых, многие алгоритмы сталкиваются с проблемой «проклятия размерности». Показано [1], что большие данные могут стать непреодолимым препятствием для традиционных методов анализа данных. Более того, измерения часто собираются из различных источников и после объединения могут повторять друг друга. Вектор переменных, как правило, является неоднородным: он может сочетать в себе численные и категориальные значения, а также неструктурированную текстовую информацию. Наборы данных могут содержать пропуски и аномальные измерения, вследствие чего требуется привлечение экспертов для построения робастных моделей [2]. Далее, в [3] авторы акцентируют внимание на том, что эффективность модели может снижаться с включением все большего и большего количества нерелевантных переменных. Несмотря на то, что некоторые модели (такие как наивный байесовский классификатор) робастны к нерелевантным данным, их эффективность снижается, если в данных присутствуют коррелированные переменные [4]. Эти факты свидетельствуют о необходимости тщательной предобработки данных и реализации новых методов, которые могли бы работать с большими данными.

В статье рассматривается три аспекта проблемы анализа больших данных, которые включают: селекцию обучающих примеров, отбор информативных признаков и обработку пропусков. Первые два аспекта имеют отношение к большому числу измерений и переменных в выборке, что является типичным для больших данных. Один из широко применяе-

[1] Работа выполняется при поддержке Министерства образования и науки РФ в рамках проектной части государственного задания, проект № 2.1680.2017/ПЧ.

мых экстенсивных методов для больших размерностей заключается в использовании параллельных вычислений и современных параллельных архитектур. Однако данный подход не всегда применим вследствие ограниченности ресурсов. В этом случае должны быть использованы методы снижения объемов данных, которые представляют собой более современный подход для решения задач с большим числом измерений и признаков. В то же время, при решении реальных задач исследователи и аналитики либо не используют методы снижения размерности данных, либо используют простейшие: в качестве отбора информативных признаков они осуществляют анализ корреляций для исключения сильно коррелированных переменных, а для селекции обучающих примеров – применяют подходы, основанные на методе k-ближайших соседей. Третий аспект также очень важен: с одной стороны, многие эффективные методы не могут обрабатывать пропуски, что накладывает ограничения на выбор инструментов анализа, и, с другой стороны, исключение признаков или измерений, содержащих пропуски из набора данных, может привести к потере полезной информации.

Открытые вопросы данной области свидетельствуют о необходимости отыскания новой эффективной парадигмы, которая могла бы быть использована для решения перечисленных проблем. Одной из таких парадигм являются эволюционные вычисления и эволюционные алгоритмы (ЭА) [5, 6], которые в последние годы показали свою эффективность в применении к анализу больших данных. В сравнении с другими алгоритмами оптимизации у них есть ряд важных преимуществ: ЭА (в частности, генетические алгоритмы (ГА)) могут эффективно справляться с проклятием размерности; они используются в различных поисковых пространствах (например, это может быть пространство векторов релевантных признаков или наборов информативных измерений); они могут одновременно оптимизировать несколько различных критериев. Эти и некоторые другие свойства ЭА являются во многом уникальными, и, вследствие этого, мы ожидаем дальнейшее развитие инструментов эволюционного машинного обучения и их широкое применение для анализа больших данных.

В работе приводятся положительные эффекты использования ЭА при отборе информативных признаков, селекции обучающих примеров и работе с пропусками. Описывается ряд применений ЭА для реальных и тестовых задач большой размерности. Наша основная цель состоит в демонстрации того, что подходы, основанные на ЭА, могут быть использованы в качестве эффективной альтернативы традиционным методам.

1. Эволюционные алгоритмы и их полезные свойства для анализа больших данных

ЭА являются стохастическими популяционными оптимизационными алгоритмами, которые всегда включают три главных компонента: набор кандидатов – решений, правило сравнения кандидатов и операторы для генерации новых решений [7]. Несколько десятилетий исследователи разрабатывали различные модификации ЭА, направленные на устранение недостатков эволюционного поиска [8, 9]: было предложено множество адаптивных [10-12] и кооперативных [13, 14] алгоритмов, реализованы эффективные модификации применяемых операторов [15], и было продемонстрировано большое количество успешных применений ЭА в реальных задачах [16-21]. Все это стало возможным вследствие уникальных свойств ЭА. Зачастую эти свойства позволяют справляться со сложными проблемами, которые не могут быть решены какими-либо другими алгоритмами (Табл. 1).

Во-первых, ЭА часто используются для обучения моделей. В данном процессе пространство поиска есть набор различных моделей, что довольно необычно для традиционных алгоритмов оптимизации. И основным преимуществом здесь является возможность комбинирования процесса обучения модели и селекции обучающих примеров: в ходе работы алгоритма эффективность модели оценивается на определенном подмножестве обучающих примеров. Помимо этой, так называемой адаптивной селекции, существует еще один подход. ЭА легко могут быть распараллелены в силу их популяционной природы, и данное свойство эффективно используется в методе ротации обучающих примеров [22]. Основная идея заключается в разделении набора данных на несколько подвыборок и в обучении модели не на всей выборке сразу, а лишь на одной ее части в рамках

Табл. 1. Основные преимущества использования ЭА при анализе больших данных

Аспект проблемы больших данных		Полезное свойство ЭА
Селекция обучающих примеров	Ротация обучающих данных	- Возможность распараллеливания вычислений вследствие популяционной природы ЭА. - В кооперативные модификации ЭА может быть встроена ротация обучающих данных.
	Адаптивная селекция обучающих примеров	- Селекция обучающих примеров была разработана для популяционных методов. - Селекция обучающих примеров может быть успешно встроена в процесс построения модели.
Отбор информативных признаков		- Возможность сравнивать подмножества признаков благодаря удобству бинарного представления. - ЭА эффективно работают в признаковых пространствах большой размерности. - ЭА могут оптимизировать несколько критериев одновременно.
Обработка пропусков		- Некоторые модели, способные обрабатывать пропуски, могут быть сгенерированы только ЭА.

текущего этапа. На следующем этапе подвыборки подвергаются «ротации», и модели обучаются на других измерениях. Ещё одно преимущество относится к онлайн-обучению и возможности настраивать модель после получения порции новых данных. В данном случае нет необходимости запускать процесс обучения сначала и тратить значительные объемы ресурсов впустую. Таким образом, ЭА могут быть использованы в динамически изменяемой среде: популяция решений-кандидатов (например, моделей) быстро адаптируется к новым данным [23].

Во-вторых, ЭА могут работать в пространствах высоких размерностей, в то время как другие алгоритмы могут не справиться с этим препятствием [24]. В смысле отбора информативных признаков указанное свойство очень полезно: существуют наборы данных с сотнями и тысячами переменных. Вектор признаков может быть закодирован при помощи бинарной строки: 1 значит релевантный признак, в то время как 0 – нерелевантный. Нам нужно лишь определить критерий оптимизации и запустить алгоритм. В противоположность классическому отбору информативных признаков, таких как *stepwise selection*, ЭА позволяют нам сравнивать различные подмножества признаков, в то время как многие другие методы работают итерационно, добавляя или убирая по одной переменной за раз (пошаговое добавление или пошаговое исключение). Более того, ЭА способны оптимизировать не только один, а сразу несколько критериев, что означает высокую гибкость данного алгоритмического аппарата для отбора информативных признаков [25].

Кроме того, в последнее время ЭА стали широко использоваться для восстановления пропущенных данных, и причина заключается в том, что ЭА являются методами глобальной оптимизации, способными к созданию моделей, описывающих зависимость между десятками переменных, таких как нейронные сети или системы на нечеткой логике. Несколько примеров таких приложений может быть найдено в [26, 27]. Наконец, ЭА могут быть использованы для генерации баз нечетких правил, т.е. моделей, которые могут работать с пропущенными данными напрямую без восстановления пропусков.

В следующих разделах мы продемонстрируем эти преимущества использования ЭА при решении реальных задач с большими данными.

2. Методы селекции обучающих примеров

В данном подразделе мы остановимся на методах селекции обучающих примеров [28], включающих селекцию обучающего множества (адаптивная селекция) и ротацию обучающих данных. Селекция обучающих примеров базируется на создании подвыборки из набора имеющихся обучающих примеров, т.е. подготовке данных для обучающего алгоритма. Методы селекции обучающих примеров также включают активное обучение, бустинг и другие методы изменения выборки.

Использование ГА и ЭА при обучении моделей на больших данных требует большого количества итераций. Из-за этого применение методов селекции обучающих примеров в комбинации с ЭА, которые создают меньшую выборку, снижая тем самым время вычислений, выглядит многообещающе. Однако в данном случае возникает проблема релевантности подвыборки по отношению к начальной выборке. Создание подвыборки может повлечь потерю информации, но в то же самое время, оно может привести к лучшим результатам вследствие фильтрации неинформативных измерений. По этой причине создание метода селекции обучающих примеров, эффективно использующего начальную выборку для создания подвыборок, является важной задачей для анализа больших данных.

2.1. Ротация обучающих данных

Идея ротации обучающих данных была представлена в [22]. Популяция индивидов разделяется на N подпопуляций, и инициализируется такое же количество потоков N. Таким образом, оценивание функции пригодности для различных подпопуляций происходит в параллельных потоках. Для дополнительного снижения временных затрат Ишибучи предложил разделить все обучающие данные на N подвыборок и использовать каждую часть для разных подпопуляций на определенных этапах оптимизации. Термин «ротация» означает обмен обучающими примерами между подпопуляциями.

Работа кооперативного ЭА основана на использовании нескольких популяций, и, следовательно, ротация данных может быть легко встроена в оптимизационный процесс. Реализацию данной идеи будем рассматривать на примере многокритериального ГА. В [13] кооперативный алгоритм был реализован как гетерогенная система «островов»: три различные эвристики (SPEA2 [29], NSGA-II [30] и PICEA-g [31]) были использованы в качестве параллельно работающих островов, которые обменивались лучшими индивидами после определенного числа поколений (процесс, называемый «миграцией»). В основу функционирования SPEA2, NSGA-II и PICEA-g заложены различные эвристические концепции и, как было показано в [13], интеграция полезных свойств эволюционных методов в одну модель приводит к повышению надежности эвристической оптимизационной процедуры. В Табл. 2 приведены особенности каждого из рассмотренных эволюционных многокритериальных алгоритмов.

Реализация островной модели сводится к запуску параллельных процессов, каждый из которых ассоциирован с работой одного из методов SPEA2, NSGA-II или PICEA-g. Для синхронизации островов используется дополнительный процесс, который во время миграции сначала получает от основных процессов набор индивидов, подлежащих обмену, а затем пересылает их нужным компонентам. Данный вспомогательный процесс формирует финальное решение – аппроксимацию множества Парето.

Табл. 2. Основные отличия используемых алгоритмов

Алгоритм	Назначение пригодности	Сохранение разнообразия	Элитизм
SPEA2	Принцип Парето-доминирования используется в совокупности с механизмом оценки плотности решений (расстояние до k-ого ближайшего соседа в пространстве критериев)	Механизм основан на определении расстояний до ближайших соседей	Сохранение недоминируемых решений в архивном множестве
NSGA-II	Принцип основан на идее Парето-доминирования (механизм введения ниш) и оценке разнообразия решений (вычисление расстояний в пространстве критериев)	В рамках подхода оценивается степень сгущения точек (crowding distance)	Копирование лучших индивидов в следующее поколение
PICEA-g	Принцип Парето-доминирования реализуется с помощью генерирования целевых векторов	Механизм основан на определении расстояний до ближайших соседей	Сохранение недоминируемых решений в архивном множестве; копирование лучших индивидов в следующее поколение

Далее островная модель была объединена с ротацией данных [32]: выборка делилась на три части случайным образом, и в дополнении к миграции индивидов выполнялся обмен подвыборками между островами. На каждом третьем этапе миграции (после того, как все части выборки были использованы всеми популяциями), обучающие данные разбивались на три новые части (Рис. 1).

Данный метод селекции обучающих примеров был успешно применен к задаче распознавания эмоций человека по речи в сочетании с отбором информативных признаков и был подробно описан в [32]. В эксперименте базы данных содержали от 480 до 4836 измерений (использовались четыре набора данных на различных языках). В качестве критериев выступали внутри- и межклассовое расстояния, оптимизация которых проводилась на множестве подсистем информативных признаков. В результате было показано, что данный метод селекции обучающих примеров (ротация данных) позволил снизить время, потраченное на вычисления (в среднем в 2.55 раза), без каких-либо потерь в точности модели. Наибольшее снижение времени было получено для самого большого набора данных.

2.2. Адаптивная селекция обучающих примеров

В данном разделе рассмотрим адаптивный подход к селекции обучающих примеров в сочетании с нечетким классификатором, база правил которого генерируется ГА [33, 34]. Метод адаптивной селекции обучающих примеров был предложен в [35]. В статье приведем краткое описание подхода.

Адаптивный метод селекции обучающих примеров был разработан для задач классификации, хотя он может быть расширен и применен для задач регрессии. Этот метод не требует каких-либо предобработок: данные выбираются в обучающее подмножество на основании того, насколько хорошо они классифицируются.

На первом шаге создается подвыборка посредством селекции случайных измерений с равными вероятностями. Далее запускается процедура обучения классификатора на полученной подвыборке. Счетчик U_i устанавливается для каждого измерения и обозначает количество раз, когда измерение *i* было использовано и корректно

Рис. 1. Ротация данных в кооперативном ЭА [32]

классифицировано. По окончании периода адаптации, который обычно длится десятки или сотни итераций, лучшее текущее решение используется для обновления счетчиков. Обновляются только счетчики измерений, которые присутствовали в подвыборке, т.е. если измерение *i* было классифицировано корректно, то $U_i = U_i + 1$, в противном случае, $U_i = 1$. Таким образом, измерения, которые были классифицированы корректно несколько периодов адаптации подряд, получают большие значения счетчиков. По окончании каждого периода адаптации создается новая подвыборка с использованием новых счетчиков. Вероятность того что измерение *i* будет выбрано для подвыборки вычисляется как:

$$p_i = \frac{1/U_i}{\sum_{j=1}^{n} 1/U_j}. \quad (1)$$

Увеличение значения счетчика снижает вероятность, в то время как измерения со счетчиками, которые были недавно сброшены до 1, получают большие вероятности.

Адаптивный метод селекции обучающих примеров ставит меньшие вероятности в соответствие измерениям, которые легко классифицировать, отфильтровывая их из подвыборки. В то же самое время измерения, которые сложно классифицировать, а также неиспользованные ранее измерения получают большие вероятности. Данный метод следует двум основным стратегиям: исследование новых областей пространства измерений и использование информации о точности классификации для того, чтобы построить лучшее разделение между классами.

В случае если данный метод используется с популяционным алгоритмом, таким как ГА, генетическое программирование (ГП) или другим ЭА, лучшее решение на подвыборке может оказаться не лучшим для всей обучающей выборки, но популяция все ещё может содержать решение, которое имеет лучшие обобщающие свойства. Поэтому после каждого периода адаптации все индивиды в популяции должны быть протестированы на всех данных для нахождения лучшего решения.

Кроме того, селекция обучающих примеров позволяет строить сбалансированную, если это возможно, либо более сбалансированную подвыборку из изначально несбалансированной выборки, применяя предложенный алгоритм селекции обучающих примеров отдельно для каждого класса и фиксируя число измерений, которое необходимо выбрать для каждого класса.

Описанный алгоритм селекции обучающих примеров был применен к гибридному эволюционному алгоритму нечеткой классификации HEFCA, ранее описанному в [35]. Алгоритм был протестирован на 9 базах данных с репозиториев машинного обучения KEEL [36] и UCI [37]. Целью тестирования было показать возможность обработки больших объемов данных с использованием ЭА, при этом не снижая, а в некоторых случаях даже повышая качество классификации, в основном за счет использования балансирующей стратегии.

В ходе тестирования объем подвыборки устанавливался в качестве параметра и варьировался от 5% до 30% с 5% шагом и различной длиной периода адаптации в 50, 100, 200 и 400 поколений. Число индивидов в популяции было равно 210, число поколений 50000, а максимальное число нечетких правил в базе устанавливалось равным 40. Результаты применения балансирующей стратегии (instance selection, IS) сравнивались с результатами оригинального алгоритма HEFCA без селекции обучающих примеров, а также с алгоритмом HEFCA с вычислительным ресурсом, увеличенным в 10.5 раз. Результаты применения других схожих методов нечеткой классификации также представлены в Табл. 3. Значения в таблице – относительные ошибки классификации в процентах. В качестве альтернативных алгоритмов были выбраны только эволюционные методы нечеткой классификации. Среди них на работу с большим объемом данных нацелен только

Табл. 3. HEFCA с адаптивной селекцией обучающих примеров в сравнении с другими методами

База данных	HEFCA_IS балансирующая	HEFCA оригинальный	HEFCA увеличенный ресурс	Fuzzy GBML [24]	Parallel Fuzzy GBML [24]
Magic	15.08	15.73	15.87	15.42	**14.89**
Page-blocks	**3.25**	3.96	3.78	3.81	3.62
Penbased	3.81	7.46	5.02	**3.07**	3.30
Phoneme	16.88	16.48	**15.36**	15.43	15.96
Ring	**5.08**	5.82	5.52	6.73	5.25
Satimage	**12.93**	14.22	12.86	15.54	12.96
Segment	5.19	6.45	5.62	5.99	5.90
Texture	**4.45**	7.75	4.90	4.64	4.77
Twonorm	4.81	6.06	5.57	7.36	3.39
База данных	FARC-HD [38]	Bio HEL [39]	GP-Coach [40]	IVFS-Amp [41]	IVFS-Coop [42]
Magic	15.49	-	20.18	20.82	19.82
Page-blocks	4.99	-	8.77	5.84	6.57
Penbased	3.96	6.00	17.80	21.73	17.00
Phoneme	17.86	-	-	-	-
Ring	5.92	-	-	16.89	12.57
Satimage	12.68	11.60	27.50	-	-
Segment	-	**2.90**	24.04	-	-
Texture	7.11	-	-	-	-
Twonorm	**4.72**	-	15.17	-	-

Parallel Fuzzy GBML, использующий 7 вычислительных потоков для обучения и в 10 раз больший объем вычислительных ресурсов. При этом алгоритм с балансирующей стратегией показал лучшие результаты для 4 задач из 9.

Классификатор в комбинации с адаптивной селекцией обучающих примеров и балансирующей стратегией оказался лучшим методом для 3 из 9 задач, превосходя не только оригинальный алгоритм, но также и алгоритм с увеличенным в 10.5 раз вычислительным ресурсом на 7 задачах из 9. Ускорение, полученное с адаптивной селекцией обучающих примеров, было от 3.09 до 21.5 раз, в зависимости от размера подвыборки. Значения ошибки, которые не хуже, чем у оригинального метода без селекции обучающих примеров, могут быть получены уже при размере подвыборки в 10-15% от изначального, давая среднее ускорение от 6 до 12 раз.

3. Методы отбора информативных признаков

Отбор информативных признаков является важной частью предобработки данных: атрибуты могут иметь низкий уровень вариации, коррелировать друг с другом или быть измерены с ошибками. Для задач большой размерности классические методы отбора информативных признаков (пошаговое добавление или пошаговое исключение) работают неэффективно. Их возможности ограничиваются итерационной схемой реализации: невозможно оценить общий вклад нескольких переменных, так как только одна переменная добавляется или удаляется за раз.

В общем случае, если для оценивания качества сокращенного набора признаков обучается модель и ее эффективность определяется на данной подсистеме признаков, то говорят о *wrapper*-подходе (встроенный подход). Однако для больших данных это потребует большого количества вычислительных ресурсов. Таким образом, использовать данный подход неразумно. В свою очередь, альтернативный подход называется *filter*-подход (фильтр). Он основан на применении статистических критериев (например, корреляция, энтропия, расстояния) для сравнения различных наборов признаков и ищет наиболее подходящий в смысле выбранных критериев [43].

При использовании *filter*-подхода ЭА применяются в качестве оптимизатора. Вектор признаков кодируется бинарной строкой, где 1 и 0 определяют релевантные и нерелевантные атрибуты соответственно. В [44] авторы комбинируют различные статистические критерии (внутри- и межклассовое расстояния, корреляцию признаков, оценку Лапласа) и исследуют многокритериальные модели для отбора информативных признаков. Эволюционный *filter*-подход применяется для предсказания сердечнососудистых заболеваний жителей г. Куопио (Финляндия). Из всей базы данных, собранной за период 1984-1989 гг., экспертами было отобрано 393 признака, которые могли бы влиять на возникновение сердечно-сосудистых заболеваний жителей г. Куопио. В качестве выходной переменной выступало наличие (“болен”) или отсутствие (“здоров”) сердечно-сосудистых заболеваний в период 1998-2001 гг. (учитывались и болезни, и летальные исходы). Важно помнить, что в данной предметной области снижение размерности ведет к уменьшению медицинских затрат клинических испытаний. В результате для данного набора данных стало возможным значительное снижение числа признаков: с 393 до 38 без каких-либо потерь эффективности модели.

Схожий эволюционный *filter*-подход был использован для определения состояния говорящего на основании речевого сигнала. В данной задаче число переменных, являющихся акустическими характеристиками, было равно 384. На основании этих данных можно определить человека, который говорит, его пол и даже эмоции. В [45] адаптивный ГА был встроен в процедуру отбора информативных признаков в качестве оптимизатора (оптимизировались внутри- и межклассовое расстояния). В серии экспериментов было показано, что можно найти набор признаков, которые являются релевантными для определенной задачи, при этом число переменных снижается приблизительно в два раз, что приводит к повышению точности модели. Данный подход был протестирован с применением баз данных на английском, немецком и японском языках. Во многих случаях он превзошел альтернативные методы отбора информативных признаков, такие как метод главных компонент или IGR.

Табл. 4. HEFCA с адаптивной селекцией обучающих примеров для данных с пропусками

Доля пропусков	0	0.1	0.2	0.3	0.4
Page-blocks	5.12	6.92	8.19	8.52	8.60
Phoneme	18.30	20.91	22.65	23.82	24.49
Segment	7.45	14.35	18.72	23.03	28.74
Texture	7.27	15.25	24.20	36.68	46.74

4. Обработка пропусков

На сегодняшний день разработано множество методов заполнения пропусков данных, особенно для случая MCAR (Missing Completely At Random, пропуски в совершенно случайных местах) [46]. Большинство методов фокусируется на восстановлении пропущенных данных на основании имеющихся, что требует существенного числа вычислений. При этом многие традиционные методы анализа данных не могут работать с пропусками, либо же показывают очень плохие результаты, если все переменные или признаки с пропусками удалены из базы данных.

Нередко заполнение пропусков может стать сложной задачей ввиду объемов данных и процента пропусков. По этой причине рассматривается метод анализа данных, способный строить модели на измерениях с пропусками без каких-либо предобработок.

Гибридный эволюционный алгоритм нечеткой классификации, представленный в предыдущем разделе, был модифицирован для работы с пропусками. Для нечеткой базы правил обработка пропусков схожа с работой с термом игнорирования (Don't Care condition, DC), который используется для игнорирования значения переменной в определенном правиле. Даже если в правиле q для переменной i условие не является DC, то для случая с пропуском в ходе нечеткого вывода оно рассматривается как DC – условие. Одной из причин того, что данный метод не был использован ранее является то, что он требует специфического ЭА и особых операторов для создания новых правил из выборки измерений и комбинирования их при скрещивании.

В ходе экспериментов исследовалось качество работы классификаторов в зависимости от процента пропущенных измерений. На репозитории UCI [37] было выбрано 4 базы данных, среди которых Page-blocks, Phoneme, Segment и Texture. Для каждого набора данных было проведено 5 тестов: один без пропусков и 4 с пропусками, сгенерированными случайно и равномерно с вероятностями 0.1, 0.2, 0.3 и 0.4. В Табл. 4 приведена относительная ошибка классификации в процентах для каждого теста. Тесты были проведены с меньшим ресурсом: 100 индивидов, 5000 поколений и 40 правил максимум. Адаптивный метод селекции обучающих примеров использовал 15% выборки и период адаптации длиной в 50 поколений.

Результаты эксперимента варьируются для различных баз данных. Например, для задачи Page-blocks добавление до 40% пропусков не меняет точности значительно, в то время как для задачи Texture процент ошибки существенно возрастает с ростом числа пропусков.

Заключение

Проблема больших данных привлекает многих исследователей. Уже сейчас это направление можно признать многообещающим. В последнее время были предприняты различные попытки обработки данных больших размерностей эффективным образом. Однако в этой области все ещё остается множество нерешенных проблем.

В данной работе обсуждаются ЭА и их полезные свойства при решении задач с большими данными. Мы полагаем, что многие исследователи недооценивают эффективность ЭА при анализе больших данных. Поэтому было продемонстрировано несколько различных применений ЭА и показана их эффективность при отборе информативных признаков и селекции обучающих примеров, а также показана возможность обрабатывать данные без восстановления пропусков. Представленные заключения подтверждаются экспериментальными результатами на тестовых и реальных задачах.

Постоянное развитие ЭА и эволюционного машинного обучения может привести к созданию нового эффективного инструмента анализа больших данных. Таким образом, в ближайшем будущем можно также ожидать повышение научного интереса в данной области.

Литература

1. Qiu J., Wu Q., Ding G., Xu Y., Feng Sh. A survey of machine learning for big data processing // EURASIP J. Adv. Signal Process, 67. 2016. doi: 10.1186/s13634-016-0355-x.
2. Zhang Z. Missing values in big data research: some basic skills // Annals of Translational Medicine, 3(21). 2015. 323 p. doi:10.3978/j.issn.2305-5839.2015.12.11.
3. Bagherzadeh-Khiabani F, Ramezankhani A, Azizi F, Hadaegh F, Steyerberg EW, Khalili D. A tutorial on variable selection for clinical prediction models: feature selection methods in data mining could improve the results // J Clin Epidemiol, 71. 2016. P. 76–85.
4. Kohavi R., John G.H. Wrappers for feature subset selection // Artificial Intelligence, 97. 1997. P. 273–324.
5. Holland J.H. Adaptation in natural and artificial systems // J.H. Holland – Ann Arbor. MI: University of Michigan Press. 1975.
6. Goldberg D.E. Genetic algorithms in search, optimization and machine learning // Addison-Wesley. 1989.
7. Гладков Л.А., Курейчик В.В., Курейчик В.М. Генетические алгоритмы // М.: Физматлит, 2006.
8. Карпов В.Э. Эволюционное моделирование. Проблемы формы и содержания // Новости искусственного интеллекта, №5. 2003. С.35-46.
9. Карпов В.Э. Методологические проблемы эволюционных вычислений // Искусственный интеллект и принятие решений, №4. 2012. С. 95-102.
10. Khritonenko D., Semenkin E. Distributed self-configuring evolutionary algorithms for artificial neural networks design // Vestnik SibSAU, № 4 (50). 2013. P. 112–116.
11. Курейчик В.М., Лебедев Б.К., Лебедев О.Б., Чернышев Ю.О. Адаптация на основе самообучения // РГАСХМ ГОУ, Ростов н/Д. 2004.
12. Редько В.Г. Модели адаптивного поведения – биологически инспирированный подход к искусственному интеллекту // Искусственный интеллект и принятие решений. № 2. 2008. С. 11-23.
13. Brester Ch., Semenkin E. Cooperative Multiobjective Genetic Algorithm with Parallel Implementation // Advances in Swarm and Computational Intelligence, LNCS 9140. 2015. P. 471–478.
14. Akhmedova S., Semenkin E. Co-operation of biology related algorithms // IEEE Congress on Evolutionary Computation, CEC 2013. 2013. P. 2207–2214.
15. Semenkin E., Semenkina M. Self-configuring genetic algorithm with modified uniform crossover operator // Advances in Swarm Intelligence, 2012. P. 414–421.
16. Ryzhikov I., Semenkin E. Evolutionary strategies algorithm based approaches for the linear dynamic system identification // Lecture Notes in Computer Science, T. 7824 LNCS. 2013. P. 477–484.
17. Zaloga A., Yakimov I., Burakov S., Semenkin E., Akhmedova S., Semenkina M., Sopov E. On the application of co-operative swarm optimization in the solution of crystal structures from x-ray diffraction data // Lecture Notes in Computer Science, T. 9140, 2015. P. 89–96.
18. Сергиенко Р.Б. Метод формирования нечеткого классификатора самонастраивающимися коэволюционными алгоритмами // Искусственный интеллект и принятие решений. 2010. №3. С. 98-106.
19. Семенкина М.Е. Самоадаптивные эволюционные алгоритмы проектирования информационных технологий интеллектуального анализа данных // Искусственный интеллект и принятие решений. 2013. №1. С. 13-24.
20. Бухтояров В.В. Эволюционный метод формирования общего решения в коллективах нейронных сетей // Искусственный интеллект и принятие решений. 2010. №3. С. 89-97.
21. Полковникова Н.А., Курейчик В.М. Многокритериальная оптимизация на основе эволюционных алгоритмов // Известия ЮФУ. Технические науки. 2015. № 2 (163). С. 149-162.
22. Ishibuchi H., Mihara S., Nojima Y. Parallel Distributed Hybrid Fuzzy GBML Models With Rule Set Migration and Training Data Rotation // IEEE Transactions on fuzzy systems, vol. 21, no. 2. 2013. P. 355–368.
23. Непомнящих В.А., Попов Е.Е., Редько В.Г. Бионическая модель адаптивного поискового поведения // Известия РАН. Серия теория и системы управления, №1. 2008. С. 85-93.
24. Lin Q, Liu W, Peng H, Chen Y. Efficient Genetic Algorithm for High-Dimensional Function Optimization // Proceedings of the 9th International Conference on Computational Intelligence and Security. 2013. P. 255–59.
25. Грошев С.В., Карпенко А.П., Мартынюк В.А. Эффективность популяционных алгоритмов парето-аппроксимации. Экспериментальное сравнение // Интернет-журнал Науковедение. Т. 8. № 4 (35). 2016. DOI: 10.15862/67EVN416.
26. Mahajan R., Kaur G. Neural Networks using Genetic Algorithms// International Journal of Computer Applications, 77(14). 2013. P. 6–11.
27. Thounaojam D.M., Khelchandra T., Singh K.M., Roy S. A Genetic Algorithm and Fuzzy Logic Approach for Video Shot Boundary Detection // Computational Intelligence and Neuroscience, vol. 2016. 2016. P. 11. doi:10.1155/2016/8469428.
28. Cano J. R., Herrera F., Lozano M. Evolutionary Stratified Training Set Selection for Extracting Classification Rules with trade off Precision-Interpretability // Data & Knowledge Engineering archive, volume 60, issue 1. 2007. P. 90–108.
29. Zitzler, E., Laumanns, M., Thiele, L. SPEA2: Improving the Strength Pareto Evolutionary Algorithm for Multiobjective Optimization // Evolutionary Methods for Design Optimisation and Control with Application to Industrial Problems EUROGEN 2001 3242 (103). 2002. P. 95-100.
30. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T. A fast and elitist multiobjective genetic algorithm: NSGA-II // IEEE Transactions on Evolutionary Computation 6 (2). 2002. P.182-197.
31. Wang, R. Preference-Inspired Co-evolutionary Algorithms // A thesis submitted in partial fulfillment for the degree of the Doctor of Philosophy, University of Sheffield. 2013. P. 231.
32. Brester C., Semenkin E., Sidorov M. Multi-objective heuristic feature selection for speech-based multilingual emotion recognition // Journal of Artificial Intelligence and Soft Computing Research, vol. 6, no. 4. 2016. P. 243–253.
33. Fernandez A., Garcia S., Luengo J., Bernado-Mansilla E., Herrera F. Genetics-Based Machine Learning for Rule Induction: State of the Art, Taxonomy, and Comparative Study // IEEE Transactions on Evolutionary Computation, vol.14, issue 6. 2010. P. 913–941.

34. Booker L. B., Goldberg D. E., Holland J.H. Classifier systems and genetic algorithms // Artif. Intell., vol. 40, no. 1–3, 1989. P. 235–282.
35. Stanovov V., Semenkin E., Semenkina O. Self-configuring hybrid evolutionary algorithm for fuzzy classification with active learning // IEEE Congress on evolutionary computation (CEC 2015, Japan). 2015. DOI: 10.1109/CEC.2015.7257108.
36. Alcalá-Fdez J., Sánchez L., Garcia S., Jesus M. J., Ventura S., Garrell J.M., Otero J., Romero C., Bacardit J., Rivas V.M., Fernández J. C., Herrera F. KEEL: A software tool to assess evolutionary algorithms for data mining problems // Soft Comput., vol. 13, no. 3. 2009. P. 307–318.
37. Asuncion A., Newman D. UCI machine learning repository // University of California, Irvine, School of Information and Computer Sciences. URL: http://www.ics.uci.edu/~mlearn/MLRepository.html. 2007.
38. Alcala-Fdez J., Alcala R., Herrera F. A fuzzy association rulebased classification model for high-dimensional problems with genetic rule selection and lateral tuning // IEEE Trans. Fuzzy Syst., vol. 19, no. 5. 2011. P. 857–872.
39. Bacardit J., Burke E. K., Krasnogor N. Improving the scalability of rule-based evolutionary learning // Memetic Comput. J., vol. 1, no. 1. 2009. P. 55–67.
40. Berlanga F. J., Rivera A. J., del Jesus M. J., Herrera F. GP-COACH: Genetic programming-based learning of compact and accurate fuzzy rulebased classification systems for high-dimensional problems // Inf. Sci., vol. 180, no. 8. 2010. P. 1183–1200.
41. Sanz J.A., Fernandez A., Bustince H., Herrera F. Improving the performance of fuzzy rule-based classification systems with interval-valued fuzzy sets and genetic amplitude tuning // Inf. Sci., vol. 180, no. 19. 2010. P. 3674–3685.
42. Sanz J., Fernandez A., Bustince H., Herrera F. A genetic tuning to improve the performance of fuzzy rule-based classification systems with interval-valued fuzzy sets: Degree of ignorance and lateral position // Int. J. Approx. Reason., vol. 52, no. 6. 2011. P. 751–766.
43. Venkatadri M., Srinivasa Rao K. A multiobjective genetic algorithm for feature selection in data mining // International Journal of Computer Science and Information Technologies, vol. 1, no. 5. 2010. P. 443–448.
44. Brester C., Kauhanen J., Tuomainen T.P., Semenkin E., Kolehmainen M. Comparison of two-criterion evolutionary filtering techniques in cardiovascular predictive modelling // Proceedings of the 13th International Conference on Informatics in Control, Automation and Robotics (ICINCO'2016), Lisbon, Portugal, vol. 1. 2016. P. 140–145.
45. Sidorov M., Brester C., Semenkin E., Minker W. Speaker State Recognition with Neural Network-based Classification and Self-adaptive Heuristic Feature Selection // Proceedings of the 11th International Conference on Informatics in Control, Automation and Robotics (ICINCO'2014), Vienna, Austria, 2014, vol. 1. 2014. P. 699–703.
46. Schafer, J.L., Graham, J.W. Missing Data: Our View of the State of the Art // Psychological Methods, vol. 7, no. 2. 2002. P. 147–177.

Брестер Кристина Юрьевна. Старший преподаватель кафедры высшей математики Сибирского государственного университета науки и технологий имени академика М.Ф. Решетнева (СибГУ), Красноярск. Кандидат технических наук. Окончила СибГУ в 2014 году. Количество печатных работ: 45. Область научных интересов: эволюционные алгоритмы, многокритериальная оптимизация, нейросетевое моделирование. E-mail: christina.brester@gmail.com.

Становов Владимир Вадимович. Старший научный сотрудник Сибирского государственного университета науки и технологий имени академика М.Ф. Решетнева (СибГУ), Красноярск. Кандидат технических наук. Окончил СибГУ в 2014 году. Количество печатных работ: 29. Область научных интересов: эволюционные алгоритмы, нечеткая логика. E-mail: vladimirstanovov@yandex.ru.

Семенкина Ольга Эрнестовна. Профессор кафедры высшей математики Сибирского государственного университета науки и технологий имени академика М.Ф. Решетнева (СибГУ), Красноярск. Доктор технических наук, профессор. Окончила Кемеровский государственный университет в 1982 году. Количество печатных работ: 55, в том числе три монографии. Область научных интересов: моделирование и оптимизация сложных систем, дискретная оптимизация, анализ данных. E-mail: semenkina.olga@mail.ru.

Семенкин Евгений Станиславович. Профессор кафедры системного анализа и исследования операций Сибирского государственного университета науки и технологий имени академика М.Ф. Решетнева (СибГУ), Красноярск. Доктор технических наук, профессор. Окончил Кемеровский государственный университет в 1982 году. Количество печатных работ: 320, в том числе 5 монографий. Область научных интересов: моделирование и оптимизация сложных систем, вычислительный интеллект, анализ данных. E-mail: eugenesemenkin@yandex.ru

About the use of evolutionary algorithms in big data analysis

Ch. Yu. Brester, V.V. Stanovov, O.E. Semenkina, E.S. Semenkin

Abstract. This article is a survey: several examples demonstrate the expediency of using evolutionary algorithms in Big Data analysis. Evolutionary algorithms have evident advantages: their high scalability and flexibility, ability to solve global optimization problems and optimize several criteria simultaneously are essential for feature selection, instance selection and missing-data imputation problems. Moreover, we

illustrate the use of evolutionary algorithms in combination with such machine learning tools as neural networks and fuzzy systems. Our examples show that Evolutionary Machine Learning is getting more and more applicable in data processing and we anticipate seeing the further development of this area especially in the sense of Big Data.

Keywords: evolutionary algorithms, Big Data, feature selection, instance selection, missing-data imputation.

References

1. Qiu, J., Wu, Q., Ding, G., Xu, Y., Feng, Sh. 2016. A survey of machine learning for big data processing. EURASIP J. Adv. Signal Process. 67. doi: 10.1186/s13634-016-0355-x.
2. Zhang, Z. 2015. Missing values in big data research: some basic skills. Annals of Translational Medicine. 3(21). 323 p. doi:10.3978/j.issn.2305-5839.2015.12.11.
3. Bagherzadeh-Khiabani, F., Ramezankhani, A., Azizi, F., Hadaegh, F., Steyerberg, E.W., Khalili, D. 2016. A tutorial on variable selection for clinical prediction models: feature selection methods in data mining could improve the results. J Clin Epidemiol. 71: 76–85.
4. Kohavi, R., John, G.H. 1997. Wrappers for feature subset selection. Artificial Intelligence. 97: 273–324.
5. Holland, J.H. 1975. Adaptation in natural and artificial systems. J.H. Holland – Ann Arbor. MI: University of Michigan Press.
6. Goldberg, D.E. 1989. Genetic algorithms in search, optimization and machine learning. Addison-Wesley.
7. Gladkov, L.A., Kureychik, V.V., Kureychik, V.M. 2006. Geneticheskie algoritmy [Genetic algorithms]. Moscow: Fizmatlit.
8. Karpov, V.E. 2003. Evolyutsionnoye modelirovaniye. Problemy formy i soderzhaniya [Evolutionary modeling. Problems of form and content]. Novosti iskusstvennogo intellekta. №5. 35-46.
9. Karpov, V.E. 2013. Methodological problems in evolutionary computation. Scientific and Technical Information Processing. 40(5): 286-291.
10. Khritonenko, D., Semenkin, E. 2013. Distributed self-configuring evolutionary algorithms for artificial neural networks design. Vestnik SibSAU. 4 (50): 112–116.
11. Kureychik, V.M., Lebedev, B.K., Lebedev, O.K., Chernyshev, Yu., O. 2004. Adaptatsiya na osnove samoobucheniya [Adaptation based on learning]. Rostov-on-Don: RGASKhM GOU.
12. Redko, V. G. 2008. Modeli adaptivnogo povedenija – biologicheski inspirirovannyj podhod k iskusstvennomu intellektu [Models of adaptive behavior – biologically inspired approach to artificial intelligence]. Iskusstvennyj intellekt i prinjatie reshenij. 2: 11-23.
13. Brester, Ch., Semenkin, E. 2015. Cooperative Multiobjective Genetic Algorithm with Parallel Implementation. Advances in Swarm and Computational Intelligence, LNCS 9140: 471–478.
14. Akhmedova, S., Semenkin, E. 2013. Co-operation of biology related algorithms. IEEE Congress on Evolutionary Computation, CEC 2013: 2207–2214.
15. Semenkin, E., Semenkina, M. 2012. Self-configuring genetic algorithm with modified uniform crossover operator. Advances in Swarm Intelligence: 414–421.
16. Ryzhikov, I., Semenkin, E. 2013. Evolutionary strategies algorithm based approaches for the linear dynamic system identification. Lecture Notes in Computer Science, T. 7824 LNCS: 477–484.
17. Zaloga, A., Yakimov, I., Burakov, S., Semenkin, E., Akhmedova, S., Semenkina, M., Sopov, E. 2015. On the application of co-operative swarm optimization in the solution of crystal structures from x-ray diffraction data. Lecture Notes in Computer Science, T. 9140: 89–96.
18. Sergienko, R.B. 2010. Method of fuzzy classifier design with self-tuning coevolutionary algorithms. Scientific and Technical Information Processing. 3: 98-106.
19. Semenkina, M.E. 2013. Effectiveness investigation of adaptive evolutionary algorithms for data mining information technology design. Scientific and Technical Information Processing. 1: 13-24.
20. Bukhtoyarov, V.V. 2010. Evolutionary method for design of neural networks ensemble prediction. Scientific and Technical Information Processing. 3: 89-97.
21. Polkovnikova, N.A., Kureichik, V.M. 2015. Mnogokriterialnaya optimizatsiya na osnove evoliutsionnykh algoritmov [Multiobjective optimization on the base of evolutionary algorithms]. Izvestiya SFedU. Engineering Sciences. 2(163): 149-162.
22. Ishibuchi, H., Mihara, S., Nojima, Y. 2013. Parallel Distributed Hybrid Fuzzy GBML Models With Rule Set Migration and Training Data Rotation. IEEE Transactions on fuzzy systems. 21(2): 355–368.
23. Nepomnyashchikh, V.A., Popov, E.E., Redko, V.G. 2008. A bionic model of adaptive searching behavior. Journal of Computer and Systems Sciences International. 47(1): 78-85.
24. Lin, Q., Liu, W., Peng, H., Chen, Y. 2013. Efficient Genetic Algorithm for High-Dimensional Function Optimization. Proceedings of the 9th International Conference on Computational Intelligence and Security: 255–59.
25. Groshev, S.V., Karpenko, A.P., Martynyuk, V.A. 2016. Effektivnost populiatsionnykh algoritmov pareto-approksimatsii. Eksperimentalnoe sravneniye [The effectiveness of population-based Pareto-approximation algorithms. Experimental comparison]. On-line Journal "Naukovedenie". 8(4). doi: 10.15862/67EVN416.
26. Mahajan, R., Kaur, G. 2013. Neural Networks using Genetic Algorithms. International Journal of Computer Applications. 77(14): 6–11.
27. Thounaojam, D.M., Khelchandra, T., Singh, K.M., Roy, S. 2016. A Genetic Algorithm and Fuzzy Logic Approach for Video Shot Boundary Detection. Computational Intelligence and Neuroscience. 2016: 11. doi:10.1155/2016/8469428.

28. Cano, J.R., Herrera, F., Lozano, M. 2007. Evolutionary Stratified Training Set Selection for Extracting Classification Rules with trade off Precision-Interpretability. Data & Knowledge Engineering archive, 60(1): 90–108.
29. Zitzler, E., Laumanns, M., Thiele, L. 2002. SPEA2: Improving the Strength Pareto Evolutionary Algorithm for Multiobjective Optimization. Evolutionary Methods for Design Optimisation and Control with Application to Industrial Problems EUROGEN 2001 3242 (103): 95-100.
30. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T. 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Transactions on Evolutionary Computation. 6 (2): 182-197.
31. Wang, R. 2013. Preference-Inspired Co-evolutionary Algorithms. A thesis submitted in partial fulfillment for the degree of the Doctor of Philosophy, University of Sheffield. 231 p.
32. Brester, C., Semenkin, E., Sidorov, M. 2016. Multi-objective heuristic feature selection for speech-based multilingual emotion recognition. Journal of Artificial Intelligence and Soft Computing Research. 6(4): 243–253.
33. Fernandez, A., Garcia, S., Luengo, J., Bernado-Mansilla, E., Herrera, F. 2010. Genetics-Based Machine Learning for Rule Induction: State of the Art, Taxonomy, and Comparative Study. IEEE Transactions on Evolutionary Computation. 14(6): 913–941.
34. Booker, L.B., Goldberg, D.E., Holland, J.H. 1989. Classifier systems and genetic algorithms. Artif. Intell. 40(1–3): 235–282.
35. Stanovov, V., Semenkin, E., Semenkina, O. 2015. Self-configuring hybrid evolutionary algorithm for fuzzy classification with active learning. IEEE Congress on evolutionary computation (CEC 2015, Japan). doi: 10.1109/CEC.2015.7257108.
36. Alcalá-Fdez, J., Sánchez, L., Garcia, S., Jesus, M.J., Ventura, S., Garrell, J.M., Otero, J., Romero, C., Bacardit, J., Rivas, V.M., Fernández, J.C., Herrera, F. 2009. KEEL: A software tool to assess evolutionary algorithms for data mining problems. Soft Comput. 13(3): 307–318.
37. Asuncion, A., Newman, D. 2007. UCI machine learning repository. University of California, Irvine, School of Information and Computer Sciences. URL: http://www.ics.uci.edu/~mlearn/MLRepository.html.
38. Alcala-Fdez, J., Alcala, R., Herrera, F. 2011. A fuzzy association rulebased classification model for high-dimensional problems with genetic rule selection and lateral tuning. IEEE Trans. Fuzzy Syst. 19(5): 857–872.
39. Bacardit, J., Burke, E.K., Krasnogor, N. 2009. Improving the scalability of rule-based evolutionary learning. Memetic Comput. J. 1(1): 55–67.
40. Berlanga, F.J., Rivera, A.J., del Jesus, M.J., Herrera, F. 2010. GP-COACH: Genetic programming-based learning of compact and accurate fuzzy rulebased classification systems for high-dimensional problems. Inf. Sci. 180(8): 1183–1200.
41. Sanz, J.A., Fernandez, A., Bustince, H., Herrera, F. 2010. Improving the performance of fuzzy rule-based classification systems with interval-valued fuzzy sets and genetic amplitude tuning. Inf. Sci. 180(19): 3674–3685.
42. Sanz, J., Fernandez, A., Bustince, H., Herrera, F. A genetic tuning to improve the performance of fuzzy rule-based classification systems with interval-valued fuzzy sets: Degree of ignorance and lateral position. Int. J. Approx. Reason. 52(6): 751–766.
43. Venkatadri, M., Srinivasa, Rao K. 2010. A multiobjective genetic algorithm for feature selection in data mining. International Journal of Computer Science and Information Technologies, 1(5): 443–448.
44. Brester, C., Kauhanen, J., Tuomainen, T.P., Semenkin, E., Kolehmainen, M. 2016. Comparison of two-criterion evolutionary filtering techniques in cardiovascular predictive modelling. Proceedings of the 13th International Conference on Informatics in Control, Automation and Robotics (ICINCO'2016), Lisbon, Portugal. 1: 140–145.
45. Sidorov, M., Brester, C., Semenkin, E., Minker, W. 2014. Speaker State Recognition with Neural Network-based Classification and Self-adaptive Heuristic Feature Selection. Proceedings of the 11th International Conference on Informatics in Control, Automation and Robotics (ICINCO'2014), Vienna, Austria. 1: 699–703.
46. Schafer, J.L., Graham, J.W. 2002. Missing Data: Our View of the State of the Art. Psychological Methods. 7(2): 147–177.

Brester Christina Yuryevna. Senior lecturer of the Department of Higher Mathematics, Reshetnev Siberian State University of Science and Technologies (SibSU), 31 Krasnoyarskiy rabochiy ave., Krasnoyarsk, Russia. Candidate of technical sciences. In 2014 she graduated from SibSU. Published papers: 45. Scientific interests: evolutionary algorithms, multi-objective optimization, neural network modeling. E-mail: christina.brester@gmail.com

Stanovov Vladimir Vadimovich. Senior research fellow, Reshetnev Siberian State University of Science and Technologies (SibSU), 31 Krasnoyarskiy rabochiy ave., Krasnoyarsk, Russia. Candidate of technical sciences. In 2014 he graduated from SibSU. Published papers: 29. Scientific interests: evolutionary algorithms, fuzzy logic. E-mail: vladimirstanovov@yandex.ru

Semenkina Olga Ernestovna. Professor of the Department of Higher Mathematics, Reshetnev Siberian State University of Science and Technologies (SibSU), 31 Krasnoyarskiy rabochiy ave., Krasnoyarsk, Russia. Doctor of technical sciences, professor. In 1982 she graduated from Kemerovo State University. Published papers: 55 (including 3 monographs). Scientific interests: modeling and optimization of complex systems, discrete optimization, data analysis. E-mail: semenkina.olga@mail.ru

Semenkin Eugene Stanislavovich. Professor of the Department of System Analysis and Operation Research, Reshetnev Siberian State University of Science and Technologies (SibSU), 31 Krasnoyarskiy rabochiy ave., Krasnoyarsk, Russia. Doctor of technical sciences, professor. In 1982 he graduated from Kemerovo State University. Published papers: 320 (including 5 monographs). Scientific interests: modeling and optimization of complex systems, computational intelligence, data analysis. E-mail: eugenesemenkin@yandex.ru