ОТЗЫВ

официального оппонента Кельманова Александра Васильевича
на диссертационную работу Сташкова Дмитрия Викторовича
«Системы автоматической группировки объектов на основе разделения смеси распределений»,
представленную на соискание ученой степени кандидата технических наук по специальности 05.13.01 — системный анализ, управление и обработка информации (космические и информационные технологии)

Актуальность темы работы. *Объектом* исследования работы являются прикладные проблемы классификации технических объектов, заданных наборами числовых характеристик. *Предмет* исследования – модели, алгоритмы и технологии кластеризации, основанные на известных математических подходах к решению задачи разделения смеси вероятностных распределений. *Цель* исследования – усовершенствование существующих моделей, алгоритмов и технологий кластеризации, ориентированное на повышение качества (точности) решения прикладной задачи выделения в партии промышленной продукции семейства непересекающихся однородных групп за счет более тонкого учета особенностей этой задачи. Исследование *мотивировано*, во-первых, актуальностью задачи улучшения качества электронной продукции для обеспечения повышенной надежности функционирования сложных технических систем космического назначения, а во-вторых, слабой изученностью математических моделей рассматриваемой содержательной проблемы.

Структура диссертационной работы. Диссертация изложена на 172 страницах, включает 208 библиографических источников и состоит из введения, трех глав, заключения, списка литературы и двух приложений.

Во **введении** обоснована актуальность темы диссертации, сформулированы цель и задачи исследований, дана общая характеристика работы, а также сформулированы положения, выносимые на защиты.

В **первой главе** приведен обзор постановок задач разбиения объектов на однородные группы, а также подходов, моделей и алгоритмов для решения этих задач. В результате обзора соискатель останавливает свой выбор на оптимизационной модели (задаче) разделения смеси вероятностных распределений и алгоритмах локального поиска для отыскания решения задачи. В качестве базового (для последующих модификаций) выбран EM-алгоритм.

Во **второй главе** предложены несколько эвристических процедур (жадных, генетических, чередующихся окрестностей) – алгоритмов, которые позже по отдельности и в комбинациях встраиваются в предложенные алгоритмы (п.1 и 2 результатов, выносимых на защиту) для поиска приближенного решения оптимизационной задачи (индуцированной моделью разделения смеси вероятностных распределений). В этой же главе в результате численных экспериментов установлены области предпочтительного применения

предложенных эвристических алгоритмов (по размерности задачи, числу кластеров, точности решения, времени работы).

В **третьей главе** приведено описание содержательной проблемы, а также общей схемы принятия классификационных решений, сформулированы критерии адекватности аппроксимирующих кластеризационных моделей. Здесь же описаны дополнительные эвристические процедуры («нормировки» и «отсеивания выбросов»), направленные, по сути, на компенсацию имеющегося несоответствия принятой в работе модели (разделения смеси гауссовских распределений) и содержательной проблемы. Эти дополнительные процедуры вместе с процедурами, представленными во второй главе, составляют основу предложенных автором компьютерных технологий кластеризации. Результатами экспериментов установлено, что новые технологии позволяют снизить процент ошибочных решений при выделении в партии электронных изделий однородных групп.

В **заключении** диссертации представлен расширенный (более подробный, чем во введении) вариант основных результатов работы. **Приложение А** содержит совокупность таблиц, отражающих результаты экспериментов. В **Приложении Б** представлен акт о внедрении (на одном из предприятий) созданных компьютерных технологий в опытную эксплуатацию в рамках системы автоматизированного контроля радиоэлектронных изделий.

Основные результаты и их **новизна**. Новизну результатов подтверждаю. К числу оригинальных разработок относятся:

- эвристические генетические алгоритмы для задачи разделения смеси распределений;
- эвристические алгоритмы поиска с чередующимися окрестностями для этой же задачи;
- компьютерные технологии, в основе которых предложенные алгоритмы, а также рекомендации по их применению для решения проблемы выделения однородных групп в партии радиоэлектронных изделий с целью оценивания их качества и отбраковки.

Предложенные алгоритмы в численных экспериментах продемонстрировали улучшенные показатели по точности в сравнении с аналогами, а компьютерные технологии в практических задачах позволили вдвое сократить процент ошибок принятия решения.

Достоверность и обоснованность научных положений, выводов и рекомендаций. Справедливость сформулированных положений, рекомендаций и выводов опирается на результаты экспериментов, которые подтверждают заявленные преимущества предложенных алгоритмов и проведены с соблюдением критериев достоверности измерений и статистических испытаний. Результаты, выносимые на защиту, достоверны и получены на основе общепринятой методологии научных исследований.

Значимость полученных результатов для науки и практики. Работа носит экспериментально-прикладной (технический) характер. Полученные результаты не имеют принципиального значения для математики. Однако для технических наук они имеют несомненную ценность, поскольку дополняют совокупность успешных эвристических приемов (техник, средств), обеспечивающих повышение точности решения некоторых важных прикладных задач. Практическая значимость результатов подтверждена актом об их внедрении в опытную эксплуатацию на промышленном предприятии.

По работе имеются следующие замечания и вопросы.

1. На мой взгляд, в методологическом плане изложение диссертации не совсем удачно. Содержательная прикладная проблема описана лишь в третьей главе, фактически, в конце диссертации. Общепринятым правилом считается изложение в иной последовательности: содержательная проблема – модель для нее и подходы к решению – индуцированная задача – исследование задачи и ее упрощений (частных случаев) – алгоритмы – результаты, технологии –… Иными словами, научный труд (статью, диссертацию) принято начинать с формулировки проблемы, а не с изложения инструментов для решения еще не сформулированной задачи. Это правило при изложении материала, к сожалению, проигнорировано.

2. Классическая модель разделения смеси вероятностных распределений не соответствует (не адекватна) рассматриваемой прикладной проблеме, которая индуцирует более сложную математическую задачу. Судя по тексту, в работе следует говорить не о задаче разбиения входного множества, а о задаче поиска наилучшего (в определенном смысле) семейства непересекающихся подмножеств, объединение которых может не покрывать входное множество (причина тому – наличие так называемых «выбросов», про которые в постановочной части диссертации ничего не сказано). Либо (если все-таки формулировать задачу как разбиение) в целевой функции задачи должен фигурировать аддитивный член, соответствующий совокупности «выбросов» (т.е. вырожденных одноэлементных выборок из произвольных вероятностных распределений).

3. В тексте диссертации нет четких математических формулировок задач, которыми оперирует соискатель. Имеются в виду формулировки в виде – дано:…, найти: оптимум определенной целевой функции при ограничениях… По этой причине текст работы плохо воспринимается, а встречающиеся в тексте несоответствия приведенных формул формулировкам классических оптимизационных задач (k-средних, p-медиана и др.) вводят читателя в заблуждение.

Приведу лишь пару (из ряда) примеров такого несоответствия на стр. 85 и 86. В обеих задачах должна минимизироваться либо сумма по всем кластерам

внутрикластерных сумм, либо при иной (эквивалентной) формулировке задач для переменных должны быть четко указаны области пространства, по которым ведется оптимизация. К тому же, приведенные формулы на стр. 38 и 86 – не задачи и не минимумы, а совокупности аргументов, доставляющих минимум; F(*) (в левой части формулы на стр.85) – не минимум, а функция, аргументы которой и их области определения обязаны фигурировать в правой части формулы. Вероятно и там, и там опечатки.

На стр. 96 вероятность почему-то измеряется в процентах.

Математическая культура изложения не на высоком уровне, но этот факт, с моей точки зрения, не критичен для диссертаций по техническим наукам.

4. В главе 1 автор навесил неуместные ярлыки строго обоснованным математическим методам и алгоритмам, а также голословно (без ссылок) приписал факты труднорешаемости некоторым задачам. Математические методы и алгоритмы предназначены для решения именно математических, а не прикладных задач. Говорить о том, плох или хорош тот или иной математический метод (алгоритм) уместно лишь по отношению к математической задаче, но никак не по отношению к прикладной, четко не сформулированной проблеме. Можно сравнивать подходы к одной и той же проблеме, а также методы и алгоритмы решения одной и той же (фиксированной) математической задачи. Сравнивать алгоритмы и методы без четкой фиксации проблемы, как и говорить о том, что тот или иной математический метод (алгоритм) плохо решает ту или иную прикладную (математически не сформулированную) проблему некорректно. Некорректно сравнивать и подходы с алгоритмами. Эти замечания напрямую связаны с замечанием 1.

5. В силу того, что формулировка содержательной проблемы приведена лишь в третьей главе, обзор (глава 1) выглядит нецеленаправленным. Он не соответствует проблеме в целом, а лишь одному из этапов ее эвристического решения. Если делать обзор к существующим решениям задачи этого этапа в соответствии с названием главы, то он по объему многократно превысит объем диссертации. В литературе имеются тысячи (скорее – десятки тысяч) публикаций и постановок самых разнообразных задач кластеризации и предлагаемых алгоритмов их решения. Не ясно, почему автор не упоминает давно известные и современные методы кластеризации с теоретическими гарантиями по точности, а делает свой выбор в сторону алгоритмов локального поиска с применением эвристик, для которых установление каких-либо теоретических гарантий весьма проблематично. К тому же, в обзоре цитируются «древние» публикации. Причины такого выбора выясняются лишь в главе 3, где ситуация полностью проясняется – ни один из существующих подходов, строго обоснованных методов и алгоритмов (в том числе упомянутых в обзоре) напрямую не пригоден для решения рассматриваемой содержательной

проблемы и индуцированной задачи. В этом смысле обзор выглядит весьма странно – лишь о некоторых (не лучших) существующих подходах и алгоритмах для решения одного из этапов задачи и ничего о решении всей прикладной задачи, которая и является предметом исследования. Если бы изложение начиналось с формулировки проблемы (как указано в замечании 1), то из приведенного обзора можно было бы смело удалить большую часть цитированных работ.

В то же время выбор подхода автора к решению проблемы оправдан и очевиден, если начать чтение диссертации с третьей главы. Действительно, тщательный учет специфики (особенностей) любой индивидуальной задачи и ее допустимых входов почти всегда дает шанс на построение модифицированного алгоритма локального поиска с улучшенными показателями качества (точности или быстродействия) решения по сравнению с алгоритмами, которые не адаптированы к решению рассматриваемой индивидуальной задачи. Кроме того, автором с самого начала подразумевается (но это «скрыто» от читателя), что для решения слабоизученной проблемы он будет использовать двухэтапный подход: первый этап – предобработка, включающая отсев данных с «выбросами», второй – кластеризация. Многоэтапные подходы решения оптимизационных задач, как правило, позволяют путем применения различных эвристик и модификаций известных методов добиться неплохих и улучшенных практических результатов в решении прикладных задач. По сути, именно это и сделано в работе.

6. Утверждение автора, сформулированное в заключении (первый абзац на стр. 128) о том, что предложенные алгоритмы способны успешно решать широкий круг задач с повышенной точностью некорректно. Такие утверждения принято обосновывать теоретически (на кончике пера, т.е. строго). В работе же отсутствуют какие-либо теоретические результаты с гарантированными (доказуемыми) оценками качества (точности, временной сложности, надежности) эвристических алгоритмов. Здесь следует говорить более скромно – лишь о нескольких (конкретных) индивидуальных задачах, для которых лишь в конечном числе экспериментов получена повышенная точность решения. Уместно, видимо, напомнить, что любой репозиторий (для какой-либо труднорешаемой задачи) содержит лишь конечное число «трудных» примеров.

7. Слова «…повысить точность ... за ограниченное время», многократно встречающиеся в тексте, не очень понятны рецензенту. Время счета на практике всегда ограничено. Речь, видимо, идет о сравнении точности алгоритмов и технологий при фиксированном (т.е. одном и том же) времени работы? Точно также не очень понятно, о чем идет речь, когда говорится об «устойчивости», т.к. автор нигде определяет сами области устойчивости.

8. Не очень качественно раскрыта тема «большой» размерности (наборов данных или пространства). Эта нашумевшая тема затрагивается во многих

работах и к месту и, к сожалению, не к месту. Автор считает своим достижением увеличение размерности. По мнению рецензента это, скорее, не достижение, а факт, говорящий о том, что у автора отсутствует один из важнейших для корректного анализа данных элемент – модель порождения данных (или хоть какая-то информация о структуре данных). Если этой модели нет, то можно намерять не только сотни, но и миллиарды характеристик. Вопрос в том – надо ли? Если к нескольким значимым для классификации числовым признакам (характеристикам) добавить десяток (даже не сотни, тысячи и т.д.) незначимых (шумящих) характеристик, то значимые признаки просто «утонут» при вычислении евклидовых расстояний и функций от них. Последствия понятны и очевидны… К тому же, думается, что ставить сотни и тысячи измерительных датчиков для некоторого электронного изделия никто попросту не будет.

На самом деле, в задачах с объектами, заданными точками евклидова пространства, надо стремиться не увеличить размерность, а найти наименьшее число признаков (характеристик), достаточных для классификации объектов, т.е., наоборот, снизить размерность, удалив незначимые (никчемные) и зависимые характеристики. И здесь особую важность имеет случай, когда размерность пространства не фиксирована, но ограничена величиной ClogN, где C – некоторая константа, а N – число входных точек. Причина в том, что размерность O(logN) пространства является минимальной, при которой возможно существование N-элементного множества точек с координатами из фиксированного конечного набора значений. Мне представляется, что на это направление автору следует обратить серьезное внимание, а не гнаться за неуместными для рассматриваемой прикладной задачи «рекордами» по размерности пространства.

Высказанные замечания и вопросы не снижают **общей положительной оценки** работы и в целом могут рассматриваться как пожелания к дальнейшим исследованиям.

Результаты диссертационной работы Д.В. Сташкова прошли апробацию на нескольких международных и всероссийских конференциях прикладной направленности. Основные **результаты** диссертации **своевременно и полно опубликованы** в рецензируемых научных изданиях, в том числе изданиях, рекомендованных ВАК РФ. Содержание диссертации и полученные результаты соответствуют п.4 и п.5 паспорта специальности 05.13.01 — «Системный анализ, управление и обработка информации». Автореферат достаточно полно и правильно отражает содержание диссертационной работы. Оформление автореферата и диссертации соответствует требованиям ВАК РФ.

Заключение. Рецензируемая работа представляется **завершенной научно-квалификационной работой**, содержащей улучшенное решение актуальной прикладной задачи. Диссертация **удовлетворяет** всем требованиям и критериям п. 9 «Положения о порядке присуждения ученых степеней», утвержденного

Постановлениями Правительства РФ от 24.09. 2013 г. № 842 и от 21.04.2016 г. № 335, предъявляемым к кандидатским диссертациям, а ее автор Сташков Дмитрий Викторович заслуживает присуждения ученой степени кандидата технических наук по специальности 05.13.01 — «Системный анализ, управление и обработка информации».

Официальный оппонент,
главный научный сотрудник
лаборатории анализа данных ИМ СО РАН
д. ф.-м. н., с.н.с. А.В. Кельманов

7 ноября 2017 г.

Федеральное государственное бюджетное учреждение науки
Институт математики им. С.Л. Соболева
Сибирского отделения Российской академии наук (ИМ СО РАН)

630090, Российская Федерация, г. Новосибирск, пр. Коптюга, 4;
телефон: +7 (383) 333-28-92
тел./факс: +7 (343) 333-25-98
e-mail: im@math.nsc.ru
Web: http://www.math.nsc.ru/

Подпись
удостов
Зав. ор
ИМ СО РАН
«07» 11 2017 г.