

МЕТОДЫ УСТРАНЕНИЯ НЕПОЛНОТЫ ОТЧЕТНЫХ ДАННЫХ В СЛОЖНЫХ СИСТЕМАХ В СФЕРЕ ЗДРАВООХРАНЕНИЯ, КАЧЕСТВО КЛАСТЕРИЗАЦИИ, РАСПОЗНАВАНИЕ ОБРАЗОВ

В.В. Мочалкин,

студент магистратуры факультета Информационных бизнес систем, МФТИ.

Адрес: Москва, ул. Керченская д. 1А, корп. 1,

e-mail: mochalkin.viktor@gmail.com

В данной статье рассмотрена проблема неполноты отчетных данных в сложных системах в сфере здравоохранения. Проанализированы основные причины возникновения данной проблемы на примере системы мониторинга ПНПЗ (Приоритетный национальный проект «Здоровье»). Предложено решение по устранению неполноты отчетных данных с помощью применения методов интеллектуального анализа данных. Получены результаты применения предложенного метода к отчетным данным системы мониторинга ПНПЗ. Проведен анализ точности предложенного решения.

Ключевые слова: национальный проект, анализ данных, здравоохранение, сложные системы, качество кластеризации, распознавание образов.

Некорректные, неполные и неточные данные негативно сказываются на функционировании любой системы. Для систем комплексного мониторинга и управления деятельностью органов государственной власти эта проблема не является исключением. Назначением данной систем является обеспечение руководителей высших исполнительных ОГВ субъектов РФ эффективным инструментом мониторинга и анализа показателей эффективности деятельности. С учетом постоянного роста объема собираемых данных, а также критичностью требований к корректности и точности данных в этих системах для предоставления объективной и достоверной управленческой информации, неудивительно, что улучшение качества данных становится действительно актуальной проблемой для систем этого класса.

Неполнотой отчетных данных называется ситуация частичного отсутствия или некорректности отчетных данных, в силу ряда причин:

- ◆ ошибки ввода первичной информации;
- ◆ недостаточно времени для сбора данных;
- ◆ недостаточно средств финансирования для сбора данных;
- ◆ недостаточно первичной информации.

Более подробно причины возникновения неполноты отчетных данных рассмотрим на примере системы мониторинга Приоритетного национального проекта в сфере здравоохранения (далее Система) [1]. Данный программно-аппаратный комплекс мониторинга относится к рассмотренному выше классу систем. Целью внедрения Системы является обеспечение высоко-результативной информационно-аналитической поддержки эффек-

тивного управления, координации и мониторинга реализации ПНП «Здоровье» для деятельности органов государственной власти, на всех уровнях выполнения Проекта, что в свою очередь накладывает дополнительные требования к точности и корректности данных информационной системы.

Основные процессы поддержки управления, координации и мониторинга ПНП «Здоровье», автоматизируемые с помощью информационной системы ПАК, представлены на *рисунке 1*.

Отчетными данными в Системе являются данные по различным показателям в сфере здравоохранения за различные периоды времени в различных разрезах (по субъектам РФ, по федеральным округам, по ведомствам и т.д.).

Собираемая отчетность разделяется на два типа:

Типовая отчетность – собираемая в соответствии с Типовой инструкцией [2] по всем мероприятиям календарного плана (сетевое графика) Проекта и предоставляемая органами исполнительной власти субъектов Федерации. Эта отчетность характеризуется тем, что собирается с помощью ограниченного числа шаблонов отчетных форм установленного и зафиксированного на весь срок выполнения Проекта формата, учитывающих все выполняемые

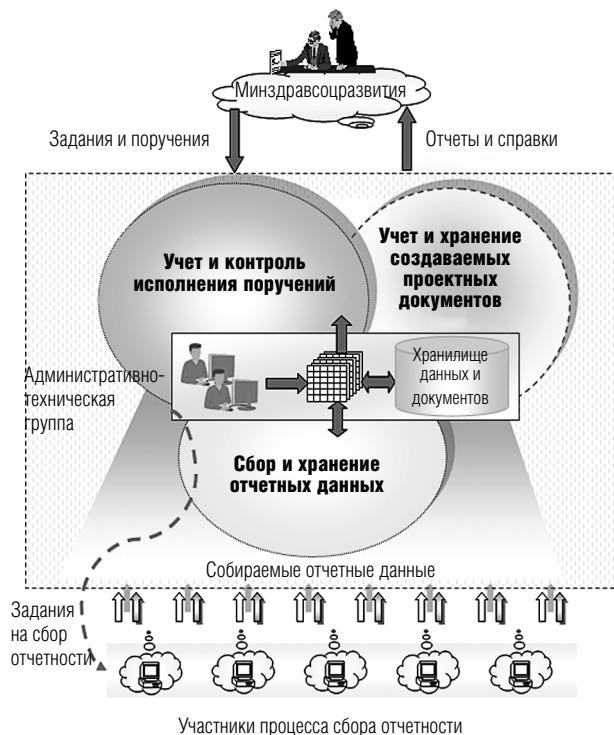


Рис. 1. Основные процессы поддержки управления, координации и мониторинга ПНП «Здоровье»

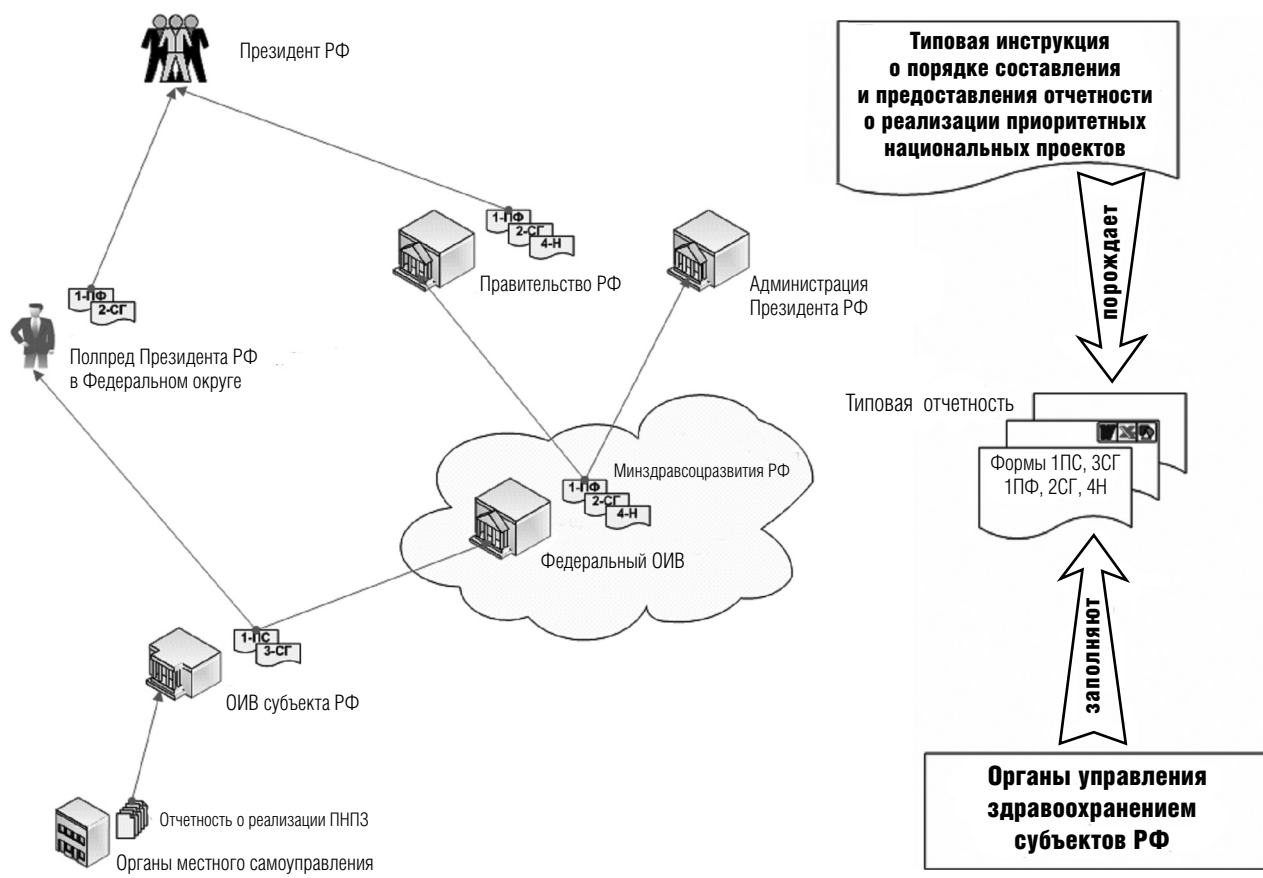


Рис. 2. Схема процесса сбора типовой отчетности

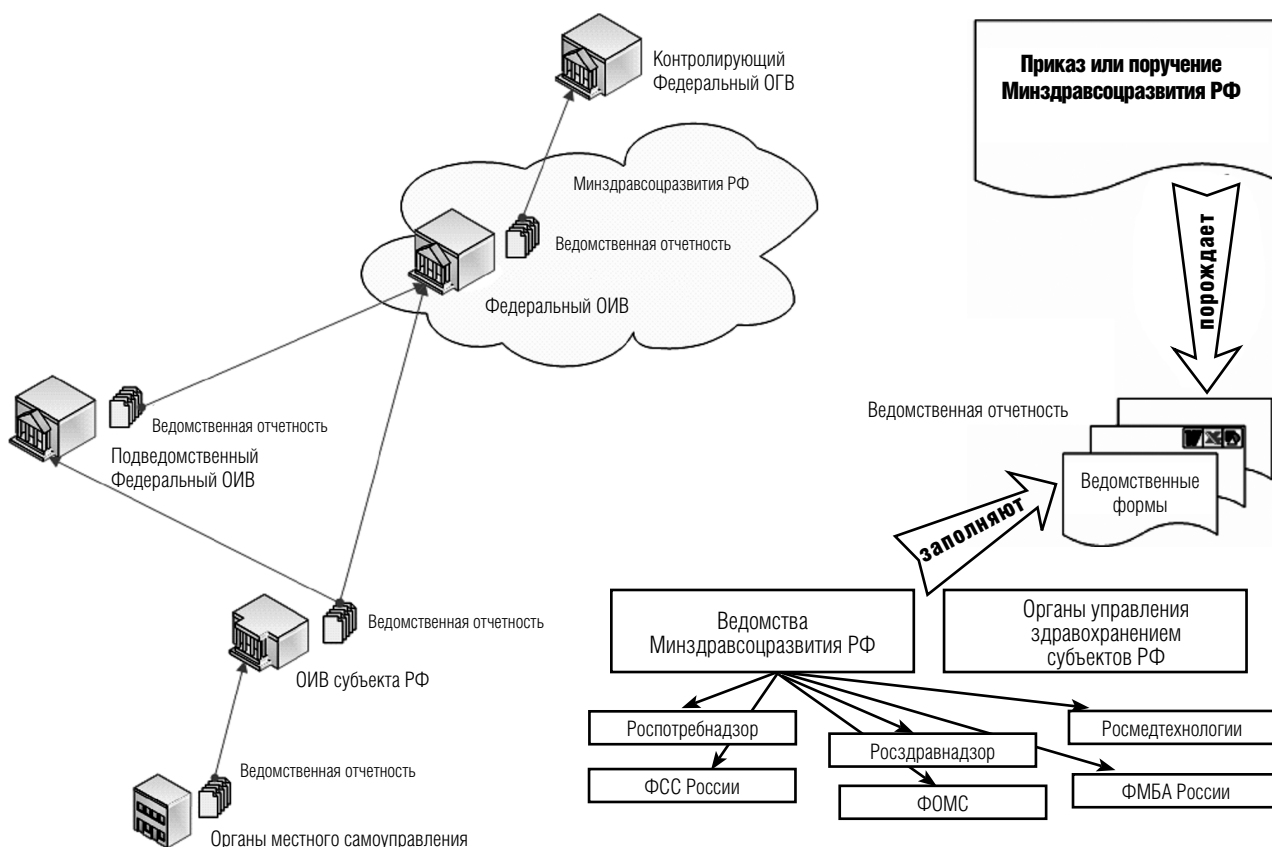


Рис. 3. Схема процесса сбора ведомственной отчетности

за отчетный период направления и мероприятия ПНП «Здоровье». Сроки предоставления отчетности строго регламентированы. Схема процесса сбора типовой отчетности представлена на рисунке 2.

Ведомственная отчетность – виды отчетности, собираемые в соответствии со специальными нормативными актами и поручениями Минздравсоцразвития России от подведомственных Министерству здравоохранения и социального развития Российской Федерации организаций. Такой вид отчетности предполагает сбор данных по конкретным направлениям календарного плана реализации Проекта, ответственными за реализацию которых на всей территории РФ являются выделенные ведомства и департаменты Минздравсоцразвития России. Ведомственная отчетность собирается с помощью различных шаблонов отчетных форм, предполагает гибкость настройки параметров сбора отчетных данных и возможность введения новых отчетных форм при необходимости. Схема процесса сбора ведомственной отчетности представлена на рис. 3.

Из рисунков 1, 2 и 3 видим, что процесс сбора отчетных данных является сложным, комплексным и крупномасштабным процессом, в котором участвует огромное количество людей и организаций. Этот процесс охватывает различные уровни управления и географически все субъекты РФ.

Таким образом, в процессе сбора отчетных данных одной из основных причин частичного отсутствия значений или некорректности данных по отдельным показателям являются ошибки ввода первичных данных операторами или администраторами на уровне ОИВ Субъектов РФ или органов местного самоуправления.

Второй существенной причиной является отсутствие первичных данных по какому-либо показателю у отчитывающихся организаций. Данная ситуация характерна для ведомственной отчетности, которая подразумевает введение новых собираемых показателей. Организации, которые ранее никогда не отчитывались по нововведенному показателю, могут не успеть предоставить данные вовремя.

Так же причиной неполноты отчетных данных может являться недостаточность финансирования процесса сбора отчетности в результате сокращения бюджетных средств.

Кроме того вследствие жестко регламентированных сроков предоставления типовой отчетности и сложной многоуровневой схемы ее сбора на повторный сбор недостающих показателей как правило не хватает времени.

В результате наличие неполноты отчетных данных приводит к некорректному функционированию Системы: неверные значения сводных и

интегральных показателей, невозможность расчета некоторых зависимых показателей и даже в некоторых случаях к неработоспособности системы. Таким образом, основная цель Системы – предоставлять объективную, достоверную и полную информацию, необходимую для обеспечения информационно-аналитической поддержки управления, координации и мониторинга реализации ПНП «Здоровье» – не достигается в полном объеме.

Устранить неполноту отчетных данных организационными методами и дополнительными сбором первичной информации не возможно по причине сложности процесса сбора отчетных данных. Следовательно, отсутствующие данные будут рассчитываться на основе имеющихся отчетных данных в Системе. Кроме того, рассчитывая прогноз отдельного значения показателя на основе остальных данных в Системе, можно проверять на корректность имеющееся в системе значение по этому показателю, что позволит отлавливать ошибки ввода первичных данных.

На данный момент в системе мониторинга ПНП «Здоровье» расчет отсутствующих значений показателей не производится, а методы расчета, применяемые в предыдущих системах в сфере здравоохранения, не эффективны и имеют значительную погрешность. Разработка эффективных методов прогнозирования позволила бы, с одной стороны, полностью устранить проблему отсутствия данных, с другой, повысить точность прогнозируемых значений показателей, а так же предоставить инструмент проверки корректности отчетных данных.

Основной гипотезой при поиске решения является предположение о возможности восстановить (спрогнозировать) недостающую информацию о состоянии системы связанных показателей, имея частичную информацию о состоянии системы в данный период времени и полную информацию о состояниях этой системы в предыдущие периоды времени [3].

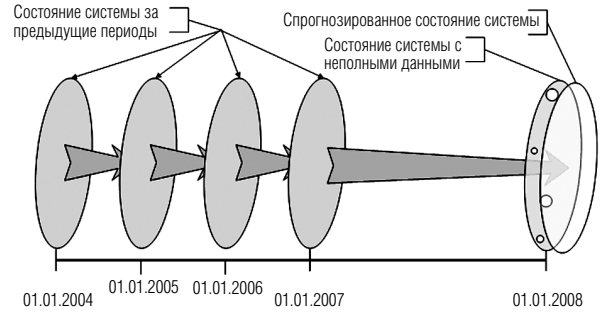


Рис. 4. Основная гипотеза исследования

Рассмотрим применяемые в сфере здравоохранения методы расчета отсутствующих значений. Первый метод заключается в усреднении известных значений показателя по субъектам федерального округа. То есть, если неизвестно значение показателя по некоторому субъекту РФ некоторого федерального округа, то это значение рассчитывается как среднее от известных значений показателя по остальным субъектам данного округа. В случае если показатель абсолютный, его пересчитывают в относительный (например, если пересчет по численности, то на 1 тыс. чел. или на 100 тыс. чел.) Второй метод заключается в применении методов регрессионного анализа и расчета значения показателя на основе известных значений этого показателя за несколько предыдущих периодов (рис. 5А).

Рассмотрим недостатки рассмотренных выше методов. Точность первого метода, во-первых, зависит от количества известных значений показателя по остальным субъектам рассматриваемого федерального округа. Во-вторых, зависит от дисперсии этих значений, которая может быть существенной в пределах субъектов данного федерального округа. Точность второго метода зависит, прежде всего, от количества периодов, за которые имеются данные по значениям показателя (рис. 5Б). Кроме того в случае линейной аппроксимации погрешность может быть значительной, если имело место перемена в тенденции изменения показателя (рис. 5В).

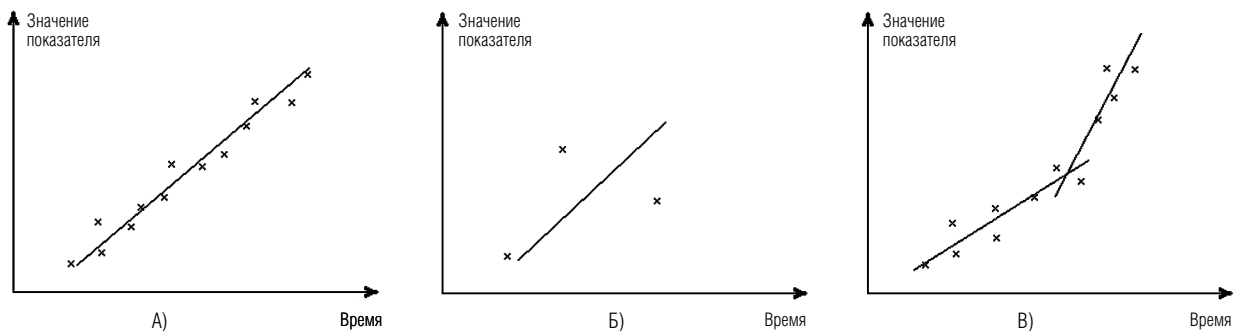


Рис. 5. Прогнозирование на основе данных за предыдущие периоды

В соответствии с вышесказанным необходимо разработать комплексный метод прогнозирования, с улучшенной точностью прогнозируемых значений, учитывающий структуру отчетных данных в виде взаимосвязанных показателей, и применяемый в случае наличия данных за небольшое количество отчетных периодов.

Для начала рассмотрим структуру хранимых отчетных данных. Данные хранятся в виде куба данных, основными измерениями которого являются: Временное измерение – иерархическое измерение значениями, которого являются временные периоды (года, кварталы, месяцы).

Территориальное измерение – иерархическое измерение значениями, которого являются территориальные единицы (РФ, Федеральные округа РФ, Субъекты РФ)

Измерение показателей – иерархическое измерение значениями, которого являются различные показатели (показатели заболеваемости, показатели здравоохранения, показатели реализации ПНП «Здоровье»).

Фактически разбиение субъектов по федеральным округам это кластеризация субъектов по признаку принадлежности к тому или иному федеральному округу. Однако не всегда субъекты, относящиеся к одному федеральному округу, схожи по прогнозируемому показателю. Следовательно, необходимо изменить метод кластеризации субъектов РФ так, чтобы субъекты, попадающие в один кластер, были схожими по состоянию дел, оцениваемых прогнозируемым показателем.

В соответствии с терминами кластерного анализа, будем называть субъекты РФ – объектами кластеризации, значения показателей – свойствами объектов кластеризации, а отчетные периоды – плоскостями разбиения (рис. 6). Причем в качестве свойств будем выбирать не любые показатели, а только связанные (коррелирующие) с прогнозируемым показателем для того чтобы объекты в кластере были схожи по прогнозируемому показателю. Выбор показателей может осуществляться процедурами экспертного оценивания (например, с помощью процедуры многовариантной экспертизы [4]) или в соответствии с коэффициентами корреляции между показателями. После выделения свойств производится разбиение объектов на кластеры (рис. 6) одним из методов кластерного анализа [5] (так же известных как алгоритмы автоматической классификации [6]).

Выполнив кластеризацию объектов в каждой плоскости разбиения, можно отследить перемещение объекта по кластерам с течением времени (рис. 7).

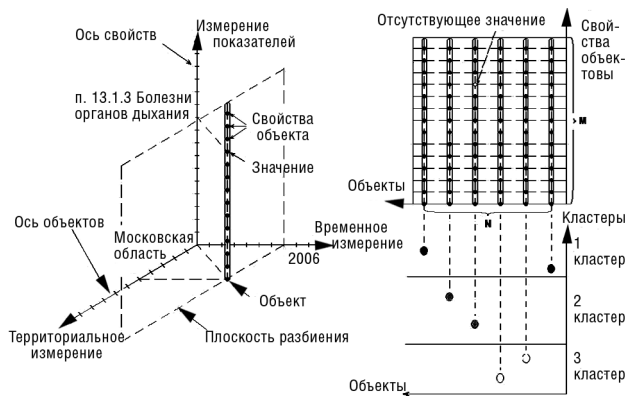


Рис. 6. Модель кластеризации

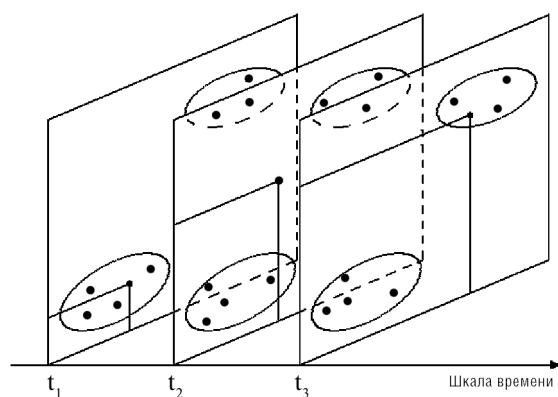


Рис. 7. Кластеризация в нескольких плоскостях разбиения

Далее можно рассчитать прогнозируемое значение одним из двух способов в зависимости от количества имеющихся отчетных периодов. Первый способ (в случае наличия только одного отчетного периода) – усреднение значений прогнозируемого показателя по объектам кластера. Так как множество объектов структурировано (разделено на классы, так чтобы объекты были схожи по прогнозируемому показателю), то дисперсия (диапазон) изменения значений показателя в пределах объектов одного кластера, как правило, будет существенно меньше, чем дисперсия изменения параметра по всем объектам. Таким образом, если по выбранным свойствам (значениям показателей) удастся произвести кластеризацию, то заполнить пропущенное значение показателя для объекта из некоторого кластера можно средним по известным значениям этого показателя для остальных объектов, попавших в этот кластер. Исходя из сделанного предположения, отклонение полученного значения от «истинного» должно быть существенно меньше (в среднем), чем обычная схема заполнения по общему среднему или среднему по округу. Расчетное значение будет находиться по формуле (1).

$$X_{ik} = \frac{\sum_{j=1}^{k-1} X_{ij} + \sum_{j=k+1}^{N_i} X_{ij}}{N_i - 1} \quad (2)$$

где X_{ik} – прогнозируемое значение показателя для k -го объекта в i -ом кластере, а N_i – количество объектов в i -ом кластере.

Второй способ – расчет с использованием усредненной тенденции изменения значений показателя по объектам отдельного кластера. Расчетное значение будет находиться по формуле (2).

$$X_{ik} = X_{ik}' + \frac{\sum_{j=1}^{k-1} [X_{ij} - X_{ij}'] + \sum_{j=k+1}^{N_i} [X_{ij} - X_{ij}']}{N_i - 1}$$

где X_{ik} – прогнозируемое значение показателя для k -го объекта в i -ом кластере, N_i – количество объектов в i -ом кластере, а X_{ik}' – значение показателя за предыдущий период для k -го объекта в i -ом кластере из разбиения в текущем периоде.

В качестве методов кластеризации были выбраны два алгоритма. Первый алгоритм, описанный в [7,8], непосредственно реализует содержательную постановку задачи автоматической классификации, а именно: если две точки расположены достаточно близко друг к другу, то они заведомо относятся к одному и тому же классу. Каждая группа определяется своим центром C и порогом T . В качестве центра первой группы выбирается точка x_p , т. е. $C_1 = x_p$. Точка x_2 относится к первой группе, если расстояние до нее от центра первой группы меньше T , т. е. если $R(C_1, x_2) \leq T$. В противном случае x_2 принимается за центр второй группы $C_2 = x_2$ и т. д. На l -м шаге, когда уже имеется r групп, текущая точка либо становится центром $(r + 1)$ -й группы, если $R(C_j, x_l) > T$ для всех $J = 1, \dots, r$, либо относится к той из r групп, для которой $R(C_j, x_l) \leq T$ (если таких групп несколько, то выбирается та, к центру которой x_l ближе всего). В качестве расстояния между двумя точками берется Евклидово расстояние в многомерном пространстве.

В качестве второго алгоритма был выбран алгоритм, описанный в [9], являющийся упрощенным вариантом алгоритма «объединение» [10]. Алгоритм работает следующим образом. Вначале строится матрица $S = \| S_{ij} \|$ мер близостей между точками x_i, x_j ($i, j = 1, \dots, N$). В качестве такой меры близости в [6] предлагается использовать величину скалярного произведения $S_{ij} = (x_i, x_j)$, в нашем же случае в качестве меры близости было взято Евклидово расстояние между двумя точками в многомерном пространстве.

Затем производится поиск максимального (не-

диагонального) элемента S_{ip} матрицы S , который соответствует ближайшим точкам x_i и x_p . Эти точки объединяются в одну группу. Пусть для определенности $p > i$, тогда из матрицы S вычеркиваются строка и столбец с номером p . Такая процедура объединения продолжается до тех пор, пока в матрице не останется заданное число строк, соответствующее числу различных групп точек.

В первом алгоритме задается пороговое значение T , во втором – количество кластеров после разбиения.

Были выбраны несколько показателей и для каждого из них по 2-3 показателя, связанных с ними, которые будут использоваться в качестве свойств объектов. Была произведена кластеризация обоими методами при различных параметрах алгоритмов разбиения и разном количестве свойств, по которым производилась кластеризация. Результаты показали, что объекты хорошо поддаются кластеризации и не сбиваются в одну кучу. Для оценки эффективности кластеризации были введены два показателя качества кластеризации $Q1$ и $Q2$, рассчитываемые по формуле (3) и (4) соответственно.

$$Q1 = \frac{1}{r} \sum_{i=1}^r \frac{1}{N_i(N_i - 1)} \sum_{j=1}^{N_i} \sum_{k=1, k \neq j}^{N_i} R(X_{ij}, X_{ik}) \quad (3)$$

где $Q1$ – среднее расстояние между объектами в кластере, r – количество кластеров, а N_i – количество объектов в i -ом кластере, $R(X_{ij}, X_{ik})$ – расстояние между объектами j и k в i -ом кластере.

$$Q2 = \frac{1}{r(r-1)} \sum_{i=1}^r \sum_{j=1, j \neq i}^r \frac{1}{N_i \cdot N_j} \sum_{l=1}^{N_i} \sum_{k=1}^{N_j} R(X_{il}, X_{jk}) \quad (4)$$

где $Q2$ – среднее расстояние между кластерами, r – количество кластеров, а N_i, N_j – количество объектов в i -ом и j -ом кластерах, $R(X_{il}, X_{jk})$ – расстояние между l -ым объектом и k в i -ом кластере.

Чем больше среднее расстояние между кластерами и чем меньше среднее расстояние между объектами в кластере, тем компактнее объекты в группах и тем больше расстояние между этими группами и, следовательно, тем качественнее произведена кластеризация. Поэтому введем параметр – единый показатель качества кластеризации Q , рассчитываемый по формуле (5).

$$Q = \frac{Q2}{Q1} \quad (5)$$

Точно таким же формулам можно воспользоваться для расчета качества кластеризации при разбиении объектов по округам, если объекты заданы в пространстве свойств (т.е. для объектов указан

перечень свойств и их значения) и, следовательно, можно рассчитать расстояние между объектами (табл. 1).

Таблица 1

Значения показателя качества кластеризации при разбиении разными методами

Метод кластеризации	Количество кластеров	Q1	Q2	Q
Кластеризация по округам	7 кластеров	15,1	17,9	1,2
Алгоритм центров	7 кластеров	4,1	39	9,5
Алгоритм объединения	7 кластеров	3,9	37,4	9,6

Видим, что качество кластеризации значительно улучшилось. Кроме того, проведя расчет качества кластеризации для разбиений по различному количеству кластеров (от 1 до 40 групп) была установлена следующая тенденция – с ростом количества кластеров показатели Q1 и Q2 монотонно уменьшаются, однако показатель Q монотонно растет.

Далее для того чтобы установить взаимосвязь между качеством кластеризации и точностью прогнозируемых значений были произведены при различных вариантах кластеризации объектов расчеты прогнозируемых значений и оценка точности этих расчетов (относительное отклонение от «истинного» значения). Так как качество кластеризации Q монотонно возрастает с увеличением кластеров, то прогнозирование проводилось при монотонно увеличивающемся количестве кластеров разбиения. Кроме того для отдельного разбиения, чтобы набрать статистику, проводился не один расчет, а целая серия расчетов – по отдельности для каждого объекта на основе данных по остальным объектам. Затем рассчитывались среднее и дисперсия относительных отклонений серии прогнозирования для каждого разбиения. На рисунке 8 представлена зависимость этих параметров от количества кластеров разбиения. Видим что, точность прогнозирования возрастает с ростом числа кластеров. Однако при увеличении количества кластеров разбиения увеличивается количество одиночных объектов (только один объект в кластере), для которых нельзя выполнить прогнозирование значения показателя (рис. 9).

В результате были внесены изменения в алгоритм прогнозирования. Количество кластеров перестало быть фиксированным числом и стало автоматически выбираться для каждого объекта, для которого

производится прогнозирование значения показателя. Делается это следующим образом: выбирается из диапазона от 1 до 40 наибольшее количество кластеров, при котором в кластере с объектом, для которого делается прогнозирование, количество объектов больше порогового значения (анализ показал, что оптимальным является порог равный 4). Таким образом, для каждого прогноза из серии расчетов выполняется наиболее оптимальное разбиение, никакие объекты не отбрасываются и, следовательно, метод можно применять к любым данным.



Рис. 8. Зависимость среднего и дисперсии относительных отклонений от количества кластеров

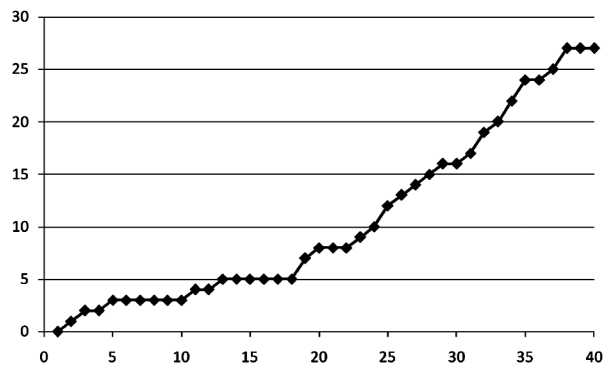


Рис. 9. Зависимость числа одиночных объектов от количества кластеров

Разработанный метод был применен к различным показателям, при различных свойствах кластеризации. В результате среднее относительное отклонение было в пределах от 45% до 50%, а дисперсия – от 60% до 70%. Анализ показал, что эти величины превышались небольшой группой прогнозов, относительные отклонения которых были от 300% до 700%.

Дальнейший анализ показал, что в среднем количество таких прогнозов было около 10% от общего числа.

Столь высокие отклонения объяснялись нарушением корреляции прогнозируемого показателя

с зависимыми показателями, по которым осуществлялась кластеризация. Т.е. хотя объект и попадал в некоторый кластер по свойствам кластеризации (следовательно, был схож с объектами кластера по зависимым показателям), однако существенно отличался по прогнозируемому показателю.

Таким образом, точность прогнозирования данного метода зависит от корреляции между прогнозируемым и зависимыми показателями. Дальнейшее улучшение невозможно без наличия данных хотя

бы еще по одному периоду. Однако следует отметить, что количество неточных прогнозов составляет порядка 10% (применительно к выбранным данным Системы). Для оставшихся 90% прогнозов при любых прогнозируемых и зависимых показателях среднее относительное отклонение порядка 30%, а дисперсия – 25%. Распределение относительных отклонений прогнозируемых значений одной из серий при использовании старого и нового методов к данным за один период представлено на рисунке 10.

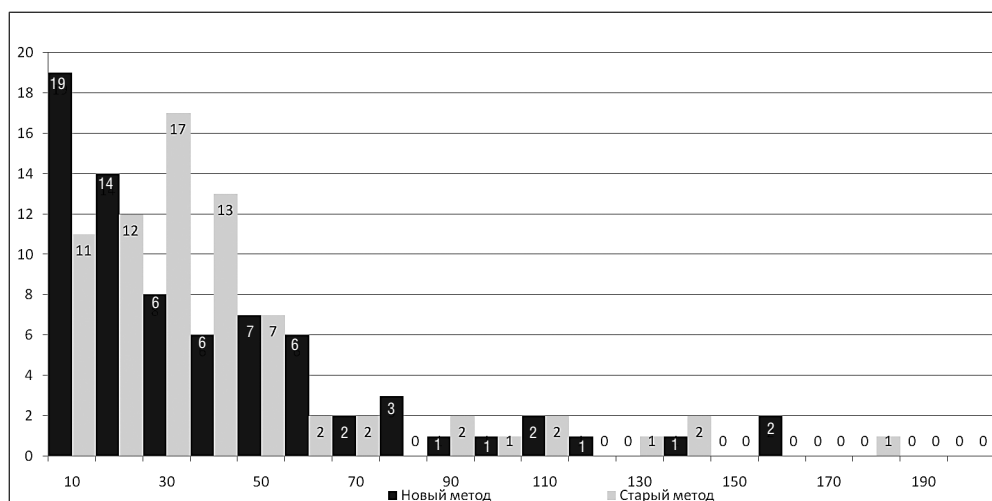


Рис. 10. Распределение отклонений при прогнозировании (1 период)

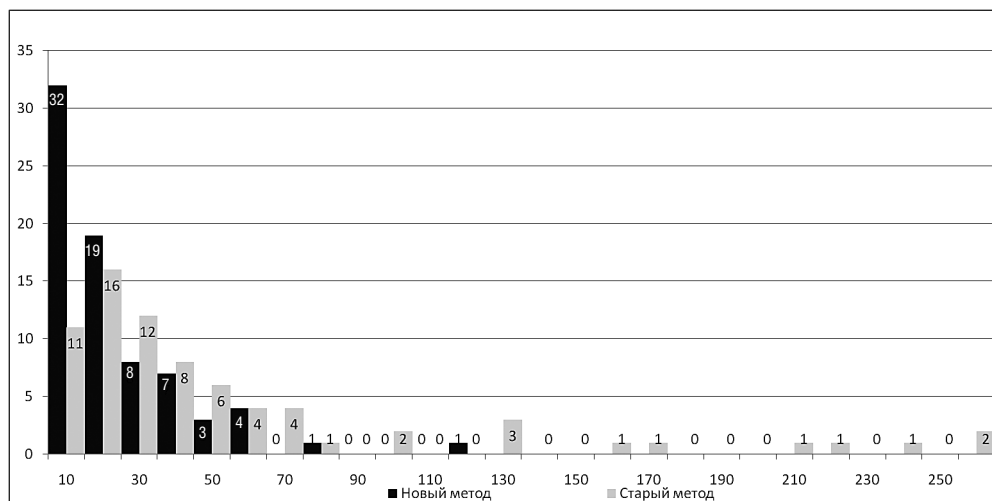


Рис. 11. Распределение отклонений при прогнозировании с учетом динамики

В случае если имеются данные, хотя бы за один предыдущий период, точность прогнозирования может быть увеличена за счет использования для расчета прогнозируемого значения формулы (2). Данный способ расчета должен устранить проблему высоких отклонений 10% прогнозов. Результаты показали, что для разных прогнозируемых показателей при различных свойствах кластеризации среднее

относительное отклонение составило около 20%, а дисперсия – 25%. Относительные отклонения почти 97% прогнозов попали в интервал от 0% до 100%. Распределение относительных отклонений прогнозируемых значений одной из серий, рассчитываемых новым методом с учетом динамики в сравнении с относительными отклонениями значений, рассчитанных старым методом, представлено на рисунке 11.

Обобщим полученные данные о результатах применения рассмотренных методов прогнозирования к имеющимся отчетным данным (таблица 2).

Таблица 2

Обобщение данных о применении рассмотренных методов расчета

МЕТОД	РЕЗУЛЬТАТ		
	Группы	Среднее относительное отклонение серий	Дисперсия относительного отклонения серий
Разбиение объектов по федеральным округам	По всем прогнозам	60-70%	70-80%
Разбиение объектов по схожести на основе их свойств (имеется только один период)	По всем прогнозам	45-60%	60-70%
	По 90% прогнозам, за исключением случаев нарушения корреляции	25-30%	20-25%
Разбиение объектов по схожести на основе их свойств (с динамикой)	По всем прогнозам	18-22%	20-25%

Полученные результаты показали большой потенциал предложенного комплексного метода прогнозирования значений показателей. Он значительно превосходит методы, использовавшиеся ранее для решения задачи устранения неполноты отчетных данных. На основе данного метода может быть разработан интегрированный в систему мониторинга ПНП «Здоровье» инструмент для устранения неполноты или верификации отчетных данных.

Дальнейшие работы могут проводиться как в исследовательской области, так и в прикладной. В исследовательской области работы могут вестись в направлении повышения точности предложенного метода. Подбор наиболее оптимальных параметров, исследование различных зависимостей, таких как точность прогнозирования от дисперсии значений показателя, от процедур экспертной оценки или от корреляции связанных показателей. Применение более сложных алгоритмов кластеризации. В прикладной области – в направлении оптимизации алгоритмов кластеризации; в направлении разра-

ботки и оптимизации возможности инструмента подстраиваться под структуру отчетных данных на основе заполненной части данных для повышения точности прогнозирования. Изучение возможности апробации данного метода для создания унифицированного метода прогнозирования, который мог бы применяться в системах мониторинга Приоритетных национальных проектов. ■

Литература

1. Технорабочая документация проекта «Создание программно-аппаратного комплекса мониторинга Приоритетного национального проекта в сфере здравоохранения».
2. Типовая инструкция о порядке составления и представления отчетности о реализации приоритетных национальных проектов. Одобрено Президиумом Совета при Президенте Российской Федерации по реализации приоритетных национальных проектов (протокол № 4 от 28 февраля 2006 г., протокол № 8 от 27 июня 2006 г.)
3. Дорофеюк А.А., Дорофеюк Ю.А. Методы структурно-классификационного прогнозирования многомерных динамических объектов / Искусственный интеллект, № 2, 2006. - с.138-141.
4. Дорофеюк А.А., Покровская И.В., Чернявский А.Л. Экспертные методы анализа и совершенствования систем управления / Автоматика и телемеханика. 2004, №10.-С. 172 – 188.
5. Мандель И.Д.. Кластерный анализ. М.: Финансы и статистика, 1988. – 176 с. (10)
6. Айвазян С.А., Бухштабер В.М., Енюков И.С. Прикладная статистика: Классификация и снижение размерности. М.: Финансы и статистика, 1989. – 344 с.
7. Sebestyen G. Pattern recognition by an adaptive process of sample set construction. Trans. IRE, IT-8, No. 5, 1962. (10)
8. Waltz M.D., Fu K.S. A heuristic approach to reinforcement learning control systems. IEEE Trans., v. AC-10, No. 4, 1965. (11)
9. King B. Step-wise clustering procedures. J. Amer. Stat. Assoc, 62, No. 317, 1967.
10. Дорофеюк А.А. Алгоритмы обучения машины распознаванию образов без учителя, основанные на методе потенциальных функций. Автоматика и телемеханика, № 10, 1966.