УДК 004.9:004.8

ТЕХНОЛОГИИ DATA MINING ДЛЯ ОЦЕНКИ РЕГИОНАЛЬНОГО УРОВНЯ РАЗВИТИЯ ОТРАСЛИ ИНФОРМАЦИОННО-КОММУНИКАЦИОННЫХ ТЕХНОЛОГИЙ

Ионис А.Г., Сметанина О.Н., Юсупова Н.И., Сазонова Е.Ю.

ФГБОУ ВО «Уфимский государственный авиационный технический университет», Уфа, e-mail: solomonarmkeys@gmail.com

В статье приведены результаты исследования в области управления уровнем развития отрасли информационно-коммуникационных технологий в регионах России. Обоснована актуальность исследования, обусловленная тем, что согласно Стратегии развития информационного общества на 2017–2030 гг. основным приоритетом развития стала цифровая экономика. Одним из определяющих факторов цифровой экономики являются информационно-коммуникационные технологии. В данной отрасли за последние десятилетия наблюдается бурное развитие, однако в целом страна до сих пор сильно отстает от уровня мировых держав. Показана одна из причин такой ситуации, а именно значительный разрыв между значениями показателей данной отрасли в различных регионах страны. Авторами предлагается методика, основанная на технологии Data Mining, которая позволит получить неявные знания о регионах, схожих по показателям. Полученные знания в дальнейшем позволяют осуществлять эффективное управление уровнем развития отраслью информационно-коммуникационных технологий. Предложенная методика основана на использовании методов факторного и кластерного анализа, а также построении системы нечетких продукционных правил. С помощью факторного анализа были выделены наиболее значимые для данной отрасли факторы. Кластерный анализ позволил выделить группы регионов, обладающих схожими показателями. Интерпретация результатов с учетом семантики предметной области и экспертные знания позволяют построить систему нечеткого вывода для формирования рекомендаций.

Ключевые слова: кластерный анализ, факторный анализ, цифровой разрыв, регионы РФ, отрасль информационно-коммуникационных технологий, интеллектуальный анализ данных

DATA MINING TECHNOLOGIES FOR EVALUATING THE REGIONAL LEVEL THE DEVELOPMENT OF THE INDUSTRY OF INFORMATION-COMMUNICATION TECHNOLOGIES

Ionis A.G., Smetanina O.N., Yusupova N.I., Sazonova E.Yu.

Federal State Budgetary Educational Institution of Higher Education Ufa State Aviation Technical University, Ufa, e-mail: solomonarmkeys@gmail.com

The article presents the results of research in the field of management of the level of development of information and communication technologies in the regions of Russia. The relevance of the study due to the fact that according to the Strategy of development of the information society for 2017-2030 the main priority of development was the digital economy. One of the defining factors of the digital economy is information and communication technologies. In this industry, in recent decades, there has been rapid development, but in General, the level of the country is still far behind the level of the world powers. One of the reasons for this situation is shown, namely, a significant gap between the values of indicators of this industry in different regions of the country. The authors propose a technique based on Data Mining technology, which will provide implicit knowledge about the regions similar in terms of indicators. The knowledge gained in the future allows for effective management of the level of development of the information and communication technologies industry. The proposed method is based on the use of methods of factor and cluster analysis, as well as the construction of a system of fuzzy production rules. With the help of factor analysis, the most important factors for the industry were identified. Cluster analysis made it possible to identify groups of regions with similar indicators. Interpretation of the results taking into account the semantics of the subject area and expert knowledge allow us to build a system of fuzzy inference for the formation of recommendations.

Keywords: cluster analysis, factor analysis, digital divide, Russian regions, information and communication technologies industry, data mining

Согласно программе цифровой экономики, сама цифровая экономика представляется тремя взаимосвязанными между собой уровнями: рынки и отрасли экономики, платформы и технологии и среды развития для платформ и технологий. При этом основными технологиями развития являются: big data, нейротехнологии, промышленный интернет и многие другие современные информационные технологии, являющиеся частью отрасли информационно-коммуникационные технологии (ИКТ). Реализация со стороны государства и региональных властей конкретных мероприятий в рамках стратегии должна поддержать темп роста ИКТ отрасли, путем должного развития национальной инфокоммуникационной инфраструктуры, создания научно-технической базы для развития инноваций и обеспечения доста-

точного комплекса доступных и надежных услуг на базе ИКТ для всех отраслей экономики, что само по себе невозможно без развитых платформ и технологий, институциональной и инфраструктурной сред.

Именно поэтому ключевой задачей при исследовании информационно-коммуникационной деятельности является оценка и прогнозирование текущего состояния, которое характеризуется степенью развития ключевых факторов, определяющих внутреннюю структуру рынка ИКТ, а также построение прогноза развития данной отрасли.

Однако решение данной задачи осложняется не только тем, что отрасль ИКТ является на сегодняшний день самой динамично развивающейся, именно в этой отрасли чаще появляются новые технологии, но и тем, что Россия состоит из 85 регионов, значения показателей уровня развития ИКТ в которых зачастую просто несопоставимы. При этом имеется разница в уровнях развития не только в рамках федеральных округов, например в Центральном и Приволжском, но и в рамках одного федерального округа. Так, существует значительный цифровой разрыв между значениями показателей Московской и Брянской областей, республикой Башкортостан и Пермским краем.

В связи с вышеизложенным целью данного исследования является анализ информационно-коммуникационных технологий в 85 регионах России, который позволит выявить основные тенденции развития, взаимосвязи между ключевыми показателями, группы сходных по уровню развития регионов, а также причинно-следственные связи между уровнем развития и основными показателями. Математическая постановка задачи формулируется следующим образом:

Дано: $X_i = \{X_{i1}, X_{i2}, \ldots, X_{in}\}$ – множество характеристик отрасли ИКТ в регионах РФ, где $X_{i1}, X_{i2}, \ldots, X_{in}$, $j = \overline{1,n}$, $n = 34$ – совокупность показателей развития отрасли, $i = \overline{1,85}$ – номер региона. Необходимо определить функцию принадлежности региона определенному кластеру F: $X_i \rightarrow Z_i \rightarrow Y_c$, где Y_c, $l = \overline{1,c}$ – группы сходных по уровню развития регионов.

Материалы и методы исследования

В качестве теоретической базы за основу были взяты научные труды ученых, посвященные использованию информационно-коммуникационных технологий: С.А. Москальонова, А.Г. Львова [1], Д. Йоргенсона, а также различные отечественные и зарубежные публикации исследований [2, 3].

В исследовании используются данные (4 класса, 34 показателя) [4] Федеральной службы государственной статистики за 2016 г. по регионам Российской Федерации [5] (рис. 1).

Для пересчета экономических показателей, исчисленных в денежном выражении, с целью избавления от уровня инфляции, была произведена нормализация по формуле

$$\text{Значение\_текущего\_года}_i = \frac{\text{Значение\_в\_текущем\_году}_i}{\text{Значение\_ВПП}} \cdot 100\%.$$

Кроме того потребовалось проведение другого вида предварительной обработки данных (рис. 2), которая предшествует Data Mining и моделированию системы нечеткого вывода. В частности, выявление аномальных отклонений, корреляционный и факторный анализ.

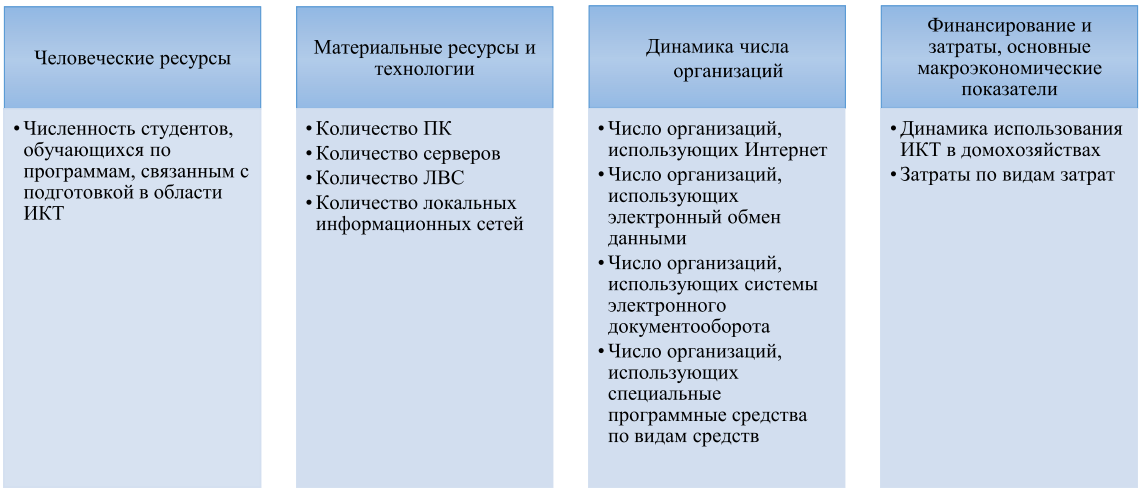

Рис. 1. Классификация показателей

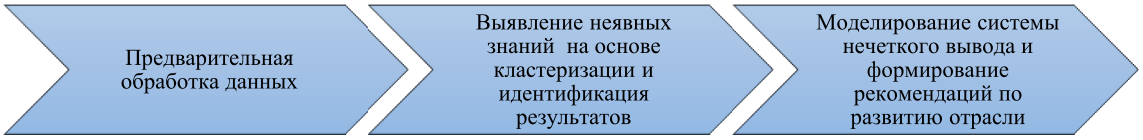

Рис. 2. Методика анализа данных и организации информационной поддержки принятия решений

В качестве инструментальной базы использовались программное средство MS Excel и аналитические платформы Deductor, SPSS. Перечисленные аналитические платформы позволяют использовать для анализа методы статистического и интеллектуального анализа данных, в частности методы корреляционного, факторного и кластерного анализа [6, 7].

На основании показателей использования ИКТ можно предложить следующий рейтинг по информационно-коммуникационной активности федеральных округов (ФО): Центральный (2); Приволжский (3); Сибирский (4); Северо-Западный (5); Уральский (6); Южный (7); Дальневосточный (8); Северо-Кавказский (9). При этом следует отметить огромное различие показателей между лидирующими округами и отстающими, а также – их аномальные «выбросы», характерные для г. Москвы. В связи с чем возникает вопрос о необходимости включения этих показателей в дальнейший анализ. Для выявления аномальных наблюдений в переменных были построены диаграммы размаха с использованием пакета SPSS, которые показали отсутствие аномальных наблюдений в показателях развития отрасли ИКТ.

Исследование взаимосвязи показателей. Корреляционный анализ проведен с целью обнаружения зависимости между показателями. При проведении анализа используются выборочные коэффициенты для проверки на значимость. В частности, проверяется нулевая гипотеза о том, что коэффициент корреляции незначим, то есть на уровне значимости α.

Результаты проведенного корреляционного анализа показали наличие высокого и очень высокого уровня корреляции между показателями. Так, например, имеется высокая корреляционная связь между показателями «Использование ПК» и «Организации, использовавшие системы электронного документооборота» (коэффициент корреляции 0,93), «Затратами на приобретение вычислительной техники и оргтехники» и «Затратами на обучение сотрудников, связанное с развитием и использованием ИКТ» (коэффициент корреляции 0,99).

Причинно-следственные связи между показателями. При возникновении существенных корреляционных связей между показателями необходимо выделить обобщенные некоррелирующие между собой факторы. Наиболее подходящим для этого выступает компонентный анализ, позволяющий снизить размерность признакового пространства без потери информативности, а также выделить факторы, которые будут упорядочены по убыванию дисперсии с целью получения возможности оценки вклада каждого фактора в объясняющую способность. Таким образом, задачей анализа будет выделение наиболее существенных факторов из совокупности признаков, характеризующих рассматриваемый объект. Поставленная цель достигается заменой исходных признаков меньшим числом нормированных и ортогональных факторов.

Для определения количества факторов воспользуемся критерием каменистой осыпи (критерий осыпи Кэттеля). Согласно [8] при использовании метода Монте-Карло это количество определяется точкой, в которой непрерывное падение собственных значений замедляется и после которой уровень остальных собственных значений отражает только случайный «шум». Согласно полученным результатам, эта точка может соответствовать фактору 5 или 6.

Для повышения качества интерпретации, к полученным результатам применяется процедура вращения, которая позволяет выделить некий набор «истинных» факторов, каждый из которых будет представлять собой изолированную группу показателей, имеющих некий общий «смысл» (рис. 3). Результаты факторного анализа при использовании метода варимакс позволили выделить пять наиболее значимых факторов с совокупной объясняющей способностью в 84%. Интерпретация фактором может быть следующей: фактор 1 – связан с материальными и техническими ресурсами, а также с их использованием в организациях (количественный), в том числе использование электронного документооборота; фактор 2 – связан с затратами на отрасль по видам деятельности (экономический); фактор 3 – связан с распространением ИКТ технологий в домохозяйствах; факторы 4 – связан с человеческими ресурсами в данной отрасли; факторы 5 – связан с распространением специальных средств, локальных сетей и серверного обеспечения.

Характеристика ИКТ-отрасли по территориальной расположенности. При выполнении исследований авторы выявили регионы, имеющие схожие показатели. Для этого использован кластерный анализ на основе карт Кохонена. Алгоритм функционирования самообучающихся карт является одним из вариантов кластеризации многомерных векторов. Отличием алгоритма является то, что в нем все нейроны упорядочены в некоторую структуру.

В ходе обучения модифицируется не только нейрон-победитель, но и его соседи, но в меньшей степени. За счет этого SOM можно считать одним из методов проецирования многомерного пространства в пространство с более низкой размерностью. Алгоритм позволяет векторам, схожим в исходном пространстве, оказаться рядом на полученной карте (рис. 4).

Результаты кластеризации демонстрируют 13 кластеров (таблица). Кластер 9 включает в себя преимущественно Центральный ФО. Кластер 3 можно охарактеризовать как кластер с наиболее высокой активностью в сфере информационных и коммуникационных технологий, в него вошла Москва. Как было показано выше, именно в этом субъекте использование ИКТ-технологий являлось аномальным по сравнению со средними значениями наблюдений. Практически по всем исследуемым показателям в субъекте наблюдались максимальные значения признака, намного превышающие его средние значения. Все остальные кластеры являются неоднородными по территориальной расположенности, однако использование ИКТ-технологий здесь находится на среднем или низком уровне.

Переменные	Окончательные факторы (Варимакс метод)				
	Фактор 1	Фактор 2	Фактор 3	Фактор 4	Фактор 5
пк	0,8597				
серверов					0,6220
локал					0,7685
глобал	0,9262				
Интернет	0,9367				
широкополосный	0,8585				
веб-сайт	0,6959				
спецсред	0,7511				
спецсреднауч					
спецсредпроект					0,6889
спецсредасу					0,8097
спецсредэконом					0,7960
спецсредфин	0,6122				
спецсредбд	0,6480				
спецсредРИдС					
спецсредобуч					0,8450
спецсредCRM					
спецсредЭППС					0,7873
спецсредПС					0,7919
Затраты		0,9844			
приобртех		0,9737			
приобртеле		0,9706			
приобрПО		0,9878			
оплатуэлектро-связи		0,9685			
оплатудоступаИнтернет		0,9250			
обучениеИКТ		0,9703			
оплатуспециалИКТ		0,9787			
прочиеЗатраты		0,9849			
использэлекдок	0,7341				
использовэлекобмен	0,7082				
домохозяйствапк			0,6174		
домохозяйстваинтернет			0,9338		
домохозяйстваширокопол...					
использованиеИнтернет			0,8470		
студенты				0,8658	

Рис. 3. Результаты факторного анализа

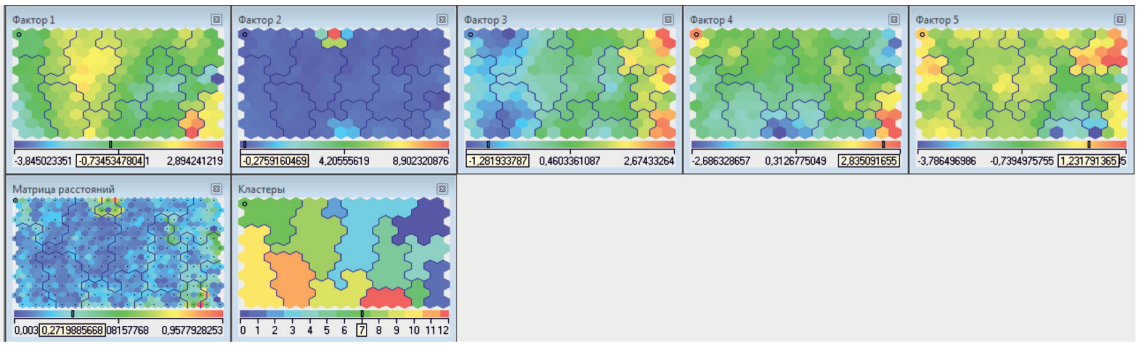

Рис. 4. Кластеризация с помощью карт Кохонена

Также стоит отметить, что в большинстве случаев большая часть регионов, принадлежащих одному кластеру, имеет примерно одинаковый уровень инвестиционной привлекательности.

В статье представлены результаты двух этапов методики проведения анализа и поддержки принятия решений. Фрагмент заключительного этапа по моделированию системы нечеткого вывода и формированию рекомендаций по развитию отрасли описан ниже.

Результаты исследования и их обсуждение

На сегодняшний день необходимо комплексное развитие и внедрение отрасли ИКТ во все отрасли экономики. Применение технологий, продуктов и услуг данной отрасли приводит к повышению производительности труда и позволяет повысить эффективность работы организаций различных отраслей, а также увеличить вклад этих отраслей в ВВП. При этом ключевые отрасли, использующие ИКТ, могут быть представлены в виде когнитивной диаграммы (рис. 5), где f_1 – отрасль ИКТ, f_2 – ВВП, f_3 – НИОКР, f_4 – сетевые организации, f_5 – рыночные услуги, f_6 – обрабатывающее производство.

Описание кластеров

Кластер	Регионы
1	Калининградская, Мурманская, Магаданская области, Ханты-Мансийский АО, Ямало-Ненецкий АО, Камчатский край, Чукотский АО
2	г. Санкт-Петербург, республика Татарстан, Хабаровский край
3	г. Москва
4	Воронежская, Тульская, Псковская, Астраханская, Волгоградская, Удмуртская, Свердловская, Челябинская, Иркутская области, Красноярский край, республики Карелия и Коми
5	Ростовская область, г. Севастополь, Приморский край, республика Саха
6	Курская, Новосибирская, Омская области, республики Калмыкия, Карачаево-Черкесская, Северная Осетия, Мордовия
7	Республика Крым, республика Ингушетия
8	Орловская, Рязанская, Кировская, Нижегородская, Тюменская, Томская области, республика Бурятия
9	Белгородская, Брянская, Владимирская, Ивановская, Смоленская, Тамбовская, Ярославская, Новгородская, Оренбургская области, Ставропольский и Пермский край, Чувашская республика и республика Башкортостан
10	Московская, Тверская, Архангельская, Ленинградская области, Ненецкий АО, Краснодарский край
11	Костромская, Самарская, Саратовская, Курганская, Кемеровская, Амурская, Сахалинская области, Еврейская АО, Алтайский край, республика Марий Эл
12	Калужская, Липецкая, Вологодская, Пензенская, Ульяновская области, республики Адыгея, Алтай и Хакасия, Забайкальский край
13	Республика Дагестан, Кабардино-Балкарская республика, Чеченская республика, республика Тыва

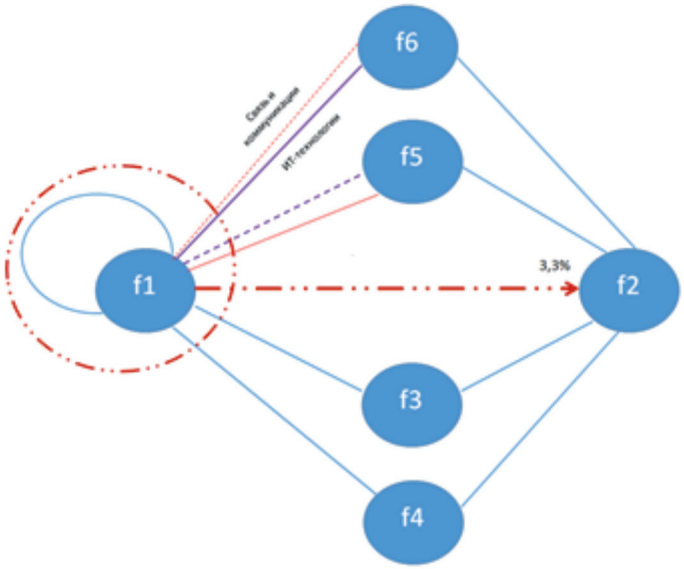

Рис. 5. Влияние ИКТ на отрасли экономики и ВВП

Как видно из диаграммы, продукты и услуги, полученные в отрасли, развивают саму отрасль ИКТ, оказывают влияние на экономический рост, выраженный в ВВП. Также стоит отметить, что для отдельно взятых областей могут требоваться продукты и услуги, полученные конкретным сегментом рынка. Так, для обрабатывающего производства в большей степени оказывают влияние непосредственно ИТ-технологии, в то время как для рыночного производства важнее сектор связи и телекоммуникаций.

Необходимость развития отрасли и ее конкретных областей находит свое отражение и в Стратегии развития информационного общества и в Программе цифровой экономики, где отмечаются ключевые аспекты, необходимые для ускорения темпов развития. К примеру, могут быть выделены такие аспекты, как повышение эффективности

государственного участия в развитии отрасли ИКТ; повышение привлекательности сектора для внешних инвестиций; усиление конкурентной борьбы; изменение социально-демографических условий.

При этом выделяются три варианта развития сектора ИКТ в России [9]: инерционного импорт-ориентированного технологического развития (1); догоняющего развития и локальной технологической конкурентно-способности (2); лидерства в ведущих научно-технических секторах и фундаментальных исследованиях (3).

На основе полученных результатов можно говорить о необходимости увеличения числа инвестиций в качественные показатели развития отрасли, а также увеличение численности студентов, обучающихся по программам, связанным с подготовкой в области ИКТ; переобучение персонала или повышение квалификации в области ИКТ и пр.

Обоснование метода дальнейшего моделирования

С учетом динамики развития отрасли, а также условий неопределенностей внешней и внутренней среды, эффективность управления развитием рынка ИКТ во многом будет зависеть от результативности, качества и скорости принятия решений. Именно поэтому наиболее важным является построение адекватной системы, которая позволит оперативно реагировать и подстраиваться под изменчивость внешней среды.

На основе имеющихся результатов авторами предложено использовать аппарат нечеткой логики и когнитивных карт.

Для моделирования развития отрасли ИКТ нечеткие модели представляются в виде нечетких продукционных сетей. Анализ показателей с привлечением экспертов позволил идентифицировать лингвистическую переменную, характеризующую показатели отрасли ИКТ с использованием следующего терм-множества, определяющего уровень развития: $T = \{$*ОН – очень низкий, Н – низкий, С – средний, В – высокий, ОВ – очень высокий*$\}$. Согласно анализу мнений экспертов, показатели имеют следующий вид функции принадлежности (рис. 6).

В процессе анализа результатов факторного и кластерного анализа, стало возможным выделить терм-множества, определяющие группу регионов со схожими характеристиками: Т1 = {Очень Низкий (ОН), Низкий (Н), Ниже среднего (НС), Средний (С), Выше среднего (ВС), Высокий (В), Очень Высокий (ОВ)}.

Согласно выделенным термам, было проведено объединение кластеров:

Терму Очень Высокий определен кластер 3, Высокий – кластер 2, Выше среднего – кластер 10, Средний – кластеры 9 и 4, Ниже среднего – кластеры 1, 6, 8, 11, Низкий – 5 и 12 кластеры, Очень низкий – 7 и 13 кластеры.

Объединение кластеров в сходные группы объясняется примерно схожим уровнем развития, а выделенные 7 терм укладываются как в полученные аномальные значения кластера 3 (г. Москва), так и в кластеры с очень низким уровнем развития, куда вошли, к примеру, Республика Крым и г. Севастополь, недавно вошедшие в состав Российской Федерации.

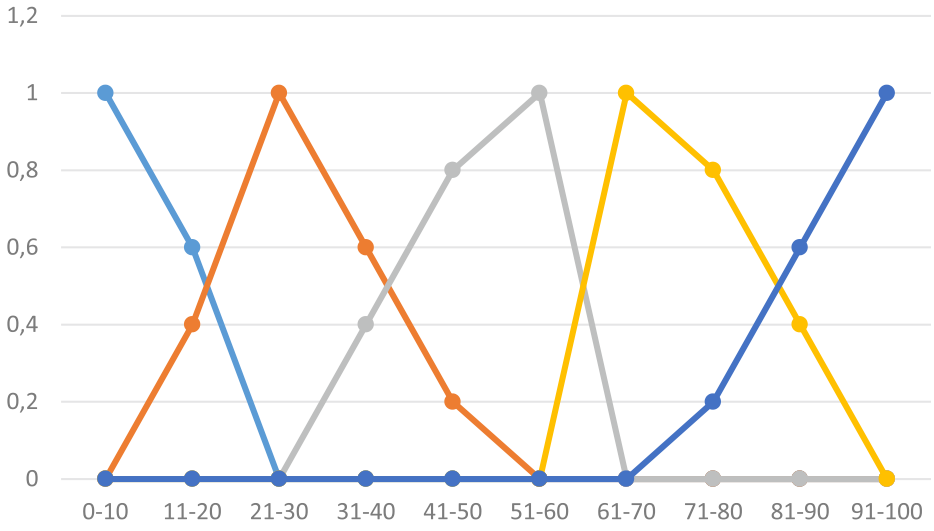

Рис. 6. Пример функции принадлежности показателя использование ПК

При создании системы принятия решений при управлении отраслью информационно-коммуникационных технологий, важнейшим аспектом является формирование базы знаний, состоящей из продукционных правил. Существует несколько способов разработки правил, основанных на различных методах и алгоритмах технологий искусственного интеллекта и нейро-нечеткого моделирования. Так, к примеру, одним из решений может быть нейро-сетевая продукционная модель отрасли с самообучением на имеющейся обучающей выборке. Однако при таком подходе велик риск проявления слишком большого количества продукционных правил (34 лингвистических переменных на одно четкое число с семью термами). В связи с этим для решения поставленной задачи формирования базы нечетких продукционных правил использовался метод формализации представления эмпирических знаний экспертов и когнитологов в области развития рынка информационно-коммуникационных технологий по схеме «если …, то …».

Предложенная нечеткая модель позволит в дальнейшем построить систему поддержки принятия решений, которая поможет лицу, принимающему решение выявить приоритеты развития в конкретном регионе, а также выработать план мероприятий.

Выводы

Одной из ключевых проблем развития рынка ИКТ в России сегодня является значительный разрыв между значениями показателей в рамках 85 регионов страны. Вторая проблема связана с вопросами выбора и сбора, характеризующими отрасль показателей с целью моделирования и анализа. Для решения проблемы потребовалось провести предварительную обработку данных.

Факторный анализ, проведенный с использованием метода главных компонент, позволил выделить и интерпретировать 5 наиболее значимых факторов с совокупной объясняющей способностью в 84%. Кластеризация по субъектам позволила разделить регионы на группы согласно уровню развития информационных технологий. Специфика показателей и имеющиеся результаты, а также опыт создания нечетких систем логического вывода позволили авторам использовать этот аппарат и создать систему нечетких продукционных правил, с использованием которой ЛПР будет получать рекомендации.

Результаты исследований, приведенные в статье, частично поддержаны грантами РФФИ 18-07-00193, 19-07-00709.

Список литературы

1. Москальонов С.А., Львов А.Г. Анализ инновационного потенциала российской экономики: метод производственных функций // Материалы XII Международной научной конференции по проблемам развития экономики и общества (г. Москва, 5–7 апреля 2011 г.). М.: Издательский дом Высшей школы экономики, 2011. С. 417–428.

2. Зимин К.В., Маркин А.В., Скрипкин К.Г. Влияние информационных технологий на производительность российского предприятия; методология эмпирического исследования // Бизнес-информатика. 2012. № 1. С. 40–48.

3. Silja Baller, Soumitra Dutta, and Bruno Lanvin The global Information Technology Report 2016 [Электронный ресурс] URL: http://www3.weforum.org/docs/GITR2016/WEF_GITR_Full_Report.pdf (дата обращения: 16.06.2019).

4. The Economist intelligence Unit, Industry Analysis [Электронный ресурс]. URL: http://www.eiu.com/home.aspx (дата обращения: 16.06.2019).

5. Федеральная служба государственной статистики. [Электронный ресурс]. URL: http://www.gks.ru/ (дата обращения: 16.06.2019).

6. Иванова Е.И., Фаттахов Р.В., Сметанина О.Н. О роли информационных ресурсов при поддержке принятия управленческих решений на региональном уровне // Вестник УГАТУ. 2007. Т. 9. № 2. С. 82–87.

7. Сметанина О.Н., Ионис А.Г., Максименко З.В. Модель развития отрасли информационно-коммуникационных технологий // Информационные технологии и системы: труды Пятой Международной научной конференции (Банное, 24–28 февраля 2016 г.). Челябинск: Издательство Челябинский государственный университет, 2016. С. 230–236.

8. StatSoft, Электронный учебник по статистике // [Электронный ресурс]. URL: http://statsoft.ru/home/textbook/modules/stfacan.html (дата обращения: 16.06.2019).

9. Прогноз долгосрочного социально-экономического развития Российской Федерации на период до 2036 года // Министерство экономического развития Российской Федерации [Электронный ресурс]. URL: http://economy.gov.ru/minec/about/structure/depMacro/201828113 (дата обращения: 16.06.2019).