Алгоритмы нечеткой кластеризации для интеллектуального анализа данных при экономической диагностике ИТ-проектов

Е. В. Чертина[1], В. Ф. Шуршев[2], А. Е Квятковская[3]
ФГБОУ ВО «Астраханский государственный технический университет»
[1]saprikinae_1912@mail.ru, [2]v.shurshev@mail.ru, [3]zima00@list.ru

***Аннотация*. Рассмотрены возможности распределения инновационных ИТ- проектов с использованием алгоритмов нечеткой кластеризации. Проведена сравнительная характеристика двух базовых алгоритмов – алгоритма FCM и Густафсона – Кесселя. Представлена процедура кластеризации по каждому алгоритму, а также графически изображены результаты работы каждого алгоритма. Проведена оценка качества кластеризации с использованием коэффициента распределения, энтропии классификации и показателя Хие – Бени. Сделан вывод, что использование алгоритма Густафсона-Кесселя позволяет достичь более качественных результатов для решения задачи разбиения ИТ – проектов для цели их экономической диагностики.**

Ключевые слова: ИТ-проект; экономическая диагностика; нечеткая кластеризация; алгоритм Густафсона–Кесселя; алгоритм FCM

I. Введение

В условиях динамичного развития цифровой экономики особое место занимает информатизация процессов. Мы можем наблюдать стремительный рост стартапов, которые предлагают современные прикладные ИТ-решения, ускоряющие экономические, технологические, сервисные и другие процессы как хозяйствующих субъектов, так и человека. Высокая концентрация стратапов в ИТ индустрии привела к развитию системы венчурных фондов, большая часть инвестиций которых распределяется в ИТ-проекты. Это обусловлено тем, что для осуществления научно-исследовательских и опытно-конструкторских работ в ИТ-проекте сейчас используется технологии быстрого результата, позволяющие значительно сократить период вывода конечного продукта на стадию коммерциализации. Все это делает ИТ-сегмент наиболее привлекательным как с точки зрения разработчиков, так и финансовых вложений. Однако, на процессе разработки и реализации нового ИТ-проекта могут воздействовать различные внешние и внутренние факторы, порождающие неопределенность конечного результата и успешности его коммерческой реализации. Для венчурного инвестора важным аспектом является степень инвестиционного риска, на который он мог бы пойти. В связи с этим, перед венчурными фондами стоит задача тщательной экономической диагностики ИТ-проектов в части определения уровня их инвестиционной привлекательности и риска невозврата инвестиций для принятия финансовых решений.

Практика и обзор работ [1, 2] показывает, что наиболее часто используемыми показателями при экономической диагностике инвестиционных показателей различных проектов являются чистый дисконтированный доход (NPV), индекс рентабельности (PI), внутренняя норма доходности (IRR), срок окупаемости (PP). Использование подобных показателей для экономической диагностики ИТ-стартапа затруднительно, так как для принятия решений по инвестированию необходимо учитывать не только финансовую составляющую проекта, а также риски, актуальность, маркетинг и т.д. Это означает, что ИТ – проект необходимо оценивать по определенным группам критериев. Оценка проектов по критериям осуществляется экспертным путем при условии согласованности экспертных мнений. Экспертные мнения имеют лингвистические описания типа «высокий», «средний», «низкий», которые выражаются в количественной мере на шкале 0 до 1. Полученные агрегированные экспертные мнения могут быть использованы как признаки классификации множества рассматриваемых ИТ-проектов. Таким образом, выборочная совокупность ИТ-проектов может быть разбита на группы проектов с определенным набором схожих характеристик, позволяющих судить об инвестиционной привлекательности того или иного ИТ-проекта. Такая процедура может быть осуществлена с использованием методов кластерного анализа.

II. Постановка задачи исследования

Для целей настоящего исследования определим следующее. Пусть имеется множество ИТ-проектов $P = \{p_1,...,p_n\}$, с оценочными характеристиками, $L_1 - L_6$, (L_1 *–новизна проекта,* L_2 *– характеристика риска,* L_3 *– характеристика создаваемой научно-технической продукции,* L_4 *– рыночный потенциал и маркетинг,* L_5 *– оценка реализуемости проекта,* L_6 *– экономическая эффективность)*. Оценка проектов производится

Работа выполнена при финансовой поддержке РФФИ, проект № 18-37-00130

экспертной группой в дискретные моменты времени $t_1,...,t_l$. Постановку задачи исследования представим в виде следующих этапов.

1. Необходимо разбить имеющееся множество ИТ-проектов P, каждый из которых обладает характеристиками $\{L_1^j,...,L_6^j\}$, на три непересекающихся кластера (группы по степени инвестиционной привлекательности (ИП)) $K=\{K_1,...,K_3\}$ (K_1 – ИТ-проекты, с высокой степенью ИП; K_2 – ИТ-проекты, со средней степенью ИП, рекомендованные к доработке; K_3 – ИТ-проекты, с низкой степенью ИП, рекомендованные к отказу в финансировании).

2. Выбрать наиболее подходящий алгоритм кластеризации (1), путем проведения оценки качества кластеризации:

$$\forall P, L, K \exists \Lambda_C : P \to K. \quad (1)$$

Следует отметить, что нечетко-множественный характер экспертных суждений при осуществлении процедуры экспертного оценивания порождает неопределенность, которая в дальнейшем будет влиять на структуру кластера. Кроме того, отнести *j*-й ИТ-проект только к одному из кластеров $\{K_1,...,K_3\}$ будет крайне сложно. Для устранения этой проблемы предлагается использование метода нечеткой кластеризации [3], отличающийся в определении степени принадлежности проекта p_j к каждому кластеру и основанном на основанной на теории нечетких множеств Л.Заде [4].

III. Анализ алгоритмов нечеткой кластеризации

Проведя анализ алгоритмов нечеткой кластеризации в исследованиях [5, 6], мы пришли к выводу, что представленные алгоритмы можно условно разделить на две основные группы. Первая группа – алгоритмы формирующие кластеры сферической формы. Вторая группа – алгоритмы, формирующие кластеры в виде гиперэлипсоидов различной ориентации. В качестве базовых алгоритмов этих групп выберем алгоритм нечетких с-средних (FCM) и алгоритм Густафсона-Кесселя соответственно. Все остальные алгоритмы нечеткой кластеризации являются их производными [7].

Обозначим выше перечисленные группы ИТ-проектов через нечеткие кластеры $\{\tilde{K}_1,...\tilde{K}_3\}$. Тогда, нечеткое разбиение ИТ-проектов по кластерам опишем матрицей следующего вида (2) [8]:

$$F=\left[\mu_{k,i}\right], \quad (2)$$

где $\mu_{k,i}\in[0;1], k=\overline{1,n}$ – функция принадлежности k-го ИТ-проекта с набором характеристик $(L_1^k,...,L_6^k)$ к кластерам $\tilde{K}_1,...\tilde{K}_3, c=\overline{1,3}$. Отсюда можно сделать вывод, что каждый ИТ-проект, имеющий различные степени принадлежности может быть отнесен к каждому из трех кластеров. При этом, необходимо выполнение следующих условий (3):

$$\begin{cases}\sum_{i=1}^{l}\mu_{k,i}=1, k=\overline{1,n}\\ 0<\sum_{k=1}^{n}\mu_{k,i}<n, i=\overline{1,l}\end{cases}. \quad (3)$$

Теперь покажем основные отличительные характеристики рассматриваемых алгоритмов.

В методе FCM минимизация функционала имеет вид (4) [9]:

$$\Im=\sum_{i=1}^{l}\sum_{k=1}^{n}\left(\mu_{k,i}\right)^m\left\|p_k-v_i\right\|_A^2 \quad (4)$$

где $V=[v_1,...,v_l], v_i\in R^n$ - вектор центров кластеров, а $D_{ikA}^2=\left\|p_k-v_i\right\|_A^2=\left(p_k-v_i\right)^T A\left(p_k-v_i\right)$.

Величины, входящие в (4) можно определить из выражений (5, 6):

$$\mu_{k_i}=\frac{1}{\sum_{j=1}^{l}\left(D_{ikA}/D_{jkA}\right)^{2/(m-1)}}; \quad (5)$$

$$v_i=\frac{\sum_{k=1}^{n}\mu_{k,i}^m p_k}{\sum_{k=1}^{n}\mu_{k,i}^m}, \quad (6)$$

где m – экспоненциальный вес.

Условие останова этого алгоритма нечеткой кластеризации $\left\|F-F^*\right\|<\varepsilon$, где ε задается лицом, принимающим решение.

В отличии от алгоритма нечетких с – средних алгоритм Густафсона-Кесселя обладает собственной матрицей A_i, т.е. в соответствие с [10] имеем выражение (7):

$$D_{ikA}^2=\left\|p_k-v_i\right\|_{A_i}^2=\left(p_k-v_i\right)^T A_i\left(p_k-v_i\right). \quad (7)$$

Тогда функционал $\Im$ будет иметь вид (8):

$$\Im = \sum_{i=1}^{l}\sum_{k=1}^{n}\left(\mu_{k,i}\right)^{m}\left(p_k - v_i\right)^{T} A_i \left(p_k - v_i\right) . \quad (8)$$

Функционал в форме (8) не может быть минимизирован по A_i, т.к. он линеен по A_i. Поэтому, чтобы получить приемлемое решение, необходимо, чтобы $\|A_i\| < \rho_i, \rho > 0$, т.е. следует ограничить определители матриц A_i. Тогда нечеткая ковариационная матрица для *i*-го кластера будет определяться следующим образом (9):

$$F_i = \frac{\sum_{k=1}^{n}\left(\mu_{k,i}\right)^{m}\left(p_k - v_i\right)\left(p_k - v_i\right)^{T}}{\sum_{k=1}^{n}\left(\mu_{k,i}\right)^{m}} . \quad (9)$$

Для следующего этапа исследования были оценены 50 ИТ-проектов, полученные экспертные оценки согласованы, аффилированность между экспертами отсутствует. Исходные данные для реализации алгоритмов следующие: $m = 2$, $c = 3, \varepsilon = 1 \cdot e^{-6}$, матрица P представляет собой агрегированные экспертные оценки по рассмотренным выше критериям $\left\{L_1^j, ..., L_6^j\right\}$.

IV. Реализация алгоритмов нечеткой кластеризации

A. Алгоритм FCM

Наглядно последовательность этапов реализации алгоритма FCM представим в виде следующей блок – схемы на рис. 1.

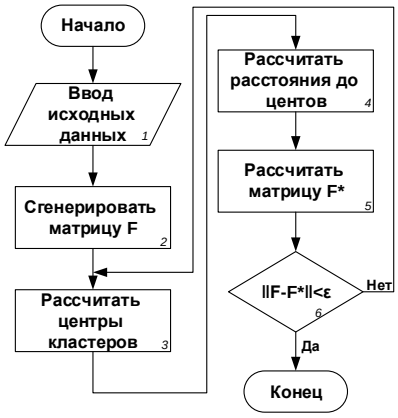

Рис. 1. Блок-схема алгоритма кластеризации ИТ-проектов (FCM)

На рис. 2 представлена визуализация полученных результатов, с применением метода главных компонент (Principal Component Analysis – PCA, реализованный в SOMToolbox среды инженерных расчетов Matlab) [11].

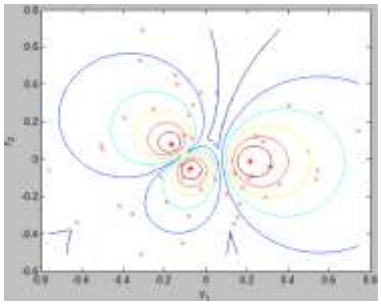

Рис. 2. Кластеризация алгоритмом FCM с использованием PCA

B. Алгоритм Густафсона–Кесселя

После этого реализован алгоритм Густафсона–Кесселя, блок-схема которого представлена на рис. 3.

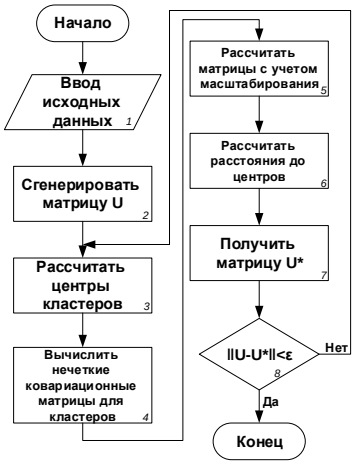

Рис. 3. Блок-схема алгоритма кластеризации ИТ-проектов (алгоритм Густафсона–Кесселя)

При реализации алгоритма Густафсона–Кесселя осуществлена 141 итерация до момента, как «сработала» точка останова алгоритма.

На рис. 4 отображены результаты кластеризации методом Густафсона–Кесселя с применением PCA.

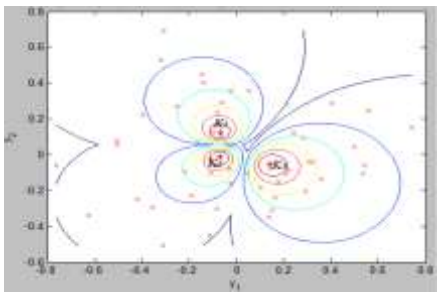

Рис. 4. Кластеризация алгоритмом Густафсона–Кесселя, с использованием PCA

C. Оценка качества кластеризации

В работах [12] предлагается использование следующих показателей оценки качества кластеризации.

1. Коэффициент распределения (10):

$$R_1 = \frac{1}{n}\sum_{i=1}^{l}\sum_{k=1}^{n}\left(\mu_{k,i}\right)^2 \quad . \qquad (10)$$

Он используется в качестве меры нечеткости (чем он выше, тем лучше с точки зрения оценка нечеткости и косвенно кластеризации), однако он не учитывает попарные расстояния, необходимые для оценки компактности и разделения. Поэтому предложен другой показатель.

2. Энтропия классификации (11):

$$R_2 = \frac{1}{n}\sum_{i=1}^{l}\sum_{k=1}^{n}\mu_{k,i}\log\left(\mu_{k,i}\right) \quad . \qquad (11)$$

Этот показатель изменяется в пределах $0 \le R_2 \le \ln l$. Основная цель применения показателей R_1 и R_2 – поиск приемлемого числа кластеров в нечетком разбиении. Оба показателя зависят от числа кластеров (l), поэтому подходят для сравнения разбиений только с одинаковым числом кластеров.

3. Коэффициент Хие-Бени (12):

$$R_3 = \frac{\sum_{i=1}^{l}\sum_{k=1}^{n}\left(\mu_{k,i}\right)^m \left\|p_j - v_i\right\|^2}{n\min_{i,j}\left\|p_j - v_i\right\|^2} \quad . \qquad (12)$$

Коэффициент является наиболее подходящим для оценки компактности и хорошей «разделяемости» кластеров в нечетком разбиении, а также позволяет оценить адекватность полученных результатов. В таблице приведены результаты оценки качества кластеризации по двум алгоритмами с использованием рассмотренных показателей.

ТАБЛИЦА I РЕЗУЛЬТАТЫ ОЦЕНКИ КАЧЕСТВА КЛАСТЕРИЗАЦИИ

Показатель	Алгоритм FCM	Алгоритм Густафсона–Кесселя
R_1	0,405	0,623
R_2	1,022	0,751
R_3	1,038	1,183

Из таблицы видно, что FCM обладает меньшим значением R_1, большим значение энтропии и его коэффициент Хие-Бени R_3 превышает аналогичный показатель алгоритма Густафсона–Кесселя. Таким образом, для решение поставленной в исследовании задачи наиболее предпочтительным является алгоритм нечеткой кластеризации Густафсона–Кесселя. Кроме того, достоинством алгоритма Густафсона–Кесселя является то, что он формирует адаптивную форму для каждого кластера, что позволяет упорядочивать объекты по кластерам более корректно.

V. ЗАКЛЮЧЕНИЕ

Проведенное исследование позволило достичь следующих результатов:

- обоснована необходимость использования нечеткой кластеризации для решения задачи экономической диагностики ИТ-проектов в части определения уровня их инвестиционной привлекательности;
- проведен анализ двух базовых алгоритмов нечеткой кластеризации Густафсона–Кесселя и FCM, а также рассмотрены особенности их функционала;
- осуществлена практическая реализация рассматриваемых алгоритмов для 50 ИТ-проектов, имеющих агрегированные экспертные оценки;
- проведена оценка качества кластеризации и сделан вывод о предпочтительности использования алгоритма Густафсона–Кесселя.

Предложенный подход позволит формализовать неопределенность и риск при экономической диагностике ИТ-проектов, а также повысить эффективность принимаемых финансовых решений венчурными фондами, а также другими инвестиционными компаниями.

СПИСОК ЛИТЕРАТУРЫ

[1] Куликов Д.Л., Кучеров А.А. Становление и развитие методов оценки эффективности инновационных проектов [Электронный ресурс] // Современные проблемы науки и образования. 2015. № 1. Режим доступа: https://www.science-education.ru/ru/article/view?id=19451 (дата обращения 14.05.2018)

[2] Малова О.Т. Подходы к оценке инновационных проектов // Журнал Educatio. 2015. № 3 (10). С.140-142

[3] Bezdek J.C., Ehrlich R., Full W. FCM: The Fuzzy c-Means Clustering Algorithm // Computers & Geoscience. 1984, Vol. 10. № 2-3, pp. 191-203.

[4] Заде Л.А. Понятие лингвистической переменной и его применение к принятию приближенных решений. М.: Мир, 1976. 165 с.

[5] Нейский И.М. Классификация и сравнение методов кластеризации [Электронный ресурс]. Режим доступа: http://it-claim.ru/Persons/Neyskiy/Article2_Neiskiy.pdf (дата обращения 14.05.2018)

[6] Jain A.K., Murty M.N., Flynn P.J. Data Clustering: A Review // ACM Computing Surveys. 1999, Vol. 31, no. 3, pp. 264–323.

[7] Rozilawati Binti Dollah, Aryati Binti Bakri, Mahadi Bin Bahari, Pm Dr. Naomie Binti Salim, Feasibility Study Of Fuzzy Clustering Techniques In Chemical Database For Compound Classification, 2006.

[8] Штовба С.Д. Проектирование нечетких систем средствами MATLAB. М.: Горячая линия – Телеком, 2007. 288 с.

[9] Bezdek J.C., Dunn J.C. Optimal Fuzzy Partitions: A Heuristic for Estimating the Parameters in a Mixture of Normal Dustrubutions // IEEE Transactions on Computers. 1985, pp. 835-838.

[10] Gustafson D.E., Kessel W.C. Fuzzy clustering with fuzzy covariance matrix // In Proceedings of the IEEE CDC, San Diego. 1979, pp. 761-766.

[11] Jolliffe I.T. Principal Component Analysis. Springer Series in Statistics, 2nd ed., Springer, NY, 2002,XXIX, 487 p.

[12] Xie X.L., Beni G.A. Validity measure for fuzzy clustering // In Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence. 1991, Vol. 3(8), pp. 841-846.