DOI 10.37882/2223-2966.2025.04.37

РАЗРАБОТКА МОДЕЛИ ДАННЫХ ЦИФРОВОГО СЛЕДА ПРОЕКТОВ

DEVELOPMENT OF A DATA MODEL FOR THE DIGITAL FOOTPRINT OF PROJECTS

M. Stepanov

Summary. Cluster analysis (CA) can be used in the process of studying Digital Traces (DT) of students of an educational institution, as well as other educational institutions that implement the Digital Educational Environment (DEE) in the educational process. The use of CA methods will improve the personalization of training and increase the effectiveness of educational programs. In the context of ensuring Information Security (IS) of the DEE of educational institutions, the technologies and methods of CA analysis can be useful, for example, for: monitoring students' network activity; analyzing student authorization and authentication logs; detecting malware and attacks on the DEE; analyzing information security threats to the DEE as a whole; predicting vulnerabilities. The paper presents the results of experimental studies of the level of competence of students of different specialties at the university in information security and protection of information assets of the DEE. CA methods were used in the process of studying the DT of students. Based on the DT of different groups of students registered in the university's DEE, six types of users were identified. As a result of applying the CA methods, the students registered in the university's DEE were divided into corresponding clusters according to the criteria affecting the information security risks.

Keywords: digital traces, cluster analysis, digital educational environment of the educational institution, information secure.

Степанов Михаил Александрович
аспирант, Томский политехнический университет
levacat77@gmail.com

Аннотация. Кластерный анализ (КА) может использоваться в процессе исследования Цифровых Следов (ЦС) студентов учебного заведения, а также других учебных заведений, которые внедряют в учебный процесс Цифровую Образовательную Среду (ЦОС). Применение методов КА позволит улучшить персонализацию обучения и повысить эффективность образовательных программ. В контексте обеспечения Информационной Безопасности (ИБ) ЦОС учебных заведений технологии и методы анализа ЦС могут быть полезны, например, для: мониторинга сетевой активности студентов; анализа журналов авторизации и аутентификации студентов; выявления вредоносных программ и атак на ЦОС; анализа угроз ИБ ЦОС в целом; прогнозирования уязвимостей. В работе представлены результаты экспериментальных исследований уровня компетентностей студентов разных специальностей в университете по ИБ и защите информационных активов ЦОС. Были использованы способы КА в процессе исследования ЦС студентов. На основе КА ЦС разных групп студентов, зарегистрированных в ЦОС университета, было выделено шесть типов пользователей. В результате применения методов КА студенты, зарегистрированные в ЦОС университета, были разбиты на соответствующие кластеры по критериям, влияющим на риски ИБ.

Ключевые слова: цифровые следы, кластерный анализ, цифровая образовательная среда учебного заведения, информационная сохранность.

Введение

В последнее время большинство учебных заведений во всем мире, особенно на фоне стремительного распространения пандемии Covid-19, кроме традиционных форм обучения, предлагают студентам различные формы онлайн обучения. Такой формат обучения, например, реализуют через Системы Дистанционного Обучения (СДО) или онлайн курсы по определенным дисциплинам. Однако для эффективного функционирования подобных онлайн курсов и СДО в целом, необходима соответствующая цифровая среда или Информационная Система (ИС), которая будет выполнять функции управления и организации учебного процесса. В большинстве случаев подобные задачи решаются с помощью систем управления обучением, а именно LMS, предлагающих множество функций для поддержки преподавателей в процессе создания, администрирования и управления онлайн курсами.

Однако такие системы в большинстве случаев имеют достаточно ограниченный арсенал модулей для анализа накапливаемых данных в ходе организации учебного процесса [1], [2]. В частности, в таких системах недостаточно реализован потенциал для анализа о Цифровых Следах (ЦС) студентов во время обучения [3–5]. Подобные LMS системы практически не используют накопленные данные по ЦС студентов для повышения степени защищенности и информационной безопасности университета от внешних вмешательств.

Заметим, что и в контексте обеспечения Информационной Безопасности (ИБ) ЦОС учебного заведения технологии и методы Интеллектуального Анализа Данных (ИАД) и ЦС также могут быть полезны, например, для: мониторинга сетевой активности студентов и преподавателей; анализа журналов авторизации и аутентификации студентов и преподавателей; выявления вредоносных программ и атак на ЦОС учебного заведения;

анализа угроз ИБ ЦОС учебного заведения в целом; прогнозирования уязвимостей.

Таким образом, все вышеописанное и определяет актуальность данного исследования, а также мотивирует нас к продолжению работы в данном направлении.

Анализ последних исследований и публикаций. В [4] авторами показано, что исследование процессов обучения может выиграть от изучения ЦС, оставленных студентами, в частности, при просмотре учебного контента образовательных платформ. В [5] авторы представляют результаты исследований по интеграции стилей обучения на основании анализа ЦС в образовательные вебсистемы на базе LMS Claroline, Ganesha, Chamilo, Moodle. Такая интеграция результатов анализа ЦС поможет студентам в усвоении учебного контента курса. Однако авторы ограничились лишь сравнительным анализом LMS систем, не детализируя методики применения конкретного метода анализа ЦС студентов в системе LMS.

В [6] авторы, использовав методы кластерного анализа данных, полученных в том числе цифровыми следами студентов, попытались решить задачу, связанную с путями повышения эффективности образовательного процесса. Однако работа не рассматривает многие технологические аспекты анализа ЦС студентов.

Используя методы корреляционного и кластерного анализа, авторы исследования в [7] проанализировали стратегии образовательной деятельности студентов в различных социальных сетях.

В работах [8–9] авторы подробно исследовали с помощью различных методов ИАД зависимости между успехами студентов и их активностью в онлайн-курсах на базе LMS Moodle. Такие исследования позволили выявить поведенческие стратегии студентов во время онлайн-обучения.

Динамика роста инцидентов ИБ в учебных заведениях, активно внедряющих в образовательный процесс информационные системы и технологии, свидетельствует об эволюционировании такого рода угроз.

В исследованиях [8,9] авторами рассматриваются различные аспекты проблематики применения методов ИАД для обеспечения ИБ ЦОС учебного заведения.

В работах [9,10] авторы исследуют проблематику связи — «Цифровой след» — «Интеллектуальный анализ данных» — «Информационная безопасность».

Как показал анализ предыдущих исследований, Кластерный Анализ (КА) может быть полезным инструментом для разделения студентов, зарегистрированных в ЦОС учебного заведения и, в частности, в СДО, по уровню их технических знаний об ИБ, рисках и мерах безопасной работы в ЦОС учебного заведения. Методы кластерного анализа позволят выделить группы студентов на основе подобных характеристик и потребностей, которые также могут быть использованы для формирования индивидуальных образовательных программ.

Целью работы является выявление группы в ЦОС учебного заведения с наименьшими контекстно-зависимыми характеристиками по вопросам соблюдения правил ИБ при работе в данной среде и уровней рисков ИБ при работе разных категорий пользователей в ЦОС учебного заведения.

В этой связи необходимо решить следующие задачи:
1. идентифицировать массив подобных групп по определенным критериям;
2. провести анализ ЦС студентов, отнесенных к этой группе и провести кластерный анализ на основе метода k-means.

Методика исследования

Для выявления групп студентов и преподавателей в ЦОС учебного заведения с наименьшими знаниями и опытом в вопросах ИБ используется анализ ЦС и КА. Были последовательно реализованы следующие этапы:

Этап 1. Сбор данных. Были собраны доступные данные о пользователях, зарегистрированных в ЦОС университета с использованием соответствующих логов LMS Moodle и ОС. Эти данные описывали, например, прохождение соответствующих курсов по ИБ, степень и успеваемость (для студентов) выполнения заданий по ИБ, результаты тестирований и т.п. по соответствующим курсам, связанным с ИБ. Также эти данные получены на основании: оценки результатов тестирования студентов по темам ИБ; мониторинга активности пользователей в сети; оценки мер защиты в ЦОС университета (например, насколько активно студенты используют такие меры защиты, как стойкие пароли, двухфакторная аутентификация и шифрование данных, что свидетельствует об информированности пользователей относительно угроз ИБ).

Этап 2. Подготовка данных. Собранные данные были отфильтрованы для удаления неполных записей. А затем преобразованы в числовой формат.

Этап 3. Анализ цифровых следов: ЦС пользователей ЦОС университета могут указывать на риск ИБ. К таким ЦС можно отнести: поведенческие данные (необычные попытки доступа, например, к СДО, частые неудачные попытки входа в аккаунт, неожиданные изменения местоположения или устройств для доступа к учебным

материалам; аномальная активность (необычно интенсивное скачивание или копирование исходящих материалов) за рамки привычного использования), отклонения от привычного расписания (неожиданные или необычные периоды активности, входы в систему в необычное время для конкретного студента), изменения в обычных паттернах поведения;

Используя методы анализа ЦС, также была изъята дополнительная информация о познании, а также опыте студентов и преподавателей по ИБ, например, это может быть анализ активности в ЦОС университета, стиля поведения, участия в дискуссионных форумах и т.д.

Этап 4. Реализация КА. Применение методов КА для группирования студентов на основе их знаний и опыта по ИБ.

Этап 5. Интерпретация результатов.

Заметим, что результаты анализа ЦС и КА могут быть приближены и требуют дополнительной интерпретации, а также проверки. Однако эти методы могут помочь руководителям разных уровней учебного заведения начать процесс выявления групп студентов, имеющих наименьший уровень знаний и опыта по вопросам ИБ с целью повышения степени защищенности ЦОС университета в целом.

Отметим через Ω все множество пользователей, зарегистрированных в ЦОС университета. Для классификации и кластеризации нужно сформировать однородные по признаку кластеры пользователей по уровню рисков для ИБ ЦОС университета. Для этого воспользуемся методами кластерного анализа — иерархической классификацией и методом k-means.

В самом простом варианте в качестве метрики можно использовать эвклидовое расстояние.

$$\rho_{ik} = \sqrt{\sum_{i=1}^{m} \left(z_{ji} - z_{jk} \right)^2}, \qquad (1)$$

где ρ_{ik} — расстояние между, соответственно, i, k наблюдениями, которое сформировано на основе стандартизированных данных, упомянутых выше;

z_{ji} — матрица стандартизированных данных.

Результаты

Выполним кластеризацию, используя максимально приближенные к действительности данные группы пользователей, зарегистрированных в ЦОС университета.

Пусть множество Ω шесть групп, приведенных в табл. 1. Данные получены на основе анализа ЦС пользователей, а именно студентов разных специальностей Алтайского филиала РАНХиГС.

Таблица 1.

Классификация пользователей, зарегистрированных в ЦОС университета по критерию рисков для ИБы

№	Название группы (типы пользователей)	Признаки (Описание модели поведения)
1	Знающие пользователи	К этой группе можно отнести пользователей, хорошо знающих риски в сети учебного заведения (или ЦОС учебного заведения в целом). Эти пользователи принимают меры по обеспечению ИБ своих данных и учетных записей в ЦОС учебного заведения. Они всегда придерживаются рекомендаций по созданию сложных паролей, регулярно обновляют программное обеспечение, не открывают подозрительные ссылки или вложения в электронных письмах, а также используют надежное антивирусное программное обеспечение (ПО) на своих устройствах.
2	Небрежные пользователи	В группу отнесены пользователи, которые не обращают должного внимания на мероприятия ИБ. Такие пользователи уязвимы для атак. Пользователей этой группы характеризуют использование слабых паролей, повторение паролей для различных аккаунтов, не своевременное обновление ПО, игнорирование подозрительной активности и мер по защите своих данных в ЦОС учебного заведения.
3	Незнающие пользователи	Группа содержит пользователей, не имеющих достаточного уровня знаний об ИБ, риски при работе в сети учебного заведения. Эта группа может использовать ненадежное ПО на своих устройствах при подключении к ЦОС учебного заведения, а также могут передавать конфиденциальные данные через незащищенные каналы связи
4	Равнодушные пользователи	В эту группу входят пользователи, которые не проявляют интереса к вопросам ИБ ЦОС учебного заведения. Эта категория пользователей не проверяет свои аккаунты по поводу взлома, не обращает внимание на предупреждения о возможных угрозах и т.д.
5	Безответственные пользователи	Участники образовательного процесса нарушают правила и политику ИБ в сети учебного заведения. Они могут попытаться получить несанкционированный доступ к системам ЦОС учебного заведения, распространять вредоносное ПО, нарушать конфиденциальность данных или вести недобросовестную активность в ЦОС учебного заведения.
6	Деструктивные пользователи	Эти пользователи пытаются нанести вред сети учебного заведения, в том числе распространяя вирусы, блокируя ресурсы сети и т.п.

Эти группы пользователей достаточно условны. Часто нет четких границ между группами пользователей. Также могут быть отличия между группами. По мере приобретения знаний, например, за счет соответствующих курсов пользователи могут переходить от одного типа к другому, осознавая важность ИБ в сети и принимая соответствующие меры для защиты своих данных и учетных записей.

Любую группу пользователей можно охарактеризовать, используя следующие признаки, пример приведенные на рис. 1:

1) Название специальности и ID группы пользователей (NS_IDGroup — доступно, например, с LMS Claroline, Ganesha, Chamilo, Moodle);
2) Курс/Год обучения 1–4/5 (Year_study — например, 1 (1-й курс) и т.д.) (берется, например, из LMS Claroline, Ganesha, Chamilo, Moodle);
3) Личная ответственность (Pers_resp (1–100)) — характеризуется наличием специфических признаков, например, степень аккуратности в обработке личных данных, частота смены паролей, базовые знания об ИБ, активность использования защищенных соединений при работе в ЦОС заведения и т.д. эти данные частично можно получить на основе анализа ЦС пользователей посредством LMS Claroline, Ganesha, Chamilo, Moodle, а также используя SIEM);
4) Средняя успеваемость по курсам, связанным с ИТ и ИБ (Aver_perf (0–100) — доступна, например, с LMS Claroline, Ganesha, Chamilo, Moodle));
5) Оценка компетентностей с ИБ (Comp_ass (0-10), например, наличие знаний о: фишинге, вредоносном ПО, использовании слабых/сильных паролей, обновлении ПО, включая операционную систему, безопасности сети, в том числе использование защищенного Wi-Fi; настройка брандмауэра и использование VPN; безопасное хранение данных и т.д.

Обсуждение результатов исследования

Цель КА — разбить пользователей ЦОС учебного заведения на классы. Причем каждый из таких классов отвечает своей группе риска в контексте ИБ. Попавшие в одну группу наблюдения характеризуются одинаковой вероятностью инцидента ИБ.

Исследование производилось с использованием пакета STATISTICA 12.5.

NS_IDGroup	Year_ study	Pers_ resp	Aver_ perf	Comp_ ass
Software Engineering_1501	1	21	68	4
Software Engineering_1501	2	40	74	5
Software Engineering_1501	3	56	77	6
Software Engineering_1501	4	65	84	6
Information Technology_1502	1	17	65	4
Information Technology_1502	2	37	77	5
Information Technology_1502	3	61	81	6
Information Technology_1502	4	71	96	6
Economic cybernetics_1503	1	21	71	4
Economic cybernetics_1503	2	44	76	5
Economic cybernetics_1503	3	60	80	6
Economic cybernetics_1503	4	71	74	7
Computer engineering_1504	1	22	69	4
Computer engineering_1504	2	42	75	5
Computer engineering_1504	3	61	74	7
Computer engineering_1504	4	75	78	9
Economy_1601	1	14	61	2
Economy_1601	2	29	73	3
Economy_1601	3	45	74	4
Economy_1601	4	50	75	4
Enterprise economy_1602	1	12	62	2
Enterprise economy_1602	2	25	73	3
Enterprise economy_1602	3	40	73	4
Enterprise economy_1602	4	49	75	4
Management_1701	1	11	65	2
Management_1701	2	27	72	3
Management_1701	3	39	73	3
Management_1701	4	42	75	4
Ecology_1604	1	12	66	2
Ecology_1604	2	26	73	3
Ecology_1604	3	34	72	3
Ecology_1604	4	41	80	4

NS_IDGroup	Year_ study	Pers_ resp	Aver_ perf	Comp_ ass
Software Engineering_1501	-1,32051	-0,97789	-0,87045	-0,20733
Software Engineering_1501	-0,44017	0,033262	0,032936	0,395806
Software Engineering_1501	0,44017	0,884758	0,484629	0,998939
Software Engineering_1501	1,320511	1,363725	1,538579	0,998939
Information Technology_1502	-1,32051	-1,19077	-1,32214	-0,20733
Information Technology_1502	-0,44017	-0,12639	0,484629	0,395806
Information Technology_1502	0,44017	1,150851	1,086886	0,998939
Information Technology_1502	1,320511	1,683037	3,345352	0,998939
Economic cybernetics_1503	-1,32051	-0,97789	-0,41876	-0,20733
Economic cybernetics_1503	-0,44017	0,246136	0,334065	0,395806
Economic cybernetics_1503	0,44017	1,097633	0,936322	0,998939
Economic cybernetics_1503	1,320511	1,683037	0,032936	1,602073
Computer engineering_1504	-1,32051	-0,92467	-0,71989	-0,20733
Computer engineering_1504	-0,44017	0,139699	0,1835	0,395806
Computer engineering_1504	0,44017	1,150851	0,032936	1,602073
Computer engineering_1504	1,320511	1,895911	0,635193	2,808339
Economy_1601	-1,32051	-1,35042	-1,9244	-1,41359
Economy_1601	-0,44017	-0,55214	-0,11763	-0,81046
Economy_1601	0,44017	0,299354	0,032936	-0,20733
Economy_1601	1,320511	0,565447	0,1835	-0,20733
Enterprise economy_1602	-1,32051	-1,45686	-1,77384	-1,41359
Enterprise economy_1602	-0,44017	-0,76502	-0,11763	-0,81046
Enterprise economy_1602	0,44017	0,033262	-0,11763	-0,20733
Enterprise economy_1602	1,320511	0,512229	0,1835	-0,20733
Management_1701	-1,32051	-1,51008	-1,32214	-1,41359
Management_1701	-0,44017	-0,65858	-0,26819	-0,81046
Management_1701	0,44017	-0,01996	-0,11763	-0,81046
Management_1701	1,320511	0,139699	0,1835	-0,20733
Ecology_1604	-1,32051	-1,45686	-1,17158	-1,41359
Ecology_1604	-0,44017	-0,7118	-0,11763	-0,81046
Ecology_1604	0,44017	-0,28605	-0,26819	-0,81046
Ecology_1604	1,320511	0,08648	0,936322	-0,20733

Рис. 1. Скриншоты таблицы в STATISTICA 12.5 с исходными данными для анализа и со стандартизированными переменными

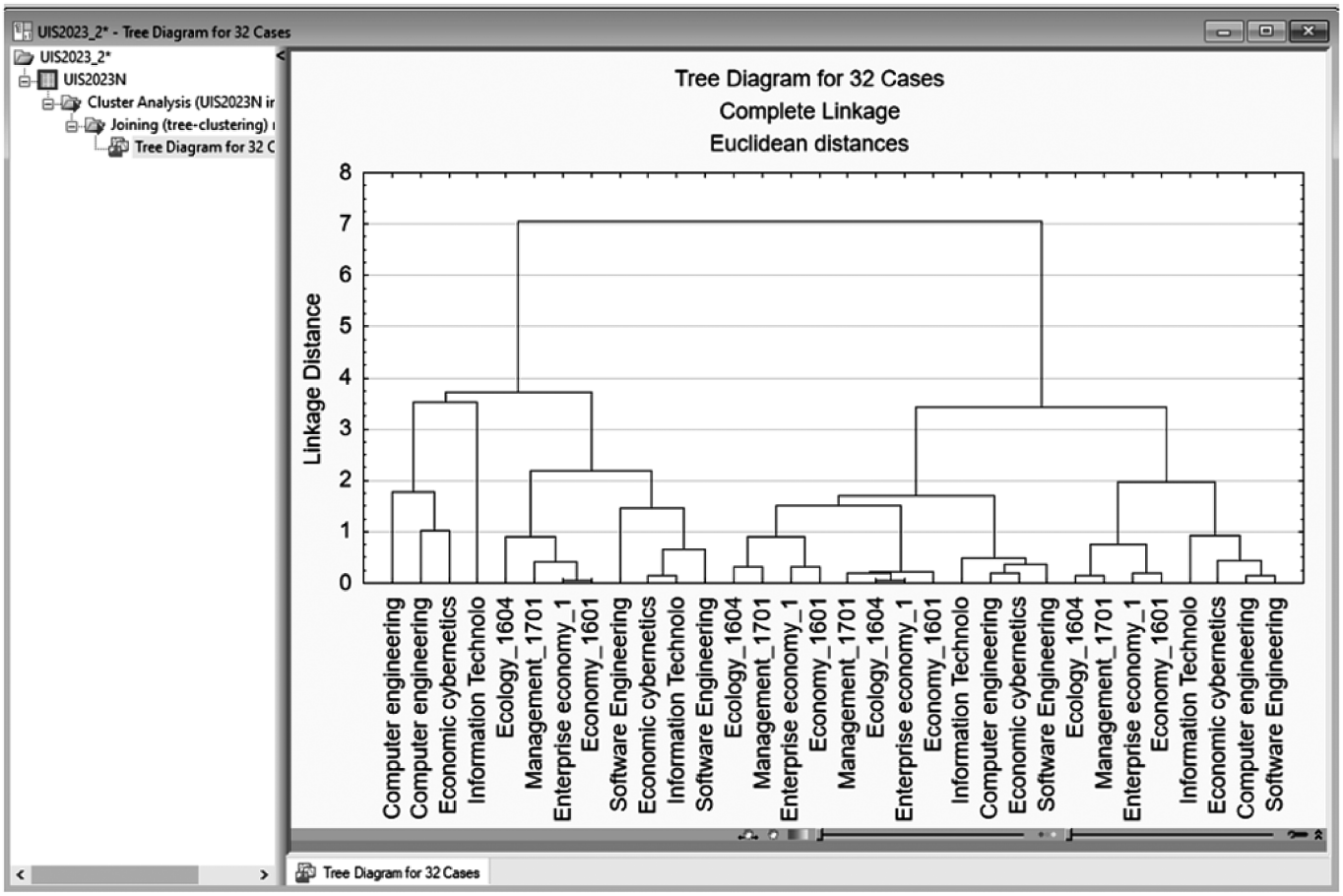

Рис. 2. Скриншот результатов иерархической классификации (пакет STATISTICA 12.5)

На первом этапе на основе иерархической классификации были получены дендрограммы, необходимые для группировки объектов в подмножества (кластеры), на основе их сходства (рис. 2).

Необходимо среди имеющегося множества групп выделить соответствующие кластеры. Например, пользователи, отнесенные к опасным с точки зрения критериев ИБ и рискам для ЦОС учебного заведения, безразличны и нейтральны по оцениваемым признакам.

Предположение. Данные обо всех вышеуказанных признаках доступны и могут быть оценены и измерены. К примеру, изменение пароля, IP и другие ЦС можно отслеживать в системах LMS, а также с помощью уже упомянутых SIEM систем.

В результате исследования на основе метода k-means были получены графики средних и доверительных интервалов для переменных в каждом кластере, характеризующих пользователей ЦОС учебного заведения в контексте соблюдения правил ИБ и влияния их стиля работы на риски для ИБ (рис. 3).

Сравнив метод k-means и иерархическую классификацию, следует отметить, что несомненным преимуществом первого метода является возможность работы с первичными данными. Это позволит, например, специалистам по ИБ учебных заведений, обрабатывать достаточно большие объемы данных, которые можно импортировать из LMS в формате *.csv.

В случае значительного информационного массива пользователей больших учебных заведений это несомненное преимущество. Более того, метод k-means может компенсировать последствия некачественного исходного разбиения исходного массива данных.

Заключение

В работе представлены результаты исследования уровня компетентностей пользователей разных специальностей Цифровой Образовательной Среды Университета (ЦОСУ) по вопросам, связанным с ИБ. Используются методы кластерного анализа и анализа ЦС пользователей ЦОС. На основе анализа ЦС разных групп зарегистрированных пользователей в ЦОС университета было исследовано поведение шести типов пользователей. Эти типы пользователей различных специальностей, рассматриваемых в тестовом наборе, в результате применения иерархической классификации и метода k-means были разбиты на соответствующие кластеры

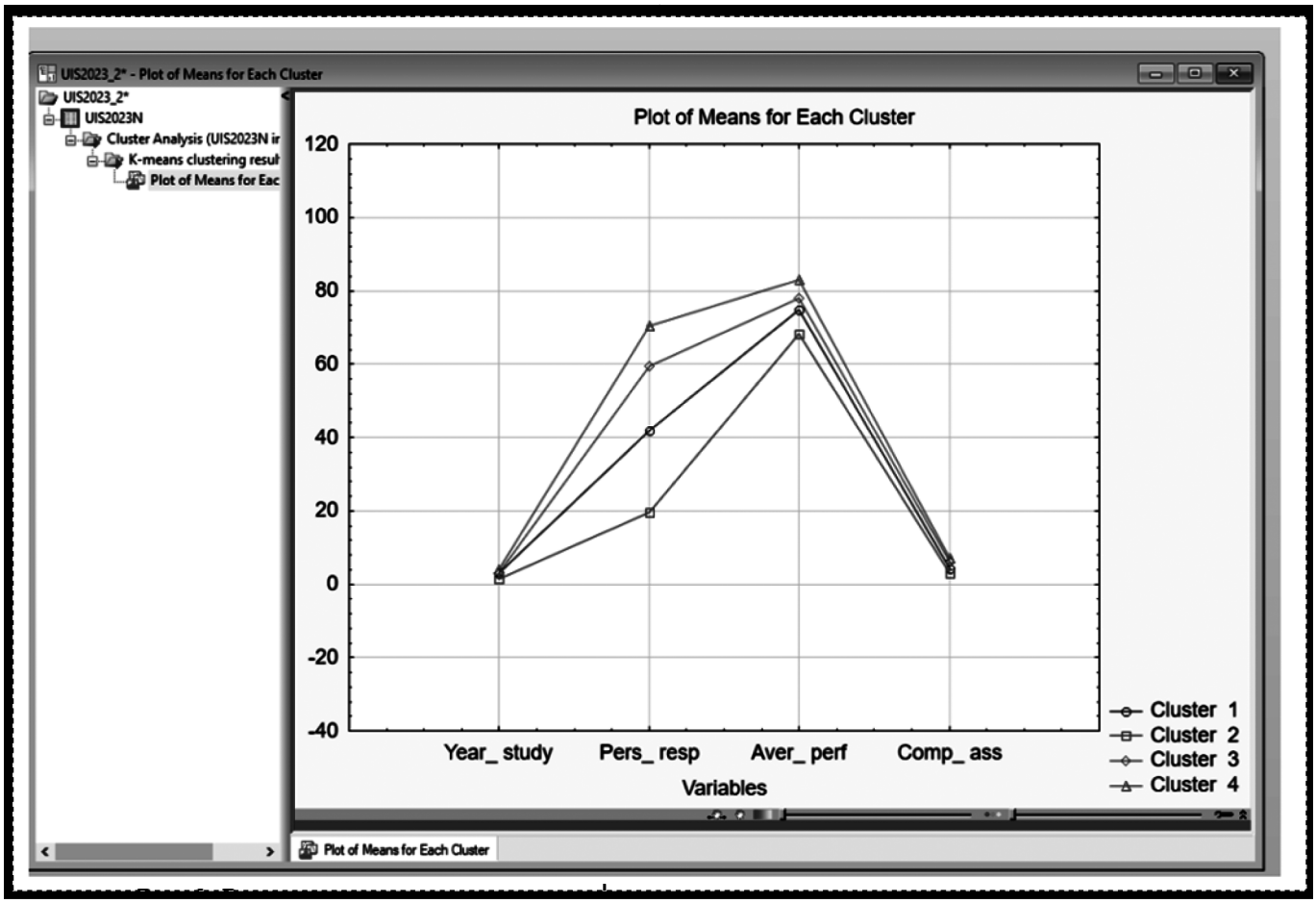

Рис. 3. Результаты кластерного анализа

по критериям, влияющим на риски ИБ ЦОС учебного заведения. Для каждого кластера экспертом из ИБ ЦОС университета может определяться вероятность наступления случаев, связанных с высоким уровнем риска ИБ, и, соответственно, могут приниматься меры по устранению причин таких случаев и разрабатываться рекомендации для пользователей.

ЛИТЕРАТУРА

1. Oliveira P., Cunha C., Nakayama M. Learning Management Systems (LMS) and E-learning Management: An Integrative Review and Research Agenda // JISTEM-J. Inf. Syst. Technol. Manag. 2016. V.13. P. 157–180.
2. Aldiab A., et al. Utilization of Learning Management Systems (LMSs) in Higher Education System: A Case Review for Saudi Arabia // Energy Procedia. 2019. V. 160. P. 731–737. https://doi.org/10.1016/j.egypro.2019.02.186
3. Azcona D., Hsiao I., Smeaton A. Detecting Students-at-Risk in Computer Programming Classes with Learning Analytics from Students' Digital Footprints. User Modeling and User-Adapted Interaction. 2019. V. 29. P. 759–788. https://doi.org/10.1007/s11257-019-09234-7
4. Nai R., et al. Process Mining on Students'Web Learning Traces: A Case Study with an Ethnographic Analysis. European Conference on Technology Enhanced Learning // Lecture Note in Computer Science. 2023. V. 14200. P. 599–604. https://doi.org/10.1007/978-3-031-42682-7_48
5. Mohssine B., et al. Adaptive Help System Based on Learners 'Digital Traces' and Learning Styles // Int. J. Emerging Technol. Learning. 2021. V.16. №10. P. 288–294. https://doi.org/10.3991/ijet.v16i10.19839
6. Ye D., & Pennisi S. Using Trace Data to Enhance Students' Self-Regulation: A Learning Analytics Perspective // The Internet and Higher Education. 2022. V. 54. P. 100855. https://doi.org/10.1016/j.iheduc.2022.100855
7. Noskova T., Pavlova T., Yakovleva O. Study of Students' Educational Activity Strategies in the Social Media Environment // E-learning and Smart Learning Environment for the Preparation of New Generation Specialists. 2018. V.10. P. 113–125.
8. Kadoić N., Oreški D. Analysis of Student Behavior and Success Based on Logs in Moodle. 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). 2018. https://doi.org/10.23919/MIPRO.2018.8400123
9. Mogus A., Djurdjevic I., Suvak N. The Impact of Student Activity in a Virtual Learning Environment on Their Final Mark // Active Learning in Higher Education. 2012. V.13. №3. P. 177–189. https://doi.org/10.1177/1469787412452985
10. Stiller K., Bachmaier R. Identifying Learner Types in Distance Training by Using Study Times // EDEN Conference Proceedings. 2018. V. 1. P. 78–86.