УДК 004.9

Informatics and Mathematical Methods in Simulation
Vol. 7 (2017), No. 3, pp. 213-219

ПРИМЕНЕНИЕ КЛАСТЕРНОГО АНАЛИЗА ДАННЫХ ДЛЯ ВЫДЕЛЕНИЯ МЕРЫ СХОЖЕСТИ ФАКТОРОВ ВЛИЯНИЯ НА РАБОТОСПОСОБНОСТЬ СЛОЖНЫХ ТЕХНИЧЕСКИХ СИСТЕМ

В.В. Вычужанин, Д.С. Шибаев, Н.Д. Рудниченко

Одесский национальный морской университет,
ул. Мечникова, 34, Одесса, 65029, Украина; e-mail: nickolay.rud@gmail.com

В статье приведены результаты применения кластерного анализа данных для выделения меры схожести факторов влияния на работоспособность сложных технических систем. В качестве примера системы рассмотрена судовая энергетическая установка. Предложена концептуальная модель взаимодействия модулей системы поддержки принятия решений по управлению работоспособностью сложных технических систем с модулем кластеризации. Созданы карты домена понятий и ER-модель базы данных модуля кластерного анализа данных, построены дендрограммы выделения иерархической взаимосвязанности кластеров на базе применения методов Nearest neighbor и Furthest neighbor

Ключевые слова: кластерный анализ данных, сложные технические системы, работоспособность технических систем, меры схожести факторов

Введение

В настоящее время наблюдается устойчивая тенденция усложнения состава технических систем. Заметным является стремительное развитие и совершенствование сложных технических систем (СТС) в различных сферах человеческой жизни. К наиболее распространенным причинам подобного тренда можно отнести необходимость в: агрегации новых функциональных возможностей, расширении первоначального назначения, повышении требований к надежности работы [1,2].

Во многих отраслях применения СТС разрабатываются и используются дополнительные компоненты, расширяющие возможности систем, что приводит к появлению новых типов взаимосвязей между их отдельными элементами. Это оказывает существенное влияние на объем вычислительных действий по обработке данных, выполнение которых необходимо для эффективного функционирования системы и выполнения поставленных перед ней задач [3].

Во избежание возникновения аварийных сценариев эксплуатации СТС и различных программно-аппаратных сбоев, приводящих к понижению их работоспособности, необходимо применение методов превентивного мониторинга и диагностики состояния отдельных компонентов технических систем.

Учет характера воздействия на систему внешних и внутренних факторов различной природы, оказывающих серьезное влияние на функционирование таких систем, представляется в данных условиях наиболее приоритетной задачей. Ее решение возможно посредством применения систем поддержки принятия решений (СППР), использующих модели и имплементирующих современные методы, технологии и средства искусственного интеллекта [4].

Однако, анализ ряда источников позволяет установить, что в настоящее время существуют две сопутствующие проблемы проведения однозначной оценки влияния разнородных факторов на работу СТС.

Первая из них заключается в том, что факторы не всегда могут быть явным образом выделены, полноценно формализованы и четко определены, вследствие их стохастической вероятностной природы [5].

Второй проблемой является необходимость обеспечения средств оперативной обработки больших объемов данных, которые генерируются в процессе отслеживания состояния компонентов системы [6-9]. Для этого необходимо проектирование масштабируемых распределенных хранилищ данных, организующих оптимальную реляционную структуру хранения информации и осуществляющих поддержку возможностей многомерного интеллектуального анализа данных [10].

Решение подобных задач возможно путем разработки модуля кластерного анализа данных (КАД) с целью выявления скрытых закономерностей в больших объемах данных по факторам, оказывающим влияние на работоспособность СТС [11,12]. Современные технологии кластеризация осуществляют разбитие входных данных на отдельные группы, каждая из которых имеет определенные признаки [13]. Их использование наиболее эффективно в случаях необходимости выделения определенных правил, корреляции и тенденции в больших наборах данных, что является актуальным в рамках рассматриваемой нами задачи [14,15].

Цель статьи и постановка задачи исследований

Целью данной работы является применение КАД для выделения мер схожести факторов, оказывающих влияние на степень работоспособности СТС.

Задачами настоящего исследования являются:

1. Разработка концептуальной модели взаимодействия модулей СППР управления работоспособностью СТС с модулем КАД.
2. Разработка карты домена понятий проектируемого модуля КАД.
3. Разработка ER-модели базы данных для проведения КАД.
4. Применение иерархического метода КАД для выделения взаимосвязанности кластеров.

Основная часть

Для объединения функций получения и предобработки входных данных, идентификации неисправностей в системе, оценки сценариев снижения степени работоспособности СТС и формирования управляющих решений для предотвращения возникновения аварийных ситуаций необходимым является формирование централизованной и масштабируемой СППР. В предлагаемой концепции построения такой СППР (рис.1.), наряду с блоками, реализующими приведенные выше функции, целесообразно размещение отдельного блока интеллектуального анализа данных (ИАД). Данный блок включает в свой состав следующие функциональные модули: КАД, многомерного анализа данных (OLAP) и обработки результатов в форме графической реализации с интерфейсом пользователя (EDV). Использование этих модулей позволяет обеспечить комплексных подход к формированию перечня альтернатив для лица принимающего решения по предотвращению аварийных ситуаций СТС. Во-первых, это достигается благодаря асинхронному созданию необходимых SQL-запросов в БД на выборку, вставку, обновление и удаление (SUID) данных КАД и OLAP-модулями. Во-вторых, полученные, посредством функционирования этих модулей, результаты ИАД формализуются и сохраняются в

базу знаний в виде отдельных продукционных правил и векторов кластеров. Это может быть использовано в дальнейшем для расширения возможностей СППР или создания модуля экспертной системы или искусственной нейронной сети с целью автоматизации процесса кластеризации и формирования итоговых альтернатив.

Каждый из модулей КАД, OLAP и EDV целесообразно реализовать в виде отдельного программного приложения, функционирующего в отдельном потоке для ускорения процесса обработки данных, посредством применения порождающего шаблона проектирования Factory. К факторам, оказывающим влияние на работоспособность СТС, согласно ряду источников [5,16], можно отнести следующие: структурный и функциональный ущербы, вероятность выхода компонента из строя, степень ремонтопригодности, количество проведенных ремонтов, структурный и функциональный риск отказов, длительность эксплуатации, режим эксплуатации, степень износа, температура окружающей среды и воздействие внешних сил (неблагоприятное влияние климатических, природных или техногенных сил).

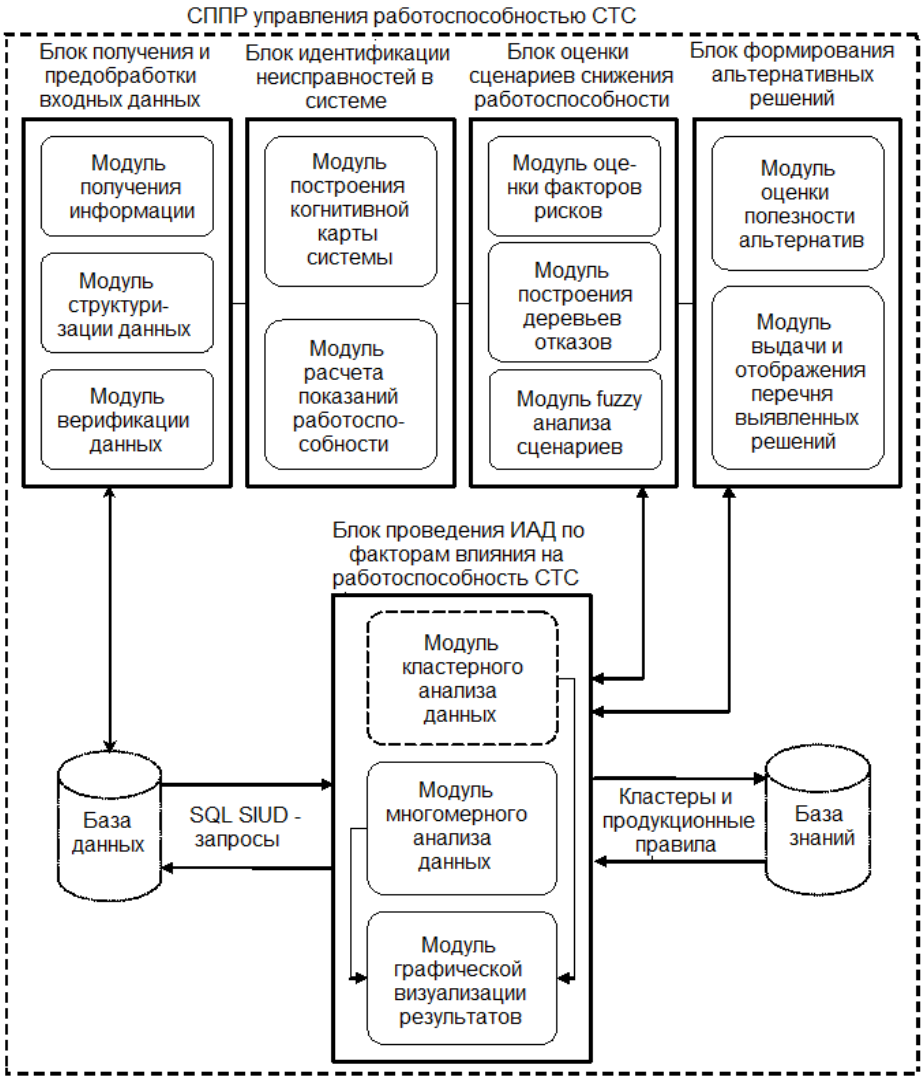

Рис. 1. Схема взаимодействия блока подсистемы КАД в рамках СППР

В качестве объекта исследования СТС в данной статье рассмотрена судовая энергетическая установка и ее статистика отказов из оффшорных и открытых источников данных (OREDA, EMSA, SRIC). Для расширяемой структуры, а также с целью обеспечения целостности и унификации процесса хранения данных разработана ER-модель БД средствами свободной СУБД MySQL WorkBench (рис.2).

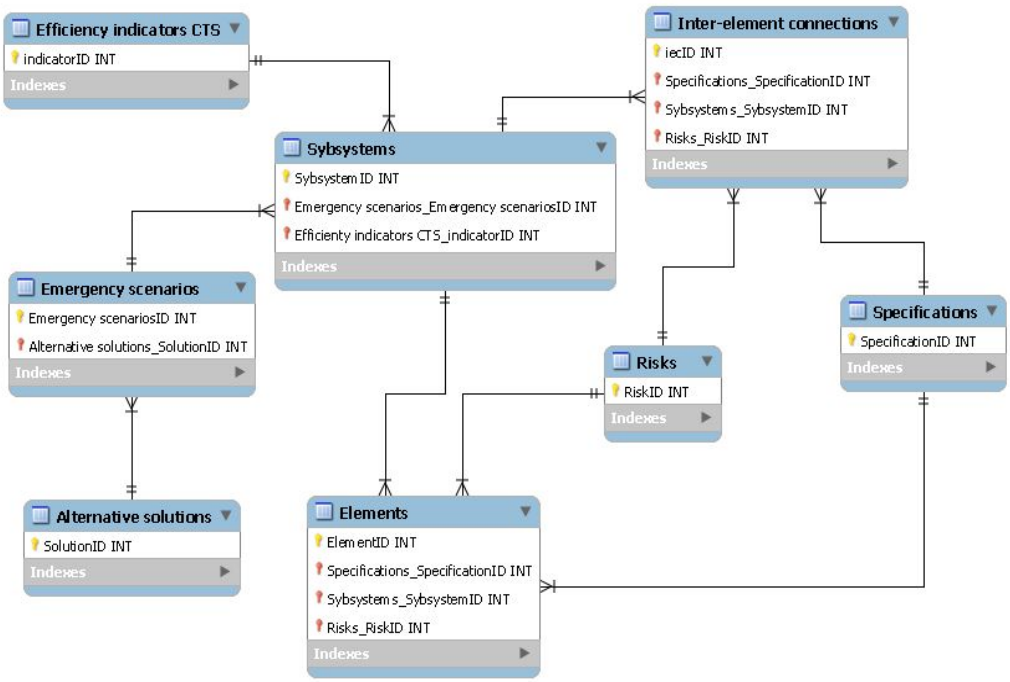

Рис. 2. ER-диаграмма связей между сущностями разработанной БД

Предусмотрено наличие связей типа «один ко многим», посредством создания внешних ключей между таблицами хранения данных, по: факторам работоспособности СТС (Efficiency Indicators CTS); элементам (Elements); межэлементным связям (Inter-element connections; подсистемам СТС (Sybsystems); техническим характеристикам компонентов (Specifications); альтернативным вариантам решений (Alternative solutions); аварийным сценариям развития (Emergency scenarios) и рискам (Risks). Для формализации процесса применения КАД для выделения мер схожести факторов, оказывающих влияние на степень работоспособности СТС, с помощью открытого облачного SaaS-сервиса mindmup.com разработана карта домена понятий, отображающая последовательность этапов кластеризации (рис.3).

Рис. 3. Карта домена понятий по этапам реализации процессов функционирования модуля КАД

Сбор входных данных из статистических БД осуществляется отдельным программным приложением (парсером), формирующим процесс переноса данных в разработанную БД. Отбор и обработка данных с последующей формализацией выполняется средствами используемой СУБД. Далее проводиться нормализация, осуществляемая на базе определения среднего значения (СЗ) и стандартного отклонения (СО) по всем признакам в отдельности. Обозначим средние значения признаков как $\overline{P_i}$, а стандартные отклонения – как $S_i, i = 1,...,M$ (где M – количество признаков) для N - объектов по X - признакам. Таким образом, определение СЗ и СО производиться по формулам (1) и (2):

$$\overline{P}_i = \frac{1}{N}\sum_{j=1}^{N} X_{ij}\ , \tag{1}$$

$$S_i = \sqrt{\frac{1}{N-1}\sum_{j=1}^{N}(X_{ij} - \overline{P_i})^2}\ . \tag{2}$$

В качестве меры различия между анализируемыми объектами применяется евклидово расстояние. Его значение между некоторыми объектами X_j, X_k

$$D(X_j, X_k) = \sqrt{\sum_{i=1}^{M}(X_{ij} - X_{ik})^2}\ . \tag{3}$$

Для осуществления КАД использованы методы Nearest neighbor и Furthest neighbor, в которых дистанция между парой выделенных кластеров вычисляется на базе определения расстояний между наиболее близкими и наиболее далекими объектами выборки соответственно. Полученные, в результате проведения КАД дендрограммы выделения взаимосвязанности кластеров приведены на рис.4.

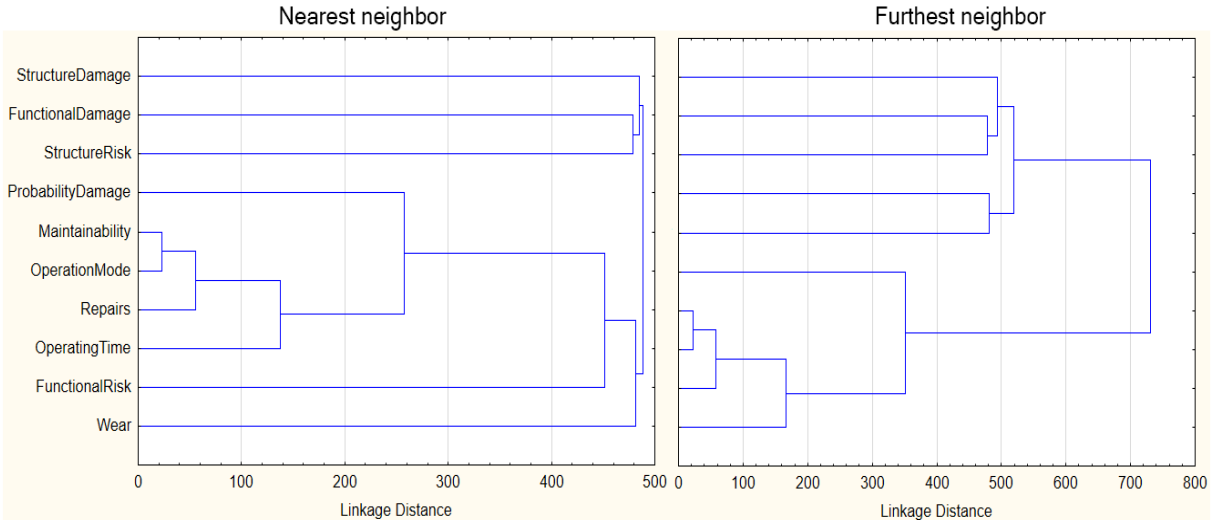

Рис. 4. Созданные дендрограммы выделения взаимосвязанности кластеров

Оценка адекватности результатов осуществлена путем определения внешних, внутренних и относительных метрик оценки качества кластеризации. Использованы методы Rand statistic, Hubert statistics и Индекс Дана.

Выводы

Полученные результаты оценки адекватности позволяют утверждать, что ошибка кластеризации находиться в допустимых пределах, однако лучшие значения метрик получено для метода Furthest neighbor. Анализируя процесс КАД можно отметить, что метод Nearest neighbor ориентируется на минимизацию числа больших по размеру кластеров, а метод Furthest neighbor направлен на увеличение числа компактных (меньших по размеру) кластеров.

Экспериментально установлено, что факторы функционального ущерба, структурного риска, вероятности выхода из строя, степени работоспособности и структурного ущерба обладают высокой степенью схожести и объединяются в один кластер через 4 итерации. Ко второму кластеру относятся факторы режима эксплуатации, количества ремонтов, времени эксплуатации, износа и функционального риска, которые также формируются в один кластер через 4 итерации. Дальнейшим развитием предложенной концепции КАД может быть уточнение принадлежности ряда объектов к нескольким кластерам путем использования нейро-нечеткой кластеризации.

Список литературы

1. Рудниченко, Н.Д. Информационная когнитивная модель технологической взаимозависимости сложных технических систем / Н.Д. Рудниченко, В.В. Вычужанин // Информатика и математические методы в моделировании. – 2013. – №3. – С. 240–247.
2. Вычужанин, В.В. Повышение эффективности эксплуатации судовой системы комфортного кондиционирования воздуха при переменных нагрузках. Монография/ В.В. Вычужанин, Одесса: ОНМУ, 2009.–206 с
3. Рудниченко, Н.Д. Оценка структурного и функционального рисков сложных технических систем / Н.Д. Рудниченко, В.В. Вычужанин // Восточно-европейский журнал передовых технологий. – 2014. – №1(67). – С. 18-22.
4. Rudnichenko, N. Concept implementation of decision support software for the risk management of complex technical system / N. Rudnichenko, V. Boyko, S. Kramskoy, Y. Hrechukha, N. Shibaeva // Advances in Intelligent Systems and Computing of the series Advances in Intelligent Systems and Computing. – 2016. – № 512. – P. 255-269.
5. Рудниченко, Н.Д. Разработка модели нейросети для прогнозирования риска отказов компонентов сложных технических систем / Н.Д. Рудниченко, В.В. Вычужанин // Информатика и математические методы в моделировании. – 2016. – №4. – С. 333-338.
6. Vychuzhanin, V.V. Big data mapping in the geopositioning systems for fishing industry / V.V. Vychuzhanin, D.S. Shibaev, V.D. Boyko, N.O. Shibaeva, N.D. Rudnichenko // Матеріали XII Міжнародної науково-технічної конференції CSIT 2017. – Т.1. – Львів: Вежа і Ко. – 2017. – С. 28-31.
7. Казиев, Г.З. Модели и методы кластеризации больших данных для их анализа и обработки / Г.З. Казиев, В.В. Курдюков // Сборник научных статей по итогам международной научно-практической конференции. НОУ ДПО "Санкт-Петербургский институт проектного менеджмента. – 2016. – С. 197-199.
8. Ахметханов, Р.С. Метод кластеризации диагностических данных при вибродиагностике технических систем / Р.С. Ахметханов, Е.Ф. Дубинин, В.И. Куксова // Вестник научно-технического развития. – 2017. – № 5(117). – С.3-16.
9. Вычужанин, В.В. Гибридные экспертные системы для противоаварийного управления сложными техническими системами/ В.В. Вычужанин, С.Н. Коновалов// Вісник Одеського національного морського університету, збірник наукових праць. – 2017. – № 2 (51). – С.165-179.
10. Vychuzhanin, V. Devising a method for the estimation and prediction of technical condition of ship complex systems / V. Vychuzhanin, N. Rudnichenko, V. Boyko, N. Shibaeva, S. Konovalov // Eastern-European Journal of Enterprise Technologies. – 2016. – №6/9 (84). – P. 4-11.
11. Григоренко, Д.В. Кластеризация систем обработки специальных данных / Д.В. Григоренко, В.Н. Ручкин // Методы и средства обработки и хранения информации. –Рязань: РГРТУ. – 2012. – С. 98 - 103.

12. Надев, А.И. Диагностика технического состояния судовых дизелей на основе интеллектуального анализа данных / А.И. Надеев, Б.Х. Нгок, Ф.В. Свирепов // Вестник АГТУ. – Сер.: Морская техника и технология. – 2011. – № 2. – С.105-110.
13. Егоров, А.В. Особенности методов кластеризации данных / А.В. Егоров, Н.И. Куприянова // Известия ЮФУ. Технические науки. – 2011. – №11. – С.174-178.
14. Суслов, С.А. Кластерный анализ: сущность, преимущества и недостатки / С.А. Суслов // Вестник НГИЭИ. – 2010. – №1. – С.51-57.
15. Kohonen, M.M. Self Organizing Map with Modified K-means clustering For High Dimensional Data Set / M.M. Kohonen // International Journal of Applied Information Systems (IJAIS). Foundation of Computer Science FCS, New York, USA.– 2012. –P. 34–39.
16. Nobuo, K. Statistical Study on Reliability of Ship Equipment and Safety Management–Reliability Estimation for Failures on Main Engine System by Ship Reliability Database System / K. Nobuo // Bulletin of the JIME. – 2011. – Vol. 29. – №2. – P.64-70.

ЗАСТОСУВАННЯ КЛАСТЕРНОГО АНАЛІЗУ ДАНИХ ДЛЯ ВИДІЛЕННЯ МІРИ СХОЖОСТІ ФАКТОРІВ ВПЛИВУ НА ПРАЦЕЗДАТНІСТЬ СКЛАДНИХ ТЕХНІЧНИХ СИСТЕМ

В.В. Вичужанін, Д.С. Шибаєв, М.Д. Рудніченко

Одеський національний морський університет,
вул. Мечнікова, 34, Одеса, 65029, Україна; e-mail: nickolay.rud@gmail.com

У статті наведені результати застосування кластерного аналізу даних для виділення заходи схожості факторів впливу на працездатність складних технічних систем. Як приклад системи розглянута суднова енергетична установка. Запропоновано концептуальну модель взаємодії модулів системи підтримки прийняття рішень з управління працездатністю складних технічних систем з модулем кластеризації. Створено карти домену понять і ER-модель бази даних модуля кластерного аналізу даних, побудовані дендрограмми виділення ієрархічної взаємозв'язку кластерів на базі застосування методів Nearest neighbor і Furthest neighbor
Ключові слова: кластерний аналіз даних, складні технічні системи, працездатність технічних систем, заходи схожості факторів
Ключові слова: кластерний аналіз даних, складні технічні системи, працездатність технічних систем, заходи схожості факторів

CLUSTER DATA ANALYSIS FOR THE SIMILARITY MEASURE IDENTIFY OF THE COMPLEX TECHNICAL SYSTEMS OPERABILITY FACTORS

V.V. Vichuzhanin, D.S. Shibaev, N.D. Rudnichenko

Odessa National Maritime University
34, Mechnikova Str., Odessa, 65029, Ukraine; e-mail: nickolay.rud@gmail.com

The article presents the results of the application of cluster data analysis to identify the measure of the similarity of the factors affecting the performance of complex technical systems. As an example of the system, a ship power plant is considered. A conceptual model of the interaction of the modules of the decision support system for managing the operability of complex technical systems with a clustering module is proposed. The concept domain maps and the ER model of the database of the cluster data analysis module have been created, dendrograms have been constructed for highlighting the hierarchical interconnection of clusters based on the Nearest neighbor and Furthest neighbor methods..
Keywords: cluster analysis of data, complex technical systems, the efficiency of technical systems, measures of similarities of factors.