Российский экономический интернет-журнал №4/2021

Научная статья

Original article

УДК 330.131.5

Анализ неполадок сетевого оборудования методами data mining для повышения экономической эффективности

Харитонова А.Е., к.э.н., доцент,

Федеральное государственное бюджетное образовательное учреждение высшего образования «Российский государственный аграрный университет – МСХА имени К.А. Тимирязева, Москва, Россия

Коломеева Е.С., к.э.н., доцент,

Федеральное государственное бюджетное образовательное учреждение высшего образования «Российский государственный аграрный университет – МСХА имени К.А. Тимирязева, Москва, Россия

Аннотация. В статье проводится анализ неполадок сетевого оборудования на основе технологии data mining и языка программирования R. В теоретическом плане описываются основные методы классификации, такие как дерево решений, случайный лес и метод опорных векторов. Построенные модели позволяют классифицировать проблемы сетевого оборудования по времени неполадок, по хосту и по месту установки, что поможет своевременно устранять и предотвращать неполадки и повысит экономическую эффективность производства.

Ключевые слова: data mining, дерево решений, случайный лес, метод опорных векторов, сетевое оборудование.

Application of data mining technology to analyze the operation of network equipment

Kharitonova A.E., Candidate of Economics, Associate Professor, Federal State Budgetary Educational Institution of Higher Education Russian State Agrarian University-Moscow Timiryazev Agricultural Academy, Moscow, Russia.
Kolomeeva E.S., Candidate of Economics, Associate Professor, Federal State Budgetary Educational Institution of Higher Education Russian State Agrarian University-Moscow Timiryazev Agricultural Academy, Moscow, Russia.

Annotation. The article analyzes the problems of network equipment based on data mining technology and the R programming language. The main classification methods are described, such as a decision tree, a random forest and a support vector machine. The constructed models make it possible to classify network equipment problems by time of failure, by host, and by place of installation. It will help to correct and prevent problems in a timely manner and increase the economic efficiency of production.

Keywords: data mining, decision tree, random forest, support vector machine, network equipment.

Введение. Информационные технологии связаны с обеспечением бесперебойной работы вычислительных сетей. В связи с чем, возникает необходимость улучшения существующих способов мониторинга работоспособности локальных вычислительных сетей. Современный рынок информационных продуктов предлагает готовое программное обеспечение, позволяющее провести мониторинг оборудования на работоспособность, однако проведённый анализ выявил недостатки современных программных средств, это:

– стандартные решения плохо адаптируются к различным сетям;

– расширение сети приводит к тому, что контроль над сетью уменьшается и выходит за рамки заложенных параметров.

В результате необходима налаженная система мониторинга сетевого оборудования для принятия оперативных мер по устранению неполадок.

Объектом исследования является данные о неполадках в оборудовании локальных вычислительных сетей.

Предметом является система методов data mining для выявления проблем в работоспособностиоборудования локальных вычислительных сетей.

Целью статьи является разработка моделей классификации проблем в сетевом оборудовании, позволяющей выявлять и прогнозировать причины возникающих проблем.

Результаты исследования. Интеллектуальный анализ данных (Data Mining) – это процесс поиска потенциально полезных шаблонов из огромных наборов данных. Это междисциплинарный навык, который использует машинное обучение, статистику и искусственный интеллект для извлечения информации для оценки вероятности будущих событий. Информация, полученная в результате интеллектуального анализа данных, используется для маркетинга, обнаружения мошенничества, научных открытий и т.д[1].

Data Mining – это обнаружение скрытых, неожиданных и ранее неизвестных, но действительных взаимосвязей между данными. Интеллектуальный анализ данных также называется обнаружением знаний в данных (KDD), извлечением знаний, анализом данных / шаблонов, сбором информации и т.д.

Интеллектуальный анализ данных – не новое изобретение, пришедшее с эпохой цифровых технологий. Эта концепция существует уже более века, но привлекла большее внимание общественности в 1930-х годах. Один из первых примеров интеллектуального анализа данных произошел в 1936 году, когда Алан Тьюринг представил идею универсальной машины, которая могла бы

[1] C. Bouveyron, G. Celeux, B. Murphy and A. Raftery, Model-based Clustering and Classification for Data Science, with Applications in R, 2019. – 386 p.

выполнять вычисления, аналогичные вычислениям на современных компьютерах[2].

В настоящее время предприятия используют интеллектуальный анализ данных и машинное обучение для улучшения всего, от процессов продаж до интерпретации финансовых показателей в инвестиционных целях. В результате специалисты по обработке данных стали жизненно важными для организаций по всему миру, поскольку компании стремятся достичь с помощью науки о данных более крупных целей, чем когда-либо прежде.

Интеллектуальный анализ данных – это процесс анализа огромных объемов данных для обнаружения бизнес-аналитики, которая помогает компаниям решать проблемы, снижать риски и использовать новые возможности. Эта отрасль науки о данных получила свое название от сходства между поиском ценной информации в большой базе данных и добычей руды в горах. Оба процесса требуют просеивания огромного количества материала, чтобы найти скрытую ценность[3].

Интеллектуальный анализ данных может дать ответ на бизнес-вопросы, которые обычно требовали слишком много времени, чтобы решить их вручную. Используя ряд статистических методов для анализа данных различными способами, пользователи могут определять закономерности, тенденции и взаимосвязи, которые в противном случае они могли бы упустить. Они могут применять эти результаты, чтобы предсказать, что может произойти в будущем, и предпринять действия, чтобы повлиять на результаты бизнеса.

Интеллектуальный анализ данных используется во многих областях бизнеса и исследований, включая продажи и маркетинг, разработку продуктов, здравоохранение и образование. При правильном использовании интеллектуальный анализ данных может обеспечить серьезное преимущество перед конкурентами, позволяя вам больше узнавать о клиентах, разрабатывать

[2] Грас, Дж. Data Science. Наука о данных с нуля. Пер. с англ. – Спб.: БХВ-Петербург, 2019. – 336 с.
[3] RAN Task View: Machine Learning & Statistical Learning [Электронный ресурс]. – Режим доступа: https://cran.r-project.org/web/views/MachineLearning.html

эффективные маркетинговые стратегии, увеличивать доходы и сокращать расходы.

Для интеллектуального анализа данных используется множество методов, но решающим шагом является выбор из них подходящей формы в соответствии с бизнесом или постановкой задачи. Эти методы помогают прогнозировать будущее и принимать соответствующие решения. Это также помогает в анализе рыночных тенденций и увеличении доходов компании.

Некоторые группы методов:

- ассоциация;
- классификация;
- кластерный анализ;
- прогнозирование;
- последовательные паттерны или отслеживание паттернов;
- деревья решений;
- анализ выбросов или анализ аномалий;
- нейронная сеть.

Существует две формы анализа данных, которые можно использовать для извлечения моделей, описывающих важные классы, или для прогнозирования будущих тенденций данных. Эти две формы следующие:

- классификация;
- прогнозирование.

Классификационные модели предсказывают категориальные обозначения классов; а модели прогнозирования предсказывают непрерывные функции. Например, мы можем построить модель классификации, чтобы классифицировать заявки на получение банковского кредита как безопасные или рискованные, или модель прогнозирования, чтобы спрогнозировать долларовые расходы потенциальных клиентов на компьютерное оборудование с учетом их доходов и занятий[4].

[4] Shobha A. Shinde.et,al. International Journal of Technology and Engineering Science[IJTES], 2015. – Volume 3[8]. – pp: 4001-4007

Методы дерева классификации (то есть методы дерева решений) рекомендуются, когда задача интеллектуального анализа данных содержит классификации или предсказания результатов, и цель состоит в том, чтобы создать правила, которые можно легко объяснить и перевести на SQL или естественный язык запросов.

Дерево классификации маркирует, записывает и присваивает переменные дискретным классам. Дерево классификации также может дать меру уверенности в правильности классификации.

Дерево классификации строится с помощью процесса, известного как двоичное рекурсивное разбиение. Это итеративный процесс разделения данных на разделы с последующим разделением их на каждую из ветвей [5].

Алгоритм случайного леса является усовершенствованием существующего алгоритма дерева решений, который страдает серьезной проблемой «переобучения». Он также считается более быстрым и точным по сравнению с алгоритмом дерева решений. Он объединяет результаты нескольких деревьев решений и классифицирует выходные данные на основе результата[5].

Алгоритмы машин опорных векторов (SVM) – это набор алгоритмов, которые можно использовать с различными проблемами и данными. Заменяя одно ядро на другое, SVM может решить множество задач интеллектуального анализа данных.

Машины опорных векторов – это набор контролируемых методов обучения, используемых для классификации, регрессии и обнаружения выбросов.

Преимущества машин опорных векторов:

- эффективен в пространствах больших размеров;
- эффективен в случаях, когда количество измерений превышает количество образцов;

[5] Грас, Дж. Data Science. Наука о данных с нуля. Пер. с англ. – Спб.: БХВ-Петербург, 2019. – 336 с.

- использует подмножество обучающих точек в функции принятия решений (называемых опорными векторами), поэтому это также эффективно с точки зрения памяти;
- универсальность: для функции принятия решения могут быть указаны различные функции ядра. Предоставляются общие ядра, но также можно указать собственные ядра.

К недостаткам опорных векторных машин можно отнести:

- если количество функций намного превышает количество выборок, избегайте чрезмерной подгонки при выборе функций ядра, и термин регуляризации имеет решающее значение;
- SVM не предоставляют напрямую оценки вероятностей, они рассчитываются с использованием дорогостоящей пятикратной перекрестной проверки[6].

Методы интеллектуального анализа данных могут быть использованы в любой области исследования. Так и при работе с неполадками в сетях системы data mining могут помочь выявить скрытые закономерности и помочь принять оперативные решения для их устранения.

В качестве исходных данных для аналитики используем выгрузку о проблемах с сетевым оборудованием

На ряде вокзалов есть особенности проблем с оборуддованием, связанных с временем совершения проблемы. Например большая часть проблем возникает в нерабочее время из-за отсутсвия работников и прочее.

Первоначально проанализируем методом дерево решений распределение проблем по рабочему-нерабочему времени в зависимости от хоста, принимаемым действиям и продолжительности неполадок. Для построения моделей используем язык программирования R. Для построяния модели дерева решений был использован пакет rpart. Для всех неполадок была добавлена

[6] C. Bouveyron, G. Celeux, B. Murphy and A. Raftery, Model-based Clustering and Classification for Data Science, with Applications in R, 2019. – 386 p.

переменная work которая принимает значение 1 если неполадка произошла в рабочее время и 0 – в нерабочее.

По результатам построения модели можно сказать, что из 181 проблемы, возникшей в нерабочее время модель может предсказать 133. Среди 150 проблем рабочего времени модель верно предсказывает 108. Точность классификации составила 73%.

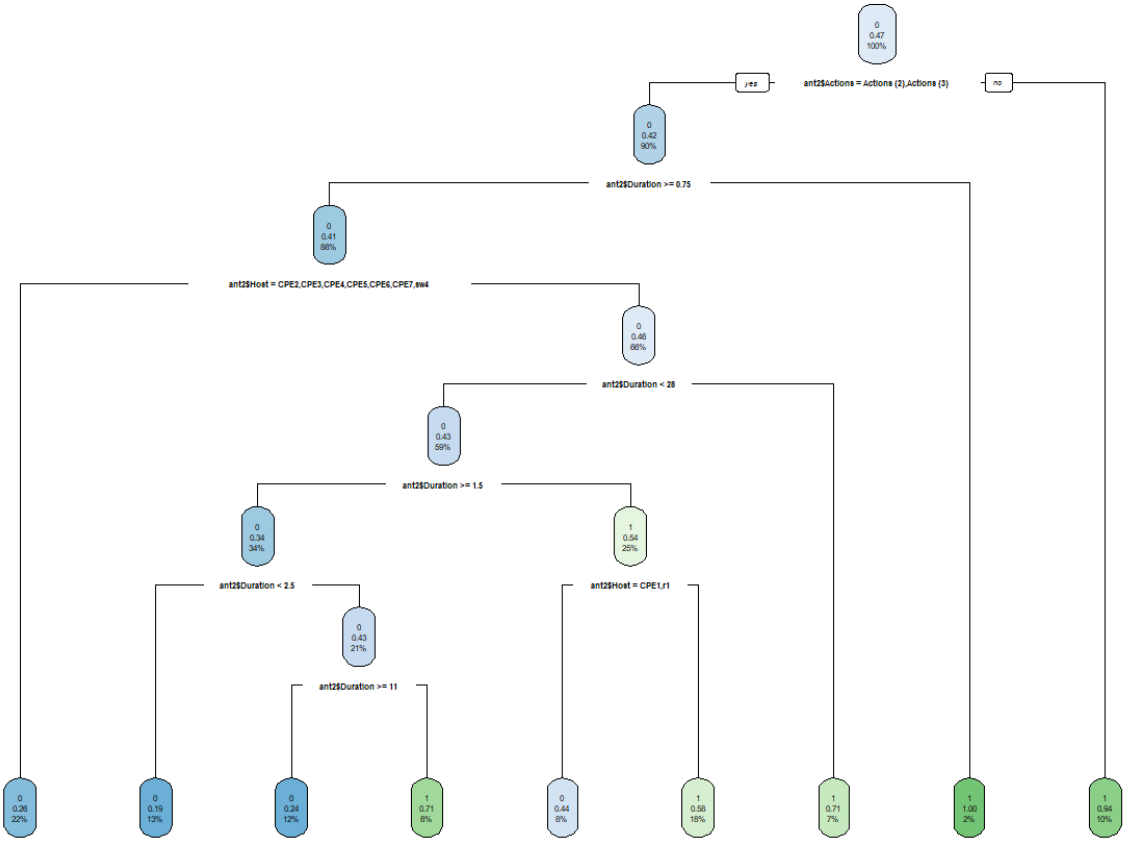

Рис. 1– Граф метода «Дерево решений»

По графу дерева решений моно заметить, что если для решения проблемы выполнялось действие 1, то неполадка происходила в рабочее время. Для остальных действий (2 и 3) играла роль продолжительность неполадки. Если проблема устранялась менее чем за минуту, то проблема была в рабочемм время. Если более минуты, то играл роль непосредственно хост. Для хостов CPE2, CPE3, CPE4, CPE5, CPE6, CPE7 и sw4 все проблемы длиннее минуты были отнесены к нерабочему времени. Для остальных хостов далее проблемы были разделены на продолительность до 28 минут и более. Если более 28 минут – рабочее время. Если менее, то продолжительность еще раз делилась на выше и

нижк 1,5 минуты. В случае менее 1,5 минут для хостов CPE1 и r1 проблемы вознилали внерабочее время, а в остальных в рабочее. Для проблем дольше 1,5 минут но короче 2,5 неполадки были в нербочее время. Также в этой подгруппе проблемы длиннее 11 минут происходили также в нерабочее время.

Аналогичная модель построенная методов «Случайный лес» (пакет RandomForest) предсказывает с точностью 89,4%. Таким образом можно для тех хостов у которых проблемы возникают в основном в нерабочее время необходимо проводить более детальный анализ проблем и принимать управленчес

Графически распределение неполадок по времени можно видеть на рис. 2. Как видно, большая часть проблем с минимальными значениями факторов относятся к рабочему времени. Таким образом видна эффективность работы персонала.

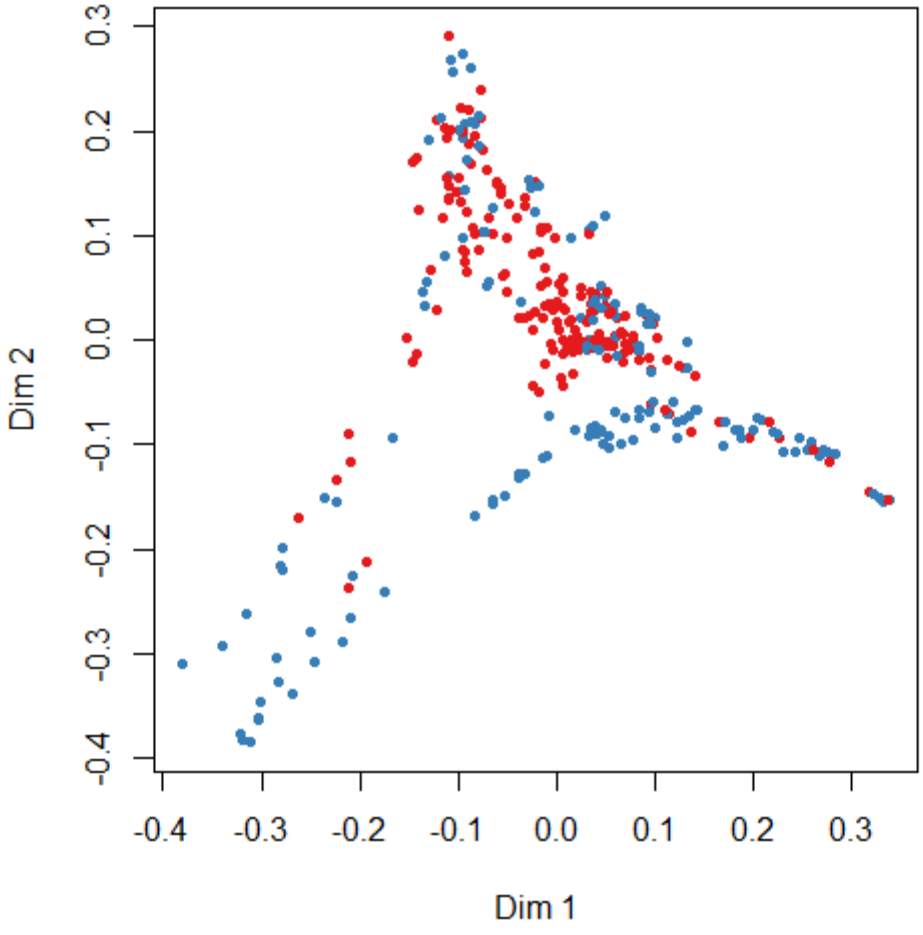

Рис. 2– Распределение неполадок по времени «Случайный лес»

Т.к. на графике наблюдается нелинейная зависимость попробуем использовать метод опорных векторов с ядерной функцией с помощью пакета e1072.

Певоначально осуществляется поиск параметров ядерной функции. Так, параметр gamma = 1, cost = 8. Таким образом исходное пространство данных преобразуется через новое соотношение паремтров. Однако точность прогноза составило лишь 77,6%.

В целом лучшее качество классификации показал метод «Случайный лес».

Далее в качестве классификационного признака используем хост. Ведь проблемы возникающие с оборудованием могут напрямую зависеть от хоста. Таким обраом можно выявить проблемные хосты и принять ряд мер по оптимизации их работы.

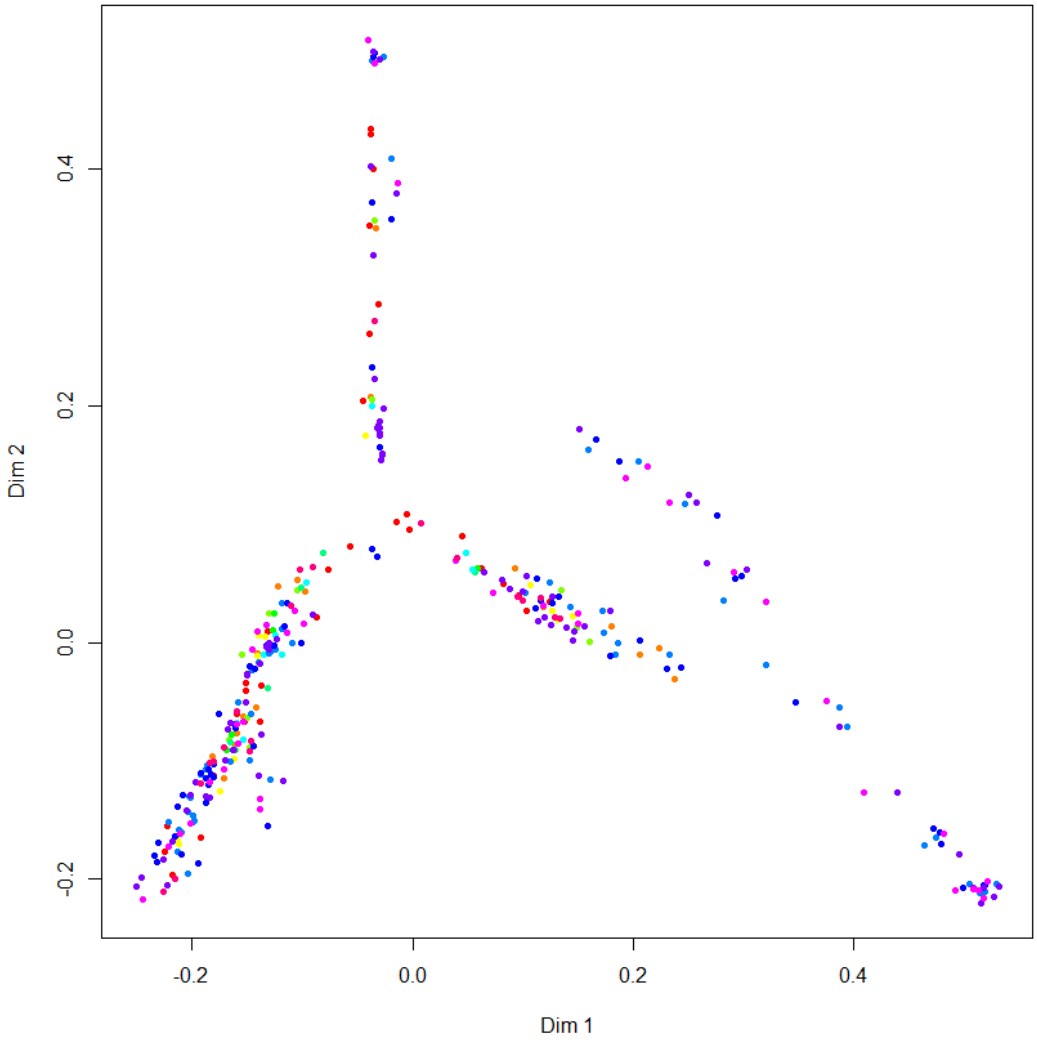

Рис. 3– График метода «Случайный лес»

По графику также видна нелинейная зависимость, таким образом в данном случае целесообразно применить метод опорных векторов с ядерной функцией.

Ядерная функция метода опорных векторов показала лучшее качество классификации хостов – 77%.

Интерес представляет также зависимость неполадок от вокзала. Метод «Случайный лес» предсказывает по неполадкам вокзал с точность 60%.

По матрице классификации можно заметить, что однозначно может классифицироваться такие вокзалы со своими особенностями, которые можно устранить.

График визуализации на двух осях показывает определенную зависимость. Однако за счет большого числа вокзалов метод опорных векторов с ядерной функцией показывает результаты хуже, чем «Случайный лес» (рис. 4).

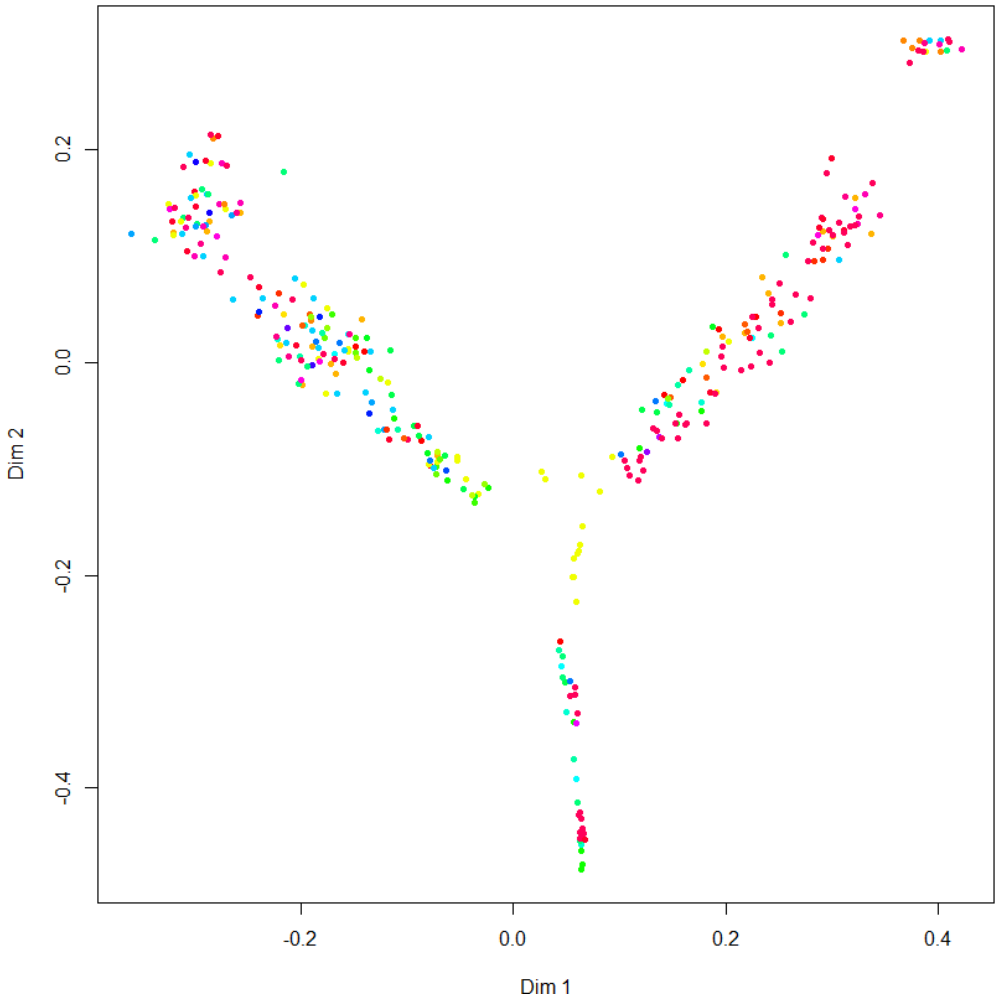

Рис. 4– Визуализация распределения вокзалов по методу «Случайный лес»

Заключение. В целом следует отметить, что применение методов интеллектуального анализа данных позволяет выявлять зависимости между неполадками, временем и продолжительностью неполадок, местом установки оборудования и хостом. Применение данных методик позволит выявить проблемные хосты и места установок и принять оперативные управленческие решения для оптимизации работы и повышения экономической эффективности деятельности.

Библиографический список:

1. C. Bouveyron, G. Celeux, B. Murphy and A. Raftery, Model-based Clustering and Classification for Data Science, with Applications in R, 2019. – 386 p.

2. RAN Task View: Machine Learning & Statistical Learning [Электронный ресурс]. – Режим доступа: https://cran.r-project.org/web/views/ MachineLearning.html

3. Shobha A. Shinde.et,al. International Journal of Technology and Engineering Science[IJTES], 2015. – Volume 3[8]. – pp: 4001-4007

4. The R Project for Statistical Computing [Электронный ресурс]. – Режим доступа: https://www.r-project.org/

5. Анализ результатов всероссийской сельскохозяйственной переписи с использованием методов машинного обучения / Харитонова А.Е., Сундупей А.А., Скачкова С.А. // Бухучет в сельском хозяйстве. – 2020. – № 12. – С. 41-48.

6. Грас, Дж. Data Science. Наука о данных с нуля. Пер. с англ. – Спб.: БХВ-Петербург, 2019. – 336 с.

7. Шитиков В.К., Мастицкий С.Э. (2017) Классификация, регрессия и другие алгоритмы Data Mining с использованием R. – 351 с.

References:

1. C. Bouveyron, G. Celeux, B. Murphy and A. Raftery, Model-based Clustering and Classification for Data Science, with Applications in R, 2019. – 386 p.

2. RAN Task View: Machine Learning & Statistical Learning [Электронный ресурс]. – Режим доступа: https://cran.r-project.org/web/views/MachineLearning.html

3. Shobha A. Shinde.et,al. International Journal of Technology and Engineering Science[IJTES], 2015. – Volume 3[8]. – pp: 4001-4007

4. The R Project for Statistical Computing [Электронный ресурс]. – Режим доступа: https://www.r-project.org/

5. Analysis of the results of the All-Russian agricultural census using machine learning methods / Kharitonova A.E., Sundupey A.A., Skachkova S.A. // Accounting in agriculture. – 2020. – № 12. – P. 41-48.

6. Grasse, J. Data Science. Data science from scratch. Per. from English - SPb.: BHV-Petersburg, 2019. – 336 p.

7. Shitikov V.K., Mastitsky S.E. (2017) Classification, Regression and Other Data Mining Algorithms Using R. – 351 p.

Для цитирования: Харитонова А.Е., Коломеева Е.С. Анализ неполадок сетевого оборудования методами data mining для повышения экономической эффективности// Российский экономический интернет-журнал. – 2021. – № 4. URL: