АНАЛИЗ ДАННЫХ МОНИТОРИНГА УСЛОВИЙ ТРУДА РАБОЧИХ МЕСТ ПРЕДПРИЯТИЙ

УДК 658.5

ДОРОВСКОЙ Дмитрий Владимирович
к.т.н., доцент кафедры «Информатики и социально-гуманитарных дисциплин» Криворожского филиала Европейского университета.
Научные интересы: маркетинг, информационные и маркетинговые технологии, мониторинг и диагностика горно-метталургического оборудования.
e-mail: postmaster@krivrig.e-u.in.ua

ДОРОВСКАЯ Ирина Александровна
старший преподаватель кафедры «Информатики и социально-гуманитарных дисциплин» Криворожского филиала Европейского университета.
Научные интересы: информационные технологии в здравоохранении, аттестация рабочих мест.

ФИЛИПЕНКО Анастасия Юрьевна
старший преподаватель кафедры «Информатики и социально-гуманитарных дисциплин» Криворожского филиала Европейского университета.
Научные интересы: информационные технологии, теория принятия решений, нейронные сети.
e-mail: nastyafilipenko@rambler.ru

ВВЕДЕНИЕ. ИССЛЕДОВАНИЕ ПРОБЛЕМЫ.

Важнейшей проблемой высокой эффективности современного предприятия является обеспечение высокого уровня механизации и автоматизации производственных процессов, повышение технологической дисциплины производственного персонала, повышение квалификации технологического персонала. Решением этих задач выше определенной проблемы является процесс мониторинга условий труда технологических параметров рабочих мест персонала (МУТТПРМП) и его автоматизация. Для проведения МУТТПРМП необходимо проведение анализа данных (АД) для разделения множества объектов не кластеры (КЛ),классы, таксоны, сгущения или группы. При проведении кластерного анализа (КА) МУТТПРМП не требуется предположений о наборе данных (НД), не вводится ограничения на представление объектов анализа и типы данных. КА используется для сокращения размерности и визуализации данных МУТТПРМП.

ПОСТАНОВКА ЗАДАЧИ

Постановка задачи КА [1-3]: Определим множество объектов данных $Q = \{q_1, q_2, ..., q_n\}$, при этом каждый объект q_j представлен набором атрибутов:

$$\tilde{x}_j = \left\langle x_{j,1}, x_{j,2}, \ldots, x_{j,m} \right\rangle. \quad (1)$$

которые принимают значения из множества действительных чисел $x_j = \{v_j^1, v_j^2, ...\}$.

Решением задачи КА является множество сформированных КЛ:

$$C = \{c_1, c_2, \ldots, c_g\}, \quad (2)$$

где

$c_k = \{q_i, q_j \,|\, q_i \in Q,\ q_j \in Q \ u\ d(q_i, q_j) < \sigma\}$ – КЛ, содержащий похожие объекты из множества Q, $d(q_i, q_j)$ – мера близости между объектами, σ – величина, определяющая меру близости между объектами. Мера близости должна отвечать следующим условиям [6]:

а) $d(q_i, q_j) \geq 0$, *для всех* q_i *и* q_j;

б) $d(q_i, q_j) = 0$, *тогда и только тогда, когда* $q_i = q_j$;

в) $d(q_i, q_j) = d(q_j, q_i)$;

г) $d(q_i, q_j) \leq d(q_i, q_k) + d(q_k, q_j)$.

Если выполняется неравенство $d(q_i, q_j) < \sigma$, объекты из множества Q рассматриваются как близкие и помещаются в один КЛ, если неравенство не выполняется объекты помещаются в разные КЛ. Меры близости в КА МУТРМП [1,4] предполагает представление объектов в виде точек m – мерного пространства – R^m и определяется расстояний между двумя точками пространства R^m.Так эвклидово расстояние, между объектами вычисляется по формуле:

$$d(q_i, q_j) = \sqrt{\sum_{k=1}^{m} (x_{i,k} - x_{j,k})^2}. \quad (3)$$

Именно эта мера придаёт больше веса более отдалённым друг от друга объектам из заданного множества Q. Расстояние по Хеммингу вычисляется следующим образом:

$$d(q_i, q_j) = \sum_{k=1}^{m} \left| x_{i,k} - x_{j,k} \right|. \quad (4)$$

Эта мера в отличие от эвклидового расстояния снижает влияние больших расхождений по отдельным атрибутам на результаты КА.

Расстояние по Чебышеву вычисляется по формуле:

$$d(q_i, q_j) = \max_{1 \leq k \leq m} \left| x_{i,k} - x_{j,k} \right|. \quad (5)$$

В нашем случае, формула Чебышева использовалась при необходимости распределения объектов по КЛ, имеющим существенное различие только по одному атрибуту (измерению).

Для случая КА МУТТПРМП использовался еще один из возможных вариантов вычислений – это расстояние Махаланобиса и его вычисляли по следующему выражению:

$$d(q_i, q_j) = (x_i - x_j) S^{-1} (x_i - x_j)^t, \quad (6)$$

где t – символ транспонирования [2,3].

ЦЕЛЬ И ЗАДАЧИ ИССЛЕДОВАНИЯ

В связи с вышеуказанной постановкой задачи целью исследований являлось повышение качества обработки и преобразования экспериментальных данных при проведении МУТТПРМП. Для достижения этой цели использовался КА данных. Задачи проведения исследований включали анализ выбор и использование алгоритмов и моделей КА при проведении МУТТПРМП.

ИЗЛОЖЕНИЕ ОСНОВНОГО МАТЕРИАЛА

Алгоритмы КА [5,6] делятся на иерархические (ИАКА) и неиерархические (НИКА). ИАКА в свою очередь разделяют на агломеративные и дивизимные. В иерархических агломеративных алгоритмах КА, исходное множество объектов Q представляется как множество кластеров C. Таким образом, для нашего случая:

– на первом шаге алгоритма имеем:

$$c_1 = \{q_1\}, c_2 = \{q_2\}, \ldots, c_g = \{q_n\} \text{ и } g = n. \quad (7)$$

– на втором шаге алгоритма, находим КЛ с наименьшим удалением друг от друга и проводим слияние КЛ c_i, c_j в общий КЛ $c_k = \{c_i, c_j\}$, используя выбранную меру близости $d()$. Повторяем процесс поиска КЛ с наименьшим удалением и их слияние. В результате этого формируются множество КЛ мощностью $g-1$, $g-2$, $g-3$, … и выполняем пересчет расстояния между КЛ c_k и КЛ $c_l, l = 1, 2, \ldots$:

$$d_{k,l} = \alpha_i d_{i,l} + \alpha_j d_{j,l} + \beta d_{i,j} + \gamma \left| d_{i,l} - d_{j,l} \right| \quad (8)$$

где $d_{i,j}$ – расстояние между КЛ c_i, c_j, $d_{i,l}$ – расстояние между КЛ c_i, c_l, $d_{j,l}$ – расстояние между КЛ c_j, c_l, $\alpha_i, \alpha_j, \beta, \gamma$ – весовые коэффициенты. При расчёте были взяты следующие значения коэффициентов:

$\alpha_i = 0.5, \alpha_j = 0.5, \beta = 0.25, \gamma = 0$ [1].

При рассмотрении дивизимных ИАКА, исходное множество представляется как единственный КЛ и на первом шаге этого алгоритма имеем:

$$c_1 = \{q_1, q_2, \ldots, q_{n_1}\}, n_1 = n \quad (9)$$

На втором шаге алгоритма выбираем объект q_r, который наиболее удален от других объектов в этом КЛ. Удаление объекта q_r определяется как наибольшее среднее расстояния до других объектов КЛ и рассчитывается по выражению:

$$d_r = 1/n_1 \times \sum d(q_r, q_k) \; \forall \; q_k \in c_1 \quad (10)$$

Новый КЛ c_2 формируем на следующем шаге алгоритма для этого выбранный объект q_r удаляется из КЛ c_1 и помещается в КЛ c_2 ($n_2 = 1$). Перенос объектов из c_1 в c_2 продолжается до тех пор, пока разности средних расстояний не станут отрицательными. В результате выполнения последовательности шагов формируются два КЛ. Выбор КЛ для разделения может осуществляться на основе оценки диаметров КЛ и выполняется с применением выражения:

$$D_k = \max d(q_i, q_j) \; \forall \; q_i, q_j \in c_k, \quad k = 1, 2, \ldots, g. \quad (11)$$

Разделение КЛ производится до тех пор, пока все члены одного КЛ не будут отвечать требованию близости или все КЛ будут содержать по одному объекту.

Использование НИКА МУТРМП обеспечивает разделение объектов таким образом, чтобы достичь максимума или минимума целевой функции. Тогда для нашего случая алгоритм КА k-means выглядит так:

– на первом шаге задаём g произвольных центров и точность кластеризации σ. В качестве центров используем объекты множества Q.

– на втором шаге все объекты разделяем по критерию близости к одному из центров на g КЛ.

– третий шаг алгоритма связан с вычислением новых центров КЛ. Вычисляем координаты центров в пространстве R^m, как средние значения атрибутов объектов, входящих в состав сформированных КЛ.

Использование алгоритмов КА Fuzzy C-Means [6], для нашего случая, только тогда когда они являются обобщением алгоритма k-means. Основное отличие этого алгоритма от ранее сформулированных это когда КЛ представляются нечёткими множествами и каждый объект принадлежит КЛ с различной степенью принадлежности.

Пример 1. Рассмотрим результаты КА когда кластерная модель (КМ) МУТРМП представляет описание КЛ и принадлежность к одному из них каждого объекта из исходного множества. В случае небольшого числа объектов, характеризующихся двумя переменными, результаты можно изобразить посредством треугольников и четырехугольников, соответствующих объектам, и прямых линий [5, 6]. На рис. 1 представлены диаграммы Вена, характеризующие разделение объектов МУТТПРМП с двумя параметрами.

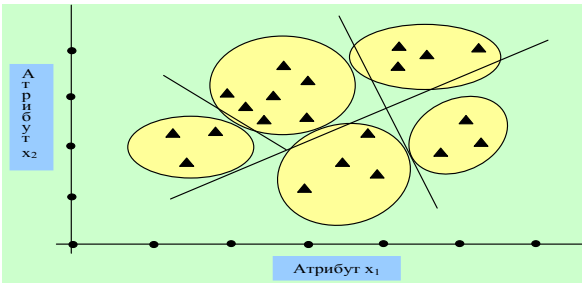

Рисунок 1 – Разделение объектов МУТТПРМП с двумя параметрами на КЛ

В случае нечёткой кластеризации (НК), принадлежность объекта к КЛ оценивали вероятностью или степенью принадлежности, а результат был представлен в виде таблицы, в которой строки соответствовали объектам, столбцы – КЛ, а в ячейках таблицы указывалась вероятность или степень принадлежности.

Существуют некоторые АК, которые строят структуры КЛ. Самый верхний уровень в структуре КЛ соответствует всему множеству объектов в виде единственного КЛ. На следующем уровне множество объектов делим на несколько КЛ, каждый из которых также делим на несколько КЛ. В принципе, построение иерархии может продолжаться до представления каждого объекта отдельным КЛ. Визуализация таких структур выполнялась в виде дендограмм [5].

АНАЛИЗ РЕЗУЛЬТАТОВ ИССЛЕДОВАНИЙ

КА данных МУТРМП представляет собой алгоритм сегментации (АС) службы AS [5, 6]. АС использует итерационные методы группировки полученных вариантов в НД, содержащих подобные характеристики, для выявления в них аномалий и создания прогнозов. По-

лученные МК определяют связи в наборе экспериментальных данных МУТРМП, который невозможно логически получить с помощью случайного наблюдения.

Пример 1. На диаграмме (рис. 2.) КЛ А соответствует рабочим, которые приезжают на работу на собственном транспорте или на транспорте предприятия, а КЛ В – рабочим, добирающимся до работы на общественном транспорте.

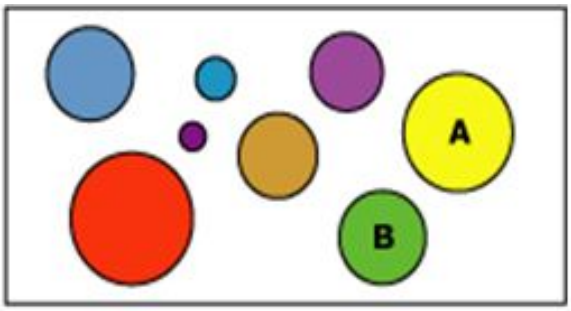

Рисунок 2 – КЛ А и КЛ В данных МУТРМП

В алгоритме кластеризации (АК) КА не назначался прогнозируемый столбец, необходимый для создания МК. АК обучает модель строго на основе связей, существующих в наборе данных(НД) и на основе КЛ, идентифицированных алгоритмом.

Пример 2. Рассмотрим НД с группой рабочих, имеющих похожие НД и представляющих собой КЛ данных. В НД базы данных, существует несколько таких КЛ. Анализируя столбцы, образующие КЛ, удалось более точно установить, какие записи в НД связаны друг с другом. АК сначала определил связи в НД МУТРМП и сформировал ряд КЛ на основе этих связей. После первого определения КЛ АК вычисляет, как КЛ представляют группирование точек, а затем повторно определил их группирование создание КЛ, которые представляют данные. АК последовательно выполняет этот процесс до тех пор, пока улучшит результаты и определение КЛ будет невозможно.

Провели настройку работы данного АК, выбирая конкретный метод объединения в КЛ, ограничивая максимальное количество КЛ или изменяя размер несущего множества, необходимого для создания МК МУТРМП. АК предназначен для использования в обучении МК и следует учитывать требования к конкретному алгоритму, в том числе к объему НД МУТРМП, и то, как эти данные используются.

Продолжая анализ данных, сформулируем требования для МК НД МУТРМП. Каждая МК должна содержать один числовой или текстовый столбец, который однозначно идентифицирует каждую запись. Каждая МК в нашем случае содержит один входной столбец и включает значение, которые используются для формирования кластеров. Ограничения на количество входных столбцов не налагаются, но, в зависимости от количества значений в каждом столбце, введение дополнительных столбцов может привести к увеличению времени на обучение МК. Необязательный прогнозируемый столбец для формирования МК этому алгоритму не потребовался, но была предусмотрена возможность добавления прогнозируемого столбца с данными любого другого типа. Значения в прогнозируемом столбце рассматривались как входные по отношению к МК. При просмотре МК в службах AS, КЛ отображались на схеме показывающей связи между КЛ, а также подробный профиль каждого КЛ, список атрибутов различий КЛ, и характеристики всего набора данных для обучения.

Для получения подробных сведения, просматривались МК с помощью алгоритма средств просмотра деревьев содержимого . Содержимое, сохраняемое для МК, включало распределение всех значений в каждом узле, вероятность каждого кластера и другую информацию. После обучения МК результаты сохранялись в виде набора закономерностей, которые исследовались и делались на их основе прогнозы с созданием запросов, возвращающих эти прогнозы для проверки соответствия новых данных обнаруженным КЛ, или предоставляющие описательные статистические данные о кластерах. Рассмотрим АК последовательностей данных(АКПД) МУТРМП, представляющий собой АКПД, предоставляемый службами MSQLS AS. АКПД КА можно использовать для просмотра данных, содержащих события, которые могут быть связаны с последовательностями. АКПД находит самые распространенные последовательности, выполняя группирование или кластеризацию идентичных последовательностей. Рассмотренный алгоритм напоминает АК, однако вместо поиска КЛ, содержащих похожие атрибуты, АКПД находит КЛ вариантов, содержащие похожие пути в последовательности.

Пример 3. Для оценки качества проведенных исследований регистрируемся на сайте оценки аттестации КА [6] МУТРМП. Применив в отношении таких данных АКПД, аттестационная группа находит КЛ технологических процессов рабочих мест (ТПРМ), для которых характерны похожие закономерности. ЛПР использует данные КЛ для анализа перемещения ТПРМ в рамках рассматриваемого предприятия, которые ближе всех связаны с другими и прогнозированными ТП РМ. АКПД МУТРМП – это гибридный алгоритм, сочетающий АК с анализом марковских цепей для определения КЛ и их последовательностей. (рис. 3)

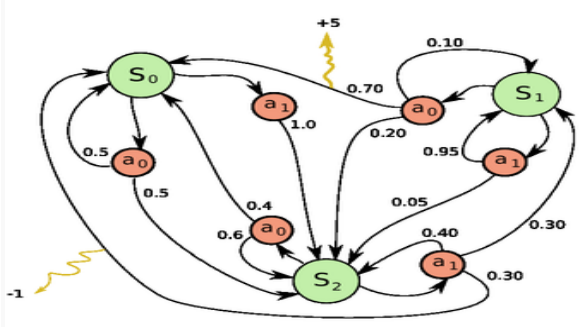

Рисунок 3 – Граф АКПД МУТРМП

Одной из особенностей АКПД является то, что используя данные последовательностей, которые представляют ряд событий или переходов между состояниями в НД. АКПД МУТРМП установил вероятность переходов, различий и расстояний, между всеми возможными последовательностями в НД и определил, какие последовательности лучше всего использовать в качестве входных данных для кластеризации. После создания АКПД списка вероятных последовательностей алгоритм использовал данные этой последовательности в качестве входных данных для EM-метода кластеризации. При подготовке входных данных для EM-метода кластеризации, рассмотрим данные, необходимые для МК последовательностей(МКПД). МКПД предназначен для использования разработки требований к конкретному алгоритму, в том числе к объему необходимых и используемых данных. Продолжая КА, рассмотрим МКПД , которая содержит описание самых распространенных последовательностей данных. При просмотре МК последовательности использовали службы AS, которые отображают КЛ, содержащие несколько переходов. После обучения МК результаты хранятся в виде набора шаблонов. Использовалось описание наиболее распространенных последовательностей в данных для прогноза следующего наиболее вероятного шага в новой последовательности. Но поскольку алгоритм включает другие столбцы, результирующую модель использовали, для определения связи между данными, включенными в последовательность, и данными, не включенными в нее и на основании этого делаем прогноз для конкретной группы РМ. Прогнозирующие запросы настраиваются для того, чтобы они возвращали переменное число прогнозов или описательные статистические данные. Продолжая анализ АК данных в КА рассмотрим алгоритм взаимосвязей Майкрософт (АВМ), который используется для механизмов выработки рекомендаций программ пользователям на основе имеющихся элементов. АВМ используют для анализа данных в имеющихся БД.

Модели взаимосвязей построены на НД, содержащих идентификаторы для отдельных вариантов и элементов этих вариантов. Идентифицируем группу элементов в этом варианте как набор элементов состоящих из рядов набора элементов и описывающих процесс группировки данных в вариантах. Правила, определяемые АВМ, были использованы для прогнозирования вероятных будущих мониторинговых аттестационных карт РМ на основе набора элементов в уже имеющихся БД.

Пример 4. На рис. 4 представлен ряд правил в наборе элементов.

Правило
Road Bottle Cage = Existing, Cycling Cap = Existing -> Water Bottle = Existing
Mountain-200 = Existing, Mountain Tire Tube = Existing -> HL Mountain Tire = Existing
Mountain-200 = Existing, Water Bottle = Existing -> Mountain Bottle Cage = Existing
Touring-1000 = Existing, Water Bottle = Existing -> Road Bottle Cage = Existing
Road-750 = Existing, Water Bottle = Existing -> Road Bottle Cage = Existing
Touring Tire = Existing, Sport-100 = Existing -> Touring Tire Tube = Existing

Рисунок 4 – Диаграмма правил описывающих процесс группировки НД в наборе элементов

Анализ диаграммы показывает, что АВМ потенциально может находить в наборе данных много правил и использует при этом только два параметра: поддержка и вероятность.

Пример 5. Если X и Y представляют два элемента, которые находятся в БД, то параметр несущего множества будет равен количеству вариантов в НД, содержащих сочетание элементов X и Y. Используя параметр несущего множества в сочетании с параметрами MINIMUM_SUPPORT, АВМ управляет количеством создаваемых наборов элементов. Параметр вероятности представляет часть вариантов в НД, содержащих X и Y. Используя параметр вероятности в сочетании с параметром MINIMUM_PROBABILITY, АВМ управляет количеством сформированных правил. АВМ прослеживает НД для поиска элементов, которые находятся в варианте совместно. Затем АВМ группирует в наборы элементов любые связанные элементы, найденные, как минимум, в количестве вариантов, определенных параметром MINIMUM_SUPPORT [6]. В дальнейшем АВМ формирует правила из наборов элементов, которые использовались для прогнозирования наличия элемента в БД. При подготовке данных для использования в АВМ, учитывались требования к конкретному алгоритму, включая то, сколько данных для него требуется и как эти данные используются. Для исследования моделей, использовали средство просмотра взаимосвязей АВМ. При просмотре АВМ в службах AS представлены корреляции под различными углами зрения, что позволяет лучше понять связи и правила, обнаруживаемые в НД. На панели средств просмотра предоставлена подробная классификация наиболее часто встречающихся сочетаний или наборов элементов. На панели правила представлен список правил, которые были выведены на основании данных, дополнительно приведены результаты вычисления вероятностей, а сами правила ранжированы по относительной важности. Средство просмотра сети зависимостей позволяет исследовать визуально, как связаны отдельные элементы. После обработки модели данных полученные правила и наборы элементов можно использовать для прогнозов. Прогнозы, выполняемые с помощью модели взаимосвязей, позволяют определить, какой элемент, скорее всего, обнаружится, если имеются сведения о присутствии указанного элемента, а сам прогноз может включать такую информацию, как вероятность, несущее множество или важность.

ВЫВОДЫ

Проведенный анализ данных мониторинга условий труда рабочих мест предприятий с применением кластерного анализа позволил выявить основные алгоритмы и модели обработки экспериментальных данных МУТРМ с целью повышения качества их обработки.

Процесс создания наборов элементов и расчеты значений корреляции могут потребовать значительных временных затрат. Несмотря на то что, в алгоритме правил взаимосвязей используются методы оптимизации, для экономии памяти ПЭВМ и ускорения обработки, следует: применять наборы данных, состоящие из небольшого количества отдельных элементов; задавать не слишком малые значения минимального размера набора элементов; группировать связанные элементы по категориям перед выполнением анализа данных.

ЛИТЕРАТУРА:

1. Algoritmy intellektual'nogo analiza dannyh (sluzhby Analysis Services – intellektual'nyj analiz dannyh). URL: http://msdn.microsoft.com/ru-ru/library/ms175595.aspx
2. Bazy znanij intellektual'nyh sistem /T.A. Gavrilova, V.F. Horoshevskij. – Spb.: Piter, 2000. – 384 s.
3. Barsegjan A.A. Analiz dannyh i processov: ucheb. posobie /A.A. Barsegjan, M.S. Kuprijanov, I.I. Holod, M.D. Tess, S.I. Elizarov. – 3-e izd., pererab. i dop. – SPb.: BHV-Peterburg, 2009. – 512 s.
4. Vvedenie v Data Mining na baze SQL Server 2008. URL: http://www.techdays.ru/videos/1376.html
5. Maklennen, Dzhemi. Microsoft SQL Server 2008: Data mining intellektual'nyj analiz dannyh: [per. s angl.] / Dzhemi Maklennen, Chzhaohujej Tang, Bogdan Krivat. – SPb.: BHV-Peterburg, 2009. –720 s.
6. Makarychev P.P, Afonin A.Ju /Operativnyj i intellektual'nyj analiz dannjah. – Penza: PGU, 2010. – 142 s.

Рецензент: *д.т.н., проф. Доровской В.А., Криворожский филиал Европейского университета.*