Использование технологии Data Mining для прогнозирования увольнения сотрудников

Predicting employee dismissal by using DataMining technology

Грабовец Р.А.
Ассистент кафедры Информационных технологий и компьютерных систем, Института информационных технологий и управления в технических системах, Севастопольского государственного университета
e-mail: roman13051991@rambler.ru

Grabovets R.A.
Assistant, Department Information Technologies and Computer Systems Department, Institute of Information Technologies and Control in Technical Systems, Sevastopol State University
e-mail: roman13051991@rambler.ru

Аннотация
В работе рассмотрена технология Data Mining. Проанализированы задачи, решаемые методами Data Mining. Выбран метод Data Mining для формирования данных, на основе которых можно создать универсальный семантический профиль уволенного сотрудника.
Ключевые слова: текучесть кадров, прогнозирование увольнения сотрудников, семантический профиль сотрудника, технология Data Mining, методы Data Mining, Data Mining.

Abstract
The article deals with the problem of reducing staff turnover. Methods for predicting the dismissal of an employee by using DataMining are described.
Key words: Staff turnover, employee dismissal forecasting, employee semantic profile, DataMining technology, DataMining methods, DataMining.

Введение

Непредвиденное добровольное увольнение сотрудников может принести предприятию большие сложности, поэтому руководители предприятий заинтересованы в снижении текучести кадров [1]. При этом, в существующих системах мониторинга работы сотрудников, проблеме снижения текучести кадров путем анализа сетевого трафика уделено слишком мало внимания [2]. В данной работе представлен вариант создания универсального семантического профиля сотрудника, содержащего предпочтения уволившихся сотрудников при использовании технологии Data Mining. Универсальный семантический профиль уволенного сотрудника является эталоном, с которым можно сравнивать семантические профили работающих сотрудников и, в случае совпадения, предупредить лицо, принимающее решение (ЛПР), о возможном намерении определенных сотрудников увольняться. Это предупреждение поможет ЛРП вовремя принять меры для удержания сотрудников на предприятии.

Постановка задачи

Найти множество R, на основе которого будет создан универсальный семантический профиль уволенного сотрудника. Множество R состоит из множества элементов семантических профилей сотрудников, уволенных по собственному желанию

(не по инициативе работодателя), не принадлежащих множеству элементов семантических профилей работающих сотрудников.

Выбор метода Data Mining

Для создания универсального семантического профиля уволенного сотрудника необходимо обработать большой объем неструктурированных данных.

Data Mining – это технология, которая помогает выявлять скрытые связи в базах данных существенных размеров. Целью Data Mining является нахождение полезных и доступных для понимания данных, которые не могут быть найдены обычными методами [3].

Data Mining используется для поиска часто встречающихся наборов товаров в корзине покупателей, для обнаружения и предотвращения новых неполадок на различных узлах телекоммуникационных сетей и во многих других сферах деятельности [4]. На рис. 1 схематично показано применение технологии Data Mining.

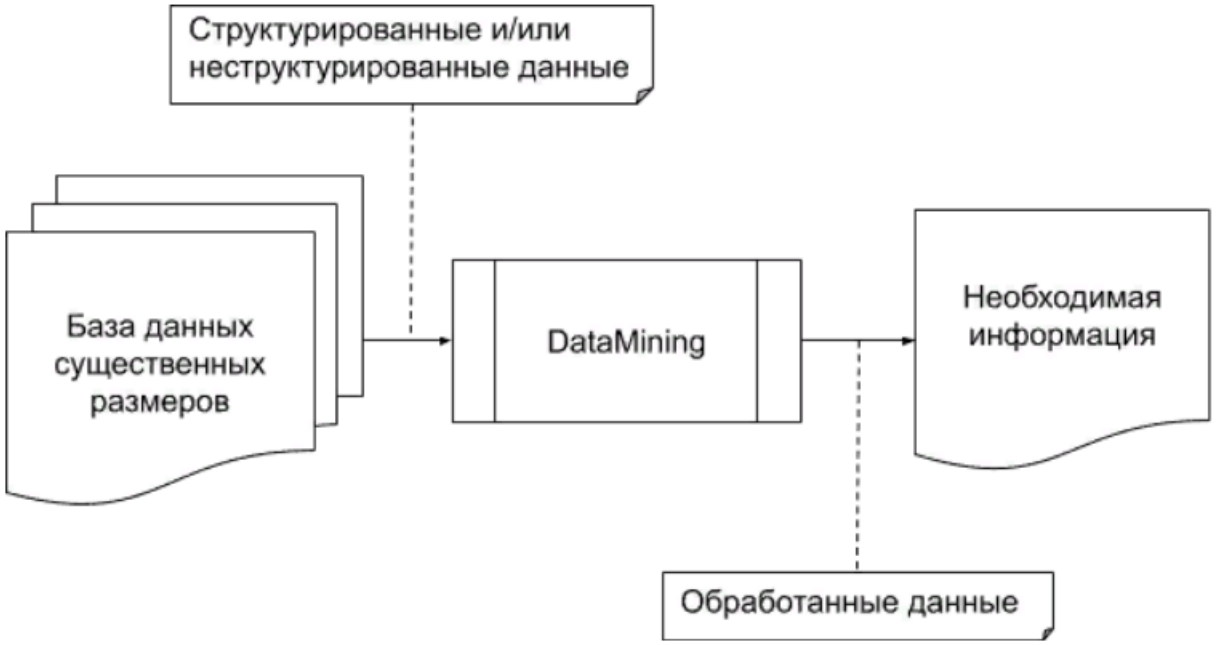

Рис. 1. Общая схема применения технологии Data Mining

Сценарии применения методов Data Mining могут быть самыми различными и могут включать сложную комбинацию разных методов. На рис. 2 показаны этапы Data Mining.

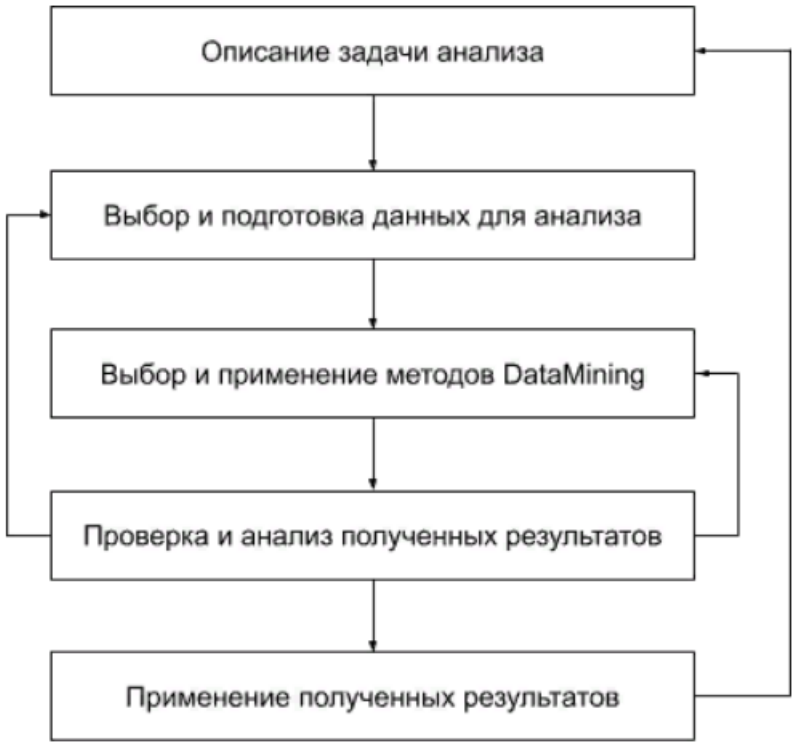

Рис. 2. Этапы Data Mining.

Согласно В.А. Дюку, технология Data Mining может найти следующие типы закономерностей:

1. Ассоциация − применяется, когда несколько событий связаны между собой. Например, 48% купивших шоколадный батончик взяли и сок, а если есть скидка на такой комплект, то сок покупают в 84% случаев.
2. Классификация − выявление черт, которые будут характеризовать группу, к которой принадлежит объект, на основе обучения на уже классифицированных объектах.
3. Кластеризация − отличается от классификации тем, что группы заранее не известны и средства Data Mining самостоятельно выявляют различные однородные группы данных.
4. Последовательность − применяется при существовании цепочки событий, связанных во времени. Например, в 53% случаев после приобретения квартиры в течение месяца приобретается ванна, а в 83% в течение года – кухонная мебель.
5. Прогнозирование − создание или нахождение шаблонов, которые будут истинно показывать тенденции поведения необходимых показателей по временным рядам. При помощи них можно предсказать поведение системы в будущем.

Согласно постановке задачи, необходимо выявить тип закономерности − классификацию, поскольку необходимо найти множество R, которое будет показывать сотрудников из множества $W_{раб.}$ ($W_{раб.}$ − множество семантических профилей работающих сотрудников), имеющих намерение увольняться, на основе обучения на уже классифицированных (уволенных) сотрудниках из множества $W_{ув.}$ ($W_{ув.}$ − множество семантических профилей уволенных сотрудников).

Основные задачи, решаемые методами Data Mining, могут быть сведены к трем направлениям:

1. Задача классификации и регрессии − позволяет определить по известным характеристикам объекта значение некоторого его параметра.
2. Задача кластеризации − применяется для поиска независимых групп (кластеров) однородных объектов и их характеристик во множестве анализируемых данных.
3. Задача поиска ассоциативных правил − применяется для обработки больших массивов неструктурированных данных.

Задачи, для решения которых целесообразно использовать методы поиска ассоциативных правил:

1. Сокращения объемов неструктурированных данных путем удаления из дальнейшего рассмотрения избыточных транзакций, исключение которых не повлияет на качество синтезируемых правил и моделей.
2. Выявления интересных правил, позволяющих извлекать новые, актуальные знания на основе имеющихся неструктурированных данных.
3. Построения моделей на основе больших массивов неструктурированных данных для решения практических задач кластеризации, классификации и прогнозирования данных [5].

Для решения поставленной задачи целесообразно использовать методы поиска ассоциативных правил, поскольку:

1. Необходимо сократить объем неструктурированных данных путем удаления избыточных элементов семантических профилей, количество вхождений которых во множества $W_{раб.}$ или $W_{ув.}$ меньше определенного значения (например, если элемент семантических профилей входит менее чем в половину семантических профилей $W_{раб.}$ или $W_{ув.}$, то этот элемент не подлежит рассмотрению).
2. Необходимо сформировать набор элементов семантических профилей, отражающих модель поведения увольняющегося сотрудника на основе больших массивов неструктурированных данных (множеств $W_{раб.}$ и $W_{ув.}$) для классификации исследуемых сотрудников, и выдаче ЛПР информации, показывающей часть сотрудников, предположительно имеющих намерение увольняться.

Соответственно, для поиска множества R наиболее целесообразно применять методы поиска ассоциативных правил, относящиеся к базовым методам Data Mining, основанным на переборе [5], в частности, алгоритм поиска ассоциативных правил − Apriori.

Основным достоинством алгоритмов поиска ассоциативных правил, к которым относится алгоритм Apriori, является простота, как с точки зрения понимания, так и реализации.

Выводы

Задачей применения методов Data Mining является поиск полезных и актуальных данных путем обработки большого объема информации (структурированных и неструктурированных данных).

Для создания универсального профиля уволенного сотрудника предложено применение алгоритма поиска ассоциативных правил − Apriori.

Литература

1. Прогнозирование увольнения сотрудников на основе анализа и обработки сетевого трафика Грабовец Р.А. Журнал технических исследований. − 2020. − Т. 6. − № 1. − С. 37−42.

2. Grabovets R., Mikhailova E. Analysis of employee internet traffic for the reduction of staff turnover // ProfMarket: Образование. Язык. Успех. —2018. — С. 351-352.
3. *Савченко Л.М., Бежитский С.С.* DataMining и области его применения // Актуальные проблемы авиации и космонавтики. − 2015. − Т. 1. − № 11. − С. 611−613.
4. Анализ данных и процессов: учеб. Пособие / А. А. Берсегян, М. С. Куприянов, И. И. Холод, М. Д. Тесс и др. Санкт-Петербург: БХВ-Петербург, 2009. − 512 с.
5. Технологии анализа данных: DataMining, VisualMining, TextMining, OLAP: учеб. пособие. 2-е изд. / А. А. Барсегян, М. С. Куприянов, В. В. Степаненко, И. И. Холод. −Санкт-Петербург: БХВ-Петербург. – 2007. − 384 с.