

Министерство образования Республики Беларусь
Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК 004.42:[004.6+614.39]

Бобровский
Константин Сергеевич

Алгоритмы и ПО системного анализа в области сервисного обслуживания
населения

АВТОРЕФЕРАТ

на соискание степени магистра технических наук
по специальности 1-40 81 02 Технологии виртуализации и облачных
вычислений

Научный руководитель

Татур Михаил Михайлович

профессор, доктор технических наук

Минск 2020

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Объект исследования

Объектом исследования являются алгоритмы системного анализа классификации большого объема данных в области сервисного обслуживания населения, в частности, здравоохранения.

Предмет исследования

Предметом исследования являются алгоритмы интеллектуальной системы классификации большого объема данных в области сервисного обслуживания населения, в частности, здравоохранения.

Цель исследования

Произвести анализ существующих алгоритмов, разработать программное обеспечение с использованием алгоритмов классификации большого объема данных для данного направления.

Задачи исследования

1. Обзор существующих аналогов и выявление сильных сторон.
2. Анализ алгоритмов классификации больших объемов данных.
3. Разработка программного обеспечения, способного классифицировать большой объем данных.

Научная новизна работы

Научная новизна работы состоит в том, что в ходе разработки были использованы на практике альтернативные способы классификации, ранее не использовавшиеся в данной области.

Апробация

Основные положения работы докладывались на 56-ой научной конференции аспирантов, магистрантов и студентов БГУИР (г. Минск, 2020 г.).

Публикации

Бобровский, К. С. Алгоритмы и ПО системного анализа в области сервисного обслуживания населения / К. С. Бобровский // Компьютерные системы и сети: 56-я научная конференция аспирантов, магистрантов и студентов, Минск, 21-22 апреля 2020 г. — Минск: Белорусский государственный университет информатики и радиоэлектроники, 2020.

По материалам диссертационной работы опубликована статья в журнале: Бобровский К. С., Придыбайло Н. А. Алгоритмы и программное обеспечение системного анализа в области сервисного обслуживания населения // Молодой ученый. – 2019. – №46. – С. 13-16.

Библиотека БГУИР

ВВЕДЕНИЕ

В настоящее время системный анализ широко применяется во многих отраслях деятельности человека. Ни одна организация не может обойтись в своей работе без применения компьютеров, которые с успехом заменяют работу, выполнявшуюся ранее вручную, повышая эффективность работы любой организации. Сфера использования системного анализа в настоящее время настолько широка, что нет такой области, где применение его было бы нецелесообразным.

В данной диссертации рассмотрена проблема быстрого и точного предоставления данных пользователю. Также произведен системный анализ и разработан интеллектуальный модуль, который способен быстро и масштабируемо классифицировать большой объем данных.

Рассмотрена такая сфера сервисного обслуживания населения, как здравоохранение, так как в современном мире здравоохранение играет огромную роль.

Система здравоохранения обязана анализировать большое количество данных о пациентах, что помогает получить представление и помочь в прогнозировании заболеваний. Эта система должна быть разумной, чтобы прогнозировать состояние здоровья путем анализа образа жизни пациента, истории его физического здоровья и социальной активности. В этом контексте интеллектуальные системы здравоохранения стали незаменимыми инструментами в процессах принятия решений в секторе здравоохранения. Их главная цель – обеспечить доступность ценной информации в нужное время, обеспечивая качество информации, достоверность и конфиденциальность.

Поскольку люди используют социальные сети для понимания состояния своего здоровья, система рекомендаций по здоровью очень важна для получения таких результатов, как рекомендации по диагностике, медицинскому страхованию, методам лечения на основе профиля здоровья пациента.

В работе отражены недавние исследования, нацеленные на использование больших объемов медицинских данных при объединении мультимодальных данных из разнородных источников, что снижает рабочую нагрузку и стоимость медицинского обслуживания. В секторе здравоохранения аналитика больших данных с использованием рекомендательных систем играет важную роль с точки зрения процессов принятия решений в отношении здоровья пациента.

Потребность в информации о здравоохранении кажется безграничной. Это неудивительно, если учесть все заинтересованные стороны, которые хотят повлиять на направление развития здравоохранения. Правительство, страховщики, руководство больницы, специалисты и пациенты имеют свои

особые интересы в сфере здравоохранения. Эти особые интересы приводят к большому разнообразию вопросов: «Каково использование ресурсов?», «Каковы результаты медицинской помощи?», «Какое качество ухода?». Чтобы ответить на эти вопросы, требуется информация, потому что без информации невозможно провести эффективный контроль. Обработка информации требует сбора данных.

Многие сотрудники больницы обычно занимаются обработкой данных, медицинские специалисты тратят много времени на запись этих данных. Более того, растет потребность в дополнительной информации: амбулаторная диагностика – это только один пример, требующий данных. Степень детализации в большинстве записей высока.

Одной из важнейших жизненных ценностей человека является здоровье, его физическое и психическое благополучие. Здоровье – это не только категория медицинской науки, но и экономическая категория, поскольку только человек с хорошими показателями здоровья может способствовать повышению производительности труда.

Здравоохранение является неотъемлемым элементом международного развития. Эффективная система здравоохранения может внести существенный вклад в экономику страны, ее развитие и индустриализацию. Здравоохранение традиционно считалось важным фактором, определяющим общее физическое и психическое здоровье и благополучие людей во всем мире.

Данная тема на сегодняшний день очень актуальна, так как каждый день человек сталкивается с проблемой выбора чего-либо (выбором поликлиники, выбором аптеки, выбором лекарства). Если предоставить человеку интеллектуальную систему корректной классификации этих данных, то это значительно сэкономит время человеку, а также упростит его жизнь.

СОДЕРЖАНИЕ РАБОТЫ

Общий объем магистерской диссертации составляет 100 страниц, включая 28 иллюстраций, 3 таблицы, библиографический список из 37 наименований, 2 приложения.

В общей характеристике работы сформулированы цель и задачи исследования, даны сведения об объекте и предмете исследования, приведены апробации и публикации результатов.

Во введении дано обоснование актуальности работы, проведен краткий обзор сферы применения разработанного программного обеспечения. Показана потребность в подобной системе для здравоохранения.

В первой главе произведен анализ предметной области. В первом разделе рассмотрены типы и представление медицинских данных, даны определения терминам кодирование, мнемонический код применительно к сфере здравоохранения. Во втором разделе показаны существующие системы классификации в здравоохранении такие, как МКБ, ICPC, DSM, ICD. Приведена таблица с кодами заболеваний. В третьем разделе проанализирована интеллектуальная система HRS основанная на использовании метода глубокого обучения, показана архитектура и основные сильные и слабые стороны данной системы.

Во второй главе приведено описание теоретической базы исследования. В первом разделе описан метод ближайших соседей kNN, приведены основные формулы, показаны рисунки с примерами классификации данных. Во втором разделе приведен метод Fuzzy kNN, основные формулы и изображения, которые характеризуют данный метод. Показана модель Hybrid-spill tree. В третьем разделе рассматривается фреймворк Apache Spark. Приводятся пояснения, почему именно данный фреймворк необходимо использовать.

В третьей главе приводится подробное описание разработанной системы. В первом разделе описывается общая архитектура системы. Во втором показан процесс выбора лучших алгоритмов для данного программного обеспечения, приводится сравнительный анализ лучших модификаций алгоритмов. В третьем разделе показаны основные составляющие выбранных алгоритмов, приведены подробные схемы работы, а также псевдокод. В четвертом разделе показываются и обосновываются результаты работы программного обеспечения. В пятом отображены инструкции для запуска ПО. В шестом разделе указано используемое аппаратное и программное обеспечение.

В заключении приведены основные выводы, а также показан вектор развития данной системы и будущая работа.

ЗАКЛЮЧЕНИЕ

В рамках работы над диссертацией проведен комплексный сравнительный анализ существующих программных средств обработки, анализа и классификации большого объема данных в сфере сервисного обслуживания населения, а именно в сфере здравоохранения, обоснована необходимость проведения собственной разработки.

Рассмотрена такая сфера сервисного обслуживания населения, как здравоохранение, так как в современном мире здравоохранение играет огромную роль.

В рамках работы над диссертацией были проанализированы и реализованы в виде ПО кастомные модификации алгоритмов классификации медицинских данных, что и является основным результатом работы.

Это два подхода MapReduce для ускорения алгоритма FkNN в задачах с большими данными. На рисунках отображены результаты работы программного обеспечения, которые показывают эффективность использования. Из-за разработки и использования технологий больших данных возможно выполнение с очень большими наборами данных.

Модификации GAHS-FkNN и LHS-FkNN достигают статистически равных результатов с точки зрения точности. С одной стороны, алгоритм LHS-FkNN демонстрирует очень высокую масштабируемость в зависимости от доступных аппаратных средств, а также высокую точность результатов. С другой стороны, алгоритм GAHS-FkNN меньше зависит от аппаратных ресурсов, но больше подвержен влиянию большого числа признаков. Таким образом, использование LHS-FkNN рекомендуется, когда у нас есть мощное оборудование в соответствии с проблемой, которую мы хотим решить, и если количество признаков велико. Использование GAHS-FkNN рекомендуется, когда количество признаков не слишком велико и у нас есть аппаратные ограничения.

Функциональность проекта была реализована на языке Scala, в качестве фреймворка использовался Apache Spark. Благодаря данному фреймворку удалось использовать преимущества операций в памяти, чтобы одновременно классифицировать большое количество данных.

Достоинством данной работы является тот факт, что спроектированные и реализованное ПО позволяет быстро и точно производить классификацию больших данных. В работе брались медицинские выборки данных с наборами до 11 миллионов экземпляров. Результаты показали улучшение времени выполнения и точности по сравнению с литературными алгоритмами.

Подводя итоги диссертационного исследования, стоит отметить, что все поставленные цели и задачи были достигнуты. При этом в данной тематике

есть еще много материала для дальнейших исследований.

В завершение, стоит отметить, что, несмотря на то, что данная работа рассматривает только одно направление классификации, алгоритмы, реализованные в ходе выполнения работы по исследованию, могут быть применены и в других областях.

Проект разработан в полном объеме. В качестве усовершенствования разработанного ПО, в качестве будущей работы, можно рассматривать решение проблемы дисбаланса классов с помощью эволюционных методов недостаточной выборки, способных обрабатывать большие наборы данных.

Библиотека БГУИР

СПИСОК ПУБЛИКАЦИЙ СОИСКАТЕЛЯ

1–А. Бобровский, К. С. Алгоритмы и ПО системного анализа в области сервисного обслуживания населения / К. С. Бобровский // Компьютерные системы и сети: 56-я научная конференция аспирантов, магистрантов и студентов, Минск, 21-22 апреля 2020 г. — Минск: Белорусский государственный университет информатики и радиоэлектроники, 2020.

2–А. Бобровский К. С., Придыбайло Н. А. Алгоритмы и программное обеспечение системного анализа в области сервисного обслуживания населения // Молодой ученый. – 2019. – №46. – С. 13-16.

Библиотека БГУИР