УДК 004.85:616-7

ПРОВЕДЕНИЕ ЭКСПЕРИМЕНТОВ ПО ДИАГНОСТИКЕ В МЕДИЦИНЕ НА ОСНОВЕ МЕТОДОВ КЛАССИФИКАЦИИ И АНАЛИЗ РЕЗУЛЬТАТОВ

Паршин И. А., Шевчук И. А.
к.т.н., профессор Рувинская Виктория Михайловна
Одесский Национальный Политехнический Университет, УКРАИНА

АННОТАЦИЯ. Проведены эксперименты в области диагностики на основе машинного обучения с использованием методов классификации *decisiontree*и *k-nearestneighbors*на основе существующего набора данных, связанного с диагностикой в области рака груди. Анализ показал высокую достоверность диагностирования и возможность использования предлагаемого подхода при разработке программных систем диагностики для конечного пользователя.

Введение. Для решения сложных медицинских проблем, таких как диагностика в случаях, когда требуется сложный комплекс анализов, содержащих множество числовых данных, от врача требуется много внимания, и значительно повышается риск врачебной ошибки. В связи с этим для диагностики целесообразно использовать методы машинного обучения и программные средства, однако многие существующие системы узко специализированы, к примеру, описанные в [1, 2].

Целью работы является повышение достоверности диагноза в медицине при решении широкого класса задач на базе использовании методов классификации.

Основная часть работы. Эксперименты проводились на базе датасета, взятого из *MachineLearningRepository*[4], с реальным набором данных (примеров), связанным с раком груди. Набор включает 8 групп (698 примеров), в каждой из которой содержится информация о пациентах, собранная в разное время. В каждом примере содержится 11 входных признаков и метка, означающая, здоров пациент или болен. При обучении и тестировании модели бинарной классификации на примерах были использованы два метода: *decisiontree* и *k-nearestneighbors*[3]. Для реализации методов на языке программирования *Python* использовались следующие библиотеки: *NumPy* – для работы с многомерными массивами, *Pandas* – для обработки и анализа данных, *Sklearn* (*scikit-learn*) – для работы с алгоритмами машинного обучения.

Результаты обучения по методу *decisiontree* в виде дерева с узлами решений показан на рис. 1. В алгоритме *k-nearestneighbors* точность зависит от выбора *k* – количества ближайших соседей. Так для рассматриваемого примера видно, что оптимальным значением является 8 ближайших соседей (рис. 2).

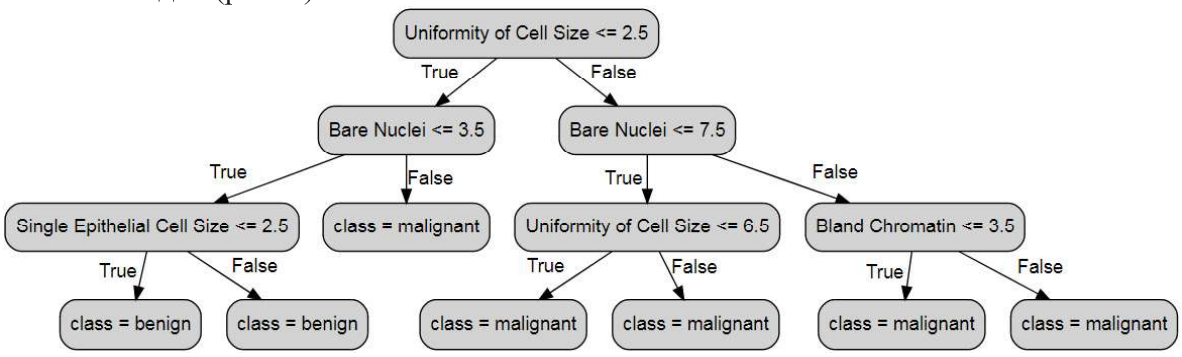

Рисунок 1. Дерево решений

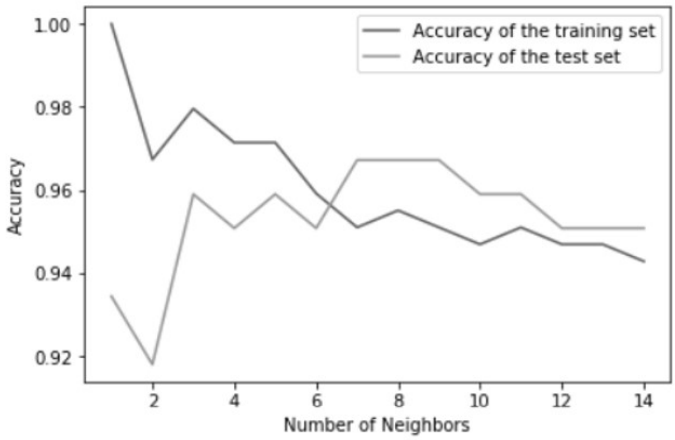

Рисунок 2. График точности классификации при разных *k*

В таблице 1 представлена матрица ошибок и достоверность классификациина тестовых данных для обоих методов. Полученные результаты показывают, что метод *k-nearestneighbors*для рассматриваемого набора данных показал лучшую достоверность и умение более правильно диагностировать случаи, когда пациенты здоровы (8 против 30 ошибок типа FN, когда пациент здоров (+), а ему ставится диагноз, что он болен (-)).

Таблица 1. Достоверность диагностирования

k-nearestneighbors			+		-	
Реальные метки (y)	+	258	TruePositive	249	FalseNegative (FN)	8
	-	74	FalsePositive (FP)	1	TrueNegative	74
Точность	0,972892					
decisiontree			+		-	
Реальные метки (y)	+	258	TruePositive	228	FalseNegative(FN)	30
	-	74	FalsePositive (FP)	0	TrueNegative	74
Точность	0,909639					

Выводы. В представленной работе были проведены эксперименты по диагностированию с помощью методов классификации на реальном наборе данных и проведен анализ результатов. На тестовом наборе метод *decisiontree*определял правильный диагноз с достоверностью 90.9%, а *k-nearestneighbors*– с достоверностью 97.2%. Таким образом, показано, что предложенный подход по диагностированию может быть использован при разработке программных систем для конечных пользователей, а именно, врачей для помощи им при постановке диагноза.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Machine Learning in Medicine [Электронныйресурс]. – Режимдоступа: URL: http://circ.ahajournals.org/content/132/20/1920.short. – Название с экрана.

2. Пример применения машинного обучения в медицинской диагностике [Электронный ресурс]. – Режим доступа: URL: https://sites.google.com/site/libvmr/home/practice/primer-primenenia-masinnogo-obucenia-v-medicinskoj-diagnostike. – Название с экрана.

3. MachineLearningRepository [Электронный ресурс]. – Режим доступа: URL: http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29. – Название с экрана.

4. Classificationalgorithms [Электронный ресурс]. – Режим доступа: URL: http://www.saedsayad.com/classification.htm. – Название с экрана.