GSM // Спецтехника и связь, 2010. – №2-3. – С.65-67.

ОЦЕНКА КАЧЕСТВА АЛГОРИТМОВ НА ОСНОВЕ МАТРИЦЫ НЕТОЧНОСТЕЙ

Прилепов Евгений Валерьевич
Аспирант кафедры компьютерных наук, ГУТ, г. Киев

АНОТАЦИЯ

Предметом изучения в статье являются процессы обработки данных, полученных в результате действия алгоритма кластеризации данных для решения прикладной задачи качественной оценки воздействия алгоритма. **Целью работы** является определение основных понятий оценки качества алгоритмов. Описание численной оценки качества алгоритмов, точности и полноты. **Задача**: качественная оценка действия алгоритмов кластеризации и обработки информации на основе диагностического тестирования. Используемым **методом** является: информационно-аналитический метод анализа предварительно обработанных данных, который представлен в виде модели табличного представления матрицы неточностей. **Выводы.** В результате было получено способ проверки действия алгоритмов обработки данных и численную оценку качества обработанной информации.

The subject of the study in this article is the processes of data processing, obtained as a result of the algorithm of data clustering to solve the application problem of qualitative evaluation of the algorithm. **The purpose** of the work is to determine the basic concepts of the quality evaluation of the algorithms. Description of the numerical quality evaluation of the algorithms, accuracy and completeness. **The task**: qualitative evaluation of the clusterization algorithms of information processing based on diagnostic testing. **The method** is: an information analytical method for analyzing pre-processed data presented in the form of a table represented in confusion matrix. **Conclusions.** As a result, were obtained a method for verifying the operation of data processing algorithms and a numerical evaluation of the quality of the processed information.

Ключевые слова: алгоритм, анализ, обработка, качество, оценка, метрика.

Keywords: algorithm, analysis, processing, quality, evaluation, metric.

Введение и постановка задачи

В области машинного обучения и, в частности, проблемы статистической классификации, матрица неточности, является специфическим макетом таблицы, которая позволяет, как правило визуализировать эффективность алгоритма. Каждый рядок матрицы представляет собой экземпляр в прогнозируемом классе, тогда как каждый столбец отображает экземпляры в фактическом классе (или наоборот). Использование этого подхода позволяет легко выяснить, перепутала система два класса или нет.

Это специальный тип таблиц для непредвиденных случаев, с двумя параметрами (фактическими и прогнозируемыми) и идентичными наборами классов в обоих измерениях (каждая комбинация размера и класса является переменной в таблице непредвиденных ситуаций).

Представление данных при их хранении и обработке требует решения трех основных задач:

- Определить способы представления элементарных данных.
- Определить способы объединения данных в структуры.
- Установить способы размещения информации.

На данный момент выделяют три уровня представления данных: концептуальный, логический и физический. На концептуальном уровне определяется общая структура информационного массива – она называется моделью данных. Сейчас используются несколько моделей данных: иерархическая, сетевая, реляционная, объектно-ориентированная. Согласно выбранной модели данных строится информационная система, в которой данные будут храниться, а также программы, ведущие их обработку. Логический уровень определяет способы представления элементарных данных, их перечень при объединении в структуру, а также характер связей между ними в рамках выбранной модели данных. Физический уровень определяет форматы размещения созданной логической структуры данных на носителях информации. Представление данных является важным фактором, обеспечивающим компактный способ записи информации при хранении и быстрый доступ к нужным данным при их использовании.

Целью работы является определение основных понятий оценки качества алгоритмов. Описание численной оценки качества алгоритмов, точности и полноты. Предоставить содержательную информацию относительно матрицы неточности и способа ее использования для оценки качества алгоритмов.

Научная новизна - использование информационно-аналитического метода анализа предварительно обработанных данных на основе матрицы неточностей, результатом действия которого является качественная оценка действия алгоритма предварительной обработки данных.

Основная часть

Численная оценка качества алгоритма.

Один из вариантов решения этого вопроса состоит в том, чтобы обучать классификатор на специально подготовленном, сбалансированном массиве данных. Недостаток этого решения в том, что классификатор теряет часть информации относительно частоты данных [1 с. 18]. Другой выход заключается в изменении подхода к формальной оценке качества.

Точность и полнота.

Точность и полнота является метриками, используемых при оценке в основном алгоритмов выборки информации [2 с. 22]. Иногда они используются сами по себе, а иногда в качестве базиса для производных метрик, таких как F-Мера или R-Точность. Точность системы в пределах класса — представляет собой долю данных, что действительно принадлежат данному классу в отношении всех данных, которые система отнесла к этому классу. Полнота системы — это доля найденных классификатором данных, принадлежащих классу относительно всех документов этого класса в тестовой выборке. Эти значения рассчитываются на основании таблицы 1, которая составляется для каждого класса отдельно.

Таблица 1. Таблица сопряженности.

Категория		Экспертная оценка	
		Положительная	Отрицательная
Оценка системы	Положительная	Истинно положительный (ИП)	Ложноположительный (ЛП)
	Отрицательная	Ложно-негативный (ЛН)	Истинно отрицательный (ИО)

В математической статистике, таблица сопряженности, имеет вид матрицы, которая показывает распределение частоты переменных. Она широко используются в научных исследованиях. Таблица позволяет увидеть основную картину взаимосвязи между двумя переменными, а также помогает найти взаимодействие между ними [3 с. 86].

В таблице содержится информация о том, сколько раз система приняла верное и сколько раз неверное решение по данным заданного класса.

Согласно с таблицей сопряженности точность метрики рассчитывается по формуле 1.

$$\text{Точность} = \frac{\text{ИП}}{\text{ИП} + \text{ЛП}} \quad (1)$$

Полнота метрики рассчитывается по формуле 2.

$$\text{Полнота} = \frac{\text{ИП}}{\text{ИП} + \text{ЛН}} \quad (2)$$

Матрица сопряженности.

Матрица неточности или матрица сопряженности — это матрица размера N на N, где N — это количество классов. Обычно используется для описания и наведения цифр относительно классификационной модели [4 с. 3]. Столбцы этой матрицы резервируются по экспертным решениям, а строки по решениям классификатора. Когда мы классифицируем документ с тестовой выборки мы увеличиваем значение числа, стоящего на пересечении строки класса, который вернул классификатор и столбца класса к которому действительно относится значения, мерная матрица и представляет собой соотношение между фактическим значением и прогнозируемым. В случае, когда количество классов относительно мала (не более 150 классов), этот подход позволяет достаточно точно представить результаты работы классификатора.

Если классификатор предусматривает наличие проблем или аномалий в данных или если фактические данные являются аномальными, то атрибут считается истинно положительным (ИП). В случае, если модель предполагает, что рабочие данные или записи являются аномальными, а данные или запись на самом деле является аномалией, то атрибут может быть использовано в качестве индикатора для показателя ИП системы.

Если классификатор предусматривает, что в данных нет проблемы и что эти данные на самом деле нормальные, то атрибут считается истинно отрицательным (ИО). В случае, если модель предполагает, что рабочие данные или записи являются нормальными, а данные или запись на самом деле являются нормальными, то атрибут может быть использовано в качестве индикатора для показателя ИО системы.

Если классификатор предусматривает, что существует проблема в данных или аномалия, в то время как данные на самом деле являются нормальными и не содержат в себе аномалии, то атрибут считается ложноположительным (ЛП). В случае, если модель предполагает, что рабочие данные или записи имеют проблему или аномалию, то атрибут может быть использовано в качестве индикатора для показателя ЛП системы. Большинство моделей выявления аномалий пытаются уменьшить показатель ЛП до максимально низкого уровня.

Наконец, если классификатор предусматривает, что в данных нет проблемы или аномалии, в то время как в данных действительно есть проблема или аномалия, тогда атрибут считается ложно-негативным (ЛН).

В случае если рабочие данные или записи не имеют проблемы или аномалии, в то время как это не так, то атрибут может быть использовано в качестве индикатора для показателя ЛН системы. В моделях выявления аномалий этот показатель указывает на уровень отказов у модели и общего качества функции обнаружения аномалий [5 с. 90].

Имея это в виду, общая точность$M_{accuracy}$ — может быть вычислена с использованием следующей формулы 3:

$$M_{accuracy} = \frac{\text{ИП} + \text{ИО}}{x} \quad (3)$$

где: — x означает общее количество записей в наборе данных тестирования.

Частота ошибочной классификации модели M_{er}, может быть вычислена по следующей формуле 4:

$$M_{er} = \frac{\text{ЛП} + \text{ЛН}}{x} \quad (4)$$

Частота модели M_f, может быть вычислена по следующей формуле 5:

$$M_f = \frac{\text{ИП}}{A_n} \quad (5)$$

где: A_n — количество фактических негативных записей x.

Коэффициент ложной позитивности модели M_{fpr}, рассчитывается по следующей формуле 6:

$$M_{fpr} = \frac{\text{ЛП}}{A_n} \quad (6)$$

Специфика модели M_s, рассчитывается по следующей формуле 7:

$$M_s = \frac{\text{ИО}}{A_n} \quad (7)$$

Положительная точность прогнозирования модели M_{ppv}, может быть рассчитана по следующей формуле 8:

$$M_{ppv} = \frac{\text{ИП}}{A_p} \quad (8)$$

где: A_p — количество фактических положительных записей x.

Негативная точность прогнозирования модели M_{npv}, может быть рассчитана по следующей формуле 9:

$$M_{npv} = \frac{\text{ИО}}{A_a} \quad (9)$$

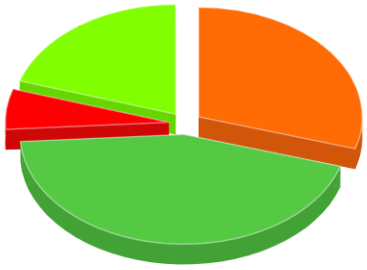
где: A_a — количество фактических аномальных записей x.

Распространенность модели M_p, может быть рассчитана по следующей формуле 10:

$$M_p = \frac{A_p}{x} \quad (10)$$

Пример использования матрицы сопряженности.

Рисунок 1. — графическое изображение аномальных данных, полученных после предварительной обработки.



Истинно положительный (ИП) 65 Истинно отрицательный (ИО) 97

Ложноположительный (ЛП) 13 Ложно-негативный (ЛН) 44

Точность (Accuracy) рассчитывается как количество всех правильных прогнозов, разделенных на количество набора данных. Лучшая точность — 1.0, худшая — 0.0.

$$M_{accuracy} = \frac{\text{ИП} + \text{ИО}}{\text{ИП} + \text{ИО} + \text{ЛП} + \text{ЛН}} = \frac{65 + 97}{65 + 97 + 13 + 44} = 0.74$$

Частота модели (Error rate) или как ее еще называют: коэффициент ошибок или коэффициент ошибочной классификации, описывает частоту ложной классификации.

$$M_{er} = \frac{\text{ЛП} + \text{ЛН}}{\text{ИП} + \text{ИО} + \text{ЛП} + \text{ЛН}} = \frac{13 + 44}{65 + 97 + 13 + 44} = 0.26$$

Специфика (True negative rate) вычисляется как количество истинных негативных прогнозов, разделенных на общее количество негативных. Лучшая специфика — 1.0, а худшая — 0.0.

$$M_s = \frac{\text{ИО}}{\text{ИО} + \text{ЛП}} = \frac{97}{97 + 13} = 0.88$$

Положительная точность прогнозирования (Positive predictive value) определяется как количество правильных положительных прогнозов, разделенных на количество положительных прогнозов. Лучшая точность — 1.0, а худшая — 0.0.

$$M_{ppv} = \frac{\text{ИП}}{\text{ИП} + \text{ЛП}} = \frac{65}{65 + 13} = 0.83$$

Негативная точность прогнозирования (Negative predictive value) определяется как количество правильных негативных прогнозов, разделенных на количество негативных прогнозов. Лучшая точность — 1.0, а худшая — 0.0.

$$M_{npv} = \frac{\text{ЛО}}{\text{ЛН} + \text{ЛО}} = \frac{97}{44 + 97} = 0.69$$

Коэффициент ложной позитивности модели (False positive rate) вычисляется как количество неправильных положительных прогнозов, разделенных на количество отрицательных. Лучший случай составляет при значении — 0.0, худший при — 1.0.

$$M_{fpr} = \frac{\text{ЛП}}{\text{ЛО} + \text{ЛП}} = \frac{44}{97 + 13} = 0.4$$

Чувствительность (Sensitivity, Recall, True positive rate) определяется как количество правильных положительных прогнозов, разделенных на количество положительных. Лучшая чувствительность — 1.0, худшая — 0.0.

$$M_{tpr} = \frac{\text{ИП}}{\text{ИП} + \text{ЛН}} = \frac{65}{65 + 44} = 0.6$$

Распространенность (Prevalence) описывает частоту получения положительного решения.

$$M_{er} = \frac{\text{ИП} + \text{ЛН}}{\text{ИП} + \text{ИО} + \text{ЛП} + \text{ЛН}} = \frac{65 + 44}{65 + 97 + 44 + 13} = 0.49$$

Таблица 2.

Результаты оценки полученных данных после предварительной обработки.

		Данные, содержащие аномалии (после предварительной классификации)		
		Позитивное состояние	Негативное состояние	
Тестовый результат	Позитивный результат	ИО = 43	ЛП = 44	Позитивная точность прогнозирования = 83%
	Негативный результат	ЛН = 13	ИО = 97	Негативная точность прогнозирования = 69%
		Чувствительность = 60%	Специфика = 88%	Точность = 74%

Выводы

Описанный подход к вопросу решения задачи оценки качества алгоритмов достаточно широко применяется и является достаточно точным при выполнении некоторых базовых правил. Для достижения максимальной точности оценки необходимо выполнить основные условия представления данных при их хранении и обработке. Также для достижения более точной оценки качества алгоритмов, количество классов не должно превышать 150. В случае неровного количества классов нужно подбирать баланс классов для обучения и метрику, которая будет корректно отображать качество классификации.

Основываясь на результаты в указанном примере, было создано таблицу 2 на основании которой можно сделать следующие выводы:

- Высокая положительная точность прогнозирования (M_{ppv} = 83%) указывает на то, что много значений тестирования являются корректными. Обычно данный тест не дает абсолютно точного результата, но тем не менее, такой тест может быть полезным своей дешевизной и удобством.
- Положительная точность прогнозирования — зависит от распространенности [6 с. 14]. Из-за высокого влияния распространенности на прогнозные значения предложено стандартизированный подход, где положительная точность прогнозирования нормализуется к распространенности 50% [7 с. 59]. В приведенном выше примере, если группа тестовых данных, включала большую долю аномальных данных, то положительная точность прогнозирования, вероятно, будет выше, а негативная точность прогнозирования ниже. Если все данные в группе будут аномальными, то M_{ppv}будет равен 100%, а M_{npv} — 0%.
- Положительная точность прогнозирования используется для обозначения вероятности того, что в случае положительного теста данные действительно имеют аномальные значения. Однако, может быть больше одной причины для аномалии, и любая отдельная потенциальная причина не всегда может привести к явному искажению данных, которое наблюдается в сенсоре.
- Результаты данного теста на прямую зависят от качества входных данных, а значит для улучшения результатов тестирования необходимо повысить качество предварительной обработки данных.

Список литературы:

1. Котов А., Красильников Н. Кластеризация данных. 2006.
2. James W. Perry Allen Kent Madeline M. Berry. Machine literature searching X. Machine language; factors underlying its design and development
3. Кендалл М., Стьюарт А. Статистические выводы и связи. — М. : Наука, 1973.f
4. Android Anomaly Detection System Using Machine. Learning Classification. Harry Kurniawan. School of Informatics and Electrical. Engineering 2015.
5. Stephen V.Stehman. Selecting and interpreting measures of thematic classification accuracy.
6. Altman, DG; Bland, JM (1994). "Diagnostic tests 2: Predictive values".
7. Heston, Thomas F. (2011). "Standardizing predictive values in diagnostic imaging research".