Технология расчета и отображения матрицы сходства объектов методами VBA Excel

Определения и понятия

Унитреугольная матрица (верхняя или нижняя) — треугольная матрица **A**, в которой все элементы на главной диагонали равны единице: a_{jj} =1.
Data Scientist — это **специалист по работе с данными для решения задач бизнеса** (аналитик баз данных). Он работает на стыке программирования, машинного обучения и математики. В основные обязанности дата-сайентиста входит сбор и анализ данных, построение моделей, их обучение и тестирование. Data Scientist должен разбираться в том, как работает компания и конкретная индустрия, в которой он занят. Профессия Data Scientist постоянно развивается и высоко оплачивается.
Коэффициент сходства Съеренсена – мера схожести двух наборов данных, равна удвоенному количеству элементов, общих для обоих наборов, деленному на сумму количества элементов в каждом наборе.

Введение

В практике вычислений, реализуемых с помощью MS Excel, встречается численная оценка похожих объектов с целью их последующего отбора по заданным критериям.

Например, при геоботанических исследованиях нередко ставятся задачи оценки сообществ по флористическому составу и определения степени их сходства друг с другом. Такие операции могут применяться при сравнении локальных флор.

Настоящая статья посвящена описанию расчетной модели матрицы коэффициентов сходства Съеренсена методами VBA Excel на примере экологической задачи поиска флористического сходства локальных флор ($Area_1$,…$Area_n$).

Описан один из подходов к поиску схожих объектов в наборе данных. Информация будет полезна аналитикам, которые изучают VBA Excel и аналитикам баз данных.

Исходные данные модели

В качестве исходных данных модели сформированы объекты в формате столбцов, состоящих из названий растений, произрастающих на площадках $Area_1$,…$Area_9$ локальных флор
(см. рис.1).

Самым простым способом вычисления схожести объектов по текстовым характеристикам является расчет коэффициента схожести Съеренсена *(Ks)*.

В обобщенном виде он выражается формулой: $$K_S = \frac{2C}{A+B} \cdot 100\%,$$

где A и B - число видов в первом и втором описаниях (столбцах), соответственно;

С - число общих видов для этой пары описаний.

	A	B	C	D	E	F	G	H	I
1	Area 1	Area 2	Area 3	Area 4	Area 5	Area 6	Area 7	Area 8	Area 9
2	*Chlorochitrium sp.*	*Prasiola sp.*	*Nostoc sp.*	*Actinotaenium cucurbita*	*Coleodesmium wrange*	*Leptolyngbya sp.*	*Leptolyngbya sp.*	*Leptolyngbya sp.*	*Actinotaenium cucurbita*
3	*Chlorococcus cf. giganteus*	*Chlorosarcinopsis cf. aggregat*	*Leptolyngbya cf. antar*	*Coleodesmium wrangelii*	*Gloeocapsosis magma*	*Microcoleus steenstrupii*	*Nostoc sp.*	*Nostoc sp.*	*Coleodesmium wrangelii*
4	*Leptolyngbya sp.*	*Scenedesmus sp.*	*Oscillatoria cf. subbrev*	*Microspora stagnorum*	*Microspora stagnorum*	*Gloeocapsopsis pleuroca*	*Stichococcus sp.*	*Gloeocapsopsis magma*	*Microspora stagnorum*
5	*Nostoc sp.*	*Leptolyngbya undosa*	*Pseudodictyochloris sp*	*Gloeocapsopsis pleurocapso*	*Microcoleus steenstru*	*Coleodesmium wrangelii*	*Heterococcus sp.*	*Fragilaria sp.*	*Oscillatoria cf. subbrevis*
6	*Chlorosarcinopsis cf. aggregat*	*Mychonastes cf. homosphaer*	*Leptolyngbya undosa*	*Chlorogloea sp.*	*Leptolyngbya sp.*	*Klebsormidium sp.*	*Klebsormidium sp.*	*Leptolyngbya nigrescens*	*Pseudodictyochloris sp.*
7				*Gloeocapsosis magma*	*Klebsormidium sp.*	Green coccoid		*Romeria cf. nivicola*	*Leptolyngbya undosa*
8					Green coccoid	*Stichococcus sp.*		*Actinotaenium cucurbita*	
9					*Stichococcus sp.*			Green coccoid	

Рис. 1. Исходные данные модели.

Для оценки схожести всех объектов модель сходства объектов формирует нижнюю унитреугольную матрицу коэффициентов Ks , в которой будут отражены сразу все объекты, находящиеся в исходных данных.

Разработка динамической модели сходства объектов

Структура модели состоит из следующих 3-х основных блоков расчета:

Блок 1. Расчет числа дубликатов в двух анализируемых столбцах (параметр С)

```
' ==============   Блок 2 расчета числа дубликатов (параметр С)  ==================

  Set InputWS = ThisWorkbook.Sheets("Lists of species")    ' лист с данными
       With Worksheets("Lists of species")
       Set InputWS = ThisWorkbook.Sheets("Lists of species")    ' лист с данными
          RowsNum1 = .Cells(Rows.Count, Compare1Column).End(xlUp).Row  ' количество строк в 1 столбце
       RowsNum2 = .Cells(Rows.Count, Compare2Column).End(xlUp).Row  ' количество строк во 2 столбце

' -------------  определяем общее число строк для будущей матрицы поиска уникальных элементов RowsNum
            If RowsNum1 > RowsNum2 Then
            RowsNum = RowsNum1
            ElseIf RowsNum2 > RowsNum1 Then
            RowsNum = RowsNum2
            Else
            RowsNum = RowsNum1
            End If
 -------------  определяем общее число строк для будущей матрицы поиска уникальных элементов RowsNum
'цикл по каждой ячейке 1-го из сравниваемых столбцов
               CountDuplicates = 0  ' начальное значение числа дубликатов
                 For i = 2 To RowsNum1   ' цикл по строкам 1 столбца, начиная со 2-й строки

                   String1 = .Cells(i, Compare1Column).Text   ' название растения в цикле по строкам 1 столбца
' если это название имеется в диапазоне 2 столбца
                    If Application.WorksheetFunction.CountIf(.Range(.Cells(2, Compare2Column), _
                    .Cells(RowsNum2, Compare2Column)), String1) > 0 Then
                     CountDuplicates = CountDuplicates + 1 ' то наращиваем счетчик совпадений
                     Else  ' иначе это название не найдено
                     End If
                  Next i  'продолжаем цикл

  CountDuplicates = CountDuplicates * 2  ' домножаем на 2
```

Примечание: Показатель Count Duplicates рассчитывается формулой Application.WorksheetFunction.CountIf в цикле, где каждый из элементов одного столбца сравнивается попарно с элементами второго столбца, создавая на выходе из цикла удвоенное число дубликатов в обоих столбцах.

Блок 2. Расчет общего числа элементов в двух анализируемых столбцах (сумма A + B)

```
' ==============   Блок  3  расчета числа значений ( сумма A + B)  ==================

 Dim InputDateRange As Range ' диапазон исходных данных
 Dim rCell As Range   ' ячейка цикла
    ' создаем матрицу - массив из несмежных диапазонов для поиска уникальных значений
 Set InputDateRange1 = .Range(.Cells(2, Compare1Column), .Cells(RowsNum, Compare1Column))   ' установить 1-й диапазон исходных данных
 Set InputDateRange2 = .Range(.Cells(2, Compare2Column), .Cells(RowsNum, Compare2Column))   ' установить 2-й диапазон исходных данных
 Set InputDateRange = Application.Union(InputDateRange1, InputDateRange2)   ' установить сводный (1+2)  диапазон исходных данных

CountTotalStrings = 0
For Each rCell In InputDateRange.Cells   ' для каждой ячейки диапазона
If rCell.Value <> "" Then
CountTotalStrings = CountTotalStrings + 1
Else
End If
Next rCell
```

Примечание: спецификой данного блока является метод Union для объединения несмежных диапазонов в единый диапазон для последующего расчета количества непустых элементов.

Показатель CountTotalStrings рассчитывается как суммарное количество элементов единого диапазона.

Блок 3. Расчет, запись и условное форматирование коэффициентов сходства

```
  ' ===============  Расчет, запись и условное форматирование коэффициентов сходства ==============
    SemblaceCoeff = Format(CountDuplicates / CountTotalStrings, "Standard")


    End With


    Worksheets("Sorensen (Dice)").Select
    Worksheets("Sorensen (Dice)").Cells(k, j).Value = SemblaceCoeff
    Worksheets("Sorensen (Dice)").Cells(k, j).HorizontalAlignment = xlCenter
    Worksheets("Sorensen (Dice)").Cells(k, j).VerticalAlignment = xlCenter
    Worksheets("Sorensen (Dice)").Cells(k, j).Font.Size = 12
'   Debug.Print " SemblaceCoeff   = " & Worksheets("Sorensen (Dice)").Cells(k, j).Value

        Next k    ' продолжаем цикл по строкам
   Next j     ' продолжаем цикл по столбцам


    Worksheets("Sorensen (Dice)").Cells(LastColumn + 1, LastColumn + 1).Value = 1 ' прописываем главную диагональ матрицы
    Worksheets("Sorensen (Dice)").Cells(LastColumn + 1, LastColumn + 1).HorizontalAlignment = xlCenter
    Worksheets("Sorensen (Dice)").Cells(LastColumn + 1, LastColumn + 1).VerticalAlignment = xlCenter
    Worksheets("Sorensen (Dice)").Cells(LastColumn + 1, LastColumn + 1).Interior.Color = vbYellow   ' заливка желтым цветом

 '  применяем цветовую легенду к значениям коэффициента Серенсена
 For Each rCell In MatrixRange.Cells   ' для каждой ячейки матрицы

 If rCell.Row > rCell.Column And IsNumeric(rCell.Value) Then    ' если ячейка находится в нижней диагонали матрицы и не пустая
    If rCell.Value >= 0# And rCell.Value <= 0.5 Then         ' если ячейка в диапазоне 0 - 0,50
      rCell.Interior.Color = vbCyan
    ElseIf rCell.Value > 0.5 And rCell.Value <= 0.99 Then       ' если ячейка в диапазоне 0,5 - 0,99
      rCell.Interior.Color = vbMagenta
 Else
 End If
```

Примечание: показатель SemblaceCoeff (коэффициент Съёренсена) рассчитывается по установленной формуле как отношение Count Duplicates к CountTotalStrings.

Технология расчета и отображения матрицы сходства объектов методами VBA Excel

В результате расчетов модель сходства объектов размещает *коэффициенты* ***Съеренсена*** в унитреугольной матрице коэффициентов (см. рис.2).
Цветом отмечены классификационные ранги сообществ: цветом cyan выделены ***слабые*** сочетания растительных сообществ, цветом magenta - ***сильные*** сочетания растительных сообществ, которые относятся к одной ассоциации.

Calculate Semblence Koeffs 0,00-0,50 - cyan; 0,50-0,99 - magenta	Area 1	Area 2	Area 3	Area 4	Area 5	Area 6	Area 7	Area 8	Area 9
Area 1	1								
Area 2	0,20	1							
Area 3	0,20	0,20	1						
Area 4	0,00	0,00	0,00	1					
Area 5	0,15	0,00	0,00	0,43	1				
Area 6	0,17	0,00	0,00	0,15	0,67	1			
Area 7	0,40	0,00	0,20	0,00	0,46	0,50	1		
Area 8	0,31	0,00	0,15	0,14	0,25	0,27	0,31	1	
Area 9	0,00	0,18	0,55	0,50	0,29	0,00	0,00	0,14	1

Рис. 2 – Унитреугольная матрица коэффициентов флористического сходства

Преимущество динамической модели сходства объектов

В технологию формирования модели сходства объектов внесены операторы VBA, создающие динамическую матрицу при изменении объемов исходных данных.
Пользователь динамической модели может как добавлять, так и удалять произвольное количество столбцов исходных данных.
Программа автоматически сформирует и оформит строки и столбцы таблицы результата для матрицы расчетного размера.

Автор: Н.Н. Дворец